Парадигма развития науки

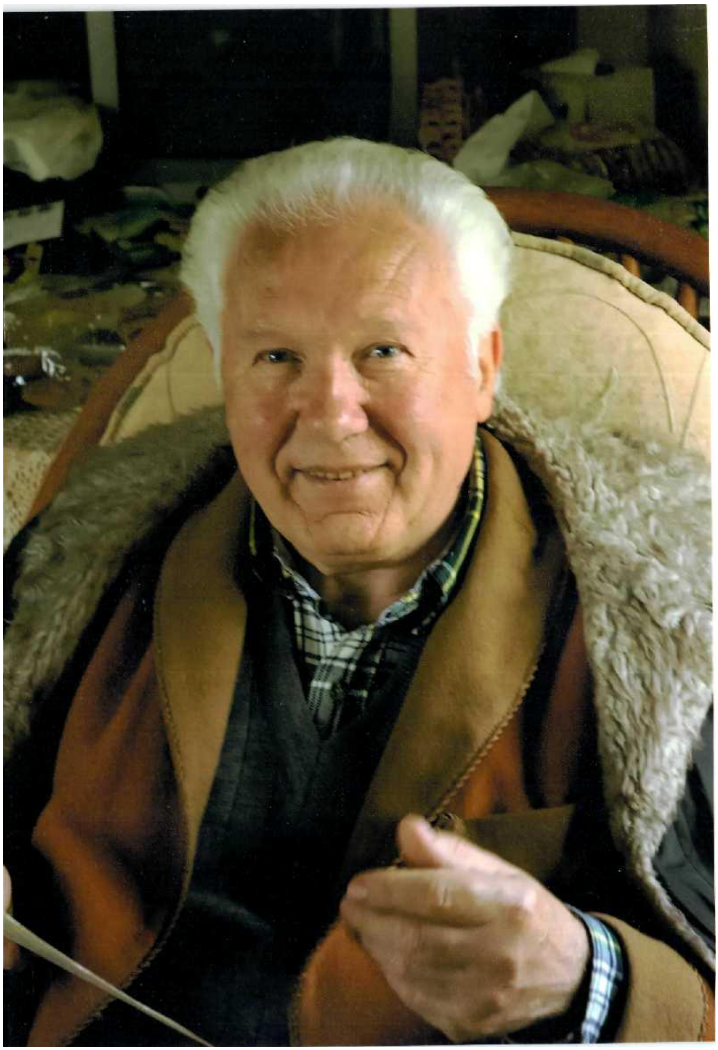
А. Е. Кононюк

Основы фундаментальной теории искусственного интеллекта

Книга 7
Меры, размерности, измерения – фундаментальные атрибуты ИИ

Часть 1
Меры, размерности, измерения, интервалы

Киев
«Освіта України»
2018



Кононюк Анатолий Ефимович

Структурная схема
парадигмы развития науки

УДК 51 (075.8)
ББК В161.я7
К65

Рецензент:
Н.К.Печурин - д-р техн. наук, проф. (Национальный авиационный университет).

Кононюк А. Е.
К213 Основы фундаментальной теория искусственного интеллекта. — В 20-и кн. Кн. 7, ч.1 — К.:Освіта України. 2018. 752с.

ISBN 978-966-373-693-8 (многотомное издание)
ISBN 978-966-373-694-15 (книга 7, ч.1)

Многотомная работа посвящена систематическому изложению общих формализмов, математических моделей и алгоритмических методов, которые могут быть исспользуемых при моделировании и исследованиях математических моделей объектов искусственного интеллекта.

Развиваются представления и методы решения, основанные на теориях эвристического поиска и автоматическом доказательстве теорем, а также процедуральные методы, базирующиеся на классе проблемно-ориентированных языков, сочетающих свойства языков программирования и автоматических решателей задач отображения искусственного интеллекта различными математическими средствами.

В работе излагаются основы теории отображения искусственного интеллекта такими математическими средствами как: множества, отношения, поверхности, пространства, алгебраические системы, матрицы, графы, математическая логика и др.

Для бакалавров, специалистов, магистров, аспирантов, докторантов всех специальностей.

УДК 51 (075.8)
ББК В161.я7

ISBN 978-966-373-693-8 (многотомное издание)
ISBN 978-966-373-694-15 (книга 7, ч.1)

Оглавление

Часть I. Меры

1. Введение в теорию мер

1.1 Определения

Первоначальные представления о площади и объеме можно описать следующими аксиомами:
1. Каждому ограниченному множеству $A \subset \mathbb{R}^n$ сопоставлено неотрицательное число *V(A)*, называемое (n-мерным) объемом этого множества.
2. Объем аддитивен: если $A \cap B = \emptyset$, то $V(A \cup B) = V(A) + V(B)$.
3. Если множества *A* и *B* конгруэнтны (совмещаются движением), то их объемы равны.
4. Объем единичного куба равен 1.

У этих аксиом есть фатальный недостаток: они внутренне противоречивы. Противоречие было предъявлено в 1914 году Хаусдорфом. Позднее (в 1926 году) Банах и Тарский оформили его в виде следующей теоремы.

Теорема (парадокс Банаха—Тарского). *Можно разбить стандартный шар* $B \subset \mathbb{R}^3$ *на 5 непересекающихся множеств* A_1, A_2, A_3, A_4, A_5 *и построить такие множества* B_1, B_2, B_3, B_4, B_5, *что*
1. Каждое множество B_i *конгруэнтно соответствующему множеству* A_i.
2. B_1 *и* B_2 *не пересекаются и их объединение равно B.*
3. B_3, B_4 *и* B_5 *попарно не пересекаются и их объединение равно B.*

Парадо́кс Ба́наха — Та́рского (также называется **парадо́ксом удвое́ния ша́ра** и **парадоксом Хаусдорфа — Ба́наха — Та́рского**)— теорема в теории множеств, утверждающая, что трёхмерный шар равносоставлен двум своим копиям.

Два подмножества евклидова пространства называются *равносоставленными*, если одно можно разбить на конечное число (не обязательно связных) попарно непересекающихся частей, передвинуть их (при этом частям не запрещается «проходить друг сквозь друга», то

есть не требуется оставаться попарно непересекающимися во всех промежуточных положениях), и составить из них второе.

Более точно, два множества A и B являются равносоставленными, если их можно представить как конечное объединение непересекающихся подмножеств $A = \bigcup_{i}^{n} A_i$, $B \bigcup_{i}^{n} B_i$ так, что для каждого i *подмножество* A_i конгруэнтно B_i.

Доказано, что для удвоения шара достаточно пяти частей, но четырёх недостаточно.

Верен также более сильный вариант парадокса:

Любые два ограниченных подмножества трёхмерного евклидова пространства с непустой внутренностью являются равносоставленными.

Ввиду того, что вывод этой теоремы может показаться неправдоподобным, она иногда используется как довод против принятия аксиомы выбора, которая существенно используется при построении такого разбиения. Принятие подходящей альтернативной аксиомы позволяет доказать невозможность указанного разбиения, не оставляя места для этого парадокса.

Удвоение шара, хотя и кажется весьма подозрительным с точки зрения повседневной интуиции (в самом деле, нельзя же из одного апельсина сделать два при помощи одного только ножа), тем не менее не является парадоксом в логическом смысле этого слова, поскольку не приводит к логическому противоречию наподобие того, как к логическому противоречию приводит так называемый парадокс брадобрея или парадокс Рассела.

Разделяя шар на конечное число частей, мы интуитивно ожидаем, что, складывая эти части вместе, можно получить только сплошные фигуры, объём которых равен объёму исходного шара. Однако это справедливо только в случае, когда шар делится на части, имеющие объём. Суть парадокса заключается в том, что в трёхмерном пространстве существуют неизмеримые множества, которые не имеют объёма, если под объёмом мы понимаем то, что обладает свойством

аддитивности, и предполагаем, что объёмы двух конгруэнтных множеств совпадают. Очевидно, что «куски» в таком разбиении не могут быть измеримыми (и невозможно осуществить такое разбиение какими-либо средствами на практике).

Для плоского круга аналогичное свойство неверно. Более того, Банах показал, что на плоскости понятие площади может быть продолжено на все ограниченные множества как конечно-аддитивная мера, инвариантная относительно движений; в частности, любое множество, равносоставленное кругу, имеет ту же площадь. Хаусдорф показал, что подобное сделать нельзя на двумерной сфере, и, следовательно, в трёхмерном пространстве, и парадокс Банаха — Тарского даёт этому наглядную иллюстрацию.

Тем не менее, некоторые парадоксальные разбиения возможны и на плоскости: круг можно разбить на конечное число частей и составить из них квадрат равной площади (квадратура круга Тарского).

1.2. Алгебра и сигма-алгебра событий

1.2.1. Алгебра событий

Пусть Ω— пространство элементарных исходов некоторого случайного эксперимента (т.е. непустое множество произвольной природы). Мы собираемся определить набор подмножеств Ω, которые будут называться событиями, и затем задать вероятность как функцию, определённую *только* на множестве событий .

Событиями мы будем называть не любые подмножества Ω, а лишь элементы некоторого выделенного набора подмножеств Ω. При этом необходимо позаботиться, чтобы этот набор подмножеств был замкнут относительно операций над событиями, т.е. чтобы объединение, пересечение, дополнение событий снова давало событие. Сначала введём понятие алгебры событий.

Определение. Множество $\mathcal{A}$, элементами которого являются подмножества множества Ω(не обязательно все) называется *алгеброй* (алгеброй событий), если оно удовлетворяет следующим условиям:

(A1) $\Omega \in \mathcal{A}$ (алгебра событий содержит достоверное событие);

(А2) если $A \in \mathcal{A}$, то $\overline{A} \in \mathcal{A}$ (вместе с любым событием алгебра содержит противоположное событие);

(А3) если $A \in \mathcal{A}$ и $B \in \mathcal{A}$, то $A \cup B \in \mathcal{A}$ (вместе с любыми двумя событиями алгебра содержит их объединение).

Из свойств **(А1)** и **(А2)** следует, что пустое множество $\varnothing = \overline{\Omega}$ также содержится в $\mathcal{A}$.

Из **(А3)** следует, что вместе с любым конечным набором событий алгебра содержит их объединение: для любого $n \geqslant 2$, для любых A_1, ..., $A_n \in \mathcal{A}$ выполнено $A_1 \cup \ldots \cup A_n \in \mathcal{A}$.

Вместо замкнутости относительно операции объединения можно требовать замкнутость относительно операции пересечения.

Свойство 1. Свойство **(А3)** в определении можно заменить на

(А4) если $A \in \mathcal{A}$ и $B \in \mathcal{A}$, то $A \cap B \in \mathcal{A}$.

Доказательство. Докажем, что при выполнении **(А1)** и **(А2)** из **(А3)** следует **(А4)**. Если $A, B \in \mathcal{A}$, то $\overline{A} \in \mathcal{A}$, $\overline{B} \in \mathcal{A}$ по свойству **(А2)**. Тогда из **(А3)** следует, что $\overline{A} \cup \overline{B} \in \mathcal{A}$, и, по **(А2)**, дополнение $\overline{\overline{A} \cup \overline{B}}$ к этому множеству также принадлежит $\mathcal{A}$. В силу формул двойственности, дополнение к объединению как раз и есть пересечение дополнений:

$$A \cap B = \overline{\overline{A} \cup \overline{B}} \in \mathcal{A}.$$

Аналогично доказывается, что при выполнении **(А1)** и **(А2)** из **(А4)** следует **(А3)**, т.е. эти два свойства в определении взаимозаменяемы.

Пример 11. Пусть $\Omega = \{\spadesuit, \clubsuit, \diamondsuit, \heartsuit\}$ — пространство элементарных исходов. Следующие наборы подмножеств Ω являются алгебрами (проверьте это по определению):

1. $\mathcal{A} = \{\Omega,\ \varnothing\} = \{\{\spadesuit, \clubsuit, \diamondsuit, \heartsuit\},\ \varnothing\}$ — *тривиальная* алгебра.

2. $\mathcal{A} = \{\Omega,\ \varnothing,\ \{\diamondsuit\},\ \Omega \setminus \{\diamondsuit\}\} = \{\{\spadesuit, \clubsuit, \diamondsuit, \heartsuit\},\ \varnothing,\ \{\diamondsuit\},\ \{\spadesuit, \clubsuit, \heartsuit\}\}$.

3. $\mathcal{A}=\{\Omega,\ \varnothing,\ A,\ \overline{A}\} = \{\{\spadesuit, \clubsuit, \diamondsuit, \heartsuit\},\ \varnothing,\ A,\ \overline{A}\}$, где A — произвольное подмножество Ω (в предыдущем примере $A = \{\diamondsuit\}$).

4. $\mathcal{A} = 2^{\Omega}$ — *множество всех подмножеств* Ω.

Упражнение. Доказать, что если Ω состоит из n элементов, то в множестве всех его подмножеств ровно 2^n элементов.

1.2.2. Сигма-алгебра событий.

В теории вероятностей часто возникает необходимость объединять счётные наборы событий и считать событием результат такого объединения. При этом свойства **(A3)** алгебры оказывается недостаточно: из него не вытекает, что объединение счётной последовательности множеств из алгебры снова принадлежит алгебре. Поэтому разумно наложить более суровые ограничения на класс событий.

Определение. Множество $\mathcal{F}$, элементами которого являются подмножества множества Ω (не обязательно все) называется σ-алгеброй (σ-алгеброй событий), если выполнены следующие условия:

(S1) $\Omega \in \mathcal{F}$ (σ-алгебра событий содержит достоверное событие);

(S2) если $A \in \mathcal{F}$, то $\overline{A} \in \mathcal{F}$ (вместе с любым событием σ-алгебра содержит противоположное событие);

(S3) если $A_1, A_2, \ldots \in \mathcal{F}$, то $A_1 \cup A_2 \cup \ldots \in \mathcal{F}$ (вместе с любым счётным набором событий σ-алгебра содержит их объединение).

Упражнение.

а)

Доказать, что вместо **(S1)** достаточно предположить непустоту множества $\mathcal{F}$.

б)

Вывести из **(S1)** и **(S2)**, что $\varnothing \in \mathcal{F}$.

Этого набора аксиом достаточно для замкнутости множества $\mathcal{F}$ относительно счётного числа любых других операций над событиями. В частности, аналогично свойству 1 проверяется следующее свойство.

Свойство 2. Свойство **(S3)** в определении 11 можно заменить на

(S4) если $A_1, A_2, \ldots \in \mathcal{F}$, то $A_1 \cap A_2 \cap \ldots \in \mathcal{F}$.

Как показывает следующее свойство, всякая σ-алгебра автоматически является алгеброй.

Свойство 3. Если $\mathcal{F}$ — σ-алгебра, то она удовлетворяет свойству **(A3)**, т.е. для любых $A \in \mathcal{F}$ и $B \in \mathcal{F}$ выполняется $A \cup B \in \mathcal{F}$.

Доказательство. Превратим пару $A,\ B$ в счётную последовательность событий так: $A,\ B,\ B,\ B,\ B,\ \ldots$, т.е. положим $A_1 = A$, $A_i = B$ при всех $i \geqslant 2$. Объединение $A \cup B$ совпадает с объединением всех множеств A_i из этой бесконечной последовательности. А так как $\mathcal{F}$ — σ-алгебра, то

$$A \cup B = \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

Упражнение. Докажите, что для любых $A, B \in \mathcal{F}$ выполнено $A \setminus B \in \mathcal{F}$.

Обратное, вообще говоря, неверно: не всякая алгебра является сигма-алгеброй. Чтобы показать это, приведём пример алгебры, не являющейся σ-алгеброй.

Пример. Пусть $\Omega = \mathbb{R}$, и пусть $\mathcal{A}$ — множество, содержащее любые конечные подмножества $\mathbb{R}$ (т.е. состоящие из конечного числа точек, в том числе пустое) и их дополнения. В частности, множество $\{0, 2, \pi\}$ принадлежит $\mathcal{A}$, множество $(-\infty, -7{,}2) \cup (-7{,}2, 5) \cup (5, \infty)$ также принадлежит $\mathcal{A}$.

Легко проверить, что множество $\mathcal{A}$ является алгеброй. Действительно, пустое множество и само $\Omega = \mathbb{R}$ там содержатся, дополнение к любому конечному подмножеству множества вещественных чисел содержится в $\mathcal{A}$ по определению, дополнение к множеству вида $\mathbb{R} \setminus A$ для конечных A совпадает с A и также принадлежит $\mathcal{A}$ по определению. Свойство **(А3)** проверяется непосредственно: объединение любых конечных множеств снова конечно и поэтому принадлежит $\mathcal{A}$. Объединение конечного множества с множеством вида $\mathbb{R} \setminus A$, где A конечно, есть снова множество вида $\mathbb{R} \setminus B$, где B конечно (или пусто). Объединение двух множеств $\mathbb{R} \setminus A$ и $\mathbb{R} \setminus B$, являющихся дополнениями до $\mathbb{R}$ конечных множеств A и B, есть снова множество такого же вида.

Однако алгебра $\mathcal{A}$ не содержит ни одного счётного множества точек. Действительно, объединяя конечные множества в конечном числе, мы можем получить только конечное множество. Например, натуральный ряд $\mathbb{N}$ не принадлежит $\mathcal{A}$. Поэтому $\mathcal{A}$ не является σ-алгеброй: для бесконечной, но счётной последовательности одноточечных множеств $A_i = \{i\}$ из $\mathcal{A}$ их объединение $\mathbb{N} = A_1 \cup A_2 \cup \ldots$ не принадлежит $\mathcal{A}$.

Все алгебры из примера являются σ-алгебрами, поскольку содержат лишь конечное число элементов. Вообще, на конечном множестве Ω понятия алгебры и σ-алгебры совпадают. Множество всех подмножеств Ω является σ-алгеброй для любого Ω.

1.2.3. Борелевская σ-алгебра в $\mathbb{R}$.

Приведём пример σ-алгебры, которая нам будет необходима в дальнейшем, — σ-алгебры *борелевских множеств* на вещественной прямой.

Борелевской сигма-алгеброй в $\mathbb{R}$ называется самая маленькая среди всех возможных σ-алгебр, содержащих любые интервалы на прямой. Разумеется, σ-алгебры, содержащие все интервалы, существуют. Например, таково множество всех подмножеств $\mathbb{R}$. Чтобы сделать эти слова — про самую маленькую σ-алгебру — понятными, поработаем с примерами.

Пример. Пусть $\Omega = \mathbb{R}$ — вещественная прямая. Рассмотрим некоторые наборы множеств, не являющиеся σ-алгебрами, и увидим, как их можно дополнить до σ-алгебр.

1. Множество $\mathfrak{A} = \{\mathbb{R}, \varnothing, [0,\ 1], \{0\}\}$ не является σ-алгеброй, так как, например, $\overline{[0,\ 1]} = \mathbb{R} \setminus [0,\ 1] = (-\infty, 0) \cup (1, \infty) \notin \mathfrak{A}$. Самый маленький набор множеств, содержащий $\mathfrak{A}$ и являющийся σ-алгеброй (минимальная σ-алгебра), получится, если включить в него всевозможные объединения, пересечения и дополнения множеств из $\mathfrak{A}$:

$$\mathcal{F} = \{\mathbb{R},\ \varnothing,\ [0,\ 1],\ \{0\},\ (-\infty, 0) \cup (1, \infty), (0, 1], (-\infty, 0] \cup (1, \infty), (-\infty, 0) \cup (0, \infty)\}$$

.

Более точно:

Определение. *Минимальной* σ-алгеброй, содержащей некоторый набор множеств $\mathfrak{A}$, называется пересечение всех σ-алгебр, содержащих $\mathfrak{A}$.

Ещё раз напомним, что пересекать в определении есть что: хотя бы одна σ-алгебра, содержащая данный набор множеств, всегда найдётся — это σ-алгебра всех подмножеств $\mathbb{R}$.

Упражнение. Доказать, что пересечение двух σ-алгебр, содержащих набор множеств $\mathfrak{A}$, снова является σ-алгеброй (невероятно!), содержащей $\mathfrak{A}$.

2. Пусть множество $\mathfrak{A}$ подмножеств вещественной прямой $\mathbb{R}$ состоит из всевозможных открытых интервалов (a,b), где $a<b$:

$$\mathfrak{A}=\{(a,b)\,|\,-\infty<a<b<\infty\}.$$

Упражнение. Проверить, что множество $\mathfrak{A}$ всех интервалов ни в коем случае не является ни алгеброй, ни σ-алгеброй! Указание: привести примеры двадцати множеств из $\mathfrak{A}$, дополнения к которым не принадлежат $\mathfrak{A}$; привести примеры пяти множеств из $\mathfrak{A}$, любые объединения которых не принадлежат $\mathfrak{A}$.

Определение. Минимальная σ-алгебра, содержащая множество $\mathfrak{A}$ всех интервалов на вещественной прямой, называется *борелевской сигма-алгеброй* в $\mathbb{R}$ и обозначается $\mathfrak{B}(\mathbb{R})$.

Перечислим некоторые множества на прямой, содержащиеся в $\mathfrak{B}(\mathbb{R})$ по определению. Таковы все привычные нам множества. Чтобы получить множество, не содержащееся в $\mathfrak{B}(\mathbb{R})$, требуются специальные построения.

Итак, мы знаем, что все интервалы на прямой принадлежат $\mathfrak{B}(\mathbb{R})$, и $\mathfrak{B}(\mathbb{R})$ — σ-алгебра. Отсюда сразу следует, что $\mathfrak{B}(\mathbb{R})$ содержит любое множество, которое можно получить из интервалов с помощью счётного числа операций объединения или пересечения, а также взятием дополнения.

В частности, $\mathbb{R}$ принадлежит $\mathfrak{B}(\mathbb{R})$.

Доказательство. Это сразу следует из свойства **(S1)** σ-алгебры, но может быть доказано и исходя из свойств **(S2), (S3)**. Интервал $(-n,n)$ принадлежит $\mathfrak{A}$, а значит, принадлежит и $\mathfrak{B}(\mathbb{R})$ при любом $n\in\mathbb{N}$, т.е. $(-n,n)\in\mathfrak{B}(\mathbb{R})$. Но $\mathfrak{B}(\mathbb{R})$ — σ-алгебра, и содержит счётное объединение любых своих элементов, поэтому

$$\mathbb{R} = \bigcup_{n=1}^{\infty} (-n, n) \ \in \ \mathfrak{B}(\mathbb{R}).$$

Далее, любой интервал вида $(a, b]$ (или $[a, b)$, или $[a,\ b]$), где $a < b$, принадлежит $\mathfrak{B}(\mathbb{R})$.

Доказательство. Интервал $(a, b + 1/n)$ принадлежит $\mathfrak{B}(\mathbb{R})$ при любом $n \in \mathbb{N}$. Тогда счётное пересечение этих интервалов

$$(a, b] = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right)$$

по свойству **(S4)** также принадлежит $\mathfrak{B}(\mathbb{R})$.

Любое одноточечное подмножество $\{b\} \subset \mathbb{R}$ принадлежит $\mathfrak{B}(\mathbb{R})$.

Доказательство. Действительно, $\{b\} = (a, b] \setminus (a, b)$, а разность $A \setminus B = A \cap \overline{B}$ двух множеств из σ-алгебры снова принадлежит σ-алгебре.

Упражнение. Докажите, что множества вида $(a_1, b_1) \cup (a_2, b_2)$ принадлежат $\mathfrak{B}(\mathbb{R})$, что множество натуральных чисел $\mathbb{N}$ принадлежит $\mathfrak{B}(\mathbb{R})$, множество рациональных чисел $\mathbb{Q}$ принадлежит $\mathfrak{B}(\mathbb{R})$.

3. Борелевская σ-алгебра в $\mathbb{R}^n$ строится совершенно так же, как в $\mathbb{R}$. Это должна быть минимальная σ-алгебра, содержащая все множества вида $(a_1, b_1) \times \ldots \times (a_n, b_n)$ — уже не интервалы, как в $\mathbb{R}$, а прямоугольники в $\mathbb{R}^2$, параллелепипеды в $\mathbb{R}^3$ и т.д. Вместе с ними $\mathfrak{B}(\mathbb{R}^n)$ содержит любые множества, являющиеся «предельными» для объединений измельчающихся прямоугольников. Например, круг в $\mathbb{R}^2$ является борелевским множеством — можно изнутри или снаружи приблизить его объединениями прямоугольников.

Итак, мы определили специальный класс $\mathcal{F}$ подмножеств пространства элементарных исходов Ω, названный σ-алгеброй событий, причём применение счётного числа любых операций (объединений, пересечений, дополнений) к множествам из $\mathcal{F}$ снова дает множество из $\mathcal{F}$, т.е. не выводит за рамки этого класса. *Событиями* будем называть *только* множества $A \in \mathcal{F}$.
Определим теперь понятие вероятности как функции, определённой на множестве событий (функции, которая каждому событию ставит в соответствие число — вероятность этого события).

А чтобы читателю сразу стало понятно, о чём пойдёт речь, добавим: вероятность мы определим как неотрицательную нормированную меру, заданную на σ-алгебре $\mathcal{F}$ подмножеств Ω.

Следующий параграф познакомит нас с понятиями меры и вероятностной меры.

1.3. Мера и вероятностная мера

1.3.1. Мера как неотрицательная σ-аддитивная функция множеств

Определение. Пусть Ω— некоторое множество и $\mathcal{F}$— σ-алгебра его подмножеств. Функция $\mu : \mathcal{F} \to \mathbb{R} \cup \{+\infty\}$ называется *мерой* на $(\Omega, \mathcal{F})$, если она удовлетворяет условиям:

(μ1) для любого множества $A \in \mathcal{F}$ его мера неотрицательна: $\mu(A) \geqslant 0$;

(μ2) для любого счётного набора попарно непересекающихся множеств $A_1, A_2, A_3, \ldots \in \mathcal{F}$ (т.е. такого, что $A_i \cap A_j = \varnothing$ при всех $i \neq j$) мера их объединения равна сумме их мер:

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

(*«счётная аддитивность»* или *«сигма-аддитивность»* меры).
Пример.

Пусть $\Omega = \{a, b, c\}$, $\mathcal{F} = 2^{\Omega}$ — множество всех подмножеств Ω. Зададим меру μ на $\mathcal{F}$ так: $\mu\{a\} = 3$, $\mu\{b\} = 17$, $\mu\{c\} = 1$, $\mu\{a, b\} = 20$, $\mu\{a, c\} = 4$, $\mu\{b, c\} = 18$, $\mu\{a, b, c\} = 21$, $\mu(\varnothing) = 0$. Для краткости записи мы вместо $\mu(\{a\})$ писали всюду $\mu\{a\}$.

Пример.

Пусть $\Omega = \mathbb{N}$, $\mathcal{F} = 2^{\mathbb{N}}$ — множество всех подмножеств натурального ряда. Зададим меру μ на $\mathcal{F}$ так: $\mu(A) = |A|$ — число элементов в множестве A (или бесконечность, если множество A не является конечным).

Пример (мера Лебега). Когда мы говорили о геометрической вероятности, мы использовали термин «мера области A в $\mathbb{R}^m$», имея в виду «длину» на прямой, «площадь» на плоскости, «объем» в трёхмерном пространстве. Являются ли все эти «длины-площади-объемы» настоящими мерами в смысле определения 14? Мы решим этот вопрос для прямой, оставляя плоскость и пространство большей размерности читателю.

Рассмотрим вещественную прямую с σ-алгеброй борелевских множеств. Эта σ-алгебра, по определению, есть наименьшая σ-алгебра, содержащая любые интервалы. Для каждого интервала $(a, b) \subset \mathbb{R}$ число $b - a$ назовём *длиной* интервала (a, b).

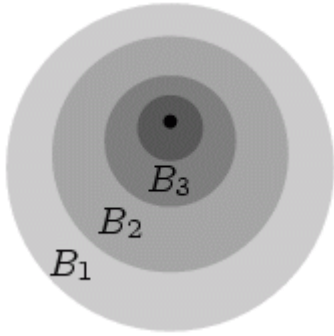
Мы не станем доказывать следующее утверждение:

Лемма 1. Существует единственная мера λ на $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, значение которой на любом интервале равно его длине: $\lambda(a,b) = b - a$. Эта мера называется *мерой Лебега.*

Замечание. Это утверждение является следствием теоремы Каратеодори о продолжении меры с алгебры на σ-алгебру, применительно к $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. Пользуясь процедурой лебеговского продолжения (пополнения) меры, можно распространить меру λ на более широкую σ-алгебру, нежели борелевская, — на σ-алгебру измеримых по Лебегу множеств. Для этого достаточно присвоить нулевую меру любым подмножествам борелевских множеств с нулевой мерой Лебега.

Нам пригодится свойство, которым обладает любая мера. Это свойство непрерывности меры иногда называют *аксиомой непрерывности*, имея в виду, что ею можно заменить (μ2) в определении .

Лемма 2 (непрерывность меры). Пусть дана убывающая последовательность $B_1 \supseteq B_2 \supseteq B_3 \supseteq \cdots$ вложенных друг в друга множеств из $\mathcal{F}$ такая, что $\mu(B_1) < \infty$ и $B = \bigcap_{n=1}^{\infty} B_n$. Тогда $\mu(B) = \lim_{n\to\infty} \mu(B_n)$.



Доказательство. Обозначим через C_n кольца: $C_n = B_n \setminus B_{n+1}$. Множества $B, C_1, C_2, C_3, \ldots$ попарно не пересекаются. Тогда из представлений

$$B_1 = B \cup \left(\bigcup_{i=1}^{\infty} C_i \right), \qquad B_n = B \cup \left(\bigcup_{i=n}^{\infty} C_i \right)$$

в силу аксиомы (μ2) следует, что

$$\mu(B_1) = \mu(B) + \sum_{i=1}^{\infty} \mu(C_i), \qquad \mu(B_n) = \mu(B) + \sum_{i=n}^{\infty} \mu(C_i).$$

Первая сумма $\sum_{i=1}^{\infty} \mu(C_i)$ в силу условия $\mu(B_1) < \infty$ есть сумма абсолютно сходящегося ряда (составленного из неотрицательных слагаемых). Из сходимости этого ряда следует, что «хвост» ряда, равный $\sum_{i=n}^{\infty} \mu(C_i)$, стремится к нулю при $n \to \infty$. Поэтому

$$\mu(B_n) = \mu(B) + \sum_{i=n}^{\infty} \mu(C_i) \underset{n \to \infty}{\longrightarrow} \mu(B) + 0 = \mu(B).$$

В полезности этого свойства легко убедиться упражнениями.
Упражнение 18.

Используя аксиому непрерывности меры для множеств $B_n = (x - 1/n, x + 1/n)$, доказать, что мера Лебега одноточечного подмножества $\{x\}$ вещественной прямой равна нулю: $\lambda\{x\} = 0$. Используя этот факт, доказать, что $\lambda(\mathbb{N}) = 0$, $\lambda(\mathbb{Z}) = 0$, $\lambda(\mathbb{Q}) = 0$, $\lambda(a, b) = \lambda[a,\ b]$.

Замечание. В отсутствие предположения $\mu(B_1) < \infty$ (или $\mu(B_n) < \infty$ для некоторого $n \geqslant 1$), заставляющего меры

вложенных множеств быть конечными, свойство $\mu(B) = \lim_{n\to\infty} \mu(B_n)$ может не выполняться.

Например, зададим меру на $\mathfrak{B}(\mathbb{R})$ так: $\mu(B) = 0$, если B не более чем счётно, иначе $\mu(B) = \infty$. Тогда для множеств $B_n = (x - 1/n, x + 1/n)$ имеем:

$$B = \bigcap_{n=1}^{\infty} B_n = \{x\}, \quad \mu(B_n) = \infty \not\to \mu(B) = 0.$$

1.3.2. Вероятность как нормированная мера.

Определение. Пусть Ω— множество и $\mathcal{F}$— σ-алгебра его подмножеств. Мера $\mu : \mathcal{F} \to \mathbb{R}$ называется *нормированной,* если $\mu(\Omega) = 1$. Другое название нормированной меры — *вероятность* или *вероятностная мера*.

То же самое ещё раз и подробно:

Определение . Пусть Ω— пространство элементарных исходов, $\mathcal{F}$— σ-алгебра его подмножеств (событий). *Вероятностью* или *вероятностной мерой* на $(\Omega, \mathcal{F})$ называется функция $\mathsf{P} : \mathcal{F} \to \mathbb{R}$, обладающая свойствами:

(P1) для любого события $A \in \mathcal{F}$ выполняется неравенство $\mathsf{P}(A) \geqslant 0$;

(P2) для любого счётного набора попарно несовместных событий $A_1, A_2, A_3, \ldots \in \mathcal{F}$ имеет место равенство

$$\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i);$$

(P3) вероятность достоверного события равна единице: $\mathsf{P}(\Omega) = 1$.

Свойства **(P1)** — **(P3)** называют *аксиомами* вероятности.

Определение. Тройка $\langle \Omega, \mathcal{F}, \mathsf{P} \rangle$, в которой Ω— пространство элементарных исходов, $\mathcal{F}$— σ-алгебра его подмножеств и P— вероятностная мера на $\mathcal{F}$, называется *вероятностным пространством.*

Докажем свойства вероятности, вытекающие из аксиом. Ниже мы не будем всякий раз оговаривать, но будем иметь в виду, что имеем дело только с событиями.

Свойство 0. $\mathsf{P}(\varnothing) = 0$.

Доказательство. События $A_i = \varnothing$, где $i \geqslant 1$, попарно несовместны, и их объединение есть также пустое множество. По аксиоме **(P2)**,

$$\mathsf{P}(\varnothing) = \sum_{i=1}^{\infty} \mathsf{P}(A_i) = \sum_{i=1}^{\infty} \mathsf{P}(\varnothing).$$

Это возможно только в случае $\mathsf{P}(\varnothing) = 0$.

Аксиома счётной аддитивности вероятности **(P2)** тем более верна для конечного набора попарно несовместных событий.

Свойство 1. Для любого конечного набора попарно несовместных событий $A_1, \ldots, A_n \in \mathcal{F}$ имеет место равенство

$$\mathsf{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathsf{P}(A_i).$$

Доказательство. Положим $A_i = \varnothing$ при любом $i > n$. Вероятности этих событий, по свойству 0, равны нулю. События $A_1, \ldots, A_n, \varnothing, \varnothing, \varnothing, \cdots$ попарно несовместны, и по аксиоме **(P2)**,

$$\mathsf{P}\left(\bigcup_{i=1}^{n} A_i\right) = \mathsf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_i) = \sum_{i=1}^{n} \mathsf{P}(A_i).$$

Сразу несколько следствий можно получить из этого свойства.

Свойство 2. Для любого события A выполнено:

$\mathsf{P}(\overline{A}) = 1 - \mathsf{P}(A)$.

Доказательство. Поскольку $A \cup \overline{A} = \Omega$, и события A и $\overline{A}$ несовместны, из аксиомы **(P3)** и свойства 1 получим

$\mathsf{P}(A) + \mathsf{P}(\overline{A}) = \mathsf{P}(\Omega) = 1$.

Свойство 3. Если $A \subseteq B$, то $\mathsf{P}(B \setminus A) = \mathsf{P}(B) - \mathsf{P}(A)$.

Доказательство. Представим B в виде объединения двух несовместных событий: $B = A \cup (B \setminus A)$. По свойству 1, $\mathsf{P}(B) = \mathsf{P}(A) + \mathsf{P}(B \setminus A)$.

Сразу же заметим, что по аксиоме **(P1)** выражение в правой части равенства $\mathsf{P}(B) = \mathsf{P}(A) + \mathsf{P}(B \setminus A)$ больше либо равно $\mathsf{P}(A)$, что доказывает следующее свойство *монотонности* вероятности.

Свойство 4. Если $A \subseteq B$, то $\mathsf{P}(A) \leqslant \mathsf{P}(B)$.

Свойство 5. Для любого события A выполнено: $0 \leqslant \mathsf{P}(A) \leqslant 1$.

Доказательство. $\mathsf{P}(A) \geqslant 0$ по **(P1)**. А так как $A \subseteq \Omega$, то $\mathsf{P}(A) \leqslant \mathsf{P}(\Omega) = 1$.

Свойство 6. Всегда $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$.

Доказательство. Имеем $A \cap B \subseteq B$, поэтому $\mathsf{P}(B \setminus (A \cap B)) = \mathsf{P}(B) - \mathsf{P}(A \cap B)$ по свойству 3 Но $A \cup B = A \cup (B \setminus (A \cap B))$, причём A и $B \setminus (A \cap B)$ несовместны. Снова пользуясь свойством 1, получим:

$$\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B \setminus (A \cap B)) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B).$$

Из этого свойства и аксиомы **(P1)** следуют два полезных свойства. Свойство 8 читатель докажет с помощью свойства 7.

Свойство 7. Всегда $\mathsf{P}(A \cup B) \leqslant \mathsf{P}(A) + \mathsf{P}(B)$.

Свойство 8. Совершенно всегда

$$\mathsf{P}(A_1 \cup \ldots \cup A_n) \leqslant \sum_{i=1}^{n} \mathsf{P}(A_i)$$

.

Следующее свойство называют *формулой включения и исключения.* Она оказывается весьма полезной в случае, когда для вычисления вероятности некоторого события A нельзя разбить это событие на удобные попарно несовместные события, но удаётся разбить событие A на простые составляющие, которые, однако, совместны.

Свойство 9. Для любого конечного набора событий $A_1, \ldots, A_n$ имеет место равенство:

$$\mathsf{P}(A_1 \cup \ldots \cup A_n) = \sum_{i=1}^{n} \mathsf{P}(A_i) - \sum_{i<j} \mathsf{P}(A_i A_j) \; + \\ + \sum_{i<j<m} \mathsf{P}(A_i A_j A_m) - \ldots + (-1)^{n-1} \mathsf{P}(A_1 A_2 \ldots A_n).$$

Доказательство. Воспользуемся методом математической индукции. Базис индукции при $n = 2$ — свойство 6. Пусть свойство 9 верно при $n = k - 1$. Докажем, что тогда оно верно при $n = k$. По свойству 6,

$$\mathsf{P}\left(\bigcup_{i=1}^{k} A_i\right) = \mathsf{P}\left(\bigcup_{i=1}^{k-1} A_i\right) + \mathsf{P}\,(A_k) - \mathsf{P}\left(A_k \cap \bigcup_{i=1}^{k-1} A_i\right).$$

По предположению индукции, первое слагаемое в правой части (3) равно

$$\mathsf{P}\left(\bigcup_{i=1}^{k-1} A_i\right) = \sum_{i=1}^{k-1} \mathsf{P}(A_i) \; - \sum_{1 \leqslant i < j \leqslant k-1} \mathsf{P}(A_i A_j) \; + \\ + \sum_{1 \leqslant i < j < m \leqslant k-1} \mathsf{P}(A_i A_j A_m) - \ldots + (-1)^{k-2} \mathsf{P}(A_1 A_2 \ldots A_{k-1}).$$

Вычитаемое в правой части (3) равно

$$\mathsf{P}\left(A_k \cap \bigcup_{i=1}^{k-1} A_i\right) = \mathsf{P}\left(\bigcup_{i=1}^{k-1} A_i A_k\right) = \sum_{i=1}^{k-1} \mathsf{P}(A_i A_k) - \sum_{1 \leqslant i < j \leqslant k-1} \mathsf{P}(A_i A_j A_k) +$$

$$+ \sum_{1 \leqslant i < j < m \leqslant k-1} \mathsf{P}(A_i A_j A_m A_k) - \ldots + (-1)^{k-2} \mathsf{P}(A_1 A_2 \ldots A_{k-1} A_k).$$

Приведём пример задачи, в которой использование свойства 9 — самый простой путь решения. Это известная *«задача о рассеянной секретарше»*.

Пример. Есть n писем и n подписанных конвертов. Письма раскладываются в конверты наудачу по одному. Найти вероятность того, что хотя бы одно письмо попадет в предназначенный ему конверт, и предел этой вероятности при $n \to \infty$.

Решение. Пусть событие A_i, $i = 1, \ldots, n$, означает, что i-е письмо попало в свой конверт. Тогда

$$A = \{\textit{хотя бы одно письмо попало в свой конверт}\} = A_1 \cup \ldots \cup A_n.$$

Так как события $A_1, \ldots, A_n$ совместны, придётся использовать формулу (2). По классическому определению вероятности вычислим вероятности всех событий A_i и их пересечений. Элементарными исходами будут всевозможные перестановки (размещения) n писем по n конвертам. Их общее число есть $|\Omega| = n!$, и событию A_i благоприятны $(n-1)!$ из них, а именно любые перестановки всех писем, кроме i-го, лежащего в своём конверте. Поэтому $\mathsf{P}(A_i) = (n-1)!/n! = 1/n$ для всех i.

Совершенно так же получим, что при любых $i \neq j$

$$\mathsf{P}(A_i A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}.$$

Вероятность пересечения любых трёх событий равна

$$\mathsf{P}(A_i A_j A_m) = \frac{(n-3)!}{n!} = \frac{1}{n(n-1)(n-2)}.$$

Аналогично посчитаем вероятности пересечений любого другого числа событий, в том числе $\mathsf{P}(A_1 \dots A_n) = 1/n!$

Вычислим количество слагаемых в каждой сумме в формуле (2). Например, в сумме по $1 \leqslant i < j < m \leqslant n$ ровно C_n^3 слагаемых — ровно столько трёхэлементных множеств можно образовать из n элементов, и каждое такое множество $\{i, j, m\}$ встречается в индексах данной суммы единожды. Подставляя все вероятности в формулу (2), получим:

$$\mathsf{P}(A) = n\frac{1}{n} - C_n^2\frac{1}{n(n-1)} + C_n^3\frac{1}{n(n-1)(n-2)} - \dots + (-1)^{n-1}\frac{1}{n!} =$$
$$= 1 - \frac{1}{2!} + \frac{1}{3!} - \dots + (-1)^{n-1}\frac{1}{n!} \longrightarrow 1 - e^{-1} \quad \text{при} \quad n \to \infty$$

Упражнение 20. Выписать разложение e^{-1} в ряд Тейлора и убедиться в том, что $\mathsf{P}(A) \longrightarrow 1 - e^{-1}$ при $n \to \infty$.

1.4. Пространства с мерой

Чтобы обойти проблемы, связанные с парадоксом Банаха—Тарского, нужно отказаться от предположения, что *все* множества имеют объем. Множества, для которых определен объем, называются *измеримыми.* Кроме того, для многих приложений необходимо, чтобы объем был *счетно-аддитивным,* то есть складывался при объединении счетного числа частей.

Определение. *Пространство с мерой* — это множество X, в котором выделена некоторая система $\mathfrak{A} \subset 2^X$ его подмножеств (называемых *измеримыми)* и задана функция $\mu : \mathfrak{A} \to [0, +\infty]$, называемая *мерой.* При этом должны выполняться следующие условия.

Класс A измеримых множеств является σ-алгеброй, то есть:

1. $X \in \mathfrak{A}$ (все пространство является измеримым множеством).
2. Объединение, пересечение и разность любых двух множеств из A тоже принадлежит A.
3. Объединение любого счетного набора множеств из A снова принадлежит A.

Функция μ аддитивна и счетно-аддитивна, то есть:

4. Если $A, B \in \mathfrak{A}$ и $A \cap B = \emptyset$, то $\mu(A \cup B) = \mu(A) + \mu(B)$. Кроме того, $\mu(\emptyset) = 0$.
5. Пусть $\{A_i\}_{i=1}^{\infty}$ — счетный набор множеств. Тогда, если все A_i измеримы и попарно не пересекаются, то то $\mu(\bigcup A_i) = \sum \mu(A_i)$.

Замечания. 1. В свойстве 5 в правой части стоит сумма ряда. Поскольку все слагаемые неотрицательны, сумма не зависит от порядка слагаемых.

2. Требование $\mu(\varnothing)$ добавлено для того, чтобы исключить единственный пример, в которм мера любого множества равна $+\infty$

3. Поскольку дополнение измеримого множества измеримо, из свойства 3 следует, что пересечение счетного набора измеримых множеств тоже измеримо. Аналогично, в свойстве 2 достаточно ограничиться только одной из операций объединения и пересечения.

Тривиальные примеры. В этих примерах можно считать, что все множества измеримы.

1. *Считающая мера.* Мера множества равна количеству его элементов.
2. *δ-мера Дирака.* Зафиксируем точку $x_0 \in X$ и положим $\mu(A) = 1$, если $x_0 \in A$ и $\mu(A) = 0$, если $x_0 \notin A$. Эта мера μ обозначается через δ_{x_0}.
3. Положим меру любого счетного множества равной 0, а любого несчетного — равной $+\infty$.
4. Измеримое подмножество пространства с мерой само является пространством с мерой.

Простейшие свойства. Любая мера μ обладает следующими свойствами:

1. Монотонность: если множества A и B измеримы и $A \subset B$, то $\mu(A) \le \mu(B)$.

2. Субаддитивность: $\mu(A \cup B) \leq \mu(A) + \mu(B)$для любых измеримых множеств *A* и *B*.
3. Счетная субаддитивность: $\mu(\bigcup A_i) \leq \sum \mu(A_i)$для любого счетного набора $\{A_i\}$ измеримых множеств.
4.Пусть — $A_1 \subset A_2 \subset \ldots$ возрастающая последовательность измеримых множеств, $A = \bigcup A_i$. Тогда $\mu(A) = \lim \mu(A_i)$.
5. Пусть $A_1 \supset A_2 \supset \ldots$ — убывающая последовательность измеримых множеств,$A = \bigcap A_i$. Предположим, что $\mu(A_1) < +\infty$. Тогда $\mu(A) = \lim \mu(A_i)$.
Замечание: условие$\mu(A_1) < +\infty$существенно.
Задача. Пусть$\{A_i\}$— последовательность измеримых множеств в пространстве *X*, причем$\mu(X) < +\infty$. Пусть *A* — *верхний предел* этой последовательности, то есть множество всех точек, принадлежащих бесконечно многим из множеств A_i. Докажите, что *A* измеримо и $\mu(A) \geq \overline{\lim} \mu(A_i)$.

1.5 Борелевские множества

Далее основное множество *X* будет всегда предполагаться метрическим пространством. В приложениях можно считать, что$X = \mathbb{R}^n$. Напоминание: множество$A \subset X$называется *открытым,* если оно вместе с каждой точкой содержит некоторую ее окрестность, где под окрестностью понимается шар с центром в этой точке. Множество$A \subset X$ называется *замкнутым,* если его дополнение $X \setminus A$ открыто (это эквивалентно тому, что *A* содержит все свои предельные точки).
Определение. *Борелевская σ-алгебра* пространства *X* — это минимальная по включению σ-алгебра в *X*, содержащая все открытые множества. Множество $A \subset X$ называется *борелевским,* если оно принадлежит борелевской σ-алгебре. *Борелевская мера* на X — мера, определенная на борелевской σ-алгебре.
Чтобы доказать, что борелевская σ-алгебра существует, рассмотрим пересечение всех σ-алгебр, содержащих все открытые множества. Аналогично определяется σ-алгебра, порожденная произвольным множеством $\mathcal{P} \subset 2^X$.
Для построения борелевской σ-алгебры не обязательно начинать с множества всех открытых (или всех замкнутых) множеств. Например, борелевская σ-алгебра на прямой порождается лучами вида $[a, +\infty)$.

1.6. Мера Лебега

Теорема (Лебег). *Существует единственная борелевская мера* $\mu_n \in \mathbb{R}^n$*, инвариантная относительно параллельных переносов и такая, что мера стандартного единичного куба* $I^n = [0,1]^n$ *равна 1.* Мера из теоремы называется *n-мерной мерой Лебега* или *n-мерным евклидовым объемом.* Теорема Лебега будет доказана позднее.

Замечание. На самом деле меру Лебега определяют на большей σ-алгебре, чем борелевская.

Примеры. Предполагая доказанным существование одномерной меры Лебега, найдем меры некоторых множеств.

1. Мера точки равна 0, так как в единичный отрезок помещается сколь угодно много точек.
2. Мера отрезка [*a, b]* равна *b — a.* Для рационального *b — a* это доказывается разбиением на отрезки длины 1/N, где *N* — знаменатель, в общем случае — приближением рациональными длинами снизу и сверху.
3. Мера множества рациональных чисел (как и любого счетного множества) равна 0.
4. Мера множества иррациональных чисел из отрезка [0,1] равна 1.
5. Мера стандартного канторовского множества равна 0.

Задача. Постройте на отрезке замкнутое множество положительной меры, не содержащее интервалов.

Объемы некоторых множеств в $\mathbb{R}^n$**.**

1. Мера *(n* — 1)-мерного линейного подпространства (и любого его измеримого подмножества) равна 0.

2. Объем параллелепипеда с ребрами $a_1,\ldots, a_n$, параллельными осям координат, равен произведению $a_1 \ldots a_n$. Доказывается аналогично вычислению меры отрезка на прямой.

Пример неизмеримого множества

Предполагая существование меры Лебега, построим неизмеримое множество на отрезке [0,1). Отождествим отрезок с окружностью *S* радиуса 1/2π (длины 1) с помощью соответствия

$$t \mapsto \frac{1}{2\pi}(\cos 2\pi t, \sin 2\pi t).$$

Мере Лебега на отрезке соответствует мера μ на окружности, инвариантная относительно поворотов.

Пусть $\alpha = \pi\sqrt{2}$ (вместо $\sqrt{2}$ можно взять любое иррациональное число). Объявим точки на окружности эквивалентными, если они получаются друг из друга поворотом на угол, кратный α. Окружность

разбивается на классы эквивалентности, каждый класс — счетное множество.

Воспользовавшись аксиомой выбора, построим множество *A*, содержащее по одной точке из каждого класса. Для каждого $k \in Z$ обозначим через A_k образ множества *A* при повороте на угол $k\alpha$. Тогда множества A_k, $k \in \mathbb{Z}$, попарно не пересекаются и покрывают окружность. Следовательно, *A* неизмеримо: если $\mu(A) = 0$, то $\mu(A_k) = 0$ при всех *k*, откуда $\mu(S) = 0$ в силу счетной аддитивности, а если $\mu(A) > 0$, то, аналогично, $\mu(S) = \infty$, μ(S)=∞, противоречие.

Замечание. Без использования аксиомы выбора построить неизмеримое множество невозможно.

1.7. Мера Хаусдорфа

Пусть *X* — метрическое пространство. Расстояние между точками $x, y \in X$ будет обозначаться через |*xy*|. *Диаметром* непустого множества $A \subset X$ называется величина

$$\operatorname{diam}(A) = \sup\{|xy| : x, y \in A\},$$

диаметр пустого множества полагаем равным 0.

Покрытием множества *A* называется любой (конечный или бесконечный) набор множеств $\{A_i\}$ такой, что $A \subset \bigcup A_i$. *Мелкостью* набора множеств $\{A_i\}$ называется число $\Delta(\{A_i\}) = \sup_i \operatorname{diam} A_i$.

Определение. Пусть $d \geq 0$, $A \subset X$. Будем называть *d-мерным весом* конечного или набора множеств $\{A_i\}$ величину

$W_d(\{A_i\}) = \sum \operatorname{diam}(A_i)^d$.

Примечание: если $d = 0$ и $\operatorname{diam}(A_i) = 0$, считаем $\operatorname{diam}(A_i)^d = 1$.

Пусть ε > 0. Определим величину

$$\mathcal{H}^d_\varepsilon(A) = \inf\{W_d(\{A_i\}) : A \subset \bigcup A_i, \Delta(\{A_i\}) < \varepsilon\},$$

где инфимум берется по всем конечным и счетным покрытиям $\{A_i\}$ мелкости меньше ε.

Величина $\mathcal{H}^d_\varepsilon$ возрастает при убывании ε, поэтому она имеет предел при $\varepsilon \to 0$. Положим

$$\mathcal{H}^d(A) = C_d \cdot \lim_{\varepsilon \to 0} \mathcal{H}^d_\varepsilon(A),$$

где C_d — нормировочная константа, которая будет определена позже. Величина $\mathcal{H}^d(A)$ называется *d-мерной мерой Хаусдорфа* множества *A*.

При *d* = 0 и *d* = 1 полагаем нормировочную константу равной 1.

Замечание. В определении можно ограничиться открытыми покрытиями $\{A_i\}$ (то есть такими, в которых все множества A_i открыты). Действительно, произвольное покрытие $\{A_i\}$ можно заменить на открытое, сколь угодно мало изменив мелкость и вес.

Примеры. 1. $\mathcal{H}^0$ — считающая мера.

2. $\mathcal{H}^2(\mathbb{R}) = 0$.

3. $\mathcal{H}^1([0,1]) = 1$ (доказывается с использованием компактности).

Теорема. $0 < \mathcal{H}^n(I^n) < \infty$.

Доказательство. Для доказательства неравенства $\mathcal{H}^n < \infty$ достаточно предъявить сколь угодно мелкое покрытие с весом, ограниченным сверху некоторой константой. Возьмем, например, разбиение I^n на кубики с ребром $1/N$, $N \to \infty$.

Для доказательства неравенства $\mathcal{H}^n > 0$ нужно проверить, что вес любого покрытия отделен от нуля некоторой константой. Покрытие можно считать открытым, а значит конечным. Увеличив диаметры не более чем в 2n раз, можно заменить покрывающие множества на кубики с ребрами, параллельными координатным осям. Вес каждого кубика в константу раз отличается от его элементарного объема (элементарный объем прямоугольного параллелепипеда — произведение ребер).

Теперь утверждение следует из леммы:

Лемма. *Пусть $P, P_1, P_2, \ldots, P_N$ — параллелепипеды с ребрами, параллельными координатным осям. Предположим, что $P \subset \bigcup P_i$. Тогда $\sum V(P_i) \geq V(P)$, где V — элементарный объем.*

Лемма доказывается по индукции.

Теперь можно определить нормировочную константу C_d при целых *d:* это такое число, что *d*-мерная мера Хаусдорфа куба I^d получается равной 1.

Информация: $C_d = \dfrac{\pi^{d/2}}{2^d \cdot \Gamma(d/2+1)}$, где $\Gamma(x) = \int_0^{+\infty} x^{d-1} e^{-x}\, dx$. Эту формулу можно использовать для определения нормировочной константы и для нецелых *d.*

Свойства меры Хаусдорфа. 1. Монотонность: если $A \subset B$, то $\mathcal{H}^d(A) \leq \mathcal{H}^d(B)$.

2. Счетная субаддитивность: $\mathcal{H}^d(\bigcup A_i) \leq \sum \mathcal{H}^d(A_i)$ для любого конечного или счетного набора множеств $\{A_i\}$.

3. Пусть $A, B \subset X$, $\operatorname{dist}(A,B) > 0$, где $\operatorname{dist}(A,B) = \inf\{|xy| : x \in A, y \in B\}$. Тогда $\mathcal{H}^d(A \cup B) = \mathcal{H}^d(A) + \mathcal{H}^d(B)$.

4. Нерастягивающие отображения не увеличивают меру. Как следствие, меры изометричных множеств равны.
5. Гомотетия с коэффициентом k умножает меру на k^d.

Этих свойств достаточно для вычисления меры Хаусдорфа в большинстве случаев. Например, площадь сферы можно вычислить, разбив ее на маленькие части и сравнив каждую часть с ее проекцией на касательную плоскость.

1.8. Хаусдорфова размерность

Теорема. *Для любого множества* $A \subset X$ *существует такое* $d_0 \in [0, +\infty]$, *что* $\mathcal{H}^d(A) = 0$ *при всех* $d > d_0$ *и* $\mathcal{H}^d(A) = \infty$ *при всех* $d < d_0$.

Число d_0 называется *хаусдорфовой размерностью* множества A и обозначается $\dim_H(A)$.

Доказательство.

Положим $d_0 = \inf\{d : \mathcal{H}^d(A) < \infty\}$. Тогда $\mathcal{H}^d(A) = \infty$ при всех $d < d_0$. Докажем, что $\mathcal{H}^d(A) = 0$ при $d > d_0$. Выберем между d_0 и d такое d', что $H^{d'}(A) < \infty$. Пусть $d = d' + a$. Тогда $\mathcal{H}^d_\varepsilon \le \varepsilon^a \mathcal{H}^{d'}_\varepsilon(A)$ для любого $\varepsilon > 0$.

Поскольку $\mathcal{H}^{d'}(A) < \infty$ и $\varepsilon^a \to 0$ при $\varepsilon \to 0$, получаем $\mathcal{H}^d(A) = 0$.

Примеры. 1. $\dim_H(\mathbb{R}^n) = n$, так как $0 < \mathcal{H}^n(I^n) < \infty$.

2. Размерность стандартного канторовского множества K равна $\frac{\ln 2}{\ln 3}$. Это число можно угадать из следующих соображений:
пусть $d = \dim_H(K)$, $A = \mathcal{H}^d(K)$. Тогда $A = 2A(1/3)^d$, так как K состоит из двух копий, подобных ему с коэффициентом 1/3.
Предполагая, что $A \ne 0$ и $A \ne \infty$, получаем $3^d = 2$, откуда $d = \frac{\ln 2}{\ln 3}$.
Это не доказательство, так как нет гарантии, что $\mathcal{H}^d(K)$ не ноль и не бесконечность.

Задача. Постройте на прямой множество размерности 1 и меры 0.

Задача. Докажите, что $\dim(K) = \frac{\ln 2}{\ln 3}$.

Задача. Пусть множество A — объединение счетного набора множеств $A_1, A_2, \ldots$. Докажите, что $\dim_H(A) = \sup\{\dim_H(A_i)\}$.

1.9. Внешние меры и критерий Каратеодори

Определение. Пусть X — произвольное множество. *Внешняя мера* на X — это функция $\mu : 2^X \to [0, +\infty]$, обладающая следующими свойствами:

1. $\mu(\emptyset) = 0$.
2. Монотонность: если $A \subset B$, то $\mu(A) \le \mu(B)$.

3. Счетная субаддитивность: $\mu(\bigcup A_i) \le \sum \mu(A_i)$ для любого конечного или счетного набора множеств $\{A_i\}$.

Примеры. 1. Мера Хаусдорфа — см. свойства.

2. Пусть μ — мера на какой-нибудь σ-алгебре $\mathfrak{A} \subset 2^X$. Ее можно продолжить до внешней меры μ^*, определенной равенством

$$\mu^*(A) = \inf\{\mu(B) : A \subset B \in \mathfrak{A}\}.$$

3. Если $\mathfrak{A} \subset 2^X$ — произвольная система множеств и $\mu : \mathfrak{A} \to [0, +\infty]$, то можно определить внешнюю меру μ^* так:

$$\mu^*(A) = \inf\{\sum \mu(A_i); A \subset \bigcup A_i, A_i \in \mathfrak{A}\},$$

где инфимум берется по всем не более чем счетным покрытиям множества A множествами из $\mathfrak{A}$.

Определение. Пусть на X задана внешняя мера μ. Будем говорить, что множество $A \subset X$ *хорошо разбивает* множество $B \subset X$, если $\mu(B) = \mu(B \cap A) + \mu(B \setminus A)$. Множество $A \subset X$ называется *хорошо разбивающим* или *μ-измеримым,* если оно хорошо разбивает любое множество $B \subset X$.

Замечание. В равенстве $\mu(B) = \mu(B \cap A) + \mu(B \setminus A)$ содержательно только неравенство "$\ge$", обратное неравенство следует из определения внешней меры.

Теорема. *Пусть μ — внешняя мера на X. Тогда класс всех μ-измеримых множеств является σ-алгеброй, и сужение μ на эту σ-алгебру является мерой.*

Доказательство. 1. Очевидно, что пустое множество и все пространство X — хорошо разбивающие.

2. Если A — хорошо разбивающее, то и его дополнение $X \setminus A$ — хорошо разбивающее. Действительно, определение симметрично относительно замены A на $X \setminus A$, так как $B \setminus B = B \cap (X \setminus A)$.

3. Пусть A_1, A_2 — хорошо разбивающие, докажем, что $A_1 \cap A_2$ хорошо разбивающее. Пусть B — произвольное множество. Оно разбивается на четыре части: $B_0 = B \setminus A_1 \setminus A_2$, $B_1 = B \cap A_1 \setminus A_2$, $B_2 = B \cap A_2 \setminus A_1$ и $B_3 = B \cap A_1 \cap A_2$. Так как A_1 хорошо разбивает B, имеем

$$\mu(B) = \mu(B \cap A_1) + \mu(B \setminus A_1).$$

Так как A_2 хорошо разбивает $B \cap A_1$, имеем

$$\mu(B \cap A_1) = \mu(B \cap A_1 \cap A_2) + \mu(B \cap A_1 \setminus A_2).$$

Таким образом,

$$\mu(B) = \mu(B \cap A_1 \cap A_2) + \mu(B \cap A_1 \setminus A_2) + \mu(B \setminus A_1).$$

Так как A_1 хорошо разбивает множество $B \setminus (A_1 \cap A_2)$, с учетом теоретико-множественных тождеств

$(B \setminus (A_1 \cap A_2)) \cap A_1 = B \cap A_1 \setminus A_2$ и $(B \setminus (A_1 \cap A_2)) \cap A_1 = \mu(B \setminus A_1)$ получаем

$$\mu(B \setminus (A_1 \cap A_2)) = \mu(B \cap A_1 \setminus A_2) + \mu(B \setminus A_1).$$

Отсюда и из предыдущего равенства следует, что

$$\mu(B) = \mu(B \cap A_1 \cap A_2) + \mu(B \setminus (A_1 \cap A_2)),$$

что и требовалось.

4. Из пунктов 2 и 3 следует, что объединение и разность любых двух хорошо разбивающих множеств тоже хорошо разбивающее.

5. Докажем вспомогательное утверждение: если $A_1, A_2,...$ — дизъюнктные хорошо разбивающие множества и $B_i \subset A_i$ при всех i, то $\mu(\bigcup B_i) = \sum \mu(B_i)$.

Для каждого $n > 2$ имеем $\mu(B_1 \cup \cdots \cup B_n) = \mu(B_n) + \mu(B_1 \cup \cdots \cup B_{n-1})$, так как A_n хорошо разбивает $B_1 \cup \cdots \cup B_n$. Отсюда по индукции получаем, что

$$\mu(B_1 \cup \cdots \cup B_n) = \mu(B_1) + \cdots + \mu(B_n)$$

для всех n. Отсюда и из монотонности внешней меры следует, что

$$\mu(\bigcup B_i) \geq \mu(B_1) + \cdots + \mu(B_n)$$

при всех n. Переходя к пределу при $n \to \infty$, получаем, что $\mu(\bigcup B_i) \geq \sum \mu(B_i)$. Обратное неравенство следует из определения внешней меры.

6. Пусть $A_1, A_2,...$ — дизъюнктные хорошо разбивающие множества. Докажем, что множество $A = \bigcup A_i$ — хорошо разбивающее.

Пусть $B \subset X$, положим $B_i = B \cap A_i$. Так как каждое конечное объединение $A_1 \cup \cdots \cup A_n$ — хорошо разбивающее (по п. 4), имеем

$$\mu(B) = \mu(B \cap (A_1 \cup \cdots \cup A_n)) + \mu(B \setminus (A_1 \cup \cdots \cup A_n)) \geq \mu(B_1 \cup \cdots \cup B_n) + \mu(B \setminus A).$$

Перейдем к пределу при $n \to \infty$. Из п. 5 следует, что, $\mu(B_1 \cup \cdots \cup B_n) \to \mu(\bigcup B_i) = \mu(B \cap A)$. Значит, $\mu(B) \geq \mu(B \cap A) + \mu(B \setminus A)$, то есть A хорошо разбивает B.

7. Объединение любых хорошо разбивающих множеств $A_1, A_2,...$ — хорошо разбивающее. Действительно, $\bigcup A_i$ можно представить в виде объединения дизъюнктных множеств A_1, $A_2 \setminus A_1$, $A_3 \setminus (A_1 \cup A_2)$, ...

8. Счетная аддитивность μ на классе хорошо разбивающих множеств следует из п. 5 (подставим $B_i = A_i$).

Теорема (критерий Каратеодори). *Пусть X — метрическое пространство, μ — внешняя мера на X, обладающая таким свойством: для любых множеств $A, B \subset X$ с $\operatorname{dist}(A, B) > 0$ верно,*

что $\mu(A \cup B) = \mu(A) + \mu(B)$. *Тогда все борелевские множества μ-измеримы.*

Как следствие, сужение любой такой внешней меры (в частности, меры Хаусдорфа) на борелевскую σ-алгебру является борелевской мерой.

Доказательство. Достаточно доказать, что открытые множества — хорошо разбивающие (так как они порождают борелевскую σ-алгебру). Пусть $U \subset X$— открытое множество, $B \subset X$ — произвольное множество. Достаточно доказать, что $\mu(B \cap U) + \mu(B \setminus U) \leq \mu(B)$. Если $\mu(B) = \infty$, то это неравенство очевидно, поэтому будем считать, что $\mu(B) < \infty$.Для каждого натурального *n* определим множества

$$U_n = \{x \in U : \operatorname{dist}(x, X \setminus U_n) > \frac{1}{n}\}.$$

Докажем вспомогательное утверждение:

Лемма. $\mu(B \cap U_n) \to \mu(B \cap U)$ *при* $n \to \infty$.

Доказательство. Для каждого *n* положим $A_n = B \cap (U_{n+1} \setminus U_n)$. Рассмотрим все непустые множества вида A_{2k}.Каждые два из них разлелены некоторым положительным расстоянием. Применяя свойство из формулировки теоремы, получаем, что для любого *n*

$$\sum_{k=1}^{n} \mu(A_{2k}) = \mu(\bigcup_{k=1}^{n} A_{2k}) \leq \mu(B)$$

в силу монотонности внешней меры. Так как $\mu(B) < \infty$,отсюда следует, что ряд$\sum \mu(A_{2k})$сходится.

Аналогично ряд $\sum \mu(A_{2k+1})$сходится, значит, ряд $\sum \mu(A_k)$сходится. Обозначим $\varepsilon_n = \sum_{k=n}^{\infty} \mu(A_n)$. Поскольку ряд сходится, имеем $\varepsilon_n \to 0$. Заметим, что $B \cap U = (B_n \cap U) \cup \bigcup_{k=n}^{\infty} A_n$. По счетной субаддитивности отсюда следует, что

$$\mu(B \cap U) \leq \mu(B_n \cap U) + \sum_{k=n}^{\infty} \mu(A_n) = \mu(B_n \cap U) + \varepsilon_n.$$

Значит, $\mu(B \cap U_n) \to \mu(B \cap U)$.

Для каждого *n* рассмотрим множества$B \cap U_n$ и $B \setminus U$.Они разделены расстоянием *1/n,* поэтому

$$\mu(B \cap U_n) + \mu(B \setminus U) = \mu((B \cap U_n) \cup (B \setminus U)) \leq \mu(B).$$

Переходя к пределу с помощью леммы получаем требуемое неравенство $\mu(B \cap U) + \mu(B \setminus U) \leq \mu(B)$.

Из теоремы следует существование меры Лебега в $\mathbb{R}^n$(в качестве меры Лебега можно взять *n*-мерную меру Хаусдорфа).

1.10. Единственность меры Лебега

Для завершения доказательстве теоремы Лебега осталось проверить единственность борелевской меры μ, удовлетворяющей условиям теоремы (т.е. инвариантной относительно параллельных переносов и нормированной единицей на стандартном единичном кубе).
Назовем *кирпичом* в $\mathbb{R}^n$ множество вида $I_1 \times I_2 \times \ldots I_n$, где $I_1, \ldots, I_n$ — ограниченные интервалы на прямой (замкнутые, открытые или полуоткрытые).
Свойства. 1. Любое открытое множество можно представить в виде объединения счетного набора кирпичей. Например, можно взять объединение всех кирпичей с рациональными координатами вершин, содержащихся в данном множестве.
Следовательно, σ-алгебра, порожденная кирпичами, — это борелевская σ-алгебра в $\mathbb{R}^n$.
2. Если μ — борелевская мера в $\mathbb{R}^n$, инвариантная относительно параллельных переносов и нормированная на стандартном кубе в $\mathbb{R}^n$, то мера кирпича $I_1 \times \ldots I_n$ равна произведению длин интервалов $I_1, \ldots, I_n$. Это доказано ранее.
Теперь единственность меры Лебега следует из теоремы о единственности продолжения меры с полукольца,

1.10.1 Продолжение меры с полукольца

Определение. Система множеств называется *кольцом,* если она замкнута относительно бинарных операций объединения, пересечения и разности.
Система $\mathcal{P}$ множеств называется *полукольцом,* если для любых $A, B \in \mathcal{P}$ верно, что (1) $A \cap B \in \mathcal{P}$;
(2) $A \setminus B$ есть дизъюнктное объединение нескольких (конечного набора) множеств из $\mathcal{P}$.
Примеры полуколец. 1. Всевозможные ограниченные интервалы на прямой.
2. Интервалы вида $[a, b)$ на прямой.
3. Произведение полуколец — полукольцо. В частности, множество кирпичей в $\mathbb{R}^n$ — полукольцо.
Замечание. Если $\mathcal{P}$ — полукольцо, то множество всех конечных дизъюнктных объединений элементов $\mathcal{P}$ — кольцо.
Теорема (о единственности продолжения меры с полукольца). *Пусть X — произвольное множество,* $\mathcal{P} \subset 2^X$ *— полукольцо,* $\mathcal{A}$ *— порождаемая им σ-алгебра. Пусть* μ *и* μ' *— две меры, определенные на* $\mathcal{A}$ *и совпадающие на* $\mathcal{P}$*. Предположим, что X покрывается*

счетным набором множеств из $\mathcal{P}$, мера каждого из которых конечна. Тогда μ и μ' совпадают.

Доказательство. Достаточно доказать теорему в предположении, что все пространство входит в полукольцо и его мера конечна. Пусть $\mathcal{K}$— кольцо, порожденное полукольцом $\mathcal{P}$. Из описания этого кольца (см. выше) ясно, что μ и μ'совпадают на $\mathcal{K}$.

Рассмотрим $\mathfrak{B} = \{A \in \mathfrak{A} : \mu(A) = \mu'(A)$. Эта система множеств обладает следующими свойствами:

1. Она содержит кольцо $\mathcal{K}$.
2. Она является *монотонным классом,* то есть замкнута относительно объединений и пересечений вложенных последовательностей.

Теперь требуемое утверждение следует из следующей *леммы о монотонном классе:*

Лемма. *Если монотонный класс $\mathfrak{B}$ содержит кольцо $\mathcal{K} \ni X$, то он содержит и порождаемую этим кольцом σ-алгебру $\mathcal{A}$.*

Доказательство. Можно считать, что $\mathfrak{B}$— минимальный монотонный класс, содержащий $\mathcal{K}$. Докажем, что тогда $\mathfrak{B}$ является σ-алгеброй. Достаточно проверить, что для любых $A, B \in \mathfrak{B}$; множества $A \cap B$, $A \cup B$ и $X \setminus A$ принадлежат $\mathfrak{B}$.

Докажем, что $X \setminus A \in \mathfrak{B}$ для любого $A \in \mathfrak{B}$. Рассмотрим множество $\mathfrak{B}' = \{A \in \mathfrak{B} : X \setminus A \in \mathfrak{B}\}$.

Оно является монотонным классом и содержит $\mathcal{K}$. Отсюда и из минимальности $\mathfrak{B}$ следует, что $\mathfrak{B}' = \mathfrak{B}$. Утверждение доказано.

Докажем, что $A \cap B \in \mathfrak{B}$ для любых $A \in \mathcal{K}$ и $B \in \mathfrak{B}$.

Зафиксируем $A \in \mathcal{K}$ и рассмотрим множество $\mathfrak{B}_A = \{B \in \mathfrak{B} : A \cap B \in \mathfrak{B}\}$. Легко видеть, что $\mathfrak{B}_A$— монотонный класс. При этом $\mathcal{K} \subset \mathfrak{B}_A$, так как $\mathcal{K}$ — кольцо и $B \in \mathcal{K}$.Отсюда и из минимальности $\mathfrak{B}$ следует, что $\mathfrak{B}_A = \mathfrak{B}$,то есть $A \cup B \in \mathfrak{B}$ для любого $B \in \mathfrak{B}$.

Теперь докажем, что $A \cap B \in \mathfrak{B}$ для любых $A, B \in \mathfrak{B}$. Зафиксируем $A \in \mathfrak{B}$ и рассмотрим множество $\mathfrak{B}_A = \{B \in \mathfrak{B} : A \cap B \in \mathfrak{B}\}$.

Аналогично предыдущему рассуждению, $\mathfrak{B}_A$ — монотонный класс. По доказанному выше, $\mathcal{K} \subset \mathfrak{B}_A$. Отсюда и из минимальности $\mathfrak{B}$ следует, что $\mathfrak{B}_A = \mathfrak{B}$, то есть $A \cup B \in \mathfrak{B}$ для любого $B \in \mathfrak{B}$.

Для объединения доказательство аналогично.

Таким образом, $\mathfrak{B}$ содержит σ-алгебру, порожденную кольцом $\mathcal{K}$,что и требовалось.

1.10.2 Мера Лебега и линейные преобразования

Следующую теорему трудно доказать прямым рассуждением, но она легко следует из единственности меры Лебега.

Теорема. *Пусть μ — мера Лебега в $\mathbb{R}^n$, $L : \mathbb{R}^n \to \mathbb{R}^n$ — невырожденное линейное отображение.*
Тогда для любого борелевского множества $A \subset \mathbb{R}^n$ верно, что $\mu(L(A)) = |\det L| \cdot \mu(A)$.

1.11. Борелевская регулярность

Пусть μ — внешняя мера. Напоминание: множество называется измеримым относительно μ (μ-измеримым), если оно хорошо разбивающее для μ.
Определение. Внешняя мера μ на X называется *борелевски регулярной,* если все открытые множества μ-измеримы и для любого множества $A \subset X$ существует борелевское множество B такое, что $A \subset B$ и $\mu(A) = \mu(B)$.
Примеры. 1. Мера Хаусдорфа. Действительно, рассмотрим последовательность открытых покрытий, реализующих меру Хаусдорфа множества A. Для каждого покрытия рассмотрим объединение его членов. Пересечение этих объединений — искомое борелевское множество.
2. Если μ — борелевская мера, то ассоциированная с ней внешняя мера μ^* борелевски регулярна. Доказательство аналогично.
Последний пример показывает, что борелевские меры находятся во взаимно-однозначном соответствии с борелевски регулярными внешними мерами. А именно, борелевски регулярной внешней мере соответствует мера, получаемая сужением на борелевскую σ-алгебру, а борелевской мере соответствует порождаемая ей внешняя мера.
Теперь теорему Лебега можно сформулировать так: существует единственная борелевски регулярная внешняя мера на $\mathbb{R}^n$, инвариантная относительно параллельных переносов и нормированная на I^n. Теорема о поведении меры при линейных преобразованиях верна и для внешней меры, она доказывается точно также с помощью единственности.
Теорема. *Пусть μ — борелевски регулярная внешняя мера, $A \subset X$ — μ-измеримое множество.*
Тогда:
1. Если $\mu(A) < \infty$, то найдутся борелевские множества B, C такие, что $C \subset A \subset B$ и $\mu(B \setminus C) = 0$.
2. Если $\mu(X) < \infty$, то для любого $\varepsilon > 0$ найдется такое замкнутое множество $F \subset X$, что $\mu(A \setminus F) < \varepsilon$.
3. Если X содержится в открытом множестве конечной меры, то для любого $\varepsilon > 0$ найдется такое открытое $G \supset A$, что $\mu(G \setminus A) < \varepsilon$.

Доказательство. 1. Возьмем *B* из определения борелевской регулярности, *A* построим, применив борелевскую регулярность к $B \setminus A$.

2. Из пункта 1, множество *A* можно считать борелевским. Можно считать, что мера всего пространства конечна (иначе рассмотрим новую борелевкую меру $\mu'(B) = \mu(B \cap A)$вместо μ). Рассмотрим класс $\mathcal{A}$ всех борелевских множеств *A*, обладающих требуемым свойством. Легко проверить, что этот класс замкнут относительно счетных объединений и пересечений. Кроме того, он содержит все замкнутые множества.

Отсюда следует, что $\mathcal{A}$ совпадает с борелевской σ-алгеброй. Действительно, он содержит все открытые множества, так как любое открытое множество можно представить в виде счетного объединения замкнутых. Значит, он содержит кольцо, состоящее из конечных объединений и пересечений открытых и закнутых множеств. По лемме о монотонном классе, $\mathcal{A}$ содержит всю борелевскую σ-алгебру.

3. Пусть $U \supset A$ — открытое множество конечной меры. Применяя утверждение 2 к множеству $U \setminus A$, получаем такое $F \subset U \setminus A$, что $\mu(U \setminus A \setminus F) < \varepsilon$. Множество $G = U \setminus F$ подходит. □

Замечания. 1. Предположения можно ослабить: в первом и втором достаточно предположить, что *A* допускает локально конечное покрытие счетным набором множеств конечной меры; во третьем — что *A* содержится в объединении счетного набора открытых множеств конечной меры. Оба свойства выполняются для любого множества, если μ — мера Лебега.

2. В $\mathbb{R}^n$ во втором утверждении слово "замкнутое" можно заменить на "компактное".

3. Теперь можно дать более конструктивное описание системы множеств, измеримых по Лебегу. Из теоремы следует, что любое множество, измеримое по Лебегу, есть объединение счетного набора замкнутых множеств и множества внешней меры ноль. Обратно, любое такое множество измеримо. Для объединений замкнутых множеств это следует из борелевской регулярности, для множеств внешней меры ноль — из определения хорошо разбивающего множества.

1.12. Теоремы о покрытиях

Теорема (Безикович). *Пусть* μ *— борелевски регулярная мера в* $\mathbb{R}^n$, $A \subset \mathbb{R}^n$, $\mu(A) < \infty$. *Пусть*

$\mathfrak{B}$ *— множество замкнутых шаров в* $\mathbb{R}^n$*, такое, что для любых* $x \in A$ *и* $\varepsilon > 0$ *существует шар с центром в x и радиусом меньше ε, принадлежащий* $\mathfrak{B}$*. Тогда можно из* $\mathfrak{B}$ *выбрать не более чем счетный дизъюнктный набор шаров* $\{B_i\}$*, такой, что* $\mu(A \setminus \bigcup B_i) = 0$.

Теорема выводится из чисто геометрического факта (который тоже называется теоремой Безиковича).

Теорема. *Для любого натурального n существует такое натуральное M = M(n), что верно следующее. Пусть* $A \subset \mathbb{R}^n$ *— произвольное множество, и каждой точке* $x \in A$ *сопоставлен замкнутый шар* B_x *с центром в этой точке радиуса не больше 1. Тогда можно выбрать не более чем счетный набор шаров* $\{B_{xi}\}$*, покрывающий A, и раскрасить их в M цветов так, что шары одного цвета попарно не пересекаются.*

Вывод первой теоремы из второй. Выкинем все шары с радиусами, большими 1. Из оставшихся выберем набор $\{B_i\}$ как во второй теореме. Выберем такой швет, что шары этого цвета покрывают не менее *1/M* от меры *A*. Из них выберем конечный поднабор, покрывающий не меньше 1/2M от меры *A*. Обозначим это множество шаров через $\{B_1, \ldots, B_{N_1}\}$, а их объединение через D_1. По построению имеем $\mu(A \setminus D_1) \le (1 - \frac{1}{2M})\mu(A)$.

Теперь выкинем из $\mathfrak{B}$ все шары, пересекающиеся с множеством D_1. Из того, что D_1 замкнуто, следует, что оставшиеся шары удовлетворяют условиям первой теоремы для множества $A \setminus D_1$. Аналогичным построением выберем из них конечную систему шаров $B_{N_1+1}, \ldots, B_{N_2}$, покрывающую не менее 1/2M от меры множества $A \setminus D_1$. Теперь $\mu(A \setminus D_2) \le (1 - \frac{1}{2M})^2 \mu(A)$, где $D_2 = \bigcup_{i=1}^{N_2} B_i$. Аналогично по индукции строим последовательность попарно непересекающихся шаров $\{B_i\}$ и чисел N_k так, что $\mu(A \setminus A_k) \le (1 - \frac{1}{2M})^k \mu(A)$, где $D_k = \bigcup_{i=1}^{N_k} B_i$. Поскольку $(1 - \frac{1}{2M}^k) \to 0$ при $k \to \infty$, эта последовательность шаров — искомая.

Доказательство второй теоремы. 1. Сведем теорему к случаю, когда *A* ограничено. Разобьем $\mathbb{R}^n$ на кубики с ребром 10. Назовем соседними кубики, замыкания которых имеют общую точку. Пометим каждый кубик числом от 1 до 2^n так, чтобы соседние кубики были помечены разными числами (это легко делается с помощью индукции по размерности). Считая, что для ограниченных множеств теорема доказана, применим ее к пересечению *A* с каждым из кубиков. При этом умножим количество цветов на 2^n, разобьем цвета на 2^n групп, одинаковых количеству цветов, и в каждом кубике будем использовать для раскраски шаров группу цветов, соответствующую числу от 1 до 2^n, которым помечен этот кубик. Поскольку радиусы всех шаров

меньше 1, а ребра кубиков равны 10, такая раскраска будет удовлетворять условию.

2. Далее A предполагается ограниченным. Будем называть шар *почти самым большим* в некотором наборе, если его радиус на меньше $\frac{9}{10}$ супремума радиусов шаров набора. Выберем из $\mathfrak{B}$ почти самый большой шар B_1, выкинем все шары, центры которых лежат в B_1, из оставшихся выберем почти самый большой шар B_2 и так далее. Обозначим через r_i радиус шара B_i, через c_i — его центр. По построению, при $i > j$, имеем
$r_i \leq \frac{10}{9} r_j$ и $c_i \notin B_j$.

3. Радиусы r_i стремятся к нулю. Действительно, в противном случае они ограничены снизу числом $r > 0$. Тогда, поскольку все центры c_i лежат в ограниченной области, среди них найдутся два, c_j и c_j с $|c_i - c_i| < r/2$. Но тогда $c_i \in B_j$ и $c_j \in B_i$, противоречие.

4. Выбранные шары $B_1, B_2, \ldots$ покрывают A. Действительно, пусть точка $x \in A$ не покрыта, тогда шар B_x никогда не был выкинут. Но тогда он должен был быть выбран до того, как радиусы выбранных шаров стали меньше $\frac{9}{10}$ его радиуса.

5. Каждый шар B_k пересекает меньше M из шаров $B_1, \ldots, B_{k-1}$, где M — некоторая константа, зависящая только от размерности.
Действительно, пусть число M_0 таково, что среди любых M_0 ненулевых векторов в $\mathbb{R}^n$ найдутся два, образующие угол меньше 1° (такое M_0 существует в силу компактности сферы). Докажем, что $M = 2M_0$ подходит.
Обозначим $d_i = |c_i - c_k|$ — расстояние между центрами B_k и B_i. Назовем плохими шары из множества $\{B_1, \ldots, B_{k-1}\}$, пересекающиеся с B_k. Для плохого шара B_i имеем $r_i \geq d_i - r_k$. Предположим, что существует M плохих шаров. Тогда среди них найдутся либо M_0 таких, для которых соответствующее расстояние d_i меньше $\frac{3}{2} r_k$, либо M_0 таких, для которых соответствующее расстояние d_i не меньше $\frac{3}{2} r_k$. Среди этих шаров найдутся два, B_i и B_j ($i > j$), для которых угол между векторами $c_i - c_k$ и $c_j - c_k$ меньше 1°. В обоих случаях, пользуясь неравенствами $r_i \geq \frac{9}{10} r_k$, $r_j \geq \frac{9}{10} r_i$, $r_j \geq d_j - r_k$ и $r_i \geq d_i - r_k$, нетрудно проверить, что $|c_i - c_j| < r_j$, то есть $c_i \in B_j$, противоречие.

6. Пользуясь предыдущим утверждением, по индукции раскрасим шары в M цветов. Каждый очередной шар B_k красим в такой цвет,

который не использовался для шаров из $B_1, \ldots, B_{k-1}$, пересекающихся с B_k. Такой цвет найдется по утверждению предыдущего шага.

Замечания. 1. Не обязательно покрывать шарами. Достаточно, чтобы каждая точка была покрыта сколь угодно маленькими множествами *B*, звездными относительно шара радиуса $C\,\mathrm{diam}(B)$. Доказательство аналогично, но требует большего количества технических подробностей в шаге 5.

2. Теорема Витали: пусть *X* — произвольное метрическое пространство, μ — борелевски регулярная внешняя мера, удовлетворяющая услвию удвоения: существует такое $C > 0$, что для любых $x \in X$ и $r > 0$ верно, что $\mu(B(x,2r)) \le C\mu(B(x,r))$. Пусть $A \subset X$ содержится в открытом множестве конечной меры. Тогда из любого покрытия *A* замкнутыми шарами можно выбрать дизъюнктный набор шаров. покрывающих *A* с точностью до меры 0.

1.12.1. Нормировочная константа меры Хаусдорфа

Пусть ω_n — объем единичного шара в $\mathbb{R}^n$.

Теорема. *Нормировочная константа n-мерной меры Хаусдорфа равна* $\omega_n/2^n$.

1.12.2. Точки плотности

Пусть μ — локально конечная борелевски регулярная внешняя мера в $\mathbb{R}^n$.

Определение. Пусть $A \subset \mathbb{R}^n$, $x \in \mathbb{R}^n$. *Плотностью* множества *A* в точке *x* называется предел

$$\rho(A,x) = \lim_{r\to 0} \frac{\mu(A \cap B(r,x))}{\mu(B(r,x))}.$$

Теорема. *Если* $A \subset \mathbb{R}^n$ *-измеримо, то* $\rho(A,x)$ *определено и равно 1 для всех* $x \in A$*, кроме множества меры 0.*

Доказательство. От противного, пусть есть положительная мера точек $x \in A$, для которых

$$\underline{\rho}(A,x) := \liminf_{r\to 0} \frac{\mu(A \cap B(r,x))}{\mu(B(r,x))} < 1.$$

Для $\delta > 0$ обозначим

$A_\delta = \{x \in A : \underline{\rho}(A,x) < 1 - \delta\}$.

Найдется такое $\delta_0 > 0$, что $\mu(A_{\delta_0}) = m_0 > 0$. Обозначим $A' = A_{\delta_0}$. Найдем открытое $U \supset A$, такое, что $\mu(U \setminus A) < \delta_0 m_0$. Сопоставим каждой точке $x \in A'$ такую последовательность шаров $B(x, r_i(x))$,

i = 1, 2,..., что все шары соержатся в U, $r_i(x) \to 0$ и $\frac{\mu(B(r_i(x),x))\cap A}{\mu(B(r,x))} < 1 - \delta$ для всех i. Выберем из этого множества шаров дизъюнктный набор $B_j = B(x_j, r_j)$ как в теореме Безиковича. Для каждого из них имеем $\mu(B_j \setminus A) \geq \delta_0 \mu(B_j)$. Складывая, получаем

$$\mu(U \setminus A) \geq \sum \mu(B_j \setminus A) \geq \delta_0 \sum \mu(B_j) \geq \delta_0 \mu(A') = \delta_0 m_0,$$

противоречие.

Определение. Характеристической функцией множества $A \subset X$ называется функция $\chi_A : X \to \mathbb{R}$, определенная равенством

$$\chi_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

Следствие. *Пусть $A \subset \mathbb{R}^n$ μ-измеримое множество. Тогда $\rho(A, x) = \chi_A(x)$ для почти всех $x \in X$.*

1.13. Измеримые функции и интеграл Лебега

Пусть X — множество, $\mathcal{A}$ — σ-алгебра на X. Элементы этой σ-алгебры будем называть *измеримыми множествами.* Если σ-алгебра $\mathcal{A}$ не указана явно, то подразумевается область определения рассматривоемой в данный момент меры.

Определение. Функция $f : D \subset X \to \mathbb{R}^k$ называется *измеримой* (относительно $\mathcal{A}$), если для любого открытого $U \subset \mathbb{R}^k$ прообраз $f^{-1}(U)$ принадлежит $\mathcal{A}$ (в частности, $D \in \mathfrak{A}$).

Функция называется *борелевской,* если она измерима относительно борелевской σ-алгебры на X.

Свойства. 1. Функция измерима относительно $\mathcal{A}$ тогда и только тогда, когда прообраз любого борелевского множества измерим.

2. Функция $f : D \subset X \to \mathbb{R}$ измерима тогда и только тогда, когда для любого $a \in \mathbb{R}$ множество $\{x \in D : f(x) > a\}$ измеримо. Это следует из того, что лучи вида $(a, +\infty)$ порождают всю борелевскую σ-алгебру.

3. $f = (f_1, \ldots, f_k) : D \to \mathbb{R}^k$ измерима тогда и только тогда, когда каждая функция f_i измерима.

4. Если $A \in \mathfrak{A}$, то χ_A измерима.

5. Любая непрерывная функция измерима, если $\mathcal{A}$ содержит борелевскую σ-алгебру.

6. Если $f : D \to \mathbb{R}^k$ — измерима и $g : \mathbb{R}^k \to \mathbb{R}^m$ — борелевская, то $g \circ f$ изметрима.
7. Как следствие, сумма и произведение измеримых функций измеримы.
8. Пусть — $\{f_i\}$ последовательность измеримых функций. Тогда и $\sup f_i$, $\inf f_i$, $\underline{\lim} f_i$, $\overline{\lim} f_i$ $\lim f_i$ измеримы.

Определение. *Простая функция* — это измеримая функция с конечным множеством значений. *Ступенчатая функция* — измеримая функция с конечным или счетным множеством значений.

Замечание. Любая измеримая функция — равномерный предел ступенчатых. Любая ограниченная снизу измеримая функция $f : X \to \mathbb{R}$ — предел возрастающей последовательности простых функция, причем предел равномерный, если *f* ограничена.

1.13.1 Интеграл Лебега

Пусть (X, μ) — пространство с мерой. Интеграл Лебега — это функционал, ставящий в соответствие некоторым измеримым функциям $f : X \to \mathbb{R}$ величину из $[-\infty, +\infty]$, обозначаемую $\int_X f\, d\mu$, $\int_X f(x)\, d\mu(x)$, или просто $\int f$. При этом выполняются следующие свойства (определяющие интеграл однозначно).

1. У любой неотрицательной функции интеграл определен и неотрицателен.
2. Обозначим $f_+ = \max\{f, 0\}$, $f_- = \max\{-f, 0\}$ (тогда $f = f_+ - f_-$). Интеграл $\int$ определен тогда и только тогда, когда хотя бы один из интегралов $\int f_+$ и $\int f_-$ конечен. При этом $\int f = \int f_+ - \int f_-$.
Функция *f* называется *сумируемой,* если ее интеграл конечен.
3. Если *A* — измеримое множество, то $\int \chi_A = \mu(A)$.
4. Для любых функций f и g и любого $a \in \mathbb{R}$ верно, что $\int (af + g) = a \int f + \int g$, если правая часть определена.

5 (теорема Леви). Если *{fi}* — неубывающая последовательность измеримых функций, $f_1 \geq 0$, $f = \lim f_i$, то $\int f = \lim \int f_i$.

Замечание. Можно рассматривать функции, у которых некоторые значения равны $\pm\infty$. Если мера множества таких точек равна 0, то они не влияют на интеграл. Если $f \geq 0$ и $f = +\infty$ на множестве положительной меры, то $\int f = +\infty$. В общем случае интегрируемость определяется так же, как и для функций с конечными значениями.

1.13.2. Аппроксимативная непрерывность

Определение. Функция $f : D \subset \mathbb{R}^n \to \mathbb{R}$ называется *аппроксимативно непрерывной* (относительно меры μ) в точке $x \in D$, если для любого $\varepsilon > 0$ множество $\mathbb{R}^n \setminus \{y \in D : |f(y) - f(x)| < \varepsilon$ имеет нулевую плотность (относительно μ) в точке x.

В частности, x должна быть точкой плотности для D.

Замечание. $f : D \to \mathbb{R}$ аппроксимативно непрерывна в точке x тогда и только тогда, когда существует множество $A \subset D$ такое, что $x \in A$, $\mathbb{R}^n \setminus A$ имеет нулевую плотность в точке x и $f|_A$ непрерывна. (Упражнение).

Теорема. *Любая измеримая функция $f : \mathbb{R}^n \to \mathbb{R}$ аппроксимативно непрерывна почти всюду (относительно локально конечной борелевчки регулярной меры).*

Доказательство. Для каждого $q \in \mathbb{Q}$ рассмотрим множество $A_q = \{y \in D : f(y) > q\}$. Для почти всех $x \in \mathbb{R}^n$ верно, что плотность этого множества равна 1, если $x \in A_q$, и 0, если $x \notin A_q$. Поскольку $\mathbb{Q}$ счетно, для почти всех $x \in \mathbb{R}^n$ это свойство выполняется отновременно для всех $q \in \mathbb{Q}$. Любая такая точка нам подходит.

Следствие. *Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — ограниченная μ-измеримая функция, где μ — локально конечная борелевски регулярная внешняя мера. Тогда для почти любой точки $x \in \mathbb{R}^n$ верно, что*

$$\lim_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r))} |f - f(x)| \, d\mu = 0.$$

в частности,

$$f(x) = \lim_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r))} f \, d\mu.$$

Определение. Точки x, удовлетворяющие первому равенству, называются *точками Лебега* данной функции.

Замечание. Утверждение верно не только для ограниченных функций, но и для любых локально суммируемых.

1.14. Длина кривой

Определение. Пусть X — метрическое пространство. *Кривая в X* — это непрерывное отображение $\gamma : [a, b] \to X$. *Пунктир* кривой $\gamma : [a, b] \to X$ — конечная последовательность точек $\gamma(t_0), \ldots, \gamma(t_n)$, где $\{t_i\}$ — разбиение отрезка $[a, b]$, *длина пунктира* — это сумма $\sum |\gamma(t_i)\gamma(t_{i+1})|$. *Длина кривой* — это супремум длин ее пунктиров, обозначение: $L(\gamma)$.

Кривая конечной длины называется *спрямляемой.*

Свойства. 1. $L(\gamma) = L(\gamma|_{[a,c]}) + L(\gamma|_{[c,b]})$ для любой кривой $\gamma : [a,b] \to X$ и любого $c \in [a,b]$.

2. Если мелкость разбиения $\{t_i\}$ стремится к нулю, то $\sum |\gamma(t_i)\gamma(t_{i+1})|$ стремится к $L(\gamma)$.

3. Длина не меняется при замене параметра, у любой кривой конечной длины есть натуральная параметризация, то есть такая параметризация $\gamma : [0, L] \to X$, что $L(\gamma|_{[a,b]}) = b - a$ для любого отрезка $[a,b] \subset [0,L]$.

Теорема. *Пусть* $\gamma : [a,b] \to X$ *— кривая в метрическом пространстве X. Тогда*

$$L(\gamma) = \int_X \#\{\gamma^{-1}(x)\}\, d\mathcal{H}^1(x).$$

В частности, если γ *не имеет самопересечений, то* $L(\gamma) = \mathcal{H}^1(\gamma([a,b]))$.

Доказательство.

Лемма. $\mathcal{H}^1(\gamma([a,b])) \le L(\gamma)$.

Лемма. $\mathcal{H}^1(\gamma([a,b])) \ge \operatorname{diam}(\gamma([a,b]))$.

Определение. Скоростью кривой γ в момент t называется число

$$s_\gamma(t) = \lim_{t' \to t} \frac{|\gamma(t)\gamma(t')|}{|t - t'|}.$$

Теорема. *Пусть* $\gamma : [a,b] \to X$ *— липшицева кривая. Тогда скорость* s_γ *определена почти всюду и* $L(\gamma) = \int_{[a,b]} s_\gamma \, d\mu$, *где* μ *— одномерная мера Лебега.*

Доказательство. Определим верхнюю скорость $\overline{s}_\gamma$ и нижнюю скорость $\underline{s}_\gamma$, заменив в определении предел на верхний и нижний предел соответственно. Обе функции всюду определены и измеримы, так как их можно представить в виде верхнего или нижнего предела счетного набора функций. Например,

$$\overline{s}_\gamma(t) = \lim_{\varepsilon \in \mathbb{Q}, \varepsilon \to 0} \frac{|\gamma(t)\gamma(t+\varepsilon)|}{|\varepsilon|}.$$

Докажем, что

$$\int_{[a,b]} \overline{s}_\gamma = \int_{[a,b]} \underline{s}_\gamma = L(\gamma).$$

Доказательство проведем для верхней скорости, для нижней оно полностью аналогично.

Снабдим отрезок [*a, b]* одномерной мерой Лебега μ. Пусть $A \subset [a,b]$— множество всех точек, где $\overline{s}_\gamma$ аппроксимативно непрерывно. Пусть C

— константа Липшица для γ. Зафиксируем $\varepsilon > 0$, и пусть $\delta > 0$, что для любого разбиения отрезка $[a, b]$ мелкости меньше δ длина соответствующего пунктира кривой γ больше $L(\gamma) - \varepsilon$. Рассмотрим всевозможные отрезки $[t, t'] \subset [a, b]$ (допускается $t' < t$), обладающие следующими свойствами:

1. $\left| \frac{|\gamma(t)\gamma(t')|}{|t-t'|} - \overline{s}_\gamma(t) \right| < \varepsilon$.
2. $\mu\{x \in [t, t'] : |\overline{s}_\gamma(x) - \overline{s}_\gamma(t)| > \varepsilon\} < \varepsilon|t - t'|$.
3. $|t - t'| < \delta$.

Эти отрезки образуют покрытие множества A как в теореме Витали. Действительно, каждая точка $t \in A$ является концом сколь угодно короткого отрезка, удовлетворяющего первому свойству — это следует из определения $\overline{s}_\gamma$. При этом все достаточно короткие отрезки, содержащие t, удовлетворяют второму свойству, так как t — точка аппроксимативной непрерывности. Для каждого такого отрезка имеем

$$|\gamma(t)\gamma(t')| = \overline{s}_\gamma(t)|t - t'| \pm \varepsilon|t - t'| == \int_{[t,t']} \overline{s}_\gamma \, d\mu \pm (C + 2)\varepsilon|t - t'|.$$

С помощью теоремы Витали выберем из этих отрезков дизъюнктный набор, покрывающий почти все A, а значит, и почти весь отрезок $[a, b]$. Выберем конечный поднабор $\{[t_i, t_i']\}$ с суммой мер больше $|a - b| - \varepsilon$.Тогда

$$\int_{[a,b]} \overline{s}_\gamma \, d\mu = \int_{\bigcup [t_i, t_i']} \overline{s}_\gamma \, d\mu \pm C\varepsilon = \sum |\gamma(t)\gamma(t')| \pm (C + 2)\varepsilon|a - b| + C\varepsilon.$$

Дополним множество точек $\{t_i, t_i'\}$ до разбиения отрезка $[a, b]$ мелкости меньше δ (при этом между t_i и t'_i новых точек не вставляем). Длина кривой отличается от длины этого пунктира меньше, чем на ε, а длина пунктира отличается от суммы $\sum |\gamma(t)\gamma(t')|$ меньше чем на $C\varepsilon$.

1.15. Липшицевы функции

Теорема (Радемахер). *Любая липшицева функция* $f : \mathbb{R}^n \to \mathbb{R}^m$ *дифференцируема почти всюду.*

Доказательство. Можно считать, что $m = 1$. Будем использовать индукцию по n.

При $n = 1$ докажем более сильное утверждение: f дифференцируема почти всюду и для любых $a, b \in \mathbb{R}$, $a < b$, верно, что $f(b) - f(a) = \int_{[a,b]} f$. Для монотонной функции это следует из представления длины кривой как интеграла скорости, в общем случае представим функцию в виде суммы монотонной и линейной.

Переход: от *n* к *n* +1. Представим $\mathbb{R}^{n+1}$ как прямое произведение $\mathbb{R}^n \times \mathbb{R}$. Координаты в $\mathbb{R}^{n+1}$ будем обозначать через (*x, y*), где $x \in \mathbb{R}^n$, $y \in \mathbb{R}$. Для каждого $y \in \mathbb{R}$ рассмотрим функцию $f_y : \mathbb{R}^n \to \mathbb{R}$, определяемую равенством $f_y(x) = f(x, y)$. По индукционному предположению, каждая такая функция дифференцируема почти всюду, поэтому по теореме Фубини для почти всех $(x, y) \in \mathbb{R}^n \times \mathbb{R}$ функция f_y дифференцируема в точке *x*. Аналогично, частная производная $\partial f/\partial y$ определена почти всюду. Нетрудно проверить, что $\partial f/\partial y$ — измеримая функция, следовательно, она аппроксимативно непрерывна почти всюду.

Достаточно доказать, что *f* дифференцируема в любой точке (*x, y)* такой, что f_y дифференцируема в *x* и $\partial f/\partial y$ аппроксимативно непрерывна в точке (*x, y).* Зафиксируем такую точку (*x, y).* Пусть $A = d_x f_y$, $B = \partial f/\partial y(x, y)$. Обозначим через K_δ куб с ребром 2δ с центром в (*x, y*). Зафиксируем $\varepsilon > 0$ и выберем такое $\delta_0 > 0$, что для любого положительного $\delta < \delta_0$ верно, что

$$\mu_n(K_{2\delta}) \cap \{(x, y) : |\partial f/\partial x_n(x, y) - \partial f/\partial x_n(x_0, y_0)| > \varepsilon\}) < \varepsilon^n \delta^n$$

(такое существует в силу аппроксимативной непрерывности). Рассмотрим точку $(x', y') \in K_\delta(x_0, y_0)$ и обозначим $\Delta x = x' - x_0$, $\Delta y = y' = y_0$. Так как f_{y_0} дифференцируема в точке x_0, имеем $f(x', y_0) - f(x_0, y_0) = A\Delta x_0 + o(\delta)$. Теперь достаточно доказать, что $f(x', y') - f(x', y_0) = B\Delta y + o(\delta)$. Пусть *K* — куб с ребром $\varepsilon\delta$ с центром в *x*'. Для каждой такой точки $x \in K$ определеим величину

$$\phi(x) = \mu_1(\{t \in [-\delta, \delta] : |\partial f/\partial x_n(x, y_0 + t) - \partial f/\partial x_n(x_0, y_0)| > \varepsilon\})$$

Несложное вычисление показывает, что

$$f(x, y') - f(x, y_0) = \int_{y_0}^{y'} \frac{\partial f}{\partial y} = B\Delta y \pm 2C\phi(x) \pm \varepsilon\delta,$$

где *C* — константа Липшица для *f*. По теореме Фубини и выбору δ_0, имеем $\int_K \phi \, d\mu_n < \varepsilon^n \delta^n$. Значит, найдется такая точка $x'' \in K$, что $\phi(x'') < \varepsilon\delta$. Для этой точки имеем

$$f(x'', y') - f(x'', y_0) = B\Delta y \pm 2C\varepsilon\delta \pm \varepsilon\delta.$$

Из липшицевости следуют неравенства

$$|f(x'', y_0) - f(x', y_0)| < \varepsilon\delta$$

и

$$|f(x'', y') - f(x', y')| \le \varepsilon\delta$$

Складывая, получаем, что

$$|f(x', y') - f(x', y_0)| < (2C + 3)\varepsilon\delta,$$

что и требовалось.

Теорема. *Пусть X — метрическое пространство, $Y \subset X$, $f : Y \to \mathbb{R}$ — липшицева функция. Тогда существует функция $\tilde{f} : X \to \mathbb{R}$ с той же константой Липшица, продолжающая f.*

Доказательство. Можно считать, что константа Липшица для *f* равна 1. Тогда положим $\tilde{f}(x) = \inf_{y \in Y}(f(y) + |xy|)$.

Теорема (о приближении липшицевой функции). *Пусть $f : A \subset \mathbb{R}^n \to \mathbb{R}^m$ — липшицева функция. Тогда для любого $\varepsilon > 0$ существует C^1 функция $g : \mathbb{R}^n \to \mathbb{R}^m$ такая, что $f = g$ всюду на A, кроме множества меры ε.*

1.15.1 Доказательство теоремы о приближении

Теорема (Лузин). *Пусть μ — борелевски регулярная внешняя мера на метрическом пространстве X, такая, что X покрывается счетным набором открытых множеств конечной меры. Пусть $f : X \to \mathbb{R}$ — μ-измеримая функция. Тогда для любого $\varepsilon > 0$ найдется непрерывная функция $\tilde{f} : X \to \mathbb{R}$ такая, что $\mu(\{\tilde{f}(x) \neq f(x)\}) < \varepsilon$.*

Доказательство. 1. Пусть $f = \chi_A$, где A — измеримое множество. Найдем замкнутое F и открытое G такие, что $F \subset A \subset G$ и $\mu(G \setminus F) < \varepsilon$. По стандартной лемме Урысона из топологии можно построить такую непрерывную функцию $\tilde{f} : X \to [0, 1]$, что $\tilde{f}|_F \equiv 1$ и $\tilde{f}|_{X \setminus G} \equiv 0$.

Примечание: для метрического пространства X функцию легко предъявить явно:

$$\tilde{f}(x) = \frac{\operatorname{dist}(x, X \setminus G)}{\operatorname{dist}(x, X \setminus G) + \operatorname{dist}(x, F)}.$$

2. Если *f* — простая функция, представим ее в виде линейной комбинации характеристических и применим доказанный случай к каждому слагаемому.

3. Если *f* ограничена, представим ее как равномерный предел простых функций f_k так, что $\sup|f_k - f| < 1/2^k$. Положим $g_1 = f_1$, $g_{k+1} = f_{k+1} - f_k$, тогда $f = \sum g_k$, причем g_k простые и $\sup|g_k| \leq 1/2^{k-1}$. Для каждой функции g_k построим непрерывную $\tilde{g}_k$ с $\mu(\{\tilde{g}_k \neq g_k\} < 1/2^k$. Можно считать, что

$\sup|\tilde{g}_k| \le \sup|g_k| \le 1/2^{k-1}$ (иначе обрежем). Тогда $\tilde{f} = \sum \tilde{g}_k$ подходит.

4. Общий случай сводится к случаю ограниченной функции заменой *f* на arctg *f*.

Теорема (Егоров). *Пусть (X, μ) — пространство с мерой и* $\mu(X) < \infty$. *Пусть* f_n *— последовательность измеримых функций,* $f_n \to f$ *поточечно. Тогда для любого* $\varepsilon > 0$ *существует такое множество* $A \subset X$, *что* $\mu(A) < \varepsilon$ *и* f_n *сходятся к f равномерно на* $X \setminus A$.

Доказательство. Зафиксируем $\delta > 0$ и для каждого *n* рассмотрим множество

$$A_{n,\delta} = \{x \in X : \exists k > n \ |f_k(x) - f(x)| > \delta\}.$$

Это невозрастающая последовательность множеств.

Поскольку $f_n(x) \to f(x)$, каждая точка $x \in X$ принадлежит лишь конечному числу из них. То есть $\bigcap A_{n,\delta} = \emptyset$. Поскольку мера конечна, отсюда следует, что $\mu(A_{n,\delta}) \to 0$ при $n \to \infty$.

Для каждого $\delta = 1/k$ найдем такое n_k, что $\mu(A_{n,\delta}) < \varepsilon/2^k$. Тогда $A = \bigcup A_{n_k,1/k}$ подходит.

Следствие. *Пусть* $f_n : \mathbb{R}^n \to \mathbb{R}^m$ *— последовательность измеримых функций,* $f_n \to f$ *почти всюду относительно меры Лебега. Тогда для любого* $\varepsilon > 0$ *существует множество* $A \subset \mathbb{R}^n$ *такое, что* $\mu(\mathbb{R}^n \setminus A) < \varepsilon$ *и* f_n *сходятся к f равномерно на любом компакте* $K \subset A$.

Теперь пусть $f : \mathbb{R}^n \to \mathbb{R}$ — липшицева функция. По теореме Радемахера, она дифференцируема почти всюду.

Пусть $L_x : \mathbb{R}^n \to \mathbb{R}$ — дифференциал *f* в точке *x*, если он существует, и нулевая линейная функция в противном случае. Тогда

$$\frac{|f(x+h) - f(x) - L_x(h)|}{|h|} \to 0, \qquad h \to 0_{\mathbb{R}^n}$$

для почти всех $x \in \mathbb{R}^n$. Выкинув множество малой меры, сделаем эту сходимость равномерной на компактах. А именно, для каждого $\delta > 0$ определим функцию $\alpha_\delta : \mathbb{R}^n \to \mathbb{R}$ равенством

$$\alpha_\delta(x) = \sup_{h \in \mathbb{R}^n : 0 < |h| < \delta} \frac{|f(x+h) - f(x) - L_x(h)|}{|h|}.$$

Легко видеть, что эта функция измерима (супремум можно брать только по счетному множеству значений *h*). Для почти всех $x \in \mathbb{R}^n$ имеем $\alpha_\delta(x) \to 0$ при $\delta \to 0$. Подставляя $\delta = 1/k$, $k \to \infty$, и применяя следствие из теоремы Егорова, найдем множество $A \subset \mathbb{R}^n$ с $\mu(\mathbb{R}^n \setminus A) < \varepsilon$ и сходимость $\alpha_\delta \to 0$ равномерна на компактах в *A*. По теореме Лузина, найдется измеримое множество $B \subset A$, на котором функция $x \mapsto L_x$ непрерывна и $\mu(\mathbb{R}^n \setminus B) < \varepsilon$.

Наконец, в силу регулярности меры Лебега найдется замкнутое множество $D \subset B$ с $\mu(\mathbb{R}^n \setminus D) < \varepsilon$. Докажем, что существует C^1 функция $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$, совпадающая с f на D. Для этого достаточно доказать соответствующий частный случай теоремы Уитни о продолжении:

Теорема (теорема Уитни для C^r продолжений). *Пусть $D \subset \mathbb{R}^n$ — замкнутое множество, $f : D \to \mathbb{R}$ — непрерывная функция. Пусть каждой точке $x \in D$ сопоставлена линейная функция $L_x : \mathbb{R}^n \to \mathbb{R}$ так, что соответствие $x \mapsto L_x$ непрерывно и*

$$f(y) - f(x) - L_x(y - x) = o(|y - x|), \qquad x, y \in A, |x - y| \to 0,$$

где o равномерно на компактах.

Тогда существует C^1 функция $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$, такая, что $\tilde{f}(x) = f(x)$ и $d_x\tilde{f} = L_x$ для всех $x \in A$.

Замечание. Аналогичный критерий есть и для C^r продолжений, но его формулировка сложнее.

Доказательство. Сначала построим специальное разбиение единицы для $\mathbb{R}^n \setminus A$. Положим $h(x) = \max\{1, \operatorname{dist}(x, A)\}$. Для каждой точки $x \in \mathbb{R}^n \setminus A$ обозначим через B_x шар с центром в x радиуса $r(x) = \frac{1}{100}h(x)$. Выберем из этих шаров такое подпокрытие $\{B_{x_i}\}$, что соответствующие шарырадиуса $\frac{1}{5}r(x_i)$ не пересекается. Тогда шары удвоенных радиусов покрывают $R^n \setminus A$ с кратностью не больше $C(n)$. Для каждого шара B_{x_i} построим колоколообразную функцию u_i, равную 1 на этом шаре и нулю вне удвоенного шара. Это можно сделать так, что производная не превосходит $C/h(x_i)$. Теперь положим $\sigma(x) = \sum u_i(x)$ и заметим, что $\sigma(x) \geq 1$ и $\|d_x\sigma\| \leq C/h(x)$. Положим $\phi_i(x) = u_i(x)/\sigma(x)$. Тогда $\sum \phi_i \equiv 1$ и $\|d_x\phi_i\| \leq C/h(x)$.

Пусть a_i — ближайшая к x_i точка в A, $P_i(x) = f(a_i) + L_{a_i}(x - a_i)$. Положим

$$\tilde{f}(x) = \sum \phi_i(x)P_i(x).$$

при $x \in \mathbb{R}^n \setminus A$ и $\tilde{f}(x) = f(x)$ при $x \in A$. Докажем, что эта функция подходит. Ясно, что она гладкая на $\mathbb{R}^n \setminus A$. Надо проверить дифференцируемость и значение производной на A, а также непрерывность производной при стремлении к A.

Пусть $a \in A$. Определим $P(x) = f(a) + L_a(x - a)$. Равенство $d_a\tilde{f} = L_a$ эквивалентно соотношению

$$\tilde{f}(x) = P(x) + o(|x - a|), \qquad x \to a,$$

а непрерывность производной $\tilde{f}$ в точке a — соотношению

$d_x\tilde{f} = L_a + o(1), \qquad x \to a.$

Достаточно проверить эти соотношения только для точек $x \in \mathbb{R}^n \setminus A$. Пусть $x \in \mathbb{R}^n \setminus A$ — точка, близкая к a, $\delta = |x - a|$. Заметим, что $h(x) \le \delta$. Пусть I — множество таких индексов i, что $\phi_i(x) \ne 0$. Для каждого $i \in I$ точка x лежит в шаре радиуса $2r(x_i) = h(x_i)/50$, откуда нетрудно вывести, что $|xx_i| < 5r(x) = h(x)/20$ и $h(x_i) < 2h(x)$. Значит, соответствующая точка a_i лежит на расстоянии меньше $3h(x)$ от x, откуда $|aa_i| < \delta + 3h(x) < 4\delta$. Отсюда по условию $\|L_{a_i} - L_a\| = o(1)$ и $P_i(a) - f(a) = o(\delta)$. Заметим, что

$$P_i(x) = P_i(a) + L_{a_i}(x - a) = f(a) + L_a(x - a) + (P_i(a) - f(a)) + (L_{a_i} - L_a)(x - a) = P(x) + o(\delta).$$

Складывая с весами по всем $i \in I$, получаем

$$\tilde{f}(x) = \sum \phi_i(x)P_i(x) = \sum \phi_i(x)P(x) + o(\delta) = P(x) + o(\delta),$$

откуда следует равенство $d_a\tilde{f} = L_a$.

Чтобы доказать непрерывность производной, проведем такие же вычисления в случае, когда a — ближайшая к x точка из множества A. В этом случае $h(x) = |ax| = \delta$. Дифференцируя $\tilde{f}$, получаем

$$d_x\tilde{f} = \sum P_i(x)d_x\phi_i + \sum \phi_i(x)d_xP_i = \sum (P(x) + o(\delta))d_x\phi_i + \sum \phi_i(x)L_{a_i}.$$

Первое слагаемое есть $o(1)$, так как $\sum d_x\phi_i = 0$ и $\|d_x\phi_i\| \le C/\delta$, второе равно $L_a + o(1)$, так как $L_{a_i} = L_a + o(1)$ и $\sum \phi_i(x) = 0$. Таким образом, $d_x\tilde{f} = L_a + o(1)$, откуда следует непрерывность производной.

1.16. Якобианы

Определение. Пусть $L : \mathbb{R}^n \to \mathbb{R}^m$ — линейное отображение. Для $k \in \mathbb{N}$ определим k-мерный якобиан L (обозначение: J_kL) как максимальный k-мерный объем образа единичного k-мерного куба.

Пусть $f : \mathbb{R}^n \to \mathbb{R}^m$ — функция, дифференцируемая в точке $x \in \mathbb{R}^n$. Для $k \in \mathbb{N}$ определим k-мерный якобиан f в точке x (обозначение: $J_kf(x)$) равенством $J_kf(x) = J_k(d_xf)$.

Примеры. 1. $m = n = k$. Тогда $J_kL = |\det L|$.

2. $k = n < m$. Тогда $J_kL = \sqrt{\det LL^T}$. Замечание: LL^T — матрица Грама набора векторов $\partial f/\partial x_i$.

3. $k = m < n$. Тогда $J_k L = \sqrt{\det L^T L}$. В частности, при $m = k = 1$ якобиан равен модулю градиента.

Задача. Если ранг L не превосходит k, то $(J_k L)^2$ равен сумме квадратов миноров $k \times k$ матрицы отображения.

1.16.1 Площадь липшицевой поверхности

Определение. Пусть $k < n$, $A \subset \mathbb{R}^k$, $f : A \to \mathbb{R}^n$ — липшицево отображение. Площадь f — это число

$$\text{area}(f) = \int_A J_k f.$$

Теорема (формула площади). *Пусть $k \le n$, $A \subset \mathbb{R}^k$ — измеримое множество, $f : \mathbb{R}^k \to \mathbb{R}^n$ - липшицево отображение. Тогда*

$$\text{area}(f) = \int_{\mathbb{R}^n} \#f^{-1}(x)\, d\mathcal{H}^k(x).$$

Следствие. *Для любой интегрируемой функции $\phi : A \to \mathbb{R}$*

$$\int_A \phi \cdot J_k f = \int_{\mathbb{R}^n} \sum_{y \in f^{-1}(x)} \phi(y)\, d\mathcal{H}^k(y).$$

Доказательство. 1. Обе части аддитивны, поэтому можно разбивать A на части. В частности, можно считать, что A ограничено.

2. По предыдущей теореме, f совпадает с C^1 функцией всюду, кроме множества сколь угодно малой меры. Следующая лемма позволяет выкинуть это малое множество из A.

Лемма. $\int_{\mathbb{R}^n} \#f^{-1}(x)\, d\mathcal{H}^k(x) \le C^k \mathcal{H}^k(A)$, *где C — константа Липшица для f.*

Доказательство. Аналогично одномерной лемме: начнем с неравенства $\mathcal{H}^k(f(A)) \le C^k \mathcal{H}^k(A)$ и будем уточнять его, разбивая A на мелкие части.

3. Теперь можно считать, что $f \in C^1$. Сначала рассмотрим множество точек, где f невырождено (ранг равен k). Разобьем его на мелкие части, на каждой части представим f как композицию линейного отображения $\mathbb{R}^k \to \mathbb{R}^k$ и отображения $\mathbb{R}^k \to \mathbb{R}^n$, производная которого почти постоянна и близка к изометрии. У второго отображения и его обратного константы Липшица близки к 1, поэтому оно почти сохраняет площадь.

4. Теперь рассмотрим множество точек, где f вырождено. Надо доказать, что на этом множестве правая часть равна 0. Рассмотрим отображение $f_\varepsilon = f \times \varepsilon Id : \mathbb{R}^k \to \mathbb{R}^{n+k}$ при $\varepsilon \to 0$. Оно инъективно и невырождено. Имеем $Jf_\varepsilon \le C^{k-1}\varepsilon$, поэтому $\text{area}(f_\varepsilon|_A) \to 0$. По

разобранному имеем $\operatorname{area}(f_\varepsilon|_A) = \mathcal{H}^k(f_\varepsilon(A))$. Проекция $\mathbb{R}^{n+k} \to \mathbb{R}^n$ не увеличивает расстояния, поэтому по лемме

$$\int_{\mathbb{R}^n} \#f^{-1}(x)\, d\mathcal{H}^k(x) \le \operatorname{area}(f_\varepsilon)$$

1.16.2 Формула коплощади

Теорема (формула коплощади). *Пусть $k \le n$, $A \subset \mathbb{R}^n$ — измеримое множество, $f : \mathbb{R}^n \to \mathbb{R}^k$ - липшицево отображение. Тогда*

$$\int_A J_k f(x)\, dx = \int_{\mathbb{R}^k} \mathcal{H}^{n-k}(A \cap f^{-1}(y))\, dy$$

Следствие (о послойном интегрировании функции).
Пусть $\phi : A \to \mathbb{R}$ — интегрируемая функция.
Тогда

$$\int_A \phi(x) J_k f(x)\, dx = \int_{\mathbb{R}^k} \left(\int_{A \cap f^{-1}(y)} \phi \; d\mathcal{H}^{n-k} \right) dy$$

Доказательство формулы коплощади.
Определение. Верхний интеграл — инфимум верхних интегральных сумм.
Свойство. Для верхнего интеграла выполняется теорема Леви.
Лемма. *Пусть X — метрическое пространство, $f : X \to \mathbb{R}^k$ — липшицево с константой L.*
Тогда

$$\int_{\mathbb{R}^k}^{*} \mathcal{H}^{n-k}(f^{-1}(y))\, dy \le L^k C(k,n) \mathcal{H}^n(X),$$

где $\int^*$ — верхний интеграл Лебега.
Доказательство. Будем считать, что нормировочные константы всех мер Хаусдорфа равны 1 (это влияет только на значение константы *C(n,k)*). Пусть $\varepsilon > 0$. Покроем X таким счетным набором множеств $\{X_i\}$, что $\operatorname{diam} X_i < \varepsilon$ для всех i и $\sum (\operatorname{diam} X_i)^n \le \mathcal{H}^n(X) + \varepsilon$. Для каждой точки $y \in \mathbb{R}^k$ имеем

$$\mathcal{H}^{n-k}_\varepsilon(f^{-1}(y)) \le \sum_{i : y \in f(X_i)} (\operatorname{diam} X_i)^{n-k} = \sum_i (\operatorname{diam} X_i)^{n-k} \chi_{f(X_i)}(y).$$

Отсюда

$$\int_{\mathbb{R}^k}^{*} \mathcal{H}^{n-k}_\varepsilon(f^{-1}(y))\, dy \le \sum_i (\operatorname{diam} X_i)^{n-k} \mathcal{H}^k(f(X_i)) \le$$
$$\le C \sum_i (\operatorname{diam} X_i)^{n-k} (\operatorname{diam} f(X_i))^k \le C L^k \sum_i (\operatorname{diam} X_i)^n \le C L^k (\mathcal{H}^n(X) + \varepsilon).$$

Устремляя ε к нулю и пользуясь теоремой Леви для верхнего интеграла, получаем требуемое.

Следствие. *Если* $\mathcal{H}^n(A) = 0$, *то правая часть формулы коплощади тоже равна 0.*

Лемма. *Для почти всех* $y \in \mathbb{R}^k$ *множество* $A \cap f^{-1}(y)$ *измеримо относительно* $\mathcal{H}^{n-k}$. *Функция* $y \mapsto \mathcal{H}^{n-k}(A \cap f^{-1}(y))$ *измерима.*

Доказательство. В случае, если *A* имеет меру ноль, утверждение вытекает из предыдущего следствия. Отсюда следует, что добавление множества меры ноль не влияет на истинность утверждения леммы. Выкинув множество меры 0, сведем утверждение к случаю, когда *A* — счетное объединение вложенных компактов. Поскольку предел измеримых функций измерим, достаточно доказать утверждение для одного компакта.

Предположим, что *A* компактно. Тогда для любого $y \in \mathbb{R}^k$ множество $f^{-1}(y)$ замкнуто и, следовательно, измеримо. Проверим измеримость функции $y \mapsto \mathcal{H}^{n-k}(f^{-1}(y))$. Достаточно доказать, что для любого $t \in \mathbb{R}$ множество $A_t = \{y \in \mathbb{R}^k : \mathcal{H}^{n-k}(A \cap f^{-1}(y)) \le t\}$ измеримо. Заметим, что

$A_t = \bigcap_{i=1}^{\infty} A_{t,i}$, где

$$A_{t,i} = \{y \in \mathbb{R}^k : \mathcal{H}^{n-k}_{1/i}(A \cap f^{-1}(y)) < t + 1/i\}.$$

Докажем, что каждое множество $A_{t,i}$ открыто. Точка $y \in \mathbb{R}^k$ принадлежит ему тогда и только тогда, когда есть покрытие множества $A \cap f^{-1}(y)$ мелкости меньше $1/i$ и *k*-мерным весом меньше $t + 1/i$. В силу компактности, такое покрытие можно выбрать конечным и открытым. Тогда оно покрывает и множества $A \cap f^{-1}(y')$ для всех *y'*, достаточно близких к *y*.

Таким образом, A_t есть пересечение счетного набора открытых множеств. Следовательно, оно измеримо.

Левая и правая часть формулы коплощади счетно аддитивны по *A*. По теореме о приближении, *A* можно разбить на счетное число частей, одна из которых имеет меру 0, а на каждой из остальных *f* совпадает с сужением некоторой C^1 функции. Для множества меры 0 формула доказана, поэтому достаточно доказать теорему для случая, когда $f|_A$ — сужение C^1 функции. Можно считать, что $f \in C^1$.

Разобъем *A* на две части: множество точек, где $rk(df) = k$, и множество точек, где $rk(df) < k$.

В силу аддитивности, достаточно доказать теорему для каждого из множеств.

Сначала рассмотрим случай, когда *df* имеет ранг *k* всюду на *A*. В этом случае в окрестности любой точки $x \in A$ отображение *f* можно

представить в виде $f = L \circ \phi$, где $L = d_x f : \mathbb{R}^n \to \mathbb{R}^k$ - линейное отображение, $\phi : \mathbb{R}^n \to \mathbb{R}^n$ — диффеоморфизм этой окрестности на окрестность начала координат, причем $d_x\phi$ — тождественно. Для *L* формула коплощади следует из теоремы Фубини. Вблизи *x* отображение ϕ билипшицево с константой, близкой к 1, поэтому оно мало искажает якобианы и меры Хаусдорфа.

Осталось доказать формулу в случае, когда ранг *df* всюду меньше *k*, то есть $J_k f = 0$ всюду на *A*. В этом случае надо доказать, что

$$\int_{\mathbb{R}^k} \mathcal{H}^{n-k}(A \cap f^{-1}(y))\, dy = 0.$$

Зафиксируем $\varepsilon > 0$ и определим $g : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^k$ формулой $g(x, z) = f(x) + \varepsilon z$. Легко видеть, что *dg* всюду имеет ранг *k* и $J_k g(x, z) \le \varepsilon(C + \varepsilon)^{k-1}$, где C — константа Липшица для *f*. Пусть $Q = A \times B(0, 1) \subset \mathbb{R}^n \times \mathbb{R}^k$. Применяя уже разобранный случай, получаем

$$\int_{\mathbb{R}^k} \mathcal{H}^n(Q \cap g^{-1}(y))\, dy = \int_Q J_k g \le C(n)C(f)\varepsilon.$$

Определим $p : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^k$ формулой $p(x, z) = z$ и применим лемму к множеству $Q \cap g^{-1}(y)$ и отображению *p*. Получаем

$$\mathcal{H}^n(Q \cap g^{-1}(y)) \ge c \int_{\mathbb{R}^k} \mathcal{H}^{n-k}(Q \cap g^{-1}(y) \cap p^{-1}(w))\, dw = c \int_{B(0,1)\subset\mathbb{R}^k} \mathcal{H}^{n-k}(A \cap f^{-1}(y - \varepsilon w))\, dw.$$

Интегрируя по *y* и переставляя интегралы по *y* и по *w*, получаем

$$\int_{\mathbb{R}^k} \mathcal{H}^n(Q \cap g^{-1}(y)) \ge c \int_{B(0,1)} dw \int_{\mathbb{R}^k} \mathcal{H}^{n-k}(A \cap f^{-1}(y - \varepsilon w))\, dy.$$

Внутренний интеграл не зависит от *w*, поэтому это выражение равно

$$= c\alpha(n) \int_{\mathbb{R}^k} \mathcal{H}^{n-k}(A \cap f^{-1}(y))\, dy.$$

Итак,

$$\int_{\mathbb{R}^k} \mathcal{H}^{n-k}(A \cap f^{-1}(y))\, dy \le C(n) \int_{\mathbb{R}^k} \mathcal{H}^n(Q \cap g^{-1}(y)) \le C(n, f)\varepsilon.$$

Устремляя ε к нулю, получаем, что левая часть равна 0, что и требовалось.

Точки Лебега

Замечание. Если *f* суммируема, то функцию $\tilde{f}$ из теоремы Лузина можно выбрать так, что

$$\int |f - \tilde{f}|\, d\mu < \varepsilon.$$

Доказательство. Найдется такое $M > 0$, что $\int (f - M)_+ < \varepsilon/4$ и $\int (f_- - M) < \varepsilon/4$. Обрежем функцию значениями M сверху и — M снизу и применим теорему Лузина с $\varepsilon/2M$ вместо ε.

Теорема. *Пусть $f : \mathbb{R}^n \to \mathbb{R}$ — локально μ-суммируемая функция, где μ — локально конечная борелевски регулярная внешняя мера. Тогда почти любая точка $x \in \mathbb{R}^n$ является точкой Лебега, то есть*

$$\lim_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r))} |f - f(x)|\, d\mu = 0.$$

в частности,

$$f(x) = \lim_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r))} f\, d\mu.$$

Доказательство. Можно считать, что f суммируема. Предположим противное. Для $\delta > 0$ обозначим

$$A_\delta = \{x : \overline{\lim_{r \to 0}} \frac{1}{\mu(B(x,r))} \int_{B(x,r))} |f - f(x)|\, d\mu > \delta\}.$$

Тогда для некоторого δ_0 имеем $\mu(A_{\delta_0}) = m_0 > 0$. Воспользовавшись дополнением к теореме Лузина, представим f в виде $f = g + h$, где g непрерывна, $h = 0$ вюду кроме множества меры $m_0/2$ и $\int |h| < \delta_0 m_0/2$. Обозначим $A = A_{\delta_0} \cap \{x : h(x) = 0\}$. Для каждого $x \in A$ имеем

$$\overline{\lim_{r \to 0}} \frac{1}{\mu(B(x,r))} \int_{B(x,r))} |h|\, d\mu > \delta_0.$$

По теореме Безиковича существует дизъюнктный набор шаров $B_i = B(x_i, r_i)$, покрывающий почти все A, таких, что

$$\int_{B_i} |h|\, d\mu > \delta_0 \mu(B_i)$$

для всех i. Тогда

$$\int_{\bigcup B_i} |h|\, d\mu > \delta_0 \mu(A) \geq \delta_0 m_0/2,$$

противоречие с выбором h.

2. Нечеткие меры и интегралы

2.1. Методические замечания

При решении многих задач анализа сложных систем в условиях неопределенности широко используются методы теории вероятности и математической статистики. Эти методы предполагают вероятностную интерпретацию обрабатываемых данных и полученных статистических выводов. В последнее время возрастает потребность в новых подходах к математическому описанию информации, характеризующейся высоким уровнем неопределенности. Один из возможных подходов здесь может основываться на обобщении понятия меры и построении нечетких мер, свободных от ряда ограничений вероятностной меры. Существуют различные интерпретации понятия вероятности. Это — классическая частотпая интерпретация Лапласа, субъективная вероятность по Байесу, субъективная вероятность по Де Финетти, Севиджу и т. д.. Наиболее содержательной с математической точки зрения является аксиоматическая трактовка вероятности А. Н. Колмогорова с позиций теории меры.

Как известно, *мерой* называется функция множества $m\colon \mathcal{P}(X) \to \mathcal{R}$, удовлетворяющая следующим трем аксиомам:

1) $A \subseteq X \Leftrightarrow m(A) \geqslant 0$;
2) $m(\varnothing) = 0$;
3) если $A, B \in \mathcal{P}(X)$, то $m(A \cup B) = m(A) + m(B) - m(A \cap B)$.

Здесь $\mathcal{P}(X)$ — множество всех подмножеств X, а $\mathcal{R}$ — множество действительных чисел. При $\mathcal{R} = [0, 1]$ эти аксиомы определяют вероятностную меру.

Под субъективной вероятностью понимается степень уверенности в данном событии, возникающая у человека на основе известных ему данных. Эта степень уверенности всегда зависит от индивидуального опыта и поэтому различна для разных людей. Неясность суждений, основанных на субъективном анализе, обусловливает многие трудности, которые возникают при использовании субъективной вероятности.

Субъективную вероятность можно рассматривать как индивидуальный способ обработки тех аспектов субъективных данных, которые доступны индивидуальному суждению. Однако чаще всего такие суждения неаддитивны. Показано, что реальное поведение человека, как правило, противоречит предположению об аддитивности мер, которые он использует при оценке событий. В отличие от субъективной вероятности, нечеткая мера свободна от весьма

ограничительного требования аддитивности, что делает ее особенно привлекательной для решения ряда задач при наличии неопределенности типа нечеткости.

В настоящее время существует тенденция вероятностной трактовки НМ. Следует отметить, что, с точки зрения теории меры, такой подход является неоправданным, поскольку понятие вероятностной меры является сужением понятия нечеткой меры. Для сравнения рассмотрим обе теоретико-мерные трактовки вероятности и нечеткости.

Пусть $(X, \mathcal{B}, P)$— вероятностное пространство. Здесь $\mathcal{B} \subseteq \subseteq \mathcal{P}(X)$ — поле борелевских подмножеств множества X (минимальная σ-алгебра, содержащая все открытые подмножества множества *X),* а *P* — вероятностная мера, т. е. функция множества $P\colon \mathcal{B} \to [0, 1]$, удовлетворяющая условиям 1) —3). С другой стороны, нечеткое множество Л. Заде описывается функцией принадлежности μ, принимающей свои значения в интервале [0, 1]. С точки зрения теории отображений $P\colon \mathcal{B} \to [0, 1]$ и $\mu\colon X \to$ $\to [0, 1]$ — совершенно разные объекты. Вероятность *P* определяется в σ-алгебре $\mathcal{B}$ и является функцией множества, а $\mu(x)$ есть обычная функция, областью определения которой является множество *X.* Поэтому понятия вероятности и нечеткого множества не имеет смысла сравнивать на одном уровне абстрагирования.

Когда *X* — является конечным множеством, очевидно, можно сравнивать $P(\{x\})$ с $\mu_A(x)$: $\sum\limits_{x \in X} P(\{x\}) = 1$ и $\sum\limits_{x \in X} \mu_A(x) \neq 1.$

В этом случае, когда $X \subset \mathcal{R}$, приходится сталкиваться со следующими трудностями. Если

$$(a, b] \subset \mathcal{R}, \text{ то } P((a, b]) = \int_a^b p(x)\,dx,$$

где *p(x)* — плотность вероятности. При этом очевидно, что $\forall x \in \mathcal{R}\colon P(\{x\}) \neq 0$, когда $p(x) \neq 0$. Нетрудно увидеть, что понятия плотности вероятности и функции принадлежности сравнимы. В то время как вероятностная мера является шкалой для измерения неопределенности типа случайности, нечеткие меры являются субъективными шкалами для нечеткости.

2.2. Нечеткие меры

Рассмотрим основные свойства нечетких мор и интегралов, а также их содержательную связь с мерами возможности , используемыми для построения алгоритмов нечеткого вывода.

Пусть X — произвольное множество, а $\mathcal{B}$— поле борелевских множеств (σ-алгебра) для *X*.

Определение 1. Функция $g(\cdot)$, определяемая в виде $g\colon \mathcal{B} \to [0, 1]$, называется *нечеткой мерой,* если она удовлетворяет следующим условиям:

$$\left.\begin{array}{l} 1)\ g(\varnothing) = 0, \\ 2)\ g(X) = 1, \end{array}\right\}\ \text{(ограниченность)};$$
$$3)\ \text{если } A, B \in \mathcal{B} \text{ и } A \subset B, \text{ то } g(A) \leqslant g(B)\ \text{(монотонность)};$$
$$\left.\begin{array}{l} 4)\ \text{если } F_n \in \mathcal{B} \text{ и } \{F_n\} \text{ является монотонной последовательностью, то } \lim\limits_{n\to\infty} g(F_n) = g\left(\lim\limits_{n\to\infty} F_n\right)\ \text{(непрерывность)}. \end{array}\right\} \quad (1)$$

Тройка $(X, \mathcal{B}, g)$ называется *пространством с нечеткой мерой.* Для нечеткой меры в общем случае не должно выполняться условие аддитивности: $g(A \cup B) \neq g(A) + g(B)$. Таким образом, нечеткая мера является однопараметрпческим расширением вероятностной меры.

Выражение *g(A)* представляет собой меру, характеризующую степень нечеткости *A,* т. е. оценку нечеткости суждения «$X \in A$» или степень субъективной совместимости *X* с *A.* Нетрудно увидеть, что монотонность меры *g* влечет за собой

$$\forall A, B \in \mathcal{B}\colon\ g(A \cup B) \geqslant \max(g(A),\ g(B));$$
$$\forall A, B \in \mathcal{B}\colon\ g(A \cap B) \leqslant \min(g(A),\ g(B)).$$

Для построения нечетких мер используется следующее *λ-правило*.

Пусть $A, B \in \mathcal{B}$, $A \cap B = \varnothing$. Тогда

$$g_\lambda(A \cup B) = g_\lambda(A) + g_\lambda(B) + \lambda \cdot g_\lambda(A) \cdot g_\lambda(B), \quad -1 < \lambda < \infty. \quad (2)$$

В случае $A \cup B = X$ будем называть выражение (2) условием нормировки для g_λ-мер. Очевидно, что $g_\lambda(X) = 1;\ g_\lambda(\varnothing) = 0$.

Параметр $\lambda \in (-1, +\infty)$ называется параметром нормировки g_λ-меры. При $\lambda > 0$, $g_\lambda(A \cup B) > g_\lambda(A) + g_\lambda(B)$ имеем класс супераддитивных мер, а при $-1 < \lambda < 0$, $g_\lambda(A \cup B) < g_\lambda(A) + g_\lambda(B)$ получаем класс субаддитивных мер.

Легко убедиться, что если $\bar{A} = X \setminus A$, $A \in \mathcal{B}$, то из (2) следует

$$g_\lambda(\bar{A}) = \frac{1 - g_\lambda(A)}{1 + \lambda \cdot g_\lambda(A)}. \quad \textbf{(3)}$$

Формула (3) определяет класс так называемых λ-дополнений Сугено.

В общем случае, когда A и B — произвольные непересекающиеся подмножества множества X, т. е. $A, B \in \mathcal{B}, A \cap B \neq \varnothing$, выражение (2) приобретает вид

$$g_\lambda(A \cup B) = \frac{g_\lambda(A) + g_\lambda(B) - g_\lambda(A \cap B) + \lambda \cdot g_\lambda(A) \cdot g_\lambda(B)}{1 + \lambda \cdot g_\lambda(A \cap B)}. \quad \textbf{(4)}$$

Если $X = \mathcal{R}$, то g_λ-меру можно получить с помощью непрерывной функции h, удовлетворяющей следующим свойствам:

1) если $x \leqslant y$, то $h(x) \leqslant h(y)$; $x, y \in \mathcal{R}$;
2) $\lim\limits_{x \to -\infty} h(x) = 0$; $\lim\limits_{x \to +\infty} h(x) = 1$.

Функция h аналогична функции распределения вероятности и называется нечеткой функцией распределения.

Таким образом, нечеткую меру g_λ на $(\mathcal{R}, \mathcal{B})$ можно построить в виде

$$g_\lambda([a, b]) = \frac{h(b) - h(a)}{1 + \lambda \cdot h(a)} \quad \forall [a, b] \subset \mathcal{R}. \qquad (5)$$

Мера g_λ в (5) удовлетворяет λ-правилу. В частности,

$$g_\lambda((-\infty, x]) = h(x) \quad \forall \lambda \in (-1, +\infty). \quad \textbf{(6)}$$

Далее предположим, что $K = \{s_1, s_2, \ldots, s_n\}$. Мера g_λ на $(K, 2^K)$ строится следующим образом $(0 \leqslant g_n^i \leqslant 1, i \in I)$:

$$\frac{1}{\lambda}\left[\prod_{i=1}^{n}(1 + \lambda g^i) - 1\right] = 1, \quad \lambda \in (-1, +\infty). \qquad (7)$$

Если $K' \subset K$, то

$$g_\lambda(K') = \frac{1}{\lambda}\left[\prod_{s_i \in K'}(1 + \lambda g^i) - 1\right]. \qquad (8)$$

Выражение (8) также удовлетворяет λ-правилу и из (7) следует, что

$$g_\lambda(\{s_i\}) = g^i,$$
$$g_\lambda(\{s_i, s_j\}) = g^i + g^j + \lambda g^i g^j, \quad i \neq j. \quad (9)$$

Рассмотрим несколько примеров нечетких мер (см. рис.1).

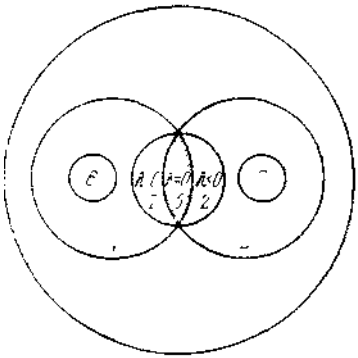


Рис. 1. Соотношения между нечеткими мерами : *1* — нечеткие меры (исключая меру Дирака); *2*— g_λ-меры, $-1 < \lambda < \infty$; *3* — функции доверия; *4* — меры правдоподобия; $5 = 3 \cap 4$ — вероятностная мера ($\lambda = 0$); 6—согласованные функции доверия (мера необходимости); 7 — мера возможности

Меры Дирака. Примитивный класс мер Дирака определяется соотношением

$$\forall A \in \mathcal{B}, \quad \mu(A) = \begin{cases} 1 & \text{при } x_0 \in A, \\ 0 & \text{в противном случае,} \end{cases} \tag{10}$$

где x_0 — заданный элемент в X. Меры Дирака — частный случай вероятностной меры, соответствующий детерминпрованной ситуации (меры полной уверенности). Все рассматриваемые далее нечеткие меры можно разделить на два класса: *супераддитивные меры* ($X > 0$) *и субаддитивные меры* $(-1 < \lambda < 0)$.

2.2.1. Супераддитивные меры. Функция доверия.

Определение *функции доверия* (belief function) предполагает, что степень доверия высказыванию $A\,(A \neq \varnothing)$, которое является истинным, не обязательно равна 1. Это означает, что сумма степеней доверия высказыванию A к его отрицанию $\bar{A}$ также не обязательно равна 1, а может быть либо равной, либо меньшей 1. Другими словами, когда высказывание $A\,(A \neq \varnothing)$ является истинным с определенной степенью $s \in [0, 1]$, его мера неопределенности выражается с помощью функции

$$\forall B \in \mathscr{B}: \quad b(B) = \begin{cases} 1, & \text{если } B = X, \\ s, & \text{если } B \supset A, \quad B \neq X, \\ 0, & \text{если } B \not\supset A, \end{cases} \quad (11)$$

которая называется *простой функцией носителя, сосредоточенной на A.*

Если *s = 1,* то получаем меру, которая называется мерой определенности, сосредоточенной на *A.* Если $|A| = 1$, то получаем меру Дирака, сосредоточенную на *A.*

Если *s* = 0 или *A = X,* то тогда *b (B)* называется пустой функцией доверия (полное незнание). В результате обобщения этих рассуждений введена *функция доверия* — мера, удовлетворяющая следующим свойствам:

$$\left.\begin{array}{l} 1)\ b(\varnothing) = 0; \quad b(X) = 1; \quad \forall A \in \mathscr{B}: 0 \leqslant b(A) \leqslant 1; \\ 2)\ \forall A_1, \ldots, A_n \in \mathscr{B}: b(A_1 \cup \ldots \cup A_n) \geqslant \sum\limits_{i=1}^{n} b(A_i) - \\ - \sum\limits_{i<j} b(A_i \cap A_j) + \ldots + (-1)^{n+1} b(A_1 \cap A_2 \cap \ldots \cap A_n). \end{array}\right\} \quad (12)$$

В случае, когда $|\mathscr{B}| = 2$, получаем:

$$\forall A, B \in \mathscr{B}: b(A \cup B) \geqslant b(A) + b(B) - b(A \cap B)$$

(свойство супераддитивпости) и

$$\forall A \in \mathscr{B}: b(A) + b(\bar{A}) \in [0, 1].$$

Возможно также другое определение этой меры. Пусть *m* — мера, удовлетворяющая следующим свойствам:

$$\left.\begin{array}{ll} 1) & m(\varnothing) = 0; \\ 2) & \sum\limits_{A \in \mathscr{B}} m(A) = 1 \quad \text{(полное доверие).} \end{array}\right\} \quad (13)$$

Тогда

$$\forall A \in \mathscr{B}: b(A) = \sum_{B \subset A} m(B) \quad (14)$$

является *функцией доверия.* Поэтому функции доверия называются также нижними вероятностями. Из (14) вытекает:

$$\forall A \in \mathscr{B}: b(A) + b(\bar{A}) = 1 - \sum_{\substack{B \not\subset A \\ B \not\subset \bar{A}}} m(B) \in [0, 1]. \quad (15)$$

Любая λ-нечеткая мера (кроме меры Дирака) является функцией доверия тогда и только тогда, когда $\lambda \geq 0$. Отсюда следует, что мера вероятности есть частный случай функции доверия.

Согласованная функция доверия. Понятие согласованной функции доверия (consonant belief function) базируется на определении ядра $C = \{B \subset X | m(B) > 0\}$, полностью упорядоченного по вложенности. Легко показать, что любая простая функция носителя является согласованной функцией доверия. Если $A \neq X$, то мера неопределенности

$$\forall B \subset \mathscr{B}:\ b(B) = \begin{cases} s, & \text{если } B = A, \\ 1 - s, & \text{если } B = X, \\ 0, & \text{если } B \neq A,\ B \neq X. \end{cases} \quad (16)$$

Согласованная функция доверия определяется с помощью следующих аксиом:

$$\begin{aligned} &1)\ b(\varnothing) = 0;\quad b(X) = 1, \\ &2)\ b(A \cap B) = \min\,(b(A),\ b(B));\ \forall A \in \mathscr{B}. \end{aligned} \quad (17)$$

При этом

$$\min\,(b(A),\ b(\bar{A})) = 0;\ \forall b\ \exists A,\ B:\ b(A \cup B) > \max(b(A),\ b(B)).$$

2.2.2. Субаддитивные меры.

Меры правдоподобия. *Мера правдоподобия* множества A из X определена как

$$\mathrm{Pl}(A) = 1 - b(\bar{A}), \quad (18)$$

где b — функция уверенности.

Мера правдоподобия удовлетворяет следующим аксиомам

$$\left.\begin{aligned} &1)\ \mathrm{Pl}(\varnothing) = 0;\ \mathrm{Pl}(X) = 1, \\ &2)\ \forall A_1, \ldots, A_n \subseteq X:\ \mathrm{Pl}(A_1 \cap \ldots \cap A_n) \leqslant \\ &\qquad \leqslant \sum_{i=1}^{n} \mathrm{Pl}(A_i) - \sum_{i<j} \mathrm{Pl}(A_i \cup A_j) + \ldots + \\ &\qquad\qquad + (-1)^{n+1}\,\mathrm{Pl}(A_1 \cup \ldots \cup A_n). \end{aligned}\right\} \quad (19)$$

Существует другой способ определения функции правдоподобия. Пусть m — мера, удовлетворяющая свойствам (13), тогда

$$\forall A \in \mathscr{B}:\ \mathrm{Pl}(A) = \sum_{B \cap A \neq \varnothing} m(B) \quad (20)$$

является *мерой правдоподобия.* Меры правдоподобия называются также верхними вероятностями .

Пусть μ и ν — две меры такие, что $\forall A \in \mathscr{B}:\ \mu(A) + \nu(\bar{A}) = 1$. В этом случае μ является функцией доверия тогда и только тогда, когда ν — мера правдоподобия.

Мера возможности. Мерой возможности называется функция $\Pi\colon \mathscr{B} \to [0,\ 1]$, удовлетворяющая следующим аксиомам:

$$\left.\begin{array}{ll} 1) & \Pi(\varnothing) = 0;\ \ \Pi(X) = 1, \\ 2) & \forall i \in \mathbf{N},\ \ A_i \subset X,\ \ \Pi\left(\bigcup\limits_{i\in\mathbf{N}} A_i\right) = \sup\limits_{i\in\mathbf{N}} \Pi(A_i). \end{array}\right\} \quad (.21)$$

Здесь **N** — множество натуральных чисел.

Мера возможности может быть построена с помощью распределения возможности $\pi(x)$, являющегося функцией $\pi\colon X \to [0,\ 1]$, такой, что $\sup\limits_{x\in X} \pi(x) = 1$ (условие нормировки). Нетрудно увидеть, что $\forall A \in \mathscr{B}\colon \Pi(A) = \sup\limits_{x\in A} \pi(x)$. Очевидно, что для счетного множества $\pi(x) = \Pi(\{x\})$.

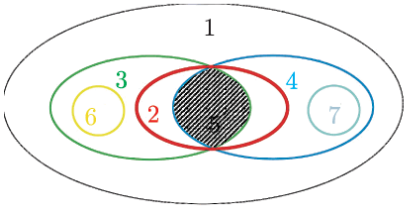
Любая мера возможности является нечеткой мерой тогда и только тогда, когда существует функция распределения f такая, что $\sup\limits_{x\in X} f(x) = 1$.

Любая мора возможности П является g_λ-мерой $(\lambda \in (-1,\ \infty))$ тогда и только тогда, когда П — мера Дирака.

Пусть μ и ν две меры такие, что $\forall A \in \mathscr{B}\colon \ \mu(A) + \nu(\bar{A}) = 1$.

Нечеткая мера μ является согласованной функцией доверия тогда и только тогда, когда ν является мерой возможности.

Мера вероятности. Вероятностная мера ($h = 0$) является частным случаем функции доверия или меры правдоподобия (см. рис. 2).



1—нечеткие меры
2—g_λ-меры
3—функции доверия
4—меры правдоподобия
6—согласованные функции доверия
7—мера возможности

Рис.2

Нечеткая мора $g \triangleq P$ является вероятностной мерой тогда и только тогда, когда выполняются условия:

1) $\forall A \in \mathscr{B}\colon\ P(A) \in [0,\ 1];\ P(\varnothing) = 0;\ P(X) = 1;$

2) если $\forall i \in \mathbf{N}\colon\ A_i \in \mathscr{B}$ и $\forall i \neq j\colon A_i \cap A_j = \varnothing$, то $P\left(\bigcup\limits_{i\in\mathbf{N}} A_i\right) = \sum P(A_i)$.

g_ν-мера. Нечеткая мера $g \triangleq g_\nu$ называется g_ν-мерой, если она удовлетворяет следующим аксиомам:

$$\left.\begin{array}{l} 1)\ g_\nu(X)=1;\ g_\nu(\varnothing)=0; \\ 2)\ \text{если}\ \forall i\in \mathbf{N}:\ A_i\in\mathscr{B}\ \text{и}\ \forall i\neq j:\ A_i\cap A_j=\varnothing, \\ \text{то} \\ \quad g_\nu\left(\bigcup\limits_{i\in N} A_i\right)=(1-\nu)\bigvee\limits_{i\in\mathbf{N}} g_\nu(A_i)+\nu\sum\limits_{i\in\mathbf{N}} g_\nu(A_i), \\ \text{где}\ \nu\geqslant 0; \\ 3)\ \forall A,\ B\in\mathscr{B}:\ A\subseteq B,\ g_\nu(A)\leqslant g_\nu(B). \end{array}\right\} \quad (22)$$

Нетрудно увидеть, что g_ν-мера является расширением меры Цукамото, для которой $\nu \in [0, 1]$. Очевидно, что при $\nu = 0$, g_ν -мера является мерой возможности, а при $\nu = 1$ — вероятностной мерой. Если $\nu > 1$, то g_ν-мера описывает неопределенность, отличающуюся по своим свойствам от вероятности или возможности.

Условие нормировки для g_ν-меры в случае счетного множества X имеет вид

$$g_\nu(X)=(1-\nu)\bigvee_{i\in\mathbf{N}} g_i+\nu\sum_{i\in\mathbf{N}} g_i=1, \quad (23)$$

где $g_i=g_\nu(\{x_i\}),\ \forall i\in\mathbf{N},\ x_i\in X.$

Если $X=\mathscr{R}$, то нетрудно увидеть, что для нечеткой плотности $f_\nu(x):\ X\to[0,\ 1]$ можно получить

$$g_\nu(X)=(1-\nu)\sup_{x\in X} f_\nu(x)+\nu\int_X f_\nu(x)\,dx.$$

Утверждение 1. Пусть X — произвольное множество, $A\subset X$, а $g_\nu:\ \mathscr{B}\to[0,\ 1]$ является g_ν-мерой. Тогда для $\bar{A}=X\setminus A$ мера нечеткости примет вид

$$g_\nu(\bar{A})=\begin{cases}\max\left((1-g_\nu(A))/\nu,\ 1-\nu g_\nu(A)\right), & \text{если}\ \nu>1;\\ \min\left((1-g_\nu(A))/\nu,\ 1-\nu g_\nu(A)\right), & \text{если}\ \nu\in[0,\ 1].\end{cases}$$

Доказательство. Поскольку $\forall a,\ b\in\mathscr{R}:\ a\vee b=(a+b)/2+(a-b)/2$, то условие нормировки для $A\subset X$ и $A=X\setminus A$ примет вид:

$$(1-\nu)\left((g_\nu(A)-g_\nu(\bar{A}))/2+(g_\nu(A)+g_\nu(\bar{A}))/2\right)+\nu(g_\nu(A)+g_\nu(\bar{A}))=1.$$

Если $g_\nu(A)\geqslant g_\nu(\bar{A})$, тогда $g_\nu(\bar{A})=(1-g_\nu(A))/\nu$, а при $g_\nu(A)<g_\nu(\bar{A})$, $g_\nu(\bar{A})=1-\nu g_\nu(A)$. Для случая $\nu>1$ условие нормировки имеет силу, если $g_\nu(\bar{A})=\max\left((1-g_\nu(A))/\nu,\ 1-\nu g_\nu(A)\right)$, а для $\nu\in[0,\ 1]$ $g_\nu(\bar{A})=\min\left((1-g_\nu(A))/\nu,\ 1-\nu g_\nu(A)\right)$.

Утверждение 2. Пусть X — произвольное множество, $\mathscr{B}$ — борелевская σ-алгебра, $g_\nu:\ \mathscr{B}\to[0,\ 1]$ — нечеткая g_ν-мера. Тогда $\forall A,\ B\in\mathscr{B}:\ A\cap B\neq\varnothing$,

$$g_\nu(A\cup B)=(1-\nu)\left(g_\nu(B)\vee g_\nu(C)\right)+\nu\left(g_\nu(B)+g_\nu(C)\right),$$

где $C = A \setminus (A \cap B)$;

$$g_\nu(C) = \begin{cases} \min\left((g_\nu(A) - g_\nu(A \cap B))/\nu,\ g_\nu(A) - \nu g_\nu(A \cap B)\right), & \text{если } \nu \in [0, 1]; \\ \max\left((g_\nu(A) - g_\nu(A \cap B)/\nu,\ g_\nu(A) - \nu g_\nu(A \cap B)\right), & \text{если } \nu > 1. \end{cases}$$

Доказательство. Условие нормировки для $A,\ B \in \mathcal{B}$ относительно $g_\nu(A)$ имеет вид

$$g_\nu(A) = (1-\nu)\left(g_\nu(A \setminus (A \cap B)) \vee g_\nu(A \cap B)\right) + \nu\left(g_\nu(A \setminus (A \cap B)) + g_\nu(A \cap B)\right),$$

если $g_\nu(A \setminus (A \cap B)) \geq g_\nu(A \cap B)$, тогда $g_\nu(A \setminus (A \cap B)) = (g_\nu(A) - g_\nu(A \cap B))/\nu$, аесли $g_\nu(A \setminus (A \cap B)) < g_\nu(A \cap B)$, то $g_\nu(A \setminus (A \cap B)) = g_\nu(A) - \nu g_\nu(A \cap B)$.

Нетрудно увидеть, что если $\nu > 1$, то

$$g_\nu(A \setminus (A \cap B)) = \max\left((g_\nu(A) - g_\nu(A \cap B))/\nu,\ g_\nu(A) - \nu g_\nu(A \cap B)\right),$$

а при $\nu \in [0, 1]$

$$g_\nu(A \setminus (A \cap B)) = \min\left((g_\nu(A) - g_\nu(A \cap B))/\nu,\ g_\nu(A) - \nu g_\nu(A \cap B)\right),$$

что доказывает утверждение.

Утверждения 1 и 2 справедливы только для конкретного разбиения множества на подмножества.

В завершение рассмотрения нечетких мер приведем их параметрическое представление, которое необходимо для общей систематизации пространств с нечеткой мерой.

Пространством с λ-нечеткой мерой называется математическая структура (X, φ, g_λ), функция нечеткой меры g_λ которой удовлетворяет условиям (4), (5), (6), (13). Пространство с λ-нечеткой мерой (X, φ, g_λ) в свою очередь допускает несколько конкретизации в зависимости от значения параметра λ. Наиболее важными с точки зрения классификации пространств с нечеткой мерой являются случаи $\lambda \in (-1;0)$, $\lambda=1$ и $\lambda \in (0;+\infty)$, для которых получаются конкретизации введенных в рассмотрение пространств с нечеткой мерой.

Пространством с λ-нечеткой мерой доверия называется математическая структура (X, φ,), функция нечеткой меры которой удовлетворяет условиям (4), (5), (6) и дополнительному условию (13) при $\lambda \in (0,+\infty)$. Пространством с λ-нечеткой мерой правдоподобия называется математическая структура (X, φ, g^*_λ), функция нечеткой меры g^*_λ которой удовлетворяет условиям (4), (5), (6) и дополнительному условию (13) при (-1, 0). Следует заметить, что вероятностное пространство (X, φ, p) является также пространством с λ-нечеткой мерой при $\lambda=0$. Рассмотренные выше пространства с

различными вариантами нечетких мер представляют собой математические структуры, каждую из которых можно считать некоторым абстрактным классом. Важной особенностью подобной точки зрения является определенная взаимосвязь классов пространств с неопределенностью, основанная на их конкретизации посредством усиления соответствующих аксиом.

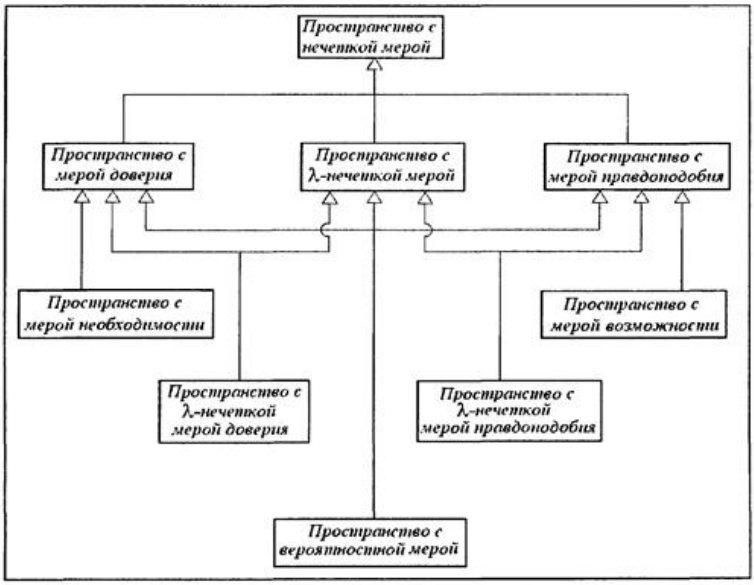


Рис. 3

Таким образом, может быть получена классификация различных классов пространств с нечеткой мерой, которая представлена на рис.3 в форме диаграммы классов языка UML. На этой диаграмме базовая математическая структура пространства с нечеткой мерой может использоваться в качестве своеобразного шаблона, параметрами которого являются аксиомы рассмотренных нечетких мер. При этом отношение обобщения соответствует усилению отдельных аксиом. Если между двумя математическими структурами пространств с нечеткой мерой имеется отношение обобщения, то математическая

структура нижнего уровня является конкретизацией соответствующей математической структуры верхнего уровня. При этом все математические структуры нижних уровней наследуют аксиоматику соответствующих математических структур верхних уровней. Так, например, математическая структура пространства с мерой необходимости является конкретизацией или частным случаем математической структуры пространства с мерой доверия. Математическая структура вероятностного пространства является конкретизацией как математической структуры пространства с мерой доверия, так и математической структуры пространства с мерой правдоподобия, тем самым реализуется так называемое множественное наследование . Указанные взаимосвязи имеют существенное значение при рассмотрении математических структур пространств с нечеткой мерой.

2.3. Особенности аппроксимации нечетких мер

При решении практических задач моделирования нечетких систем с использованием аппарата теории нечетких мер возникает необходимость оперирования большими объемами нечетких данных. Поэтому для упрощения вычислительных алгоритмов на ЭВМ удобно аппроксимировать нечеткие меры. Для этой цели можпо использовать *(L — R)*-функции.

Определение 2. Функция, обозначаемая *L* (или *R),* является *функцией (L — R)-типа* тогда и только тогда, когда

$$\forall x \in \mathscr{R}^{+} \triangleq [0, +\infty): \; L(-x) = L(x); \; L(0) = 1,$$

$L(\cdot)$ монотонно убывает на $\mathscr{R}^{+}$.

Пример: $L_1(x) = \max(0, \; 1 - |x|^{p}); \quad L_2(x) = \exp(-|x|^{p}); \; p \geq 1.$

Особенно удобно использовать *(L — R)*-функции в случав g_λ -меры Сугено.

При этом функция *h(x)* может быть представлена как

$$h(x) = L((a - x)/\beta \vee 0); \quad x \in X \subset \mathscr{R}, \quad (24)$$

где *a* — параметр, при котором *h(x)=1,* β — коэффициент нечеткости. Пример функции *(L — R)* -типа, аппроксимирующей функцию распределения нечеткости, приведен на рнс. 4.

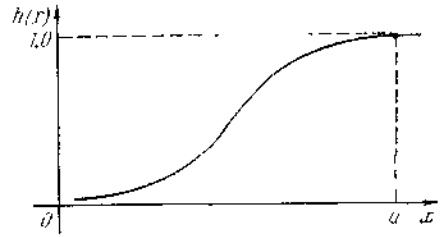


Рис. 4. Функция (*L — R*)-типа, аппроксимирующая функцию распределения нечеткости

Рассмотрим особенности процедуры приближения экспериментальных функций распределения нечеткости функциями (*L — R*)-типа. Пусть в результате формализации некоторой выборки нечетких данных получен ряд экспериментальных значений плотности распределения нечеткости $g_1', g_2', \ldots, g_n'$, которым соответствуют

$$x_1, x_2, \ldots, x_n; \; x_i \in X \subset \mathcal{R}, \; t \in I = \overline{(1, n)}.$$

Все множество *X* можно разбить на подынтервалы таким образом, что $X = \bigcup_{i \in I} \Delta_i; \; x_i \in \Delta_i; \; \Delta_i = [x_i - \Delta/2, \; x_i + \Delta/2]$, где $\Delta \in \mathcal{R}$,

Δ — длина подынтервала; $\Delta = (b - a)/(n - 1); \; a = \inf X$,

$b = \sup X$.

Значение плотности распределения нечеткости в *i*-й точке интервала [a, *b]*, определенное экспериментально, можно приближенно определять как $g_\lambda(\Delta_i) \approx g_\lambda(\{x_i\})$. Нечеткие меры для подиптервалов Δ_i можно вычислять, используя *(L — R)* -аппроксимацию функции распределения нечеткости. При этом

$$g_\lambda(\Delta_i) = L\left(\frac{2(a - x_i) + \Delta}{2\beta}\right) - L\left(\frac{2(a - x_i) - \Delta}{2\beta}\right)\left(1 + \lambda L\left(\frac{2(a - x_i) - \Delta}{2\beta}\right)\right).$$

Пусть $S \subset \{\Delta_i | \Delta_i \subset [a, \; b]\}$ — множество подынтервалов множества X, $\mathcal{P}(S)$ — множество всех подмножеств множества подынтервалов. Нетрудно увидеть, что $\forall A \in \mathcal{P}(S)$: $\Delta_i \in A \subset S$;

$$g_\lambda(A) = \frac{1}{\lambda}\left[\prod_{i \in \theta}(\lambda g_\lambda(\Delta_i) + 1) - 1\right],$$

где $\theta = \{i | x_i \in \Delta_i \in A\}$; $\Delta_i = [x_i - \Delta/2, \; x_i + \Delta/2]$. Нечеткой мере $g_\lambda(A)$ будет соответствовать нечеткая мера $g_\lambda'(A)$, полученная из эксперимента при формализации нечеткой плотности:

$$g_\lambda'(A) = \frac{1}{\lambda}\left[\prod_{i \in \theta}(\lambda g_i' + 1) - 1\right]; \quad A \in \mathcal{B}.$$

Параметр λ определяется из условия нормировки:

$$g'_\lambda(X) = \frac{1}{\lambda}\left[\prod_{i\in\theta}(\lambda g'_i + 1) - 1\right] = 1.$$

Таким образом, задача *(L — R)* -аппроксимации функции распределения нечеткости сводится к оценке параметров *a* и β *(L — R)* -функции по минимуму функционала качества

$$\mathscr{H} = \left\{\sum_{A\in\mathscr{P}(S)}\left(g_\lambda(A) - g'_\lambda(A)\right)^2\right\}^{1/2} \to \min. \quad (25)$$

При большем количество экспериментальных точек минимизация функционала (25) становится затруднительной. В этом случае можно воспользоваться приближенной процедурой, смысл которой заключается в использовании только части множества подмножеств подынтервалов $\mathscr{P}(S)$ для оценки *a* и β. При этом

$$\mathscr{H} = \frac{1}{n-1}\left(\sum_{i=1}^{n-1}\left(\left(\frac{1}{\lambda}\left(\prod_{k=1}^{i}(\lambda g_\lambda(\Delta_i) + 1) - 1\right) - \right.\right.\right.$$

$$\left.\left.\left. - \frac{1}{\lambda}\left(\prod_{k=1}^{i}(\lambda g'_i + 1) - 1\right)\right)^2\right)\right)^{1/2} \to \min. \quad (26)$$

Задачу можно упростить, если параметр *a* определять непосредственно по результатам эксперимента. Можно показать, что если $\lambda \to -1$, то функция множества $g_\lambda(\{(-\infty,\ x]\}) = 1$ при $x = x^*$, где $x^* = \arg\sup_{i\in N} g(\{x_i\})$. Таким образом, для определения *a*, при $\lambda = -1$, достаточно найти минимальное значение $x \in [a,\ b]$, при котором нечеткая плотность равна 1. Если $\lambda > -1$, тогда $a = \sup X$. В этом случае параметр β может быть легко найден при помощи любой процедуры численной минимизации.

Решение многих задач нахождения значения g_ν-меры для случаев множества действительных чисел может выглядеть сравнительно просто, если применять аналитическую аппроксимацию нечетких плотностей, с помощью которых задаются g_ν-меры. Такая аппроксимация может быть сделана с помощью аналога *(L — R)* -функций — функций *(S — L)* -типа.

Определение 3. Функция, обозначаемая $SL(\cdot)$, является функцией *(S — L)* -типа тогда и только тогда, когда

$$\forall x \in \mathscr{R}^+:\ SL(-x) = SL(x);\quad SL(0) = S,$$

причем $SL(\cdot)$ — монотонно убывает на $\mathscr{R}^+$: $S \in [0,\ 1]$.

Пример: $SL(x) = S\max(0,\ 1 - |x|^p);\ SL(x) = S\exp(-|x|^p);$ *p≥1.*

Определение 4. Нечеткой плотностью *SL-типа.* называется нечеткая плотность $g'\colon X \to [0,\ 1]$ такая, что

$$g'(x) = \begin{cases} SL'\left(\frac{a'-x}{\underline{a}}\right) & \text{при } x \leqslant a',\ \underline{a} \geqslant 0; \\ SL''\left(\frac{x-a''}{\bar{a}}\right) & \text{при } x > a'',\ \bar{a} \geqslant 0; \\ S, & \text{если } x \in [a', a''] \subset \mathcal{R}; \end{cases}$$

где $\bar{a}$, $\underline{a}$ — правый и левый коэффициенты нечеткости, $L'(\cdot)$, $L''(\cdot)$ — функции $(L-R)$-типа.

Очевидно, что если $L' \triangleq L'' = L$, то

$$g'(x) = SL\left(\frac{a'-x}{\underline{a}} \vee \frac{x-a''}{\bar{a}} \vee 0\right) \triangleq L(a(x)).$$

Можно показать, что $\forall [a, b] \subset X \subset \mathcal{R}$

$$g_\nu([a, b]) = \sup_{x \in [a,b]} g'(x) \cdot (1-\nu) + \nu \int_a^b g'(x)\,dx =$$
$$= S\left((1-\nu)L\left(\inf_{x \in [a,b]}\left(\frac{a'-x}{\underline{a}} \vee \frac{x-a''}{\bar{a}} \vee 0\right)\right) + \nu \overline{L}_a^b\right),$$

где $\overline{L}_a^b \triangleq \int_a^b L((a'-x)/\underline{a} \vee (x-a'')/\bar{a} \vee 0)dx$. Нетрудно увидеть, что

$$\inf\left(\frac{a'-x}{\underline{a}} \vee \frac{x-a''}{\bar{a}} \vee 0\right) = \begin{cases} 0, & \text{если } [a', a''] \cap [a, b] \neq \varnothing; \\ \frac{a'-b}{\underline{a}}, & \text{если } b \leqslant a'; \\ \frac{a-a''}{\bar{a}}, & \text{если } a \geqslant a''. \end{cases}$$

Рассмотрим особенности приближения экспериментальных g_ν-мер аналитическими выражениями с помощью функций $(S-L)$-типа. Аналогично вышеизложенному будем предполагать, что имеется экспериментальная последовательность значений нечеткой плотности $g'_1, g'_2, \ldots, g'_n$. Используя аналогичные обозначения для подынтервалов, можно предположить, что нечеткая мера на элементарном подынтервале равна значению нечеткой плотности в точке, принадлежащей этому подынтервалу, т. е. $g_\nu(\Delta_i) \approx \approx g'_\nu(\{x_i\})$, где $g_\nu(\Delta_i)$ — нечеткая мера, задаваемая аналитически для Δ_i, $g_\nu(\{x_i\}) = g'_i$. В случае $(S-L)$ -аппроксимации получаем

$$g'_\nu(\Delta_i) = S\left((1-\nu)\sup_{x \in \Delta_i} L(a(x)) + \nu \int_{\Delta_i} L(a(x))\,dx\right).$$

Параметр S определяется как

$$S = \arg \sup_{x \in X} g'_v(\{x_i\}).$$

Параметр нормировки g_v-меры v может быть найден из условия нормировки (23) по формуле

$$v = \left(1 - \bigvee_{i=1}^{n} g'_i\right) \Big/ \left(\sum_{i=1}^{n} g'_i - \bigvee_{i=1}^{n} g'_i\right).$$

Оценка параметров *(L — R)* -функций может быть проведена аналогично (25). При этом

$$\mathcal{H} = \left(\sum_{A \in \mathcal{P}(S)} \left(g_v(A) - g'_v(A)\right)^2\right)^{1/2} \to \min, \tag{27}$$

где

$$g_v(A) = S\left((1 - v) \bigvee_{i \in \theta} g_v(\Delta_i) + v \sum_{i \in \theta} g_v(\Delta_i)\right),$$

$$g'_v(A) = S\left((1 - v) \bigvee_{i \in \theta} g'_i + v \sum_{i \in \theta} g'_i\right),$$

$$\theta = \{i \mid x_i \in \Delta_i \in A\}.$$

Когда минимизация функционала (27) затруднительна, можно воспользоваться приближенной процедурой, аналогичной (26). При этом

$$\mathcal{H} = \frac{1}{n-1}\left(\sum_{i=1}^{n-1}\left(S\left((1 - v) \sup_{\Delta_k \in D_i} L(a(x)) + v \int_{D_i} L(a(x))\,dx\right) - \right.\right.$$

$$\left.\left. - (1 - v) \sup_{k \in \overline{(1,i)}} g'_k + v \sum_{k=1}^{i} g'_k\right)^2\right)^{1/2} \to \min,$$

где

$$D_i = \bigcup_{k \in \overline{(1,i)}} \Delta_k.$$

Впростейшем случае, оценивание параметров *(S — L)* -функций следует производить, используя функционал следующего вида:

$$\mathcal{H} = \left(\sum_{x_i \in X} \left(S L(a(x_i)) - g'_i(\{x_i\})\right)^2\right)^{1/2}.$$

Рассмотренные методы аппроксимации позволяют значительно упростить процедуры вычисления нечетких мер при определении значений нечетких интегралов в различных алгоритмах. (Здесь и далее под нечеткими интегралами понимаются некоторые монотонные и, вообще говоря, нелинейные функционалы, определяемые на базе нечетких мер.)

Кроме того, при использовании *SL*- и *(L — R)*-аппроксимаций можно значительно сократить объем памяти ЭВМ, необходимый для хранения информации о функциях распределения нечеткости.

2.4. Нечеткие интегралы

Определение 5. *Нечеткий интеграл* от функции $h;\ X \to [0, 1]$ на множество $A \subseteq X$ по нечеткой мере g определяется как

$$\int_A h(x) \circ g = \sup_{\alpha \in [0,1]} (\alpha \wedge g(A \cap H_\alpha)), \quad (28)$$

где $H_\alpha = \{x | h(x) \geqslant \alpha\}$. Нечеткий интеграл принято также называть нечетким ожиданием или FEV (fuzzy expected value) . Пусть $\mathcal{F}(X)$— множество нечетких подмножеств базового множества *X*. Поскольку понятие нечеткого подмножества включает в себя понятие обычного иодмпожостиа, то $\mathcal{F}(X)$ является нечетким расширением $\mathcal{B}$: $\mathcal{F}(X) \supset \mathcal{B}$.

Определение 6. Функция множества $\tilde{g}$, определяемая в виде

$$\tilde{g}(A) = \int \mu_A \circ g \quad (29)$$

для $A = \{(x,\ \mu_A(x))\},\ \mu_A \in \mathcal{F}(X)$, называется расширением g на $\mathcal{F}(X)$.

Определение 7. Нечеткий интеграл от функции h: $X \to$ $\to [0,\ 1]$ на нечетком множестве $\mu_A \in \mathcal{F}(X)$ по нечеткой мере g определяется как

$$\int_{\mu_A} h(x) \circ g = \int_X (\mu_A(x) \wedge h(x)) \circ g. \quad (30)$$

Для описания различных видов неопределенности в теории нечетких мер используется общее понятие «степень нечеткости». В общем случае это понятие включает в себя «степень важности», «степень уверенности» и как отдельный случай «степень принадлежности» в теории НМ. Нечеткая мера, таким образом, может интерпретироваться различными способами в зависимости от конкретного применения. Пусть необходимо оценить степень принадлежности некоторого элемента $x \in X$ множеству $E \subset X$.

Очевидно, что для пустого множества эта степень принадлежности равна 0, а для $x \in F$ $(F \supset E)$ равна 1, т. е. степень принадлежности для $x \in F$ будет больше, чем для $x \in E$, если $E \subset F$. Если степень принадлежности $x_0 \in E$ равна $g(x_0,\ E)$, а вместо *E* задано нечеткое подмножество $\mu_A \in \mathcal{F}(X)$, то

$$g(x_0, A) = \int\limits_X \mu_A(x) \circ g(x_0, \cdot) = \mu_A(x_0). \quad (31)$$

Это говорит о том, что степень нечеткости суждения «$x_0 \in A$» равна степени принадлежности x_0 нечеткому подмножеству μ_A. Таким образом, понятие степени нечеткости в теории нечетких мер включает в себя понятие степени принадлежности теории НМ. Отметим основные свойства нечетких интегралов (НИ).

Пусть $a \in [0, 1]$, $(E, F) \subseteq X$. Тогда, если $h: X \to [0, 1]$, то:

$$\int\limits_E (a \vee h) \circ g = a \vee \int\limits_E h \circ g;$$

$$\int\limits_E (a \wedge h) \circ g = a \wedge \int\limits_E h \circ g;$$

$$\int\limits_E (h_1 \wedge h_2) \circ g \leqslant \int\limits_E h_1 \circ g \wedge \int\limits_E h_2 \circ g;$$

$$\int\limits_E (h_1 \vee h_2) \circ g \geqslant \int\limits_E h_1 \circ g \vee \int\limits_E h_2 \circ g;$$

$$\int\limits_{E \cup F} h \circ g \geqslant \int\limits_E h \circ g \vee \int\limits_E h \circ g;$$

$$\int\limits_{E \cap F} h \circ g \leqslant \int\limits_E h \circ g \wedge \int\limits_F h \circ g.$$

Кроме того,

$$\int\limits_A h \circ g = M$$

тогдаи только тогда, когда $g(A \cap F_M) \geqslant M \geqslant g(A \cap F_{M+0})$, где $F_M = \{x | h \geqslant M\}$ и $F_{M+0} = \{x | h > M\}$.

Можно показать, что понятие НИ сходно с понятием интеграла Лебега. Для этого рассмотрим разбиение множества X на непересекающиеся подмножества E_i: $X = \bigcup\limits_{i=1}^{n} E_i$, $E_i \cap E_j = \varnothing$, $i \neq j$, $i = 1, ..., n$. Пусть

$$h(x) = \sum_{i=1}^{n} \alpha_i \cdot f_{E_i}(x),$$

где $\alpha_i \in [0, 1]$, $E_i \in \mathscr{B}$, а f_{E_i} — характеристическая функция обычного множества E_i, т. е. $f_{E_i}(x) = 1$, если

$x \in E_i$, и $f_{E_i}(x) = 0$, если $x \notin E_i$. Пусть l есть мера Лебега. Интеграл Лебега от функции h по множеству A определяется как

$$\int_A h\, dl = \sum_{i=1}^{n} \alpha_i l(A \cap E_i), \quad (32)$$

где $i \in I = (1, 2, \ldots, n)$; $\alpha_1 \leqslant \alpha_2 \leqslant \ldots \leqslant \alpha_n$. Предположим, что $F_i = E_i \cup E_{i+1} \cup \ldots \cup E_n$. Тогда, определяя h и виде $h(x) = \max\limits_{i=\overline{1,n}} \min(\alpha_i, f_{F_i}(x))$, полупаем следующее выражение для НИ:

$$\int_A h(x) \circ g(\cdot) = \max_{i=\overline{1,n}} \min(\alpha_i, g(A \cap F_i)). \quad (33)$$

Оба интеграла — лебегов и нечеткий — можно сравнить, используя вероятностную меру. Если $(X, \mathcal{B}, P)$— вероятностное пространство, а $h: X \to [0, 1]$ есть $\mathcal{B}$-измеримая функция, то имеем, что

$$\left| \int_X h(x) \circ P(\cdot) - \int_X h(x)\, dP \right| \leqslant \frac{1}{4}. \quad (4.34)$$

В теории НИ имеет место следующая теорема.

Теорема 1. Пусть $(Y, \mathcal{B}_Y, g_Y)$ и $(X, \mathcal{B}_X, g_X)$_ пространства с нечеткими морами g_Y и g_X соответственно; $h: X \times Y \to [0, 1]$, $x \in X$, $y \in Y$. Тогда если $g = g_X \times g_Y$, то для $Z = X \times Y$:

$$\int_Z h(x,y) \circ g = \int_Y \left(\int_X h(x, y) \circ g_X \right) \circ g_Y. \quad (35)$$

Данная теорема является аналогом теоремы Фубини из теории меры и называется теоремой Сугено — Фубини.

Пусть $\{h_n\}$ — монотонная последовательность $\mathcal{B}$-измеримых функций, тогда

$$\int \lim_{n \to \infty} h_n \circ g = \lim_{n \to \infty} \int h_n \circ g.$$

Если h_n — монотонно возрастающая (убывающая) последовательность $\mathcal{B}$-измеримых функции и $\{a_n\}$— монотонно убывающая (возрастающая) последовательность вещественных чисел, то

$$\int \left[\bigvee_{n=1}^{\infty} (a_n \wedge h_n) \right] \circ g = \bigvee_{n=1}^{\infty} \left[a_n \wedge \int h_n \circ g \right].$$

На рис. 5 дан пример графической интерпретации НИ для $X = \mathcal{R}$, где $S = \int_A h(x) \circ g = \sup_{\alpha \in [0,1]} [\alpha \wedge g(H_\alpha \cap A)]$; $H_\alpha = \{x | h(x) \geq \alpha\}$; $f(x)$ — нечеткая плотность.

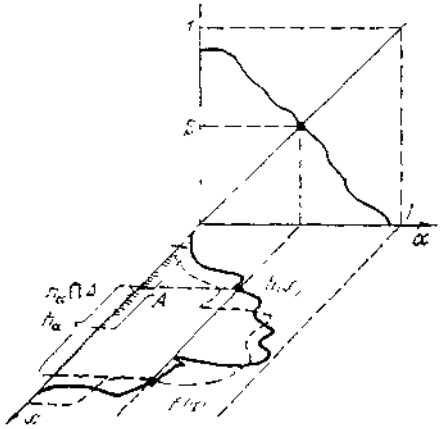


Рис. 5. Графическая пнтерпретация для нечеткого интеграла

Пусть $\varphi: X \to Y$, тогда борелевская σ-алгебра $\mathcal{B}^{(\varphi)}$ и нечеткая мера $g^{(\varphi)}$ индуцируются из X в Y. То ость $F \in \mathcal{B}^{(\varphi)}$ тогда и только тогда, когда $\varphi^{-1}(F) \in \mathcal{B}$, $g^{(\varphi)}(F) = g(\varphi^{-1}(F))$.

Пространство с нечеткой мерой $(Y, \mathcal{B}^{(\varphi)}, g^{(\varphi)})$ интерпретируется следующим образом. Если Y связано с X с помощью отображения φ, тогда нечеткая мера на Y, с помощью которой измеряется степень нечеткости в X, также связана с мерой нечеткости в X.

Пусть $E \in \mathcal{B}$ и $F \in \mathcal{B}^{(\varphi)}$. Обозначим через $\rho(E|\varphi = y)$ семейство всех функций, эквивалентных $h(y)$ по отношению g:

$$g(E \cap \varphi^{-1}(F)) = \int_F h(y) \circ g^{(\varphi)}.$$

Здесь $\rho(\cdot|\varphi = y)$ называется условной нечеткой мерой при условии $\varphi = y$.

Пусть $F = Y$, тогда $g(E) = \int_Y \rho(E|\varphi = y) \circ g^{(\varphi)}(\cdot)$.

Условная нечеткая мера обладает следующими свойствами:

1) Для фиксированных $E \in \mathcal{B}$, $\rho(E|\varphi = y)$ как функция от y яилятся $\mathcal{B}^{(\varphi)}$-измеримой.

2) Для фиксированных y, $\rho(\cdot|\varphi = y)$ являются нечеткой мерой для $(X, \mathscr{B})$ в смысле $g^{(\varphi)}$.

Если два пространства с нечеткими мерами $(X, \mathscr{B}_X, g_X)$ и $(Y, \mathscr{B}_Y, g_Y)$ связаны друг с другом, то отображение φ нельзя определить в общем случае. Далее условную нечеткую меру $\rho(\cdot|\varphi = y)$ будем обозначать как $\rho_X(\cdot|y)$.

В этом случае будет справедливо $g_X(\cdot) = \int\limits_Y \rho_X(\cdot\,|\,y) \circ g_Y$.

Если заданы нечеткие меры g_X, $\rho_Y(\cdot|x)$, g_Y, то существует нечеткая условная мера $\rho(\cdot|y)$ такая, что

$$\rho_Y(F\,|\,x) \circ g_X = \int\limits_Y \rho_X(E\,|\,y) \circ g_Y.$$

Данное уравнение соответствует байесовской формуле определения апостериорной вероятности в этом смысле $\rho_X(\cdot|y)$, которая называется апостериорной нечеткой мерой, а g_x — априорной нечеткой мерой.

В качестве примеров рассмотрим вычисления НИ для счетных множеств в случаях g_λ- и g_v-мер.

Пример. Пусть задано пятиэлементное счетное множества $X = \{x_i\}\quad i \in \overline{(1, 5)} \triangleq I$. Каждому элементу $x_i \in X$ соответствуют значения нечетких плотностей g_i из табл. 1.

Таблица 1

i	1	2	3	4	5
g_i	0,170	0,257	0,216	0,212	0,061
$h(x_i)$	0,5	0,7	0,1	0,2	0,3

Согласно условию нормировки для g_v-меры получаем $\lambda = 0{,}25$.

Значение НИ $S = \sup\limits_{\alpha \in [0,1]} [\alpha \wedge g_\alpha]$, где

$$g_\alpha = \frac{1}{\lambda}\left(\prod_{i \in \theta_\alpha}(\lambda g_i + 1) - 1\right);\ \theta_\alpha = \{i\,|\,h(x_i) \geqslant \alpha\},$$

принимает величину $S = 0{,}4379$.

Для g_v-меры из условия нормировки можно получить

$$v = \left(1 - \bigvee_{i=1}^{5} g_i\right) \Big/ \left(\sum_{i=1}^{5} g_i - \bigvee_{i=1}^{5} g_i\right) = 1{,}127.$$

При этом для g_v-меры $S = 0{,}448$.

2.5. Применение нечетких мер и интегралов для решения слабо структурированных задач

2.5.1. Процесс субъективного оценивания.

Рассмотрим задачу субъективного оценивания некоторым индивидом нечетко описываемых объектов, например, дом, лицо и т. д.. Предположим, что объект характеризуется *n* показателями.

Пусть $K = \{s_1, \ldots, s_n\}$— множество показателей. При оценивании дома такими показателями могут быть: $s_1 \triangleq$ площадь, $s_2 \triangleq$ удобства и т. д., а для лица $s_1 \triangleq$ глаза, $s_2 \triangleq$ нос и т. д.

В общем случае множество *K* необязательно должно быть множеством физических показателей, оно может быть множеством мнений, критериев и т. д. Пусть $h: K \to [0, 1]$ — частная оценка объекта, т. е. *h(s)* —оценка элемента *s.* Если речь идет о распознавании образов, то *h(s)* может рассматриваться как характеристическая функция образа. На практике *h (s)* может быть легко определена объективно или субъективно.

Например, когда объект — дом, объективно имеем оценку $h(s_1) = h$ (площадь) = 800 м^2, которая может быть нормализована числом из интервала [0, 1]. Для лица мы можем пользоваться лишь субъективной оценкой индивида; например, *h* (глаза) = 0,7.

Предположим, что нечеткая мера для $(K, 2^K)$ является субъективной мерой, выражающей степень важности подмножества из *K.*

Например, $g(\{s_1\})$ выражает степень важности элемента s_1 при оценке объекта, $g(\{s_1, s_2\})$— аналогично обозначает степень важности показателей s_1 и s_2. Необходимо отметить, что степень важности всего множества *K* равна единице.

Вычисляя НИ от *h* до *g* получаем:

$$e \triangleq \int_K h(s) \circ g, \qquad (36)$$

где *e* — обобщенная оценка объекта.

Уравнение (36) представляет собой свертку *n* частных оценок. Линейный обобщенный критерий используется обычно в том случае, когда отдельные показатели взаимно независимы. Свертка () может быть очень полезной, когда существует взаимозависимость показателей, что характерно для большинства задач выбора в нечеткой обстановке.

Процесс субъективного оценивания объектов предполагает идентификацию самой нечеткой меры.

2.5.2. Экспериментальное определение нечеткой меры.

Рассмотрим метод приближенного экспериментального определения нечетком меры . Предположим, что существует *m* объектов. Пусть $h_j\colon K \to [0, 1]$ —частная оценка *j*-го объекта, а e_j — общая оценка, получаемая из (17). Предъявляя индивиду объекты и их частные оценки, можно получить его субъективные оценки *d,* из интервала [0, 1] для всех объектов.

Обозначим $\bar{e} = \max\{e_j\}$; $\underline{e} = \min\{e_j\}$ и аналогично $\bar{d}$ и $\underline{d}$.

Производи нормализацию $e_j \forall j \in \overline{[1, m]}$, мы имеем

$$w_j = \frac{\bar{d} - d}{\bar{e} - \underline{e}} e_j + \frac{d\bar{e} - \bar{d}e}{\bar{e} - \underline{e}}.$$

Субъективная нечеткая мера может быть получена при условии минимума критерия

$$J = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(d_j - w_j)^2}. \quad (37)$$

Для простоты предполагается, что *g* в (36) удовлетворяет λ-правилу. Впервые нечеткие меры применялись для оценки сходства одномерных образов. Рассматривалось решение задачи оценки домов. При этом дома оценивались по следующим пяти показателям: площадь, удобства и обстановка, окружающая среда, стоимость, время, требуемое на дорогу до места работы. Известны применения нечетких мер для оценки прпвлекательности экскурсионных районов , которые оценивались по таким показателям как красота природы, архитектурные памятники и т. д. Результаты оцепок использовались для предсказания увеличения экскурсий в ближайшие десять лет. Интересное решение задачи информационного поиска с применением нечетких мер рассмотрено применительно к библиотечной информационно-поисковой системе.

2.5.3. Принятие решения в нечеткой обстановке.

Рассмотрим пример использования условных нечетких мер для решения задачи принятия решения в нечеткой обстановке . Процесс принятия решения описывается шестеркой

$$\langle \theta, X, A, g_\theta(\cdot), \sigma_x(\cdot|\theta), l \rangle,$$

где θ — множество показателей, характеризующих оцениваемый объект *x*;

X — множество оцениваемых объектов $x \in X$;
g_θ — нечеткая мера степени важности показателей;
$\sigma_X(\cdot|\vartheta)$ — нечеткая мера привлекательности объектов из X при их оценке с точки зрения показателя $\vartheta \in \theta$;
Y — множество действий покупателя;
l — функция принадлежности нечеткого отношения на декартовом произведении $\theta \times Y$, обозначающая нечеткие потери, когда действие $y \in Y$ выбирается для $\vartheta \in \theta$. Задача заключается в поиске стратегии, которая минимизирует нечеткое ожидание функции потерь. При этом нечеткое действие A имеет функцию принадлежности $\mu_A: Y \to [0, 1]$, а нечеткая стратегия B, являющаяся нечетким отношением па декартовом произведении $X \times A$, имеет функцию принадлежности $\mu_B: X \times Y \to [0, 1]$. Нечеткое действие A, основанное на нечеткой стратегии B, определяется с помощью функции принадлежности $\mu_{B(x)}(y) = \mu_B(x, y)$. Нечеткие потери для нечеткого действия определяются через функцию принадлежности

$$l(\vartheta, A) \triangleq 1 - \max_{y \in Y} (\mu_A(y) \wedge (1 - l(\vartheta, y))). \quad (38)$$

Если ЛПР выбирает нечеткую стратегию B, то нечеткое ожидаемое значение потерь примет вид

$$\langle l \rangle_B = \int_\theta \left[\int_X l(\vartheta \mid B(x)) \circ \sigma_X(\cdot \mid \vartheta) \right] \circ g_\theta.$$

Решением задачи принятия решения будет

$$\mu_{L^0}(x, y) = \int_\theta (1 - l(\vartheta, y)) \circ \sigma_\theta(\cdot \mid x),$$

где $\sigma_\theta(\cdot|x)$ — апостериорная нечеткая мера.

Данный подход может быть использован для широкого класса задач принятия решения в нечеткой обстановке. Следует отметить, что при небольшом количестве элементов множества θ нечеткая мера g_θ может быть идентифицирована точным методом. Для идентификации нечеткой меры в этом случае эксперимент должен дать оценки степени важности всех подмножеств из θ, т. е. необходимо иметь субъективные оценки d такие, что $d: 2^\theta \to [0, 1]$. Идентификация нечеткой меры заключается в минимизации функционала

$$J = \sqrt{\frac{1}{|2^\theta|} \sum_{E \in 2^\theta} (d_\theta(E) - g_{i,\theta}(E))^2}, \quad (39)$$

где $|2^{\theta}| \triangleq \text{card}\, 2^{\theta}$ — мощность множества 2^{θ}, а $g_{\theta}(E)$ вычисляется так же, как в п. 2.3. Результатом решения задачи (39) является значение параметра λ и нечетких плотностей $g_{\theta_1}, g_{\theta_2}, \ldots$ $\ldots, g_{\theta_n}$; $n = \text{card}\, \theta$. Опыт рассмотрения задач принятия решения показывает, что значение λ на практике бывает или положительным или отрицательным числом, но не близким или равным нулю.

Еще один из вариантов применения нечетких мер и интегралов в задаче принятия решений предложен в литературе. В этом случае предпочтения ЛПР описываются с помощью логико-лингвистической модели, т. е. схемы нечетких рассуждений вида $\mathbf{C} \Rightarrow \mathbf{U}$, где $\mathbf{C} = [\mu_{ij}]$ — матрица нечетких множеств размером $n \times m$, соответствующая n значениям m лингвистических показателей, $\mathbf{U} = (\mu_{U_1}, \ldots, \mu_{U_n})^T$ — вектор нечетких множеств, характеризующих полезность. Выбор группой ЛПР рациональной альтернативы осущостиляется по критерию максимума значений НИ вида $D = \int u_{kr} \circ \omega$, где ω — нечеткая мера, характеризующая идеальную полезность, а u_{kr} — полезность k-й альтернативы для группы ЛПР. Последняя вычисляется по формуле $u_{kr} = \mathbf{M}[u_{kt}]$. Здесь **M** — оператор вычисления обобщенной меры средних , а u_{kt} — НМ, характеризующее полезность k-й альтернативы для t-го ЛПР.

2.5.4. Процесс обучения в нечеткой обстановке.

Одной из замечательных способностей человека является его способность обучаться в нечеткой обстановке. При обучении он успешно использует нечеткую информацию, которая во многих случаях является единственно доступной. В психологии по традиции используются стохастические модели обучаемости, например, модель Буша и Мостеллера, хотя ряд авторов экспериментально показал, что способность обучаться в вероятностной обстановке, как праинло, не свойственна человеку . Исходя из этой точки зрения, предложена модель обучения, которая, являясь структурным аналогом байесовской модели обучения, позволяет учитывать нечеткую информацию. Данная модель построена с помощью нечетких мер и использовалась для нахождения экстремумов миогоэкстремальных функций.

Пусть X — множество причин и Y — множество следствий; g_X и g_Y — нечеткие меры для X и Y соответственно. Пусть g_Y выражается НИ от $\sigma_Y(\cdot|x)$ по g_X как

$$g_Y(\cdot) = \int \sigma_Y(\cdot \mid x) \circ g_X, \quad (40)$$

где $\sigma_Y(\cdot|x)$ — есть условная нечеткая мера от Y по отношению к *X*. Физический смысл этого уравнения легко установить по аналогии с теорией вероятности: $g_Y(\cdot)$ соответствует вероятности $p(y)$, $y \in Y$, для случая, когда заданы вероятностная мера *p(x)* и условная вероятность $p(\cdot|x)$. Следует отметить, что определение $g_Y(\cdot)$ и математические свойства уравнения (40) совершенно отличаются от его вероятностного аналога.

Нечеткая мера g_x называется априорной нечеткой мерой, соответствующей степени нечеткости субъективной оценки суждения «один из элементов $E \subset X$ имеет место». Нечеткая мера $\sigma_Y(F|x)$, $F \subset Y$ является мерой нечеткости суждения «один из элементов $F \subset Y$ имеет место при заданном *x*».

Рассмотрим метод, позволяющий уточнять g_x в процессе получения новой информации, которая в общем случае выражается подмножеством $F \subset Y$. Эта информация может быть трех типов. Если *F* состоит лишь из одного элемента, то информация является детерминированной, а если несколько, то недетерминированной. Если *F* — нечеткое подмножество, то информация — нечеткая.

Пусть нечеткое множество $A \subset Y$ имеет функцию принадлежности $\mu_A: Y \to [0, 1]$. Нечеткая мера для нечеткого подмножества *A* определяется как

$$g_Y(A) = \int_Y \mu_A(y) \circ g_Y.$$

Здесь $g_Y(A)$ выражает степень нечеткости информации, содержащейся в *A*. Нетрудно показать, что

$$g_Y(A) = \int_Y \mu_A(y) \circ \left[\int_X \sigma_Y(\cdot \mid x) \circ g_X \right] = \int_X \sigma_Y(A \mid x) \circ g_X,$$

где

$$\sigma_Y(A \mid x) = \int_Y \mu_A(y) \circ \sigma_Y(\cdot \mid x).$$

После получения информации *A,* нечеткая мера g_x может быть уточнена таким образом, чтобы значение $g_Y(A)$ увеличилось. Если $g_X(\cdot)$ и $\sigma_Y(\cdot|x)$ удовлетворяют λ-правилу и $\sigma_Y(A|x_i)$— убывающая функция, то

$$g_Y(A) = \bigvee_{i=1}^{n} [\sigma_Y(A \mid x_i) \wedge g_X(F_i)],$$

где $F_i = \{x_1, x_2, \ldots, x_i\}$. Из литературы следует, что

$$g_Y(A) = \sigma_Y(A \mid x_l) \wedge g_X(F_l), \quad (41)$$

где l является наибольшим индексом, для которого имеет место (41) и выполняется условие

$$\sigma_Y(A \mid x_{l-1}) \wedge g_X(F_{l-1}) \leqslant \sigma_Y(A \mid x_l) \wedge g_X(F_l),$$
$$\sigma_Y(A \mid x_{l+1}) \wedge g_X(F_{l+1}) > \sigma_Y(A \mid x_l) \wedge g_X(F_l).$$

Обучение характеризуется возрастанием нечетких плотностей g_x, что приводит к увеличению $g_Y(A)$.

Пусть g_i, $i = 1, ..., n$, являются нечеткими плотностями для g_x . Тогда легко показать , что только g_i, $1 \leqslant i \leqslant l$ влияют на значения $g_Y(A)$, поэтому алгоритм обучения имеет вид

$$(g_X^i)' = \alpha g_X^i + (1 - \alpha)\sigma_Y(A \mid x_i) \quad \forall i = 1, \ldots, l,$$
$$(g_X^i)' = \alpha g_X^i \qquad\qquad\qquad\qquad \forall i = l + 1, \ldots, n,$$

где $\alpha \in (0, 1)$ — параметр, определяющий скорость сходимости. Следует отметить, что после каждой итерации должна осуще-стиллтьоя проверка условия (40).

Рассмотренный алгоритм обучения использовался при решении задачи минимизации многоэкстремальных функций. При этом отмечались высокая эффективность данного алгоритма по сравнению со стохастическими алгоритмами.

Наиболее интересное применение алгоритма рассмотрено при решении задачи классификации. Классификация в этом случае осуществлялась роботом-исследователем. Предварительно осуществлялось обучение робота, а точнее, алгоритма классификации. Информация о параметрах предметов, которые классифицировал робот, снималась в виде сигналов с рецепторов искусственной руки и представлялась в виде лингвистических переменных.

Полученные значения лингвистических переменных, соответ-стиующие отдельным классам предметов, использовались для построения условной меры нечеткости, связывающей параметры предметов с отдельным классом.

Основными преимуществами алгоритма являются: возможность использования нечеткой информации; высокая скорость сходимости; малое время вычисления и большое число используемых классов.

Обширной областью применения нечетких мер и НИ является нечеткая статистика. В ряде работ подробно исследованы методы вычисления нечетких ожиданий (FEV) и их связь с мерами центрального расположения. Практический пример применения FEV для решения задачи предсказания погоды рассматривается в литературе.

2.5.5. Применение нечеткого интеграла для оценки неопределенности ИМ.

Для решения многих практических задач с применением теории НМ необходимо оценивать степень неопределенности, размытости нечетких подмножеств, характеризующих различные объекты. Эффективным средством оценки размытости НМ является нечеткая энтропия . В литературе описан метод вычисления нечеткой энтропии с помощью НИ.

Пусть $\Phi\colon [0, 1] \to [0, 1]$ является *N*-функцией такой, что: а) $\Phi(0) = 0$; б) $\Phi(x) = \Phi(1-x)$, $x \in [0, 1]$; в) функция Ф является неубывающей в интервале [0, 0,5] и невозрастающей в [0,5, 1]. Пусть тройка $(X, \mathscr{B}, g)$ определяет пространство с нечеткой мерой *g*. В этом случае нечеткая (Ф — *g)* - энтропия есть функционал

$$E_{\Phi,g}(\mu) = \int_X \Phi(\mu) \circ g, \quad (42)$$

где Φ — $\mathscr{B}$-измеримая функция.

Если *X* — конечное множество; и его мощность есть card X = *n,* то энтропия (42) примет вид максиминпой энтропии и будет вычисляться по формуле

$$E_{\Phi,g}(h) = \bigvee_{i=1}^{n} (a_i \wedge \Phi(\mu(x_i))),$$

где

$$x_i \in X, \quad a_i = g(\{x_i, x_{i+1}, \ldots, x_n\}) \quad \forall i = 1, \ldots, n; \quad a_i > 0.$$

Рассмотренная энтропия является очень удобным инструментом анализа неопределенности НМ в задачах распознавания, принятия решения, диагностики и управления в нечеткой обстановке.

В теории систем формируется направление, предполагающее возможность использования нечетких мер и НИ для аналитического описания систем с нечеткими возмущениями на входах. При этом предполагается, что система является детерминированной. Исследуется математический аппарат для описания переходов таких систем из одного состояния в другое на основе соотношений, аналогичных уравнениям Чепмена — Колмогорова.

При исследовании сложных систем нечеткие меры представляют особый интерес для анализа их устойчивости. В случае нечетких систем устойчивость понимается как сохранение уровня сходства нечеткого состояния системы с недопустимой областью меньше некоторого порога ε. В качестве меры сходства можно взять нечеткую меру. Тогда *m*-мерная нечеткая система $\Phi\colon \mathscr{F}(X^m) \to \mathscr{F}(Y^m)$ будет ε-устойчивой относительно некоторого

семейства нечетких соответствий $F(X^m) \subset \mathcal{F}(X^m)$ тогда и только тогда, когда для $\forall \mu_A \in F(X^m)$ имеем $\Phi(\mu_A) = \mu_B$ и $g(\mu_B) \leqslant \leqslant \varepsilon$, где $\mu_A \in \mathcal{F}(X^m)$, $\mu_B \in \mathcal{F}(Y^m)$.

В литературе рассмотрены методы коррекции ε-устойчивости динамических многокритериальных систем нечеткого целевого управления, в том числе с *(L — R)* -аппроксимацией нечетких мер.

2.6. УСЛОВНЫЕ НЕЧЕТКИЕ МЕРЫ, ФУНКЦИОНАЛЬНЫЕ ОТОБРАЖЕНИЯ И НЕЧЕТКИЕ ВЕЛИЧИНЫ В РАМКАХ ВЕРОЯТНОСТНОГО ПОДХОДА

2.6.1. Введение

В литературе подробно описаны и исследованы основные выпуклые семейства нечетких мер, которые можно интерпретировать, как нижние или верхние оценки вероятностей событий. Следующий шаг развития теории состоит во введении базовых понятий, аналогичных в теории вероятностей или аддитивной теории меры. В рамках данного исследования используется так называемый «интуитивный» подход, который в некотором смысле близок к модели неточных вероятностей, основанных на множествах вероятностных мер. Аналогичные результаты могут быть получены из совсем других аксиоматических предпосылок, например, на основе согласованных нижних (верхних) предвидений, а также принципов избежания потерь и естественного продолжения, или согласованных нижних (верхних) средних по Кузнецову. Однако, по всей видимости, предлагаемый подход является более наглядным и понятным, несмотря на то, что указанные принципы имеют экономическую интерпретацию.

Отметим, что до сих пор в теории нечетких мер или шире - в теории неточных вероятностей, нет единого мнения, как определять такие базовые понятия как условные нечеткие меры, понятие независимости, декартового произведения нечетких мер и понятий более высокого уровня. Это может объясняться до сих пор глубоко неизученным, тонким взаимодействием различных видов моделируемой неопределенности, включающих неполноту, неточность, случайность и противоречивость анализируемых данных. Кроме того, здесь нужно учитывать специфику логических построений, например, при переходе от безусловных к условным вероятностям, мы можем потерять важную информацию. Такие же свойства сопутствуют понятию независимости нечетких величин, т.е. в данном случае в силу неточности данных, мы не можем судить о независимости нечетких величин по их

совместному распределению, а можем с некоторой точностью говорить о степени их зависимости. Такая ситуация приводит к аналогичным выводам, например, при поиске подходящей меры информативности по Шеннону (энтропии Шеннона) для функций доверия, а также для меры полной неопределениости .

Неоднозначность определения, например, условной нечеткой меры может объясняться наличием дополнительной априорной информации, выбором статистической модели порождения нечеткой меры, кроме того, в задачах, которые не имеют вероятностную интерпретацию, условные нечеткие меры могут иметь другой содержательный смысл, например, условной информативности относительно частичной информации. В этих случаях условные нечеткие меры могут выбираться совсем из других принципов.

С учетом этого, в настоящем разделе обсуждаются и исследуются такие базовые понятия как условные нечеткие меры, функциональные нечеткие распределения, нечеткие величины, а также их числовые характеристики, принцип обобщения в теории нечетких мер. В основном данные понятия вводятся вначале для точных нижних вероятностей, а затем полученные результаты обобщаются на более широкие семейства нечетких мер. Также приводятся теоремы, обобщающие классические утверждения из теории вероятностей.

2.6.2. Основные понятия и определения

Ниже описываются наиболее важные определения и утверждения.

Функция множества g на конечной алгебре $\mathfrak{A} = 2^X$ конечного пространства $X = \{x_1, x_2, \ldots, x_N\}$ называется нечеткой мерой, если она удовлетворяет условиям нормировки $(g(\emptyset) = 0,\ g(X) = 1)$ и монотонности $(g(A) \leqslant g(B)$ при $A \subseteq B)$.

Множество всех нечетких мер обозначается M_0, M_p - множество всех вероятностных мер. Нечеткая мера $\neg g(A) = 1 - g(\bar{A})$ называется двойственной к нечеткой мере g. Отношение $g_1 \leqslant g_2$ означает, что $g_1(A) \leqslant g_2(A)$ для любого $A \in \mathfrak{A}$.

В литературе описаны и детально изучены следующие семейства нечетких мер, которые можно интерпретировать в качестве нижних оценок вероятностей:

$M_1 = \{g \in M_0 \mid \exists P \in M_P : g \leqslant P\}$ - множество всех нижних вероятностей на алгебре A. Здесь M_p - множество всех вероятностных мер на алгебре $\mathfrak{A}$;

$M_2 = \left\{ g \in M_0 \,|\, \forall B \in \mathfrak{A},\ g(B) \neq 0 : g_B(A) = \frac{g(A \cap B)}{g(B)} \in M_1 \right\}$ множество всех обобщенных точных нижних вероятностей на алгебре $\mathfrak{A}$;

M_3 - множество всех нечетких мер на алгебре $\mathfrak{A}$, представляемых в виде выпуклой комбинации примитивных нижних вероятностей;

$M_4 = \{ g \in M_0 \,|\, \forall A \in \mathfrak{A},\ \exists P \in M_P :\ g \leqslant P,\ g(A) = P(A) \}$ множество всех точных нижних вероятностей на алгебре $\mathfrak{A}$;

M_5- множество всех 2-монотонных мер на алгебре $\mathfrak{A}$, для которых выполняется неравенство: $g(A) + g(B) \leqslant g(A \cap B) + g(A \cup B)$ для любых $A, B \in \mathfrak{A}$;

M_6 - множество всех мер доверия на алгебре $\mathfrak{A}$.

Доказано, что данные выпуклые семейства нечетких мер являются идеалами, т.е. они замкнуты относительно операции перемножения нечетких мер.

Имеют место следующие включения:

$$M_0 \supset M_1 \supset M_2 \supset \left\{ \begin{matrix} M_3 \\ M_4 \end{matrix} \right. \supset M_5 \supset M_6,$$

при этом, однако, $M_3 \not\subseteq M_4$ и $M_4 \not\subseteq M_3$.

2.6.3. Условные нечеткие меры

Пусть g точная нижняя вероятность, определенная на конечной алгебре $\mathfrak{A} = 2^X$ конечного пространства $X = \{x_1, x_2, ..., x_N\}$. Рассмотрим, как определить условное распределение $g(A|B)$ оценок вероятностей в рамках вероятностного подхода.

Известно, что точная нижняя вероятность определяет семейство вероятностных мер $\Xi = \{P_i | g \leqslant P_i\}$, причем $g(A) = \inf\limits_{P_i \in \Xi} P_i(A)$. Далее можно ввести в рассмотрение условные вероятностные распределении $P_i(A|B)$, $P_i(B) \neq 0$, порожденные семейством вероятностных мер Ξ, а также точную нижнюю оценку вероятности $P(A|B)$ как

$$g(A|B) = \inf_{P_i \in \Xi | P_i(B) \neq 0} P_i(A|B). \quad (1)$$

С учетом этого нечеткую меру $g(A|B)$ естественно назвать условной нечеткой мерой, построенной по мере g.

Лемма 1. *Пусть* $\Xi = \{P = \alpha P_1 + (1 - \alpha) P_2 | \alpha \in [0,1]\}$, *тогда*

$g(A) = \min\{P_1(A), P_2(A)\}$ *и* $g(A|B) = \min\{P_1(A|B), P_2(A|B)\}$,

если $g(B) \neq 0$.

Доказательство. Первая формула леммы очевидным образом следует из способа определения точной нижней грани для данного случая.

Докажем, что справедлива вторая формула леммы. Будем считать, что $A \subseteq B$. Тогда

$$g(A|B) = \inf_{\alpha \in [0,1]} \frac{\alpha P_1(A) + (1-\alpha)P_2(A)}{\alpha P_1(B) + (1-\alpha)P_2(B)}.$$

Дифференцируя выражение под знаком inf, получим

$$\frac{P_1(A)P_2(B) - P_1(B)P_2(A)}{[\alpha P_1(B) + (1-\alpha)P_2(B)]^2},$$

т.е. функция, стоящая под знаком inf, является монотонной на отрезке $\alpha \in [0,1]$. С учетом этого заключаем, что наибольшее значение данная функция будет принимать на концах отрезка, т.е.
$g(A|B) = \min\{P_1(A|B), P_2(A|B)\}$, если $g(B) \neq 0$.

Теорема 1. *Пусть* $\Xi = \left\{ P = \sum_{i=1}^{m} \alpha_i P_i \middle| \sum_{i=1}^{m} \alpha_i = 1,\ \alpha_i \geqslant 0 \right\}$, *тогда*

$$g(A) = \min_{i=1,\dots,m} P_i(A) \text{ и } g(A|B) = \min_{i=1,\dots,m} P_i(A|B),\ g(B) \neq 0.$$

Теорема 1 это, очевидно, следствие леммы **1.**

Лемма 2. *Пусть* g *точная нижняя вероятность, причем* $g(B) + g(\bar{B}) = 1$, $g(B) \neq 0$, *для некоторого* $B \subseteq X$. *Тогда*
$g(A|B) = \frac{g(A \cap B)}{g(B)}$.

Доказательство. Заметим, что в данном случае $P(B) = g(B)$ для любой вероятностной меры P, $P \geqslant g$. Поэтому $g(A|B) \geqslant \frac{g(A \cap B)}{g(B)}$.
Поскольку g - точная нижняя вероятность, то для любого $A \in \mathfrak{A}$ найдется вероятностная мера P, что $P \geqslant g$ и $P(A \cap B) = g(A \cap B)$.
Таким образом, для данной вероятностной меры
$P(A|B) = \frac{g(A \cap B)}{g(B)}$.

Теорема 2. *Пусть* g *точная нижняя вероятность, определенная на конечной алгебре* $\mathfrak{A}$ *пространства* $X = \{x_1, x_2, \dots x_N\}$ *и* $\Xi = \{P_i | g \leqslant P_i\}$. *Тогда* Ξ *выпуклое множество, имеющее конечное число экстремальных точек* $P_1, P_2, \dots, P_m$, *т.е. оно представляется в виде:*

$$\Xi = \left\{ P = \sum_{i=1}^{m} \alpha_i P_i \middle| \sum_{i=1}^{m} \alpha_i = 1,\ \alpha_i \geqslant 0 \right\}.$$

Доказательство. Семейство вероятностных мер Ξ является решением следующей системы неравенств:

$$\begin{cases} \sum\limits_{x_i \in A} P\{x_i\} \geqslant g(A), \\ \sum\limits_{i=1}^{n} P\{x_i\} = 1, \end{cases} \quad \text{для всех } A \subseteq X,\ A \neq \emptyset.$$

Таким образом, теорема 2 является следствием теории, описывающей решения конечной системы линейных неравенств.

Теперь ясно, как находить условные нечеткие распределения для точных нижних вероятностей. Рассмотрим, какие конструктивные результаты можно получить для 2-монотонных нечетких мер. Заметим, чти для 2-монотонных мер множество Ξ описывается следующей теоремой **.**

Теорема 3. *Пусть нечеткая мера g является 2-монотонной и $\Xi = \{P | g \leqslant P\}$. Тогда экстремальными точками выпуклого множества Ξ являются вероятностные меры P_γ, где $\gamma = (i_1, i_2, .., i_n)$ это перестановки множества чисел $\{1,2,...,n\}$, причем $P_\gamma\{x_{i_k}\} = g(A_k) - g(A_{k-1})$. где $A_k = \{x_{i_1}, x_{i_2}, ..., x_{i_k}\}$,*

$k = 1, 2, .., n$, $A_0 = \emptyset$.

Теорема 4. *Пусть нечеткая мера g является 2-мопотонной, тогда*

$$g(A|B) = \frac{g(A)}{g(A) + 1 - g(A \cup \bar{B})}$$

при $A \subseteq B$ и $g(B) \neq 0$.

Доказательство. Из формулы (1) следует, что

$$g(A|B) = \inf_{P \geqslant g} \frac{P(A)}{P(A) + P(B \cap \bar{A})}.$$

Таким образом, вероятностную меру P, $P \geqslant g$, требуется выбрать так, чтобы выражение под знаком inf принимало наименьшее значение. Обозначим $P(A) = \alpha$, $P(B \cap \bar{A}) = \beta$, тогда необходимо минимизировать функцию $f(\alpha, \beta) = \frac{\alpha}{\alpha+\beta}$. При $\alpha, \beta > 0$, эта функция монотонно возрастает относительно α и монотонно убывает относительно β. Отсюда следует, что вероятностную меру P требуется выбрать таким образом, чтобы значение $P(A)$ было как можно меньше, а значение $P(B \cap \bar{A})$ как можно больше. С учетом этого, используя свойство 2-монотонности g, выберем вероятностную меру $P \geqslant g$, так чтобы $g(A) = P(A)$, $g(A \cup \bar{B}) = P(A \cup \bar{B})$. Тогда $P(B \cap \bar{A}) = 1 - P(A \cup \bar{B}) = 1 - g(A \cup \bar{B})$. Заметим, что $g(A)$ является точной нижней оценкой вероятности события A по семейству вероятностных мер $\Xi = \{P | g \leqslant P\}$, а $1 - g(A \cup \bar{B})$ точной верхней оценкой вероятности события $B \cap \bar{A}$ по Ξ. Поэтому доказываемая формула истинна.

Следствие 1. *Пусть нечеткая мера g является 2-альтернирующей и* $A \subseteq B$, *тогда*

$$g(A|B) = \frac{g(A)}{g(A) + 1 - g(A \cup B)}.$$

т.е. формула остается такой же, как и для 2-монотонных нечетких мер.

Доказательство. Требуется, используя отношение двойственности, построить двойственную нечеткую меру для $g(A|B)$.

$$g(A|B) = 1 - g(\bar{A} \cap B|B) = 1 - \frac{g(\bar{A} \cap B)}{g(\bar{A} \cap B) + 1 - g((\bar{A} \cap B) \cup \bar{B})}.$$

Заметим, что $(\bar{A} \cap B) \cup \bar{B} = \bar{A} \cup \bar{B} = \bar{A}$, так как $A \subseteq B$. Поэтому

$$g(A|B) = 1 - \frac{1 - g(A \cup \bar{B})}{1 - g(A \cup \bar{B}) + g(A)} = \frac{g(A)}{g(A) + 1 - g(A \cup \bar{B})}.$$

Следствие доказано.

Следствие 2. *Формулу из теоремы 4 для произвольного события* $A \subseteq X$ *можно обобщить следующим образом:*

$$g(A|B) = \frac{g(A \cap B)}{g(A \cap B) + 1 - g(A \cup \bar{B})} = \frac{g(A \cap B)}{g(A \cap B) + \neg g(\bar{A} \cap B)},$$

где $\neg g$ *это двойственная мера к нечеткой мере* g.

Теорема 5. *Пусть g 2-монотонная мера. Тогда мера $g(A|B)$ также будет 2-монотонной.*

Простое доказательство этого факта можно получить, используя теорию разностей . Аналогичную теорему можно получить для мер доверия. Подчеркнем, что приведенное ниже доказательство основано на свойствах идеалов**.**

Вспомогательная лемма. *Пусть g мера доверия, тогда функция множества* $\mu(A) = \frac{g(A \cup \bar{B}) - g(A \cap B) - g(\bar{B})}{1 - g(B) - g(\bar{B})}$ *является также мерой доверия при*

$$g(B) + g(\bar{B}) < 1.$$

Доказательство. Очевидно, что функция множества μ удовлетворяет условию нормировки. Докажем, что это мера доверия. Пусть g мера доверия, тогда $g(A) = \sum_{C \subseteq X} m(C)\eta_{(C)}(A)$. Следовательно,

$$g(A) = g(A \cup \bar{B}) - g(A \cap B) - g(\bar{B}) = \sum_{C \subseteq X} m(C)\eta_{(C)}(A \cup \bar{B}) -$$

$$- \sum_{C \subseteq X} m(C)\eta_{(C)}(\bar{B}) - \sum_{C \subseteq X} m(C)\eta_{(C)}(A \cap B).$$

Поскольку $\eta_{\langle C\rangle}(A \cup \bar{B}) = \eta_{\langle C\rangle}(\bar{B})$ при $C \subseteq \bar{B}$

$$\eta_{\langle C\rangle}(A \cup \bar{B}) = \begin{cases} 1, & C \subseteq A \cup \bar{B}, \\ 0, & C \not\subseteq A \cup \bar{B}, \end{cases} = \begin{cases} 1, & C \cap B \subseteq A, \\ 0, & C \cap B \not\subseteq A, \end{cases} = \eta_{\langle C\cap B\rangle}(A),$$

а также $\eta_{\langle C\rangle}(A \cap B) = \begin{cases} \eta_{\langle C\rangle}(A), & C \subseteq B, \\ 0, & C \not\subseteq B, \end{cases}$ получим, что

$$g(A) = \sum_{C\subseteq X | C\not\subseteq \bar{B}} m(C)\eta_{\langle C\cap B\rangle}(A) - \sum_{C\subseteq B} m(C)\eta_{\langle C\rangle}(A) = \sum_{C\subseteq X \left| \begin{smallmatrix} C\not\subseteq \bar{B}, \\ C\not\subseteq B \end{smallmatrix} \right.} m(C)\eta_{\langle C\cap B\rangle}(A).$$

Таким образом, функция множества g представляется в виде линейной комбинации примитивных мер возможности с неотрицательными коэффициентами, т.е. за счет нормировки мы получим нечеткую меру μ, являющуюся мерой доверия.

Теорема 6. *Пусть g мера доверия. Тогда мера $g(A|B)$ также будет мерой доверия.*

Доказательство. Возможно два случая. Пусть $g(B) + g(\bar{B}) = 1$,тогда для любой вероятностной меры P, $P \geqslant g$, будет выполняться $g(B) = P(B)$. Это означает, что формула для вычисления $g(A|B)$ примет вид $g(A|B) = \dfrac{g(A \cap B)}{g(B)}$.

Нетрудно заметить, что для данного случая мера $g(A|B)$ является мерой доверия.

Рассмотрим второй случай, когда $g(B) + g(\bar{B}) < 1$.

$$g(A|B) = \frac{g(A \cap B)}{g(A \cap B) + 1 - g(A \cup \bar{B})} = \frac{g(A \cap B)}{(1 - g(\bar{B}))\left(1 - \frac{g(A\cup\bar{B}) - g(A\cap B) - g(\bar{B})}{1 - g(\bar{B})}\right)} =$$

$$= \frac{g(A \cap B)}{1 - g(\bar{B})} \sum_{n=0}^{\infty} \left(\frac{g(A \cup \bar{B}) - g(A \cap B) - g(\bar{B})}{1 - g(\bar{B})} \right)^n.$$

Видно, что в пределе, используя операции выпуклой суммы и произведения для нечетких мер доверия μ из вспомогательной леммы и $g_B(A) = \dfrac{g(A \cap B)}{g(B)}$, можно выразить нечеткую меру $g(A|B)$.

Следовательно, она также является мерой доверия.

Лемма 3. *Пусть g нижняя вероятность и*

$$g(A|B) = \frac{g(A)}{g(A) + 1 - g(A \cup \bar{B})},$$

$A \subset B$, $g(B) \neq 0$. Тогда для любой вероятностной меры P, $P \geqslant g$, будет выполняться неравенство: $g(A|B) \leqslant P(A|B)$.

Доказательство. Ясно, что $P(A|B) = \dfrac{P(A)}{P(A) + P(B\backslash A)}$, при этом $P(A) \geqslant g(A)$ и $P(B\backslash A) \leqslant 1 - g(A \cup \bar{B})$. Кроме того, функция

$f(\alpha,\beta)=\frac{\alpha}{\alpha+\beta}$ монотонно возрастает по α и монотонно убывает по β при α , $\beta>0$. Из этого делаем вывод, что $g(A|B)\leqslant P(A|B)$.
Смысл леммы 3 в том, что формулу для оценок условных вероятностей можно использовать на более широком семействе нечетких мер, включающих все нижние вероятности, а также верхние вероятности. При этом в общем случае будут получаться менее точные оценки, например, для точных нижних вероятностей по сравнению с формулой (1).

Лемма 4. *Пусть* g *- точная нижняя вероятность и* $g(A|B)=\inf_{P\geqslant g}\frac{P(A)}{P(B)}$, $A\subseteq B$, $g(B)\neq 0$. *Тогда* $g(A)\geqslant g(A|B)\,g(B)$.

Доказательство. Поскольку g точная нижняя вероятность, то выберем вероятностную меру P таким образом, что $P\geqslant g$ и $P(A)=g(A)$. В этом случае $g(B)\leqslant P(B)$, а также по определению $g(A|B)\leqslant P(A|B)$. Поэтому из тождества $P(A)=P(A|B)\,P(B)$ следует неравенство $g(A)\geqslant g(A|B)\,g(B)$.

Теорема 7. *Пусть* g *точная нижняя вероятность на* $\mathfrak{A}$ *и* $\{H_1,H_2,...,H_k\}$ *это разбиение пространства X, т.е.* $H_1\cup H_2\cup ...\cup H_k=X$ *и* $H_i\cap H_j=\emptyset$, $i\neq j$. *Тогда* $g(A)\geqslant\sum_{i=1}^{k}g(A|H_i)\,g(H_i)$, *если* $g(H_i)>0$.

Доказательство. Поскольку g точная нижняя вероятность, то она является супераддитивной для несовместных событий. Поэтому $g(A)\geqslant\sum_{i=1}^{k}g(A\cap H_i)$. Далее пользуясь неравенством из леммы 4 $g(A\cap H_i)\geqslant g(A|H_i)\,g(H_i)$, получим требуемое неравенство.

Возникает вопрос, в каких случаях формула полной вероятности дает более точные оценки вероятностей. Ответ на данный вопрос дают следующие теоремы.

Теорема 8. *Пусть* $\{H_1,H_2,...,H_k\}$ *- это разбиение пространства X и* $A\in\mathfrak{A}$ *- произвольное событие. Рассмотрим также разбиение* $\{B_1,B_2,...,B_k,B_{k+1},...,B_{2k}\}$, *где* $B_i=A\cap H_i$ *при* $i\leqslant k$, *и* $B_{i+k}=\bar{A}\cap H_i$ *при* $i\leqslant k$. *Тогда*

$$\sum_{i=1}^{k}g(A|H_i)\,g(H_i)\leqslant\sum_{i=1}^{2k}g(A|B_i)\,g(B_i). \quad \textbf{(2)}$$

Доказательство. Очевидно, что $g(A|H_i)g(H_i)\leqslant g(A\cap H_i)=g(B_i)$ по лемме 4. Кроме того, $g(A|B_i)=1$ при $i=1,2,...,k$, так как $B_i\subseteq A$ и $g(A|B_i)=0$ при $i=k+1,k+2,...,2k$, поскольку $B_i\cap A=\emptyset$.

Поэтому правая часть неравенства (2) преобразуется к виду $\sum_{i=1}^{k} g(B_i)$.

Теперь, используя оценку $g(A|H_i)g(H_i) \leqslant g(B_i),\ i = 1, 2, \ldots, k,$ убеждаемся в справедливости неравенства **(2).**

Следствие теоремы 8. *Пусть* $\{H_1, H_2, \ldots, H_k\}$ - *это разбиение пространства X, причем* $A = H_1 \cup H_2 \cup \ldots \cup H_m, m \leqslant k$. *Тогда неравенство полной вероятности имеет вид:* $g(A) \geqslant \sum_{i=1}^{m} g(H_i)$.

Теорема 9. *Пусть* $\mathfrak{A}_1$ *и* $\mathfrak{A}_2$ - *это конечные алгебры событий, определенные на множестве A, и* $\Omega_1 = \{H_1, H_2, \ldots, H_k\}$, $\Omega_2 = \{B_1, B_2, \ldots, B_m\}$- *это разбиения A, порождающие данные алгебры. Тогда, если* $\mathfrak{A}_1 \subseteq \mathfrak{A}_2$, *то* $\sum_{i=1}^{k} g(H_i) \geqslant \sum_{i=1}^{m} g(B_i)$, *т.е. разбиение* Ω_1 *дает более точную оценку вероятности события A по точной нижней вероятности g.*

Доказательство. Поскольку$\mathfrak{A}_1 \subseteq \mathfrak{A}_2$,то любой элемент $H \in \mathfrak{A}_1$ является объединением элементов разбиения Ω_2. Поэтому множества B_i, можно занумеровать таким образом, чтобы

$$H_1 = \bigcup_{i=i_0+1}^{i_1} B_i,\ H_2 = \bigcup_{i=i_1+1}^{i_2} B_i, \ldots,\ H_k = \bigcup_{i=i_{k-1}+1}^{i_k} B_i,$$

где $0 < i_0 < i_1 < \ldots < i_k$. Пользуясь супераддитивностью меры g для несовмест ных событий, получим

$$g(H_n) = g\left(\bigcup_{i=i_{n-1}+1}^{i_n} B_i\right) \geqslant \sum_{i=i_{n-1}+1}^{i_n} g(B_i).$$

Поэтому

$$\sum_{i=1}^{k} g(H_i) \geqslant \sum_{i=1}^{m} g(B_i).$$

2.6.4. Условные нечеткие меры при дополнительных ограничениях

В некоторых случаях оказывается возможным уточнить условное распределение оценок вероятностей, если имеется дополнительная априорная информация. Покажем, как это уточнение производится, если при расчете $g(A|B)$ каким-то образом получена точная оценка вероятности *P(B).* Пусть g является 2-монотонной нечеткой мерой и $P(B \cup C) = g(B \cup C)$, $P(C) = g(C)$, причем $C \cap B = \emptyset$. В этом

случае можно вычислить вероятность
$P(B) = P(B \cup C) - P(C) = g(B \cup C) - g(C)$. Будем вычислять $g(A|B)$ при $A \subseteq B$, используя формулу:

$$g(A|B) = \inf_{P \left| \begin{matrix} P \geqslant g,\ P(C)=g(C), \\ P(B \cup C)=g(B \cup C) \end{matrix} \right.} \frac{P(A \cup C) - g(C)}{g(B \cup C) - g(C)}.$$

Очевидно, что $P(A \cup C) \geqslant g(A \cup C)$, поэтому

$$g(A|B) \geqslant \frac{g(A \cup C) - g(C)}{g(B \cup C) - g(C)}.$$

Покажем, что в последнем неравенстве можно поставить знак равенства. Действительно, поскольку нечеткая мера g является 2-монотонной, то можно найти такую вероятностную меру P, $P \geqslant g$, что $P(C) = g(C)$, $P(A \cup C) = g(A \cup C)$ и $P(B \cup C) = g(B \cup C)$.

Лемма 5. *Пусть g - 2-монотонная нечеткая мера на $\mathfrak{A} = 2^X$. Тогда нечеткая мера $g(A|B) = \frac{g(A \cup C) - g(C)}{g(B \cup C) - g(C)}$, на алгебре $\mathfrak{A}_B = 2^B$, является 2-монотонной.*

Доказательство. Достаточно показать, что для любых событий $A_1 \subseteq A_2 \subseteq B$ найдется такая вероятностная мера P, $P \geqslant g$, что $g(A_1|B) = P(A_1|B)$, $g(A_2|B) = P(A_2|B)$. Поскольку мера g является 2-монотонной и $C \subseteq A_1 \cup C \subseteq A_2 \cup C \subseteq B \cup C$, то найдется вероятностная мера $P \geqslant g$, что $P(C) = g(C)$, $P(A_1 \cup C) = g(A_1 \cup C)$, $P(A_2 \cup C) = g(A_2 \cup C)$ и $P(B \cup C) = g(B \cup C)$, откуда уже следует требуемое.

Замечание. Фактически в данном разделе рассматривалось правило пересчета $g(A|B) = \frac{g(A \cap B \cup C) - g(C)}{g(B \cup C) - g(C)}$ при $C \cap B = \emptyset$. Это правило объединяет в себе такие известные правила, как $g(A|B) = \frac{g(A \cap B)}{g(B)}$ при $C = \emptyset$, и $g(A|B) = \frac{g(A \cup \bar{B}) - g(\bar{B})}{1 - g(\bar{B})}$ при $C = \bar{B}$.

2.6.5. Функциональные нечеткие распределения

Далее будем рассматривать нечеткую меру g на на алгебре $\mathfrak{A} = 2^{X \times Y}$ декартового произведения $X \times Y$ конечных пространств X и Y. В этом случае значения нечеткой меры на всех подмножествах $X \times Y$, и, будем предполагать, отражают некоторую функциональную зависимость между X и Y. Эту зависимость можно исследовать, используя понятие условной нечеткой меры. Пусть произошло событие $A \in \mathfrak{A}_X$, где

$\mathfrak{A}_X = 2^X$, оценим значение условной вероятности наступления события $B \in \mathfrak{A}_Y$, где $\mathfrak{A}_Y = 2^Y$, предполагая, что g точная нижняя вероятность. Ясно, что

$$g_2(B|A) = \inf_{P \geq g} \frac{P(A \times B)}{P(A \times Y)}, A \in \mathfrak{A}_X, B \in \mathfrak{A}_Y. \quad \textbf{(3)}$$

Если нет никакой другой априорной информации, т.е. в пространстве X наступает достоверное событие, то мы получаем маргинальное распределение оценок вероятностей пространства Y:

$$g_2(B) = g_2(B|X) = g(X \times B).$$

Поскольку пространства X и Y являются равносильными, мы можем аналогичным образом получить оценку условной вероятности $g_1(A|B)$, что произошло событие $A \in \mathfrak{A}_X$ при условии наступления $B \in \mathfrak{A}_Y$.

$$g_1(A|B) = \inf_{P \geq g} \frac{P(A \times B)}{P(X \times B)}, A \in \mathfrak{A}_X, B \in \mathfrak{A}_Y, \quad (4)$$

а также маргинальное распределение

$$g_1(A) = g_1(A|Y) = g(A \times Y).$$

В том случае, когда мера g является 2-монотонной. формулы (3) и (4) преобразуются к виду:

$$g_2(B|A) = \frac{g(A \times B)}{g(A \times B) + 1 - g(A \times B \cup \bar{A} \times Y)}, \quad \textbf{(3*)}$$

$$g_1(A|B) = \frac{g(A \times B)}{g(A \times B) + 1 - g(A \times B \cup X \times \bar{B})}. \quad \textbf{(4*)}$$

Рассмотрим задачу нахождения маргинального распределения $g_2(B)$, $B \in \mathfrak{A}_Y$, если известны маргинальное распределение $g_1(A)$, $A \in \mathfrak{A}_X$, и значения $g_2(B|\{x_i\}) = g_{x_i}(B)$ только в точках $x_i \in X$. В этом случае, используя принцип точной нижней вероятности (или принцип естественного продолжения), получим

$$g_2(B) = g(X \times B) = \inf_{\substack{P_{x_i} \geq g_{x_i} \\ P_1 \geq g_1}} \sum_{i=1}^{n} P_{x_i}(B) P_1\{x_i\}, \quad \textbf{(5)}$$

где $X = \{x_1, x_2, ..., x_n\}$.

В том случае, когда нечеткие меры g_{x_i} являются точными нижними вероятностями, формулу (5), очевидно, можно преобразовать следующим образом:

$$g_2(B) = \inf_{P_1 \geq g_1} \sum_{i=1}^{n} g_2(B|\{x_i\}) P_1\{x_i\}.$$

Последнее выражение можно рассматривать, как нижнюю оценку математическою ожидания нечеткой величины $g_2(B|\{x_i\})$, $x_i \in X$, по нечеткой мере g_1. Речь об этом пойдет в следующих подразделах.

2.6.6. Принцип обобщения в рамках понятия нижней (верхней) вероятности

Принцип обобщения хорошо известен в теории меры (в частности, в теории вероятностей), а также теории возможностей . Целью данного раздела было показать, что этот принцип естественно получается в рамках вероятностного подхода в теории нечетких мер.

Пусть g_X - точная нижняя вероятность на $\mathfrak{A}_X = 2^X$ и $\Xi_X = \{P_X | P_X \geqslant g_X\}$ семейство вероятностных мер, индуцированное нечеткой мерой g_X. Кроме того, рассмотрим функциональное отображение *f* из пространства *X* в конечное пространство *Y*. Покажем, каким образом в этом случае порождается нечеткая мера g_Y с помощью функционального отображения *f*. Пусть $P_X \geqslant g_X$, тогда вероятностная мера P_x порождает вероятностную меру P_Y, удовлетворяющую условиям:

$$P_Y(B) = P_X(f^{-1}(B)). \quad \textbf{(6)}$$

где $B \in \mathfrak{A}_Y$ и $f^{-1}(B) = \{x \in X | f(x) \in B\}$. Таким образом, мы можем рассматривать функциональное отображение $P_Y = f(P_X)$ вероятностной меры P_x на вероятностную меру P_Y, удовлетворяющую условиям (6). Аналогичным образом можно рассмотреть функциональное отображение *f* семейства вероятностных мер Ξ_X на Ξ_Y: $\Xi_Y = \{f(P_X) | P_X \geqslant g_X\}$. С учетом этого можно определить нечеткую меру g_Y как точную нижнюю оценку вероятности по семейству вероятностных мер Ξ_Y:

$$g_Y(B) = \inf_{P_Y \in \Xi_Y} P_Y(B) = \inf_{P_X | P_X \geqslant g_X} P_X(f^{-1}(B)) = g_X(f^{-1}(B)),$$

где $B \in \mathfrak{A}_Y$. Таким образом, мы получили формулу обобщения

$$g_Y(B) = g_X(f^{-1}(B)), B \in \mathfrak{A}_Y,$$

которую можно использовать в общем случае для любых нечетких мер. Далее рассмотрим несколько утверждений, которые позволяют установить, какие свойства нечетких мер сохраняются после действия функциональных отображений.

Утверждение 1. *Пусть g_X нечеткая мери на $\mathfrak{A}_X$ и $f : X \to Y$ - функциональное отображение, тогда $g_Y = f(g_X)$ - также нечеткая мера.*

Доказательство. Требуется проверить аксиомы нормировки и монотонности нечеткой меры:

1) $g_Y(\emptyset) = g_X(f^{-1}(\emptyset)) = g_X(\emptyset) = 0$;

2) $g_Y(Y) = g_X(f^{-1}(Y)) = g_X(X) = 1$;

3) пусть $A \subseteq B \subseteq Y$, тогда $f^{-1}(A) \subseteq f^{-1}(B)$ и $g_Y(A) = g_X(f^{-1}(A)) \leqslant g_X(f^{-1}(B)) = g_Y(B)$, т.е. условие монотонности $g_Y(A) \subseteq g_Y(B)$ при $A \subseteq B$ выполняется. Утверждение доказано.

Утверждение 2. *Пусть g_X и q_X - нечеткие меры на $\mathfrak{A}_X$ и $f : X \to Y$ функциональное отображение, причем $g_X \leqslant q_X$. Тогда $g_Y = f(g_X) \leqslant q_Y = f(q_X)$.*

Доказательство. Требуется показать, что $g_Y(B) \leqslant q_Y(B)$ для любого $B \in \mathfrak{A}_Y$. По определению последнее неравенство записывается как: $g_X(f^{-1}(B)) \leqslant q_X(f^{-1}(B))$, т.е. $g_Y \leqslant q_Y$ следует из $g_X \leqslant q_X$.

Утверждение 3. Пусть $\mu_X = \alpha g_X + \beta q_X$, где g_X и q_X нечеткие меры на $\mathfrak{A}_X$. $\alpha + \beta = 1$, $\alpha, \beta \geqslant 0$ и $f : X \to Y$ - функциональное отображение, тогда $\mu_Y = \alpha g_Y + \beta q_Y$, где $g_Y = f(g_X)$, $q_Y = f(q_X)$, $\mu_Y = f(\mu_X)$.

Доказательство. Доказываемое тождество можно переписать в виде

$$\mu_X(f^{-1}(B)) = \alpha g_X(f^{-1}(B)) + \beta q_X(f^{-1}(B)),$$

которое выполняется для любого события $B \in \mathfrak{A}_Y$. Но это тождество, очевидно, выполняется из условия утверждения.

Утверждение 4. *Пусть g_X - нечеткая мера на $\mathfrak{A}_X$ и $\neg g_X$ - двойственная к g_X мера, $f : X \to Y$ функциональное отображение. Тогда $f(\neg g_X) = \neg f(g_X)$.*

Доказательство. Требуется доказать, что $f(\neg g_X)(B) = \neg f(g_X)(B)$ для любого $B \in \mathfrak{A}_Y$. Действительно,

$$f(\neg g_X)(B) = \neg g_X(f^{-1}(B)) = 1 - g_X\left(\overline{f^{-1}(B)}\right) =$$

$$= 1 - g_X(f^{-1}(\bar{B})) = 1 - f(g_X)(\bar{B}) = \neg f(g_X)(B).$$

Утверждение 5. *Пусть g_X - нечеткая мера и $f : X \to Y$ - функциональное отображение, $g_Y = f(g_X)$. Тогда*

1) если g_X - примитивная мера, то g_Y - примитивная мера;

2) если g_X - нижняя (верхняя) вероятность, то g_Y -нижняя (верхняя) вероятность;

3) если g_X - обобщенная точная нижняя вероятность, то g_Y - обобщенная точная нижняя вероятность;

4) если $g_X \in M_3$, то $g_Y \in M_3$;

0) если g_X - точная нижняя (верхняя) вероятность, то g_Y - точная нижняя (верхняя) вероятность;
6) если g_X- 2-монотонная (2-альтернирующая) мера, то g_Y- 2-монотонная (2-альтернирующая) мера;
7) если g_X - мера доверия (правдоподобия), то g_Y - мера доверия (правдоподобия):
8) если g_X - мера необходимости (возможности), то g_Y - мера необходимости (возможности);
9) если g_X - вероятностная мера, то g_Y -вероятностная мера.

Доказательство.

1) Если примитивная g_X мера, то $g_X(A) \in \{0.1\}, A \in \mathfrak{A}_X$, поэтому и $g_Y(B) = g_X(f^{-1}(B)) \in \{0,1\}$, т.е. g_Y также является примитивной мерой.

2) Пусть g_X - нижняя вероятность. Выберем вероятностную меру P_x таким образом, что $P_X \geqslant g_X$. Пусть $P_Y = f(P_X)$, тогда согласно утверждению 2 $P_Y \geqslant g_Y$, т.е. g_Y - это нижняя вероятность.

3) Пусть g_X - обобщенная точная вероятность. Тогда согласно определению функции множества $g_X(A|B) = \frac{g_X(A \cap B)}{g_X(B)}$ являются нижними вероятностями для любых $B \in \mathfrak{A}_X$, $g_X(B) \neq 0$. Рассмотрим функцию множества

$$g_Y(A|B) = \frac{g_Y(A \cap B)}{g_Y(B)} = \frac{g_X(f^{-1}(A) \cap f^{-1}(B))}{g_X(f^{-1}(B))} = g_X(f^{-1}(A)|f^{-1}(B)).$$

Таким образом, $g_X(*|B) = g_Y(*|f(B))$, т.е. это свойство следует из 2) и того, что

$g_Y(A) = \frac{g_Y(A \cap B)}{g_Y(B)} = \frac{g_X(f^{-1}(A) \cap f^{-1}(B))}{g_X(f^{-1}(B))}$, т.е. g_Y - обобщенная точная нижняя вероятность по свойству 2).

4) Доказывается совершенно аналогично, как и (5).

5) Пусть g_X - точная нижняя вероятность и $B \in \mathfrak{A}_Y$. Выберем вероятностную меру P_x таким образом, что $P_X \geqslant g_X$ и $P_X(f^{-1}(B)) = g_X(f^{-1}(B))$. Тогда вероятностная мера $P_Y = f(P_X)$ будет удовлетворять необходимым условиям:
$P_Y(B) = g_Y(B)$ и $P_Y \geqslant g_Y$, т.е. g_Y- точная нижняя вероятность.

6) Пусть g_X - 2-монотонная мера и $A \subseteq B \subseteq Y$. Выберем вероятностную меру P_x таким образом, что $P_X \geqslant g_X$ и $P_X(f^{-1}(A)) = g_X(f^{-1}(A)), P_X(f^{-1}(B)) = g_X(f^{-1}(B))$. Это можно сделать, так как $f^{-1}(A) \subseteq f^{-1}(B)$. Далее замечаем, что вероятностная мера $P_Y = f(P_X)$ удовлетворяет всем необходимым условиям:

$P_Y \geqslant g_Y$, $P_Y(A) = g_Y(A)$ и $P_Y(B) = g_Y(B)$, т.е. g_Y - это 2-монотонная мера.

7) Пусть g_X - мера доверия, тогда ее можно продставить в виде выпуклой комбинации примитивных мер необходимости

$g_X = \sum_{i=1}^{n} \alpha_i N_{X,i}$, где

$\alpha_i \geqslant 0$, $\sum_{i=1}^{n} \alpha_i = 1$

и $N_{x,i}$ - примитивные меры необходимости. Пусть

$N_{Y,i} = f(N_{X,i})$, $i = 1, 2, ..., n$.

Тогда, используя доказанные свойства 1) и 6) теоремы, а также утверждение 5, получим $g_Y = \sum_{i=1}^{n} \alpha_i N_{Y,i}$,т.е. g_Y является мерой доверия, поскольку представляется в виде выпуклой комбинации примитивных мер необходимости $N_{Y,i}$.

Полностью доказать утверждение для верхних вероятностей, точных верхних вероятностей, 2-альтернирующих мер, мер правдоподобия и возможности можно, используя свойства 2), 3), 4), 5), 6), отношение двойственности нечетких мер, а также утверждение 4.

8) Пусть g_X -мера необходимости, тогда для любых $A, B \in \mathfrak{A}_X$ $g_X(A \cap B) = \min(g_X(A), g_X(B))$. Проверим, будет ли это свойство выполняться для меры $g_Y = f(g_X)$. Пусть $A, B \in \mathfrak{A}_Y$, тогда $f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B)$, поскольку f - функциональное отображение. Поэтому $g_Y(A \cap B) = g_X(f^{-1}(A \cap B)) = \min[g_X(f^{-1}(A)), g_X(f^{-1}(B))] = \min(g_Y(A), g_Y(B))$. т.е. g_Y - мера необходимости.

9) Следует из принципа обобщения для вероятностных мер.

Замечание 1. Утверждение 5 фактически отражает следующий факт: если нечеткая мера g_X является примитивной, или нижней вероятностью, ... и т.д., на алгебре событий $\mathfrak{A}_1$, то она будет обладать тем же свойством и на алгебре $\mathfrak{A}_2 \subseteq \mathfrak{A}_1$. Покажем, как получается алгебра $\mathfrak{A}_2$ при помощи функционального соответствия f:

1) f индуцирует алгебру $\mathfrak{A}_Y = f(\mathfrak{A}_1)$ пространства Y;

2) рассмотрим обратное отображение $\mathfrak{A}_2 = f^{-1}(\mathfrak{A}_1)$.Поскольку f функциональное отображение, алгебра $\mathfrak{A}_2 \subseteq \mathfrak{A}_1$.

Замечание 2. Ясно, что не любое свойство сохраняется при функциональном отображении. Так, например, любую противоречивую меру можно отобразить на непротиворечивую меру, в частности, на примитивную вероятностную меру. Для этого рассмотрим функциональное отображение пространства

$X = \{x_1, x_2, ..., x_N\}$ на пространство $Y = \{y\}$ по правилу $f(x_i) = y_i$. В этом случае любая нечеткая мера g_X на $\mathfrak{A}_X$ отображается на примитивную вероятностную меру g_Y пространства *Y*.

2.6. 8. Нечеткие величины

Аналогичным образом, как и в теории вероятностей, можно определить понятие нечеткой величины. При этом случайную величину можно рассматривать как частный случай нечеткой величины (В некоторой литературе не делается различия между нечеткой величиной и случайной величиной. В некотором смысле рассматриваемые здесь нечеткие величины можно считать «неточно заданными» случайными величинами.)

Определение 1. Пусть $X = \{x_1, x_2, ..., x_N\}$ нечеткое пространство, на котором определена нечеткая мера g_X. Тогда любая вещественная функция $f(x)$, $x \in X$, называется нечеткой величиной $\xi = f(x)$.

Ясно, что нечеткая величина определяет распределение нечеткости на вещественной оси. При этом достаточно рассматривать это распределение на множестве $Y = f(X)$. Таким образом, изучать нечеткую величину можно ξ, исследуя нечеткую меру $g_Y = f(g_X)$. Пусть мера g_X является точной нижней вероятностью, тогда это g_Y тоже точная нижняя вероятность. Таким образом, значение $g_Y(A) = g_X(\xi \in A)$ дает точную нижнюю оценку вероятности $P(\xi \in A)$. Другими словами, нечеткая величина ξ отличается от обычной случайной величины тем, что точный закон ее распределения неизвестен, а известны лишь оценки вероятностей. При этом, если g_Y нижняя вероятность, то величина ξ имеет вероятностный закон распределения, который описывается вероятностной мерой из семейства вероятностных мер $\Xi = \{P | P \geqslant g_Y\}$. С учетом этого, для нечеткой величины ξ можно ввести в рассмотрение количественные оценки математического ожидания, моментов различных порядков, дисперсии следующим образом:

1) величина $\underline{M}[\xi] = \inf\limits_{P \in \Xi} \sum\limits_{i=1}^{m} P\{y_i\} y_i$ называется точной нижней оценкой математического ожидания нечеткой величины ξ. (Другой более общей моделью неточных вероятностей является задание нижних (верхних) оценок математических ожиданий функций . Хотя эта модель позволяет делать более точные выводы, но, по-видимому, наиболее оптимальная модель должна быть результатом компромисса между точностью и сложностью модели.)

2) величина $\overline{M}[\xi] = \sup_{P \in \Xi} \sum_{i=1}^{m} P\{y_i\} y_i$ называется точной верхней оценкой математического ожидания нечеткой величины ξ.

3) $\underline{M}[\xi^n]$, $\overline{M}[\xi^n]$ соответственно точные верхняя и нижняя оценки момента *n*-того порядка;

4) величина $\overline{D}[\xi] = \sup_{P \in \Xi} \left(\sum_{i=1}^{m} P\{y_i\} y_i^2 - \left[\sum_{i=1}^{m} P\{y_i\} y_i \right]^2 \right)$ называется точной верхней оценкой дисперсии нечеткой величины ξ.

Следующие результаты дают представление, как находить указанные характеристики для простейших семейств вероятностных мер Ξ.

Утверждение 6. *Пусть нечеткая величина ξ описывается семейством вероятностных мер $\Xi = \{P = \alpha P_1 + \beta P_2 | \alpha + \beta = 1, \alpha, \beta \geq 0\}$. Будем считать, что вероятностная мера P_i, $i = 1,2$, описывает вероятностное распределение случайной величины ξ_i, тогда*

$\underline{M}[\xi] = \min\{M[\xi_1], M[\xi_2]\}$

Доказательство. Пусть $P = \alpha P_1 + (1-\alpha) P_2$, где $\alpha \in [0,1]$. Тогда

$$\underline{M}[\xi] = \inf_{\alpha \in [0,1]} \sum_{i=1}^{m} (\alpha P_1\{y_i\} + (1-\alpha) P_2\{y_i\}) y_i =$$

$$= \inf_{\alpha \in [0,1]} (\alpha M[\xi_1] + (1-\alpha) M[\xi_2]) = \min\{M[\xi_1], M[\xi_2]\}.$$

Утверждение доказано.

Следствие. *Пусть нечеткая величина ξ описывается семейством вероятностных мер $\Xi = \left\{ P = \sum_{i=1}^{k} \alpha_i P_i \middle| \sum_{i=1}^{k} \alpha_i = 1, \alpha_i \geq 0 \right\}$. Введем также в рассмотрение случайные величины ξ_i, описываемые вероятностными распределениями P_i, тогда $\underline{M}[\xi] = \min_{i=1,\ldots,k} \{M[\xi_i]\}$.*

Утверждение 7. *Пусть нечеткая величина ξ описывается семейством вероятностных мер $\Xi = \{P = (\alpha + 0,5) P_1 + (0,5 - \alpha) P_2 | |\alpha| \leq 0,5\}$. Введем также в рассмотрение случайные величины ξ_i, описываемые вероятностными распределениями P_i, тогда*

1) $\overline{D}[\xi] = \max\{D[\xi_1], D[\xi_2]\}$, *если* $|D[\xi_1] - D[\xi_2]| \geq (M[\xi_1] - M[\xi_2])^2$;

2) $\overline{D}[\xi] = \frac{(D[\xi_1] - D[\xi_2])^2}{4(M[\xi_1] - M[\xi_2])^2} + \frac{1}{4}(M[\xi_1] - M[\xi_2])^2 + \frac{1}{2}(D[\xi_1] + D[\xi_2])$, *в противном, случае.*

Доказательство. По определению $\overline{D}[\xi] = \sup_{\alpha \in [-0.5, 0.5]} Q(\alpha)$, где

$$Q(\alpha) = \sum_{i=1}^{m} ((\alpha + 0,5)P_1\{y_i\} + (-\alpha + 0.5)P_2\{y_i\})\, y_i^2 =$$

$$- \left[\sum_{i=1}^{m} ((\alpha + 0,5)P_1\{y_i\} + (-\alpha + 0.5)P_2\{y_i\})\, y_i\right]^2 .$$

Упрощая *Q(α),* получим

$$Q(\alpha) = -\alpha^2 (M[\xi_1] - M[\xi_2])^2 + \alpha (D[\xi_1] - D[\xi_2]) +$$

$$+0,5 (D[\xi_1] - D[\xi_2]) + 0,25 (M[\xi_1] + M[\xi_2])^2 .$$

Ясно, что многочлен *Q(α)* в точке $\alpha = \frac{D[\xi_1] - D[\xi_2]}{2(M[\xi_1] - M[\xi_2])^2}$ имеет максимум. Учитывая ограничения, накладываемые на *α,* и подставляя α в выражение для *Q(α),* легко проверить справедливость утверждения. Формулу утверждения 7 можно упростить следующим образом.

Следствие утверждения 7. *Пусть выполняются условия, указанные в утверждении 7, причем* $D[\xi_1] \geqslant D[\xi_2]$, *тогда*

$$D[\xi] = 0,25 (k - 1)^2 (M[\xi_1] - M[\xi_2])^2 + D[\xi_1],$$

где $k = \min \left\{ \frac{D[\xi_1] - D[\xi_2]}{(M[\xi_1] - M[\xi_2])^2}, 1 \right\}$.

Рассмотрим, как находить точные нижние и верхние оценки математического ожидания, если нечеткая величина описывается 2-монотонной мерой. Для этого воспользуемся следующей вспомогательной леммой.

Лемма 6. *Пусть f(x) - неотрицательная измеримая (интегрируемая) функция на вероятностном пространстве* (X, P, Ω). *причем* $h = \sup_{x \in X} f(x)$. *Тогда*

$$\int_X f(x) dP(x) = \int_0^h P\{F(\alpha)\}\, d\alpha. \quad \textit{(7) где } F(\alpha) = \{x \in X | f(x) > \alpha\}.$$

Доказательство. Докажем эту лемму вначале для простых измеримых функций. Пусть $\{\alpha_1, \alpha_2, ..., \alpha_m\}$ - это конечное множество значений, которое принимает простая измеримая функция, $A_i = \{x \in X | f(x) = \alpha_i\}$. Тогда согласно определению интеграла по Лебегу $\int_X f(x) dP(x) = \sum_{i=1}^{m} \alpha_i P\{A_i\}$. Пусть $\alpha_0 = 0$, тогда $\int_0^h P\{F(\alpha)\}\, d\alpha = \sum_{i=1}^{m} \int_{\alpha_{i-1}}^{\alpha_i} P\{F(\alpha)\}\, d\alpha$. В последнем выражении $P\{F(\alpha)\} = P\{F(\alpha_i)\}$ при $\alpha \in (\alpha_{i-1}, \alpha_i]$.Поэтому

$$\int_0^h P\{F(\alpha)\}\,d\alpha = \sum_{i=1}^{m} P\{F(\alpha_i)\}(\alpha_i - \alpha_{i-1}) =$$

$$= \sum_{i=1}^{m} P\{F(\alpha_i)\}\alpha_i - \sum_{i=0}^{m-1} P\{F(\alpha_{i+1})\}\alpha_i =$$

$$= P\{F(\alpha_m)\}\alpha_m + \sum_{i=1}^{m}(P\{F(\alpha_i)\} - P\{F(\alpha_{i+1})\})\alpha_i - P\{F(\alpha_1)\}\alpha_0.$$

Поскольку $P\{F(\alpha_m)\} = P\{A_m\}$, $F(\alpha_i)\backslash F(\alpha_{i+1}) = A_i$ и $\alpha_0 = 0$, в результате получаем: $\int_0^h P\{F(\alpha)\}\,d\alpha = \sum_{i=1}^{m} P\{A_i\}\alpha_i$, т.е. формула (7) истинна для случая простых измеримых функций.

Рассмотрим общий случай. Пусть *f(x)* - интегрируемая функция, тогда существует последовательность простых измеримых функций $\{f_n(x)\}_{n=1}^{\infty}$, что $f_n(x) \leqslant f(x)$, $\lim_{n\to\infty} f_n(x) = f(x)$, и $\int_X f(x)dP(x) = \lim_{n\to\infty}\int_X f_n(x)dP(x)$. Ясно, что $\lim_{n\to\infty} F_n(\alpha) = F(\alpha)$, где $F_n(\alpha) = \{x \in X | f_n(x) > \alpha\}$. Поэтому

$$\int_X f(x)dP(x) = \lim_{n\to\infty}\int_X f_n(x)dP(x) = \lim_{n\to\infty}\int_0^h P\{F_n(\alpha)\}\,d\alpha = \int_0^h P\{F(\alpha)\}\,d\alpha.$$

Лемма доказана.

Теорема 10. *Пусть $\xi(\omega)$ - неотрицательная нечеткая величина, заданная на конечном нечетком пространстве (Ω, μ) и μ - монотонная мера. Тогда*

$$\underline{M}[\xi] = \int_0^h \mu\{F(\alpha)\}\,d\alpha. \quad (8)$$

где $h = \max_{\omega\in\Omega}\xi(\omega)$ *и* $F(\alpha) = \{\omega \in \Omega | \xi(\omega) > \alpha\}$.

Доказательство. Данная теорема легко следует из леммы 6. Действительно, по определению $\underline{M}[\xi] = \inf_{P\in\Xi}\int_\Omega \xi(\omega)dP(\omega)$, или, используя лемму 6, $\underline{M}[\xi] = \inf_{P\in\Xi}\int_0^h P\{F(\alpha)\}\,d\alpha$. По условию для любой вероятностной меры $P \in \Xi$, $P \geqslant \mu$, поэтому $\underline{M}[\xi] \geqslant \int_0^h \mu\{F(\alpha)\}\,d\alpha$.

Кроме того, поскольку нечеткая мера μ является 2-монотонной и множества *F(α)* линейно упорядочены отношением включения, т.е. $F(\alpha_1) \supseteq F(\alpha_2)$, если $\alpha_1 < \alpha_2$, то найдется вероятностная мера

P, $P \geqslant \mu$, что $P\{F(\alpha)\} = \mu\{F(\alpha)\}$, $\alpha \in [0,1]$, т.е. формула (8) дает точную нижнюю оценку математического ожидания нечеткой величины.

Замечания

1. Формулу (8) можно использовать и для вычисления оценок моментов болеевысокого порядка. В этом случае $h = \max_{\omega \in \Omega} \xi^n(\omega)$ и $F(\alpha) =$

$= \{\omega \in \Omega | \xi^n(\omega) > \alpha\}$.

2. Формулу (8) можно применять также и для нечетких величин, описываемых нечеткими мерами, которые называются нижними вероятностями. В этом случае получается неточная нижняя оценка математического ожидания.
3. Пусть μ - это 2-альтернирующая мера, тогда формула (8) дает точную верхнюю оценку математического ожидания нечеткой величины.
4. Замечание 2 также справедливо и для верхних вероятностей.
5. Формула (8) это ничто иное, как интеграл Шоке по нечеткой мере.

Заключение

В данном разделе была рассмотрена модель неточных вероятностей, основанная на теории нечетких мер. При этом описание основных базовых понятий, таких, как условная нечеткая мера, принцип обобщения, нечеткая величина, ее характеристики, строилось подобным образом, как и в классической теории вероятностей. Подчеркнем, что в основе данного построении лежат семейства вероятностных мер, из которых вытекает вероятностная интерпретация нечетких мер, когда мы рассматриваем их в качестве верхних или нижних оценок вероятностей. Возможны и другие варианты построения основных базовых понятий. Например, можно строить условную нечеткую меру, используя в качестве базового следующее выражение $P(A \cap B) = P(A|B)P(B)$, т.е. можно пытаться выбирать условную нечеткую меру таким образом, чтобы из условного и безусловного нечеткого распределения можно было бы получить совместное распределение. Такая модель оказывается близкой но характеру к моделям неточных вероятностей второго порядка , и является более подходящей для моделей неточного вероятностного вывода. Другой вариант обобщения рассмотренных понятий связан с рассмотрением аналогичных конструкций в рамках обобщенных точных вероятностей. В этом случае мы должны рассматривать

множества не вероятностных мер, а ненормированных аддитивных мер и числовые характеристики нечетких величин следует рассчитывать по множеству $\Xi = \{P | P \geqslant g_Y\}$ ненормированных аддитивных мер P, которые мажорируют сверху обобщенную точную нижнюю вероятность g_Y. Это позволит расширить полученные результаты на более широкий класс нижних вероятностей. Дальнейшие обобщения данных понятий на все множество нечетких мер может проводиться, например, за счет использования требования монотонности, как что делается, например, при определении условных монотонных мер. За пределами данного исследования осталось много других важных проблем. Как, например, определить независимость или декартово произведение нечетких мер? Как отмечалось выше, данные проблемы также пока еще не имеют однозначного решения. Мы склоняемся к следующему определению декартового произведения. Пусть g_X, g_Y точные нижние вероятности, определенные на конечных алгебрах $\mathfrak{A}_X$ и $\mathfrak{A}_Y$ конечных пространств X и Y, тогда декартово произведение $g_X \times g_Y$ на алгебре $\mathfrak{A}_X \times \mathfrak{A}_Y$ можно определить по формуле:

$$g_X \times g_Y = \bigwedge_{P_X \in \Xi_X, P_Y \in \Xi_Y} P_X \times P_Y,$$

где $\Xi_X = \{P_X | P_X \geqslant g_X\}$, $\Xi_Y = \{P_Y | P_Y \geqslant g_Y\}$.

Данное определение легко распространяется на обобщенные точные нижние вероятности. Другие обобщения также могут проводиться с помощью требования монотонности.

Дальнейшие исследования могут также проводиться и по практическому использованию предложенных конструкций в интервальных статистических моделях, в распознавании образов, приближенных рассуждениях. Следует подчеркнуть, что многие модели приближенных рассуждений, основанные на теории возможностей, можно обосновать в рамках понятий нижней (верхней) вероятности.

II. Размерности

1. Введение в теорию размерностей

1.1. О понятие размерности

В 1912 году Пуанкаре описал интуитивное понятие размерности, не давая его точную математическую формулировку.

«Из всех теорем analysis situs наиболее важной является та, которую мы выражаем, говоря, что пространство имеет три измерения. Именно это предложение мы собираемся рассмотреть, поставив вопрос в следующее форме: когда мы говорим, что пространство имеет три измерения, то чтб мы под этим подразумеваем ?...

«... Если для того, чтобы разбить континуум, достаточно рассматривать в качестве разбивающих множеств некоторое число различных между собой элементов, то мы скажем, что этот континуум имеет одно измерение; если, напротив, для того, чтобы разбить континуум, в качестве разбивающих множеств необходимо рассматривать систему элементов, которые сами образуют один или несколько континуумов, мы будем говорить, что этот континуум имеет несколько измерений.

«Если для того, чтобы разбить континуум *С,* достаточны множества, образующие один или несколько континуумов одного измерения, то мы будем говорить, что С — континуум двух измерений; если достаточны множества, которые образуют один или несколько континуумов самое большее двух измерений, мы будем говорить, что С—континуум трех измерений и т. д.

«Чтобы оправдать это определение, необходимо продумать, не совпадает ли этот путь с тем, которым геометры вводят понятие трех измерений с самого начала своих работ. Что же мы видим? Обычно они начинают с определения поверхностей, как границ тел или кусков пространства, линий, как границ поверхностей, точек, как границ линий, и устанавливают, что этот процесс не может быть проведен дальше.

«Это и есть как раз идея, высказанная выше; чтобы разбить пространства, необходимы множества, называемые поверх ностями; чтобы разбить поверхности, необходимы множества,, называемые линиями; чтобы разбить линии, необходимы множества, называемые точками; мы не можем двигаться дальше, и точка не может быть разбита, но точка не является континуумом. Тогда линии, которые могут быть разбиты множествами, не являющимися континуумами, будут континуумами одного измерения; поверхности, которые могут быть разбиты непрерывными множествами одного измерения, будут континуумами двух измерений; и, наконец, пространство, которое может быть разбито непрерывными множествами двух измерений, будет континуумом трех измерений».

Пуанкаре очень глубоко проник в существо вопроса, подчеркивая индуктивную природу геометри ческого смысла размерности и возможность разбиения пространства подмножествами более низкой размерности. Годом позже Брауэр, использовав идею Пуанкаре,

построил точное и топологически инвариантное определение размерности для очень широкого класса пространств (локально-связных метрических со счетной базой), эквивалентное определению, которым мы пользуемся сегодня.
В течение нескольких лет работа Брауэра оставалась почти незамеченной. Затем в 1922 году, независимо от Брауэра и друг от друга, Менгер и Урысон вновь ввели понятие размерности Брауэра с важными усовершенствованиями; и,что, пожалуй, самое главное, они оправдали новое понятие, сделав его краеугольным камнем теории, внесшей единство и порядок в большую область геометрии.
Определение размерности, которое мы примем в этой книге, восходит к Менгеру и Урысону. В формулировке Менгера оно читается так:
a) пустое множество имеет размерность —1,
b) размерность пространства X есть наименьшее целое число n такое, что каждая точка $p \in X$ обладает произвольно малыми окрестностями, границы которых имеют размерность, меньшую, чем n.

Ранние понятия размерности

До появления теории множеств математики употребляли размерность только в неопределенном смысле. Конфигурация называлась n-мерной, если наименьшее число действительных параметров, необходимых для того, чтобы описать ее точки некоторым не точно определенным способом, было равно n. Опасность и противоречивость такого подхода сделались ясными после **двух знаменитых открытий конца 19-го века: канторовского взаимно однозначного соответствия между точками линии и точками плоскости, и пеановского непрерывного отображения отрезка на весь квадрат**. Первое разрушило чувство, что плоскость богаче точками, нежели линия, и показало, что размерность может изменяться при взаимно однозначных отображениях. Второе противоречило убеждению, что размерность пространства может быть определена, как наименьшее число непрерывных действительных параметров, требуемых для того, чтобы описать пространство, и показало, чго размерность может возрастать при однозначных непрерывных отображениях.
Чрезвычайно важный вопрос оставался открытым (ответ на него был дан лишь в 1911 году Брауэром): возможно ли при $m \neq n$ между n-мерным эвклидовым пространством (обычным пространством n действительных переменных) и m-мерным эвклидовым пространством установить соответствие, соединяющее свойства конструкций Кантора и Пеано, т. е. соответствие, которое одновременно взаимно однозначно

и непрерывно? Этот вопрос был критическим, так как существование между *n*-мерным и *m*-мерным эвклидовыми пространствами соответствия указанного типа показало бы, что размерность (в том естественном смысле, что *n*-мерное эвклидово пространство имеет размерность *n)* не имеет никакого топологического значения! Класс топологических отображений оказался бы, следовательно, слишком широким, для того чтобы он мог иметь какое-либо реальное применение в геометрии.

Топологическая инвариантность размерности эвклидовых пространств

Первое доказательство того, что *n*-мерное и *m*-мерное эвклидовы пространства при $n \neq m$ не гомеоморфны, было дано Брауэром в его работе: «Beweis der Invarianz der Dimensionenzahl». Однако это доказательство ясно не обнаруживало некоторых простых топологических свойств *n*-мерного эвклидова пространства, отличающих его от *m*-мерного эвклидова пространства и являющихся причиной отсутствия гомеоморфизма между этими пространствами. Более проницательным было поэтому доказательство Брауэра в 1913 году, когда он ввел свою „размерность" — **целочисленную функцию пространства**, которая была топологически инвариантна по самому своему определению. Брауэр показал, что „размерность" *n*-мерного эвклидова пространства в точности равна *n* (и, следовательно, заслуживает своего названия).

Тем временем Лебег другим путем подошел к доказательству того, что размерность эвклидова пространства топологически инвариантна. Он заметил, что квадрат может быть покрыт произвольно малыми „кирпичами" таким образом, что никакая точка квадрата не содержится более, чем в трех из этих кирпичей; но если эти кирпичи достаточно малы, то по крайней мере три из них имеют общую точку. Подобным же образом, куб *n*-мерного эвклидова пространства может быть разложен на произвольно малые кирпичи так, чтобы не более чем $n+1$ из этих кирпичей пересекались. Лебег высказал предположение, что это число $n+1$ не может быть уменьшено, т. е., что для любого разложения на достаточно малые кирпичи должна существовать точка, принадлежащая, по крайней мере, $n+1$ кирпичу. Первое доказательство этой теоремы было дано Брауэром. Теорема Лебега также обнаруживает топологическое свойство *n*-мерного эвклидова пространства, отличающее его от *m*-мерного эвклидова пространства, и, следовательно, влечет за собой топологическую инвариантность размерности эвклидовых пространств.

Размерность множеств более общей природы

Новое понятие размерности, как мы уже видели, придает точный смысл утверждению, что *n*-мерное эвклидово пространство имеет размерность *n*, и тем самым значительно выясняет его топологическую структуру. Другой особенностью, которая сделала новое понятие размерности вехой в развитии геометрии, была общность объектов, к которым оно может быть приложено. Отсутствие точного определения размерности, неудовлетворительное с эстетической и методологической точек зрения, не вызывало, однако, никаких реальных трудностей, поскольку геометрия ограничивалась изучением относительно простых фигур, таких, как полиэдры и многообразия. Несомненно, в каждом частном случае можно было установить, какую размерность надо приписать каждой из этих фигур. Это положение радикально изменилось после открытия Кантора, с развитием теории множеств. Эта новая ветвь математики чрезвычайно расширила класс сбъектов, которые ложно рассматривать как „геометрические", и обнаружила конфигурации такой сложности, какую раньше нельзя было себе представить. Поставить в соответствие каждому из таких объектов число, которое можно было бы разумно назвать размерностью, было отнюдь не тривиальной задачей. Какое число, например, надо взять в качестве размерности неразложимого континуума Брауэра или «кривой» Серпинского, каждая из точек которой является точкой ветвления? Теория размерности дает полный ответ на эти вопросы. Каждому множеству точек эвклидова пространства (и даже каждому подмножеству гильбертова пространства), какое бы оно ни было «патологическое», она ставит в соответствие некоторое целое число, которое и на интуитивных и на формальных основаниях заслуживает быть названным его размерностью.

Различные подходы к понятию размерности

Прежде чем перейти к систематическому изучению теории размерности, остановимся на рассмотрении других возможных способов определения размерности.

Мы уже упоминали о методе Лебега доказательства теоремы об инвариантности размерности эвклидовых пространств. Этот метод может быть очень хорошо использован для установления общего понятия размерности: размерность Лебега некоторого пространства есть наименьшее целое число *n,* обладающее тем свойством, что пространство может быть разложено на произвольно малые области, не

более чем $n+1$ из которых, пересекаются. Оказывается, что введенная этим методом размерность совпадает с размерностью, которая восходит к Брауэру, Менгеру и Урысону. Приведем другие примеры, показывающие, как топологические исследования совершенно различной природы приводят к этому же понятию размерности.

А) Пусть

$$f_i(x_1, \ldots, x_n), \qquad i = 1, \ldots, m, \quad \textbf{(1)}$$

m непрерывных функций от n действительных переменных принимающих действительные значения, или, что то же самое, m действительных функций точки n-мерного эвклидова пространства. Один из основных фактов анализа состоит в том, что система m уравнений с n неизвестными,

$$f_i(x_1, \ldots, x_n) = 0; \quad \textbf{(2)}$$

вообще говоря, не имеет решения, если $m>n$. Слова «вообще говоря» следующим образом могут быть сделаны точными: очень мало изменив функции f_i, можно получить новые непрерывные: функции g_i, такие, что новая система

$$g_i(x_1, \ldots, x_n) = 0 \quad \textbf{(3)}$$

не имеет решения. С другой стороны, существуют системы n уравнений с n неизвестными, которые разрешимы и которые остаются разрешимыми при достаточно малых изменениях их левых частей. Это свойство n-мерного эвклидова пространства может быть положено, в основу общего понятия размерности. Пространство X можно было бы называть n-мерным, если n есть наибольшее целое число, такое, что существует n непрерывных действительных функций (1), определенных на X, обладающих тем свойством, что система уравнений (2) имеет решение, которое существенно в указанном выше смысле. Оказывается, что эта «размерность» снова совпадает с размерностью Брауэра, Менгера и Урысона.

В) Эта проблема является видоизменением А). Рассмотрим непрерывное отображение пространства в n-мерную сферу. Каждую точку n-мерной сферы можно рассматривать, как единичный вектор («направление») в $(n+1)$-мерном эвклидовом пространстве, так что вместо того, чтобы говорить о непрерывных отображениях в n-мерную сферу, можно говорить о непрерывном поле не равных нулю векторов $(n+1)$-мерного эвклидова пространства. Пусть C — замкнутое множество пространства X. Пусть на C определено непрерывное поле таких векторов. Можно ли тогда, не изменяя это поле на C, продолжить его в непрерывное поле неравных нулю векторов $(n+1)$-мерного эвклидова пространства, определенное на' всем X? Оказывается, что размерность X есть наименьшее

число *n*, для которого такое продолжение возможно для каждого замкнутого множества *C* и каждого определенного на *C* непрерывного поля таких векторов; в терминах отображений в *n*-мерную сферу, — наименьшее целое число *n*, обладающее тем свойством, что любое непрерывное отображение любого замкнутого подмножества $C \subset X$ в *n*-мерную сферу может быть продолжено на все *X* .

С) Другой подход к понятию размерности возникает из теории гомологии. Рассмотрим одномерные циклы (грубо говоря, непрерывные замкнутые кривые) на двумерном многообразии. Некоторые из них ограничивают двумерные части многообразия или, в терминологии теории гомологии, являются ограничивающими циклами. С другой стороны, никакой двумерный цикл (за очевидным исключением нулевого двумерного цикла, все из коэфициентов которого равны нулю) не может ограничивать на двумерном многообразии, потому что не существует ничего трехмерного, что он мог бы ограничивать. Аналогично, *n*-мерное многообразие содержит не равные нулю ограничивающие *m*-мерные циклы для каждого *m,* меньшего чем *n*, но содержит только нулевой ограничивающий *n*-мерный цикл. Далее, теория гомологии может быть применена к произвольному компактному метрическому пространству. Значит, можно определить «гомологическую размерность» компактного метрического пространства, как наибольшее целое число *n,* для которого при подходящим образом выбранных коэффициентах существует не нулеиой ограничивающий *(n*—1)-мерный цикл..

Оказывается, что так определенная гомологическая размерность также совпадает со стандартной размерностью.

Д) Интуитивное восприятие размерности связывает со словом одномерный объекты, имеющие длину (или линейную меру), со словом двумерный— объекты, имеющие площадь (или двумерную меру), со словом трехмерный — объекты, имеющие объем (или трехмерную меру) и так далее. Попытка сделать это интуитивное ощущение точным встречает препятствие, состоящее в том, что **размерность является топологическим понятием, в то время как мера — понятие метрическое.** Рассмотрим, однако, вместе с данным метрическим пространством все метрики, совместимые с его топологической структурой. Мы увидим, что размерность пространства *X* может быть охарактеризована как наибольшее действительное число *p,* для которого *X в каждой метризации* имеет положительную *p*-мерную хаусдорфову меру.

1.2. Виды размерностей

1.2.1. Размерность 0

Топология, в сущности, состоит в изучении связностной структуры пространств**. Понятие связного пространства, которое в своей форме восходит к Хаусдорфу и Леннесу, можно рассматривать как коренное понятие, от которого, прямо или косвенно, происходит значительная часть важных понятий топологии (теория гомологии или теория «алгебраической связности», локальная связность, размерность и т. д.).**

Пространство связно, если оно не может быть разбито на два непустых непересекающихся открытых множества. Другими словами, пространство связно, если, за исключением пустого множества и всего пространства, не существует никакого множества, граница которого пуста. (Заметим, что множество, граница которого пуста, одновременно открыто и замкнуто, и наоборот.)

В этом разделе мы имеем дело с пространствами, которые являются несвязными в чрезвычайно сильном смысле, а именно, имеют так много открытых множеств с пустой границей, что каждая точка содержится в произвольно малых множествах этого типа.

Определение размерности 0

Определение 1. Пространство X имеет *размерность* 0 *в точке p,* если p обладает произвольно малыми окрестностями с пустой границей, т. е., если для каждой окрестности U точки p существует окрестность V точки p такая, что

$$V \subset U,$$

$$Fr\,V = 0.$$

(Под окрестностью точки мы понимаем любое открытое множество, содержащее эту точку.)

Непустое пространство X имеет *размерность* 0, $\dim X = 0$, если X имеет размерность 0 в каждой своей точке.

А) Очевидно, что свойство пространства быть нульмерным, или быть нульмерным в точке *p,* топологически инвариантно.

В) Нульмерное пространство может быть также определено как непустое пространство, в котором существует базис, состоящий из множеств, которые одновременно открыты и замкнуты.

Пример 1. Каждое непустое конечное или счетное пространство X нульмерно. Действительно, пусть U—произвольная окрестность

некоторой точки *p.* Пусть *r*—-положительное число, такое, что сферическая окрестность точки *p* радиуса *r* (множество всех точек, расстояние которых от *p* меньше *r)* содержится в *U.* Пусть x_1, x_2 ... суть точки *X,* занумерованные в некоторой последовательности, и $\rho(x_i,p)$ — расстояние от $\rho\ x_i$ до *p*. Существует положительное число *r'*, меньшее *r* и отличное от всех $\rho(x_i, p)$. Тогда сферическая окрестность точки *p* радиуса *r'* содержится в *U,* а ее граница пуста. Следовательно, *X* нульмерно.

В частности, множество $\mathfrak{R}$ действительных рациональных чисел нульмерно.

Пример 2. Множество $\mathfrak{I}$ действительных иррациональных чисел нульмерно. Ибо, если дана произвольная окрестность *U* некоторой иррациональной точки *p,* то существуют рациональные числа ρ и σ, такие, что $\rho < p < \sigma$, и множество *V* иррациональных чисел, заключенных между ρ и σ, содержится в *U* . В пространстве $\mathfrak{I}$ иррациональных чисел множество *V* открыто и имеет пустую границу потому, что каждая иррациональная точка, которая является предельной точкой) *V,* находится между ρ и σ и, следовательно, принадлежит к *V.*

Пример 3. Канторов дисконтинуум *C* (множество всех действительных чисел, которые можно выразить в форме $\sum_{n=1}^{\infty} \frac{a_n}{3^n}$, где $a_n = 0$ или 2) нульмерен.

Пример 4. Любое множество действительных чисел, не содержащее никакого интервала, нульмерно.

Пример 5. Множество $\mathfrak{I}_2$ точек плоскости, обе координаты которых иррациональны, нульмерно. Ибо любая такая точка содержится в произвольно малых прямоугольниках, ограниченных прямыми, перпендикулярными к осям координат в пересекающими последние в точках с рациональными координатами. Границы же таких четырехугольников не пересекают $\mathfrak{I}_2$.

Пример 6. Множество $\mathfrak{R}_2^1$ точек плоскости, имеющих в точности одну рациональную координату, нульмерно. Потому что любая такая точка содержится в произвольно малых прямоугольниках, ограниченных прямыми, составляющими углы в 45° с осями координат и пересекающими их в точках с рациональными координатами. Границы таких прямоугольников не пересекают $\mathfrak{R}_2^1$.

Пример 7. Множество $\mathfrak{R}_n$ точек *n*-мерного эвклидова пространстпа E_n, все координаты которых рациональны, нульмерно. Ибо $\mathfrak{R}_n$ счетно.

Пример 8. Множество $\mathfrak{I}_n$ точек E_n, все координаты которых иррациональны, нульмерно. Это простое обобщение примера 5.

Замечание. Пусть $0 \leqslant m \leqslant n$. Обозначим через $\mathfrak{R}_n^m$ множество точек из E_n, которые имеют в точности *m* рациональных координат. В примерах 7 и 8 мы видели, что $\mathfrak{R}_n^n = \mathfrak{R}_n$ и $\mathfrak{R}_n^0 = \mathfrak{G}_n$ нульмерны. Оказывается (пример 12), что $\mathfrak{R}_n^m$ нульмерно для каждых *m* и *n,* но доказательство существенно зависит от «Теоремы сложения для нульмерных множеств» — теоремы 2. Простое же доказательство примера 6 не может быть обобщено.

Пример 9. Множество $\mathfrak{R}'_\omega$ точек гильбертова параллелепипеда I_ω, все координаты которых рациональны, нульмерно(это множество несчетно). Пусть

$$a = (a_1, \quad a_2, \ \ldots)$$

— произвольная точка I_ω, и *U* — окрестность *a* в I_ω. Взяв *n* достаточно большим, а p_i и q_i достаточно близкими к a_i, $p_i < a_i < q_i$, $i = 1, \ldots, n$, получим окрестность точки *a,* содержащуюся в *U,* состоящую из точек $x = (x_1, x_2, \ldots)$ параллелепипеда I_ω, первые *n* координат которых ограничены условием:

$$p_i < x_i < q_i \qquad (1)$$

(остальные координаты ограничены только условием

$$|x_i| \leqslant \frac{1}{i}, \qquad (2)$$

которое, конечно, всегда выполняется в I_ω).

Мы должны показать, что для каждого $\varepsilon > 0$ можно найти натуральное число *n* и положительное число δ, такие, что если $q_i - p_i < \delta$ для $i \leq n$, то для всех *x,* удовлетворяющих (1) и (2), имеет место неравенство:

$$\left[\sum_{i=1}^{\infty}(x_i - a_i)^2\right]^{\frac{1}{2}} < \varepsilon. \qquad (3)$$

Чтобы показать зто, выберем *n* так, чтобы

$$\sum_{i=n}^{\infty}\frac{1}{i^2} < \frac{1}{8}\varepsilon^2,$$

и δ так, чтобы

$$n\delta^2 < \frac{1}{2}\varepsilon^2.$$

Если $q_i - p_i$ меньше δ для $i \leq n$, то для всех *x,* удовлетворяющих (1) и (2), имеем:

$$\sum_{i=1}^{\infty}(x_i - a_i)^2 < n\delta^2 + \sum_{i=n}^{\infty}\left(\frac{2}{i}\right)^2 < \frac{1}{2}\varepsilon^2 + \frac{1}{2}\varepsilon^2 = \varepsilon^2,$$

что доказывает (3).

Предположим теперь, что $a \in \mathfrak{R}'_\omega$. Взяв p_i и q_i иррациональными, получим окрестность V точки a, каждая из граничных точек которой в I_ω, имеет по меньшей мере одну иррациональную координату. Следовательно, V имеет пустую границу в $\mathfrak{R}'_\omega$, а это доказывает, что $\mathfrak{R}'_\omega$ нульмерно.

Пример 10. Множество $\mathcal{J}'_\omega$ точек гильбертова параллелепипеда, все координаты которых иррациональны, нульмерно. Доказательство подобно доказательству примера 9.

Пример 11. Множество $\mathfrak{R}_\omega$ точек гильбертова пространства, все координаты которых рациональны, не нульмерно. (В действительности $\dim R_\omega = 1$). (Сопоставьте это утверждение с утверждением примера 9.) Достаточно показать, что любая ограниченная окрестность U начала координат в $\mathfrak{R}_\omega$ имеет непустую границу. Рассмотрим прямую линию: $-\infty < x_1 < +\infty, 0 = x_2 = x_3 = \ldots$ Она содержит точки и в U и в $\mathfrak{R}_\omega \setminus U$. Отсюда следует, что можно найти рациональное число a_1 такое, что точка

$$p^1 = (a_1, 0, 0, \ldots)$$

содержится в U и находится от $\mathfrak{R}_\omega \setminus U$ на расстоянии, меньшем, чем 1. Подобным же образом, рассматривая прямую линию: $x_1 = a_1, -\infty < x_2 < \infty, 0 = x_3 = x_4 = \ldots$, мы определим точку

$$p^2 = (a_1, a_2, 0, \ldots)$$

при рациональном a_2 такую, что p^2 принадлежит U и находится от $\mathfrak{R}_\omega \setminus U$ на расстоянии, меньшем, чем 1/2. По индукции определим последовательность $\{p^n\}$:

$$p^n = (a_1 a_2, \ldots, a_n, 0, 0, \ldots),$$

где каждое a_n рационально, каждое p^n принадлежит U и $\rho(p^n, \mathfrak{R}_\omega \setminus U) < \frac{1}{n}$.

Отсюда легко следует, что точка p

$$p = (a_1, a_2, \ldots, a_k, a_{k+1}, \ldots),$$

k-я координата которой равна a_k, является граничной точкой окрестности U.

(p, действительно, есть точка $\mathfrak{R}_\omega$, потому что $\sum_{i=1}^{n} a_i^2 < d^2$, n — 1, 2, ..., где d — диаметр U; отсюда $\sum_{i=1}^{\infty} a_i^2 < \infty$.

Теорема 1. *Непустое подмножество нульмерного пространства нульмерно.*

Доказательство. Пусть p— произвольная точка непустого подмножества X' нульмерного пространства X, и пусть U' — некоторая

окрестность точки p в X'. Тогда существует окрестность U точки p в X такая, что

$$U' = U \cap X'.$$

Так как X нульмерно, существует одновременно открытое и замкнутое множество V такое, что

$$p \in V \subset U.$$

Пусть

$$V' = V \cap X'.$$

Тогда V' одновременно открыто и замкнуто в X', и

$$p \in V' \subset U',$$

так что X' нульмерно.

1.2.2. Отделение подмножеств

Определение 1. Если A_1, A_2 и B суть попарно непересекающиеся подмножества пространства X, то мы скажем, что A_1 и A_2 *отделены в X множеством B,* если $X \setminus B$ может быть разбито на два непересекающихся множества, открытых в $X\setminus B$ и содержащих соответственно A_1 и A_2, т. е. если существуют A_1' и A_2', для которых

$$X \setminus B = A_1' \cup A_2',$$

$$A_1 \subset A_1', \quad A_2 \subset A_2',$$

$$A_1' \cap A_2' = 0,$$

причем и A'_1 и A'_2, открыты в $X\setminus B$ (или, что то же, и A'_1 и A'_2 замкнуты в $X\setminus B$). Если A_1 и A_2 отделены пустым множеством, мы скажем просто, что A_1 и A_2 *отделены* в X.

А) A_1 и A_2 отделены в том и только в том случае, если существует множество A'_1 такое, что

$$A_1 \subset A_1',$$

$$A_1' \cap A_2 = 0,$$

и A'_1 одновременно открыто и замкнуто, т. е. имеет пустую границу. Ибо A'_2 равно тогда $X\setminus A'_1$.

В) Докажем теперь, что определению 1 эквивалентно

Определение 1′. Непустое пространство имеет размерность нуль, если каждая точка p и каждое замкнутое множество С, не содержащее p, могут быть отделены .

Доказательство. Предположим, что *X* нульмерно в смысле определения 1. Тогда, так как *X\C* есть окрестность, точки *p,* существует множество *V* такое, что

$$p \in V \subset X \setminus C,$$

причем *V* одновременно открыто и замкнуто. Из того, что

$$V \cap C = 0,$$

и предложения А) следует, что *p* и *C* отделены, что и требуется по определению 1'. Обратное доказывается подобным образом.

C) Связное нульмерное пространство состоит из одной единственной точки.

Доказательство. Допустим, что нульмерное пространство *X* содержит две различные точки *p* и *q*. Определение 1′ показывает, что *p* и *q* отделены. Следовательно, *X* несвязно.

D) Нульмерное пространство вполне несвязно, т. е. никакое связное его подмножество не содержит более, чем одну точку.

Доказательство. Это следует из теоремы 1 и предложения C).

E) В силу определения 1' очевидно, что пространство нульмерно, если любые два непересекающиеся замкнутые в нем множества могут быть отделены. Докажем обратное предложение: если пространство нульмерно, то любые два непересекающиеся замкнутые множества в нем могут быть отделены.

Доказательство. Путь *X* — нульмерно. Из определения 1' мы знаем, что любая точка $p \in X$ может быть отделена от любого замкнутого множества, не содержащего *p*. Пусть *C* и К—два непересекающиеся замкнутые множеста в *X*. Нам надопоказать, что *C* отделено от *K* в *X*. Для каждой точки $p \in X$ либо $p \cap C = 0$, либо $p \cap K = 0$.

Следовательно, для каждой точки *p* существует окрестность *U(p),* одновременно открытая и замкнутая и такая, что либо $U(p) \cap C = 0$, либо $U(p) \cap K = 0$. Так как пространство *X* имеет счетный открытый базис, существует последовательность $U_1, V_2 \ldots$, состоящая из множеств *U(p),* сумма элементов которой есть *X*. Определим новую последовательность множеств V_i следующим образом:

$$V_1 = U_1$$

$$V_i = U_i \setminus \bigcup_{k=1}^{i-1} U_k = U_i \cap \left(X \setminus \bigcup_{k=1}^{i-1} U_k\right) \qquad i = 2, 3, \ldots$$

Тогда

$$X = \bigcup_{i=1}^{\infty} V_i \qquad (1)$$

$$V_i \cap V_j = 0, \quad \text{если } i \neq j, \qquad (2)$$

V_i открько, (3)

либо $V_i \cap C = 0$, либо $V_i \cap K = 0$; (4)

(1), (2) и (4) очевидны. Чтобы доказать (3) заметим, что

$$\bigcup_{k=1}^{i-1} U_k$$

замкнуто, так что

$$X \setminus \bigcup_{k=1}^{i-1} U_k$$

открыто. Следовательно, V_i, как пересечение этого открытого множества с открытым множеством U_i, само открыто.
Пусть C' — сумма всех V_i, для которых $V_i \cap K = 0$, а K' — сумма остальных V_i. Тогда
$X = C' \cup K'$ по (1), $C' \cap K' = 0$ по (2), C' и K' открыты по (3) и $(C' \cap K') \cup (K' \cap C) = 0$ по (4);
отсюда следует, что $C \subset C'$ и $K \subset K'$. Таким образом, нужное отделение C от K осуществляется множествами C′ и K′.

F) Если C, и C_2 — непересекающиеся замкнутые множества пространства X, а A — нульмерное подмножество X, то существует замкнутое множество B, отделяющее C_1 от C_2, такое,, что $A \cap B = 0$.

Доказательство. Так как X нормально, существуютоткрытые множества U_1 и U_2, для которых

$$C_1 \subset U_1, \quad C_2 \subset U_2 \text{ и } \bar{U}_1 \cap \bar{U}_2 = 0. \quad (5)$$

(Пространство S *нормально* , если для любых двух непересекающихся замкнутых множеств C_1 и C_2 существуют открытые множества V_1 и V_2 такие, что

$$C_1 \subset V_1, C_2 \subset V_2 \text{ и } V_1 \cap V_2 = 0.$$

Простым следствием нормальности является существование открытых множеств U_1 и U_2 таких, что

$$C_1 \subset U_1, C_2 \subset U_2 \text{ и } \bar{U}_1 \cap \bar{U}_2 = 0.$$

Каждое метрическое пространство нормально).

Непересекающиеся множества $\bar{U}_1 \cap A$ и $\bar{U}_2 \cap A$ замкнуты в A и, так как dim A = 0, могут быть отделены в A, по предложению E), и мы получаем непересекающиеся множества C_1' и C'_2 такие, что

$$A = C'_1 \cup C'_2,$$

$$\bar{U}_1 \cap A \subset C'_1, \quad \bar{U}_2 \cap A \subset C'_2,$$

причем C_1' и C'_2, одновременно открыты и замкнуты в A. Отсюда

$(C_1' \cap \bar{U}_1) \cup (C_2' \cap \bar{U}_1) = 0,$ (6)

$(C_1' \cap \bar{C}_2') \cup (\bar{C}_1' \cap C_2') = 0;$ (7)

(5) и (6) влекут за собой

$(C_1' \cap \bar{C}_2') \cup (C_2' \cap \bar{C}_1') = 0.$ (8)

Далее, так как U_1 и U_2 открыты, из (6) вытекает, что

$(\bar{C}_1' \cap U_2) \cup (\bar{C}_2' \cap U_1) = 0,$

а отсюда, на основании (5),

$(\bar{C}_1' \cap C_2) \cup (\bar{C}_2' \cap C_1) = 0.$

Из (7), (8), (9) и того, что $\bar{C}_1 \cap \bar{C}_2 = C_1 \cap C_2 = 0$, следует, что никакое из непересекающихся множеств $C_1 \cup C_1'$ и $C_2 \cup C_2'$ не содержит предельной точки другого. Так как X вполне нормально, существует открытое множество W такое, что

$C_1 \cup C_1' \subset W$ и

$\bar{W} \cap (C_2 \cup C_2') = 0.$

(Пространство *X вполне нормально*, если каждое подпространство X нормально. Можно показать, что для любых двух непересекающихся подмножеств X_1 и X_2 вполне нормального пространства» никакое из которых не содержит предельных точек другого, существуют открытые множества W_1 и W_2, такие, что

$X_1 \subset W_1$, $X_2 \subset W_2$ и $W_1 \subset W_2 = 0$;

ясно, что $\bar{W}_1 \cap X_2 = 0$. Каждое метрическое пространство вполне нормально потому, что каждое подпространство метрического пространства является метрическим пространством.)

Граница $B = \bar{W} \setminus W$ отделяет C_1 от C_2 и не пересекается с $C_1' \cup C_2' = A$. Предложение F), таким образом, доказано.

1.2.3. Теорема сложения для нульмерных множеств

Сумма нульмерных множеств не обязана быть нульмерной. Это видно из разложения прямой на множества рациональных и иррациональных чисел, или на отдельные точки. Однако имеет место

Теорема 2. Теорема сложения для нульмерных множеств. *Пространство, являющееся суммой счетного числа нульмерных замкнутых множеству нульмерно.*

Доказательство. Предположим, что

$X = C_1 \cup C_2 \cup \ldots \cup C_i \cup \ldots,$

где каждое C_i замкнуто и нульмерно. Пусть K и L суть два непересекающихся замкнутых множества пространства X. Мы покажем, что K и L, могут быть отделены.

$K \cap C_1$ и $L \cap C_1$ — непересекающиеся замкнутые подмножества нульмерного пространства C_1. Следовательно (2E), существуют подмножества A_1 и B_1 множества C_1 замкнутые в C_1 и, следовательно, в X, такие, что

$$K \cap C_1 \subset A_1, \qquad L \cap C_1 \subset B_1,$$
$$A_1 \cup B_1 = C_1, \qquad A_1 \cap B_1 = 0.$$

Множества $K \cup A_1$ и $L \cup B_1$ замкнуты в X и не пересекаются. В силу нормальности пространства X, существуют открытые множества G_1 и H_1 для которых

$$K \cup A_1 \subset G_1, \qquad L \cup B_1 \subset H_1,$$
$$\bar{G}_1 \cap \bar{H}_1 = 0.$$

Следовательно,

$$G_1 \cup H_1 \supset C_1,$$
$$K \subset G_1, \qquad L \subset H_1,$$
$$\bar{G}_1 \cap \bar{H}_1 = 0.$$

Теперь повторим этот процесс, заменяя K и L множествами $\bar{G}_1$ и $\bar{H}_1$, а C_1 — множеством C_2. Эга замена приводит к открытым множествам G_2 и H_2, Для которых

$$G_2 \cup H_2 \supset C_2,$$
$$\bar{G}_1 \subset G_2, \qquad \bar{H}_1 \subset H_2,$$
$$\bar{G}_2 \cap \bar{H}_2 = 0.$$

По индукции построим последовательности $\{G_i\}$ и $\{H_i\}$, открытых в X множеств, для которых

$$G_i \cup H_i \cap C_i,$$
$$\bar{G}_{i-1} \subset G_i, \ \bar{H}_{i-1} \subset H_i,$$
$$\bar{G}_i \cap \bar{H}_i = 0.$$

Пусть $G = \bigcup G_i, \qquad H = \bigcup H_i,$

тогда G и H суть непересекающиеся открытые множества,

$$G \cup H \supset \bigcup_i C_i = X$$
$$K \subset G, \qquad L \subset H;$$

это и есть требуемое отделение.

Следствие 1. *Пространство, являющееся суммой счетного числа нульмерных F_σ множеств, нульмерно.*

(Под F_σ в пространстве мы понимаем сумму счетного числа замкнутых подмножеств. В метрическом пространстве любое открытое множество есть F_σ).

Следствие 2. *Сумма двух нульмерных подмножеств пространства X, по крайней мере одно из которых замкнуто, нульмерна.*

Доказательство. Пусть A и B нульмерны, и B замкнуто. Тогда $(A \cup B) \setminus B$ открыто в $A \cup B$. Как открытое множество в метрическом пространстве оно есть F_σ . Следствие 2 вытекает поэтому из следствия 1 и равенства

$$A \cup B = B \cup [(A \cup B) \setminus B].$$

Следствие 3. *Нульмерное пространство остается нульмерным после прибавления одной точки.*

(Предполагая, конечно, что расширенное пространство является метрическим со счетным базисом.)

Доказательство. Очевидно, в силу следствия 2.

Пример 12. Пусть $0 \leqslant m \leqslant n$. Обозначим через $\mathfrak{R}_n^m$ множество точек n-мерного эвклидова пространства E_n, имеющих точно m рациональных координат. Тогда $\mathfrak{R}_n^m$ нульмерно.

Каковы бы ни были m индексов $i_1, ..., i_m$, выбранных из чисел: 1, ..., n и m рациональных чисел $r_1, ..., r_m$, система уравнений

$$x_{i_1} = r_1, \quad x_{i_2} = r_2, \quad \ldots, \quad x_{i_m} = r_m \qquad (1)$$

определяет (n— m)-мерное линейное подпространство. Подмножество этого подпространства, состоящее из точек, у которых все остальные координаты иррациональны, обозначим через C_i. Каждое C_i изометрично $\mathfrak{I}_{n-m}$ и, следовательно, нульмерно (пример 8). Ясно, что каждое C_i замкнуто в $\mathfrak{R}_n^m$ и что сумма множеств C_i и есть $\mathfrak{R}_n^m$. Так как совокупность множеств C_i счетна, то из теоремы сложения для нульмерных множеств следует, что $\dim \mathfrak{R}_n^m = 0$.

Пример 13. Пусть $0 \leq m$. Обозначим через $\mathfrak{R}_\omega^m$ множество точек гильбертова параллелепипеда, имеющих в точности m рациональных координат. Тигда $\mathfrak{R}_\omega^m$ нульмерно. (Доказательство подобно доказательству примера 12 и использует пример 10).

1.2.4. Компакты

Рассмотрим следующие четыре свойства пространства *X:*

(0) *X* вполне несвязно.

(1) Любые две различные точки в *X* могут быть отделены.

(2) Любая точка может быть отделена от любого не содержащего ее замкнутого множества, т. е. *X* нульмерно (см. определение 1′),

(3) Любые два замкнутые непересекающиеся множества могут быть отделены.

Очевидно, из (3) следует (2), из (2) следует (1), из (1) следует (0). Обратно (см. предложение 2 Е)), из (2) следует (3) (это верно для пространств со счетным базисом. Для пространств без счетного базиса из (2) не следует (3). Свойства (0), (1) и (2), однако, не эквивалентны даже для метрических пространств со счетным базисом.

Пример 14. Серпинский приводит пример плоского множества, обладающего свойством (0), но не обладающего свойством (1).

Пример 15. Мы узнали из примера 11, что $\mathfrak{R}_\omega$ не обладает свойством (2), с другой стороны, оно обладает свойством (1). В самом деле, пусть p и q две точки из $\mathfrak{R}_\omega$, а i — индекс такой, что i-я координата p_i точки p отличаете» от i-й координаты q_i точки q; p_i и q_i, конечно, рациональны. Пусть p — произвольное иррациональное число, расположенное между p_i и q_i. Разложение $\mathfrak{R}_\omega$ на замкнутые непересекающиеся множества, определенные неравенствами:

$$x_i \leqslant p, \quad x_i \geqslant p,$$

осуществляет требуемое отделение p от q.

Тем не менее эквивалентность имеет место в важном случае:

A) Для компактов условия (0) — (3) эквивалентны. Остается доказать, что из (0) следует (1), и из (1) следует (2). Сначала докажем два предложения.

B) Пусть X—компакт, C — замкнутое подмножество компакта X, и p — точка из X. Тогда, если точка p может быть отделена от каждой точки $q \in C$, то точка p может быть отделена и от C.

Доказательство. Для каждой точки q множества C существуют два непересекающиеся одновременно замкнутые и открытые множества U_q и V_q такие, что $p \in U_q$, $q \in V_q$. Так как C — замкнутое подмножество компакта, то существует конечное число $q_1, \ldots, q_k$ точек q таких, что $V_{q_1} \cup \ldots \cup V_{q_k} \supset C$. Пусть

$$U = \bigcap_{i=1}^{k} U_{q_i}, \quad V = \bigcup_{i=1}^{k} V_{q_i};$$

тогда $p \in U$, $C \subset V$; кроме того U и V не пересекаются, а каждое из них и открыто, и замкнуто. Следовательно, p и C отделены. Таким образом, предложение B) доказано.

C) Пусть X—компакт, p— точка из X, и $M(p)$ — множество всех точек, которые не могут быть отделены от p. Тогда $M(p)$ связно.

Доказательство. Сначала мы покажем, что $M(p)$ замкнуто или, что то же, что $X \setminus M(p)$ открыто. Произвольная точка x принадлежит $X \setminus M$ (p) в том и только в том случае, если существует разложение

$$X = U \cup V,$$

$$U \cap V = 0,$$

$$x \in U, \quad p \in V,$$

где *U* и *V* открыты. Легко видеть, что $U \subset X \setminus M(p)$, т. е. каждая точка $x \in X \setminus M(p)$ имеет окрестность, содержащуюся в *X\M(p)*. Таким образом, *X\M(p)* открыто.

M(р), очевидно, содержит *p*. Допустим, что *M(p)* несвязно. Тогда

$$M(p) = C \cup K, \quad C \neq 0, \quad K \neq 0, \quad C \cap K = 0,$$

C и К замкнуты в *M(p)*. Допустим, что $p \in C$. Так как *M(p)* замкнуто в *X*, то С и *K* замкнуты в *X*. Следовательно, в силу нормальности пространства *X*, существует, открытое в *X* множество *U* такое, что $C \subset U$ и $\overline{U} \cap K = 0$. Так как

$$B(U) = FrU = \overline{U} \setminus U,$$

то

$$B(U) \cap M(p) = B(U) \cap (C \cup K) = 0.$$

Это означает, что каждая точка множества *B (U)* отделена от *p*. Так как *B(U)* замкнуто, то можно, применив В), получить такое одновременно открытое и замкнутое множество *V*, что $B(U) \subset V$ и $p \cap V = 0$., Таким образом, *U\V* содержит *p*. Легко видеть, что *V\V* можно представить и в виде $U \setminus \overline{V}$, так как $V = \overline{V}$, и в виде $\overline{U} \setminus V$, так как $B(U) \subset V$. Первое представление показывает, что *U\V* открыто, а второе — что *U\V* замкнуто. Но $(U \setminus V) \cap K = 0$. Следовательно, *p* отделено от точек множества $K \subset M(p)$. Однако это противоречит определению *M(p)*.

Доказательство того, что из (0) следует (1). Предположим, что *X* вполне несвязно. Для каждой точки *p* рассмотрим множество *M (p)*. В силу С), оно связно и поэтому, так как *X* вполне несвязно, состоит из одной точки *p*. Следовательно, любые две точки пространства *X* отделены.

Доказательство того, что из (1) следует (2). Это следует из В).

D) Среди компактов нульмерные пространства тождественны с вполне-несвязными пространствами.

Замечание 1. Предложение D) сохраняется также для пространств, которые только локально-компактны.

Замечание 2. *Неверно* (см. пример 16), что если пространство обладает свойством (0) или (1), то оно сохраняет это свойство при прибавлении к нему одной точки. Сравните это со следствием 3 теоремы 2. Таким образом, теорема сложения была бы неверна для теории размерности, в которой размерность 0 была бы определена как вполне-несвязность, или как возможность отделения любых пар точек.

Пример 16. Кнастер и Куратовский, приводят следующий пример плоского множества, вполне несвязного [свойство (0)], но теряющего это свойство и становящегося связным после прибавления к нему одной точки; обозначим через С канторов дисконтинуум (см. пример 3). Пусть *P*—-подмножество дисконтинуума С, состоящее из точек

$$P = \sum_{n=1}^{\infty} \frac{a_n}{3^n}$$

таких, что, начиная с достаточного большого *a*, числа a_n либо все равны 0, либо все равны 2 *(P*— счетное множество, состоящее из концевых точек, интервалов, которые надо удалить из [0,1], чтобы получить С *).* Пусть *Q* — множество остальных точек С . Пусть *a* — точка на плоскости с координатами $\left(\frac{1}{2}, \frac{1}{2}\right)$. Обозначим через *L(c)* отрезок, соединяющий переменную точку *c* множества С с точкой *a.* Обозначим для $p \in P$ через *L** (p) множество всех точек отрезка *L* (*p*), имеющих *рациональные* ординаты, а для $q \in Q$ через L* *(q)* — множество всех точек отрезка *L(q),* имеющих *иррациональные* ординаты. Тогда, как доказывают Кнастер и Куратовский,

$$X = \bigcup_{p \in P} L^*(p) \cup \bigcup_{q \in Q} L^*(q)$$

связно, хотя *X\a* вполне несвязно.

Так как *X* связно, оно не нульмерно, и так как нульмерное пространство остается нульмерным после прибавления одной точки (следствие 3 теоремы 2), то и *X\a* не нульмерно.

(На самом деле ,

$$\dim (X \setminus a) = \dim X = 1.$$

X\a является примером пространства, имеющего размерность 1, даже несмотря на то, что оно вполне-несвязно. Более того, Мазуркевич показал, чти для каждого конечного *n* существует вполне несвязное [в действительности, обладающее тем свойством, что любые две точки могут быть отделены] пространство-размерности *n*. Таким образом, тот факт, что пространство имеет положительную размерность, очень мало говорит о его связности.)

Пример 17. Серпинский (см. ссылку в примере 14) приводит пример пространства, обладающего свойством (1), но теряющего это свойство после прибавления одной точки.

1.3. Размерность *n*

Грубо говоря, пространство имеет размерность $\leq n$, если произвольно малые куски пространства, окружающие каждую точку, могут быть ограничены подмножествами размерности $\leq n$—1. Этот метод определения размерности является индуктивным методом, причем исходной точкой индукции является принятие пустого множества в качестве (—1)-мерного пространства.

1.3.1. Определение размерности *n*

Определение 1. Пустое множество и только пустое множество имеет *размерность* —1.

Пространство X имеет *размерность* $\leq n$ $(n \geq 0)$ *в точке p,* если p обладает произвольно малыми окрестностями, границы которых имеют размерность $\leq n$—1.

X имеет *размерность* $\leqslant n$, $\dim X \leqslant n$, если X имеет размерность $\leq n$ в каждой своей точке.

X имеет *размерность* n в точке p, если верно, что X имеет размерность $\leq n$ в p и неверно, что X имеет размерность $\leq n$ — 1 в p.

X имеет *размерность* n, если $\dim X \leqslant n$ верно, а $\dim X \leq n$ — 1 неверно.

X имеет *размерность* ∞, если $\dim X \leq n$ неверно для каждого n.

A) Очевидно, что свойство иметь размерность n (или иметь размерность n в точке p) топологически инвариантно. Размерность не является, однако, инвариантом непрерывных отображений. Проекция плоскости на прямую является примером непрерывного отображения, понижающего размерность, а пеановское отображение отрезка на весь квадрат—примером непрерывного отображения, повышающего размерность.

B) Условие $\dim X \leq n$ 0 эквивалентно существованию в X базиса, состоящего из открытых множеств, границы которых имеют размерность $\leq n$ — 1.

C) При $n = 0$ очевидно, что определения II 1 и III 1 совладают.

D) Пусть $\dim X = n$, n-конечное. Тогда для каждого $m \leq n$ X содержит некоторое m-мерное подмножество.

Доказательство. Так как $\dim X > n$ — 1, существует точка $p_0 \in X$ и окрестность U_0 точки p_0, обладающая тем свойством, что если V — произвольное открытое множество такое, что $p_0 \in V \subset U_0$, то

$$\dim Fr V \geqslant n - 1.$$

С другой стороны, так как $\dim X \leq n$, существует открытое множество V_0 такое, что $p_0 \in V_0 \subset U$, для которого

$$\dim Fr V_0 \leqslant n - 1.$$

Следовательно, граница множества V_o является подмножеством пространства X, размерность которого в точности равна n—1. Окончание доказательства предложения D) теперь очевидно.

Замечание. Утверждение предложения D) не может быть распространено на бесконечномерные пространства. Действительно, если справедлива континуум-гипотеза, то существуют даже такие бесконечномерные пространства, конечномерными подпространствами которых являются только счетные множества.

Пример 1. Прямая и интервал прямой имеют размерность 1.

Пример 2. Любой многоугольник имеет размерность 1.

Пример 3. Любое двумерное многообразие имеет размерность ≤2 (доказательство следует из примера 2).

Пример 4. n-мерное эвклидово пространство E_n имеет размерность $\leq n$. Индуктивное доказательство этого предоставляется читателю. Однако доказательство того, что размерность E_n в точности равна n, отнюдь не тривиально и будет являться основной целью следующего раздела.

Пример 5. Множество $\mathfrak{R}_\omega$ точек гильбертова параллелепипеда, все ковердинаты которых рациональны, одномерно. Мы уже видели , что $\dim \mathfrak{R}_\omega \geqslant 1$. Покажем теперь, что $\dim \mathfrak{R}_\omega \leqslant 1$, т. е. что каждая точка $p \in \mathfrak{R}_\omega$ может быть заключена в произвольно малые сферические окрестности в $\mathfrak{R}_\omega$, имеющие нульмерные границы. Обозначим через S сферу в гильбертовом пространстве с центром в начале координат и радиусом d<l, т. е. множество точек, находящихся на расстоянии d от начала координат. Совершенно ясно, что достаточно доказать, что $\dim S \cap \mathfrak{R}_\omega = 0$. Чтобы доказать это, поставим в соответствие каждой точке x сферы S :

$$x = (x_1, x_2, \ldots, x_i, \ldots),$$

точку

$$x' = (x_1, \frac{x_2}{2}, \ldots, \frac{x_i}{i}, \ldots);$$

x' содержится в гильбертовом параллелепипеде. Как легко можно показать, S обладает тем свойством, что последовательность $\{p^n\}$ точек из S сходится к точке $p \in S$ в том и только в том случае, если i-е координаты точек p^n сходятся к i-й координате точки p, $i=1$, 2.....Так как I_ω обладает тем же свойством, то переход от x к x' являемся гомеоморфизмом S на подмножество I_ω . Но этот гомеоморфизм, очевидно, переводит $S \cap \mathfrak{R}_\omega$ в подмножество $\mathfrak{R}'_\omega$, а мы знаем, что $\dim \mathfrak{R}'_\omega = 0$. Следовательно, $\dim S \cap \mathfrak{R}_\omega = 0$.

Теорема 1. *Подпространство пространства размерности ≤n имеет размерность ≤n.*

Доказательство. (По индукции.) Утверждение очевидно для n =—1. Предположим теперь, что оно справедливо для n — 1. Пусть X—

пространство размерности $\leq n$, X'— подпространство пространства X, и p— произвольная точка из X'. Пусть V — окрестность точки p в X. Тогда существует окрестность U точки p в X такая, что $U = U \cap X$. Так как $\dim X \leq n$, то существует множество V, открытое в X и такое, что

$$p \in V \subset U,$$

$$\dim FrV \leq n - 1.$$

Пусть $V' = V \cap X'$. Тогда V открыто в X', $p \in V' \subset U'$. Пусть B—граница множества V в X, и B' — граница множества) V' в X. ($B = \bar{V} \setminus V$, $B' = (\bar{V} \setminus \bar{V}') \cap X'$.) Тогда, как легко видеть, B' содержится в $B \cap X'$. По индуктивному предположению, $\dim B' \leq n$—1, что и требовалось доказать.

Теперь мы докажем, что определению 1 эквивалентно

Определение 1′. X имеет размерность $\leq n$, если каждая точка $p \in X$ может быть отделена от любого не содержащего ее замкнутого множества $C \subset X$ замкнутым множеством размерности $\leq n$ —1.

Доказательство. Допустим, что $\dim X \leq n$ в смысле определения 1. Множество $X \setminus C$ является окрестностью точки p и, следовательно, в силу регулярности') пространства X, существует другая окрестность V точки p, для которой

$$\bar{V} \subset X \setminus C.$$

(Пространство *регулярно* , если каждая окрестность U точки p содержит такую окрестность V точка p, что $\bar{V} \subset U$. Очевидно, что каждое метрическое пространство регулярно и что каждое нормальное пространство регулярно).

Тогда существует окрестность W точки p такая, что $\bar{W} \subset V$, а $B = FVrW$ имеет размерность $\leq n$ —1. Легко показать, что B отделяет p и C. Это доказывает, что $\dim X \leq n$ в смысле определения 1′.

Наоборот, если $\dim X \leq n$ в смысле определения 1', то $\dim X \leq n$ в смысле определения 1. Действительно, пусть U—окрестность точки p. Тогда $X \setminus U$ есть замкнутое множество, не содержащее p, и, следовательно, $X \setminus U$ может быть отделено от p замкнутым множеством B размерности $\leq n$—1. Это означает, что

$$X \setminus B = U' \cup V', \; p \in U', \; X \setminus U \subset V', \; U' \cap V' = 0,$$

причем U' и V открыты в $X \setminus B$ и, следовательно, в X. U' является окрестностью p, содержащейся в U, и так как граница FrU' содержится в B, то из теоремы 1 следует, что граница FrU' имеет размерность $\leq n$ — 1.

1.3.2. Размерность подпространств. Размерность сумм

При исследовании размерности подпространства X' пространства X иногда бывает удобно определять размерность X' с помощью окрестностей по отношению ко всему пространству X.

А) Подпространство X' пространства X имеет размерность $\leq n$ в том и только в том случае, если каждая точка из X' обладает произвольно малыми окрестностями в X, пересечение границ которых с X' имеет размерность $\leq n$ —1.

Доказательство. Допустим, что X' удовлетворяет условиям предложения А). Пусть p — произвольная точка из X', и U'— окрестность p в X'. Тогда существует окрестность U точки p в X такая, что $U' = U \cap X'$. Следовательно, существует открытое в X множество V такое, что

$$p \in V \subset U,$$

$$\dim (X' \cap Fr V) \leqslant n - 1.$$

Пусть $V' = V \cap X'$. Тогда V′ открыто в X, $p \in V' \subset U'$.

Снова, обозначая через B и B' границы множеств V в X и V′ в X' мы имеем: $B' \subset B \cap X'$. Следовательно, dim $B' \leq n$ — 1, так что dim $X' \leq n$.

Наоборот, допустим, что dim $X' \leq n$. Пусть p—произвольная точка X', и U — окрестность точки p в X. Тогда

$$U' = U \cap X'$$

есть окрестность точки p в X'. Поэтому существует окрестность V' точки p в X', для которой

$$p \in V' \subset U' \quad \text{и}$$

$$\dim B' \leqslant n - 1,$$

где B' — граница множества V' в X'. Никакое из непересекающихся множеств V' и $X' \setminus \bar{V}'$ не содержит предельной точки другого. Поэтому, так как X вполне нормально, существует открытое множество W такое, что

$$V' \subset W \text{ и } W \cap (X' \setminus V') = 0.$$

Заменив, если нужно, W пересечением $W \cap U$, можем предположить, что $W \subset U$.

Множество $\bar{W} \setminus W = Fr W$ не содержит никаких точек из $X' \setminus \bar{V}'$ и

из V'. Отсюда следует, что пересечение X' с FrV содержится в B^1 и, следовательно (теорема 1), имеет размерность $\leq n$ — 1, так что условие предложения А) выполнено.

Теперь мы используем А) для доказательства важного предложения относительно размерности суммы двух множеств. Уже было замечено,

что размерность суммы $A \cup B$ не определяется размерностями A и B. Однако:

B) Для любых двух подпространств A, B пространства X

$$\dim (A \cup B) \leqslant 1 + \dim A + \dim B.$$

Доказательство (двойной индукцией по размерностям подпространств A и B). Предложение очевидно для случая, когда

$$\dim A = \dim B = -1.$$

Пусть теперь $\dim A = m$, $\dim B = n$, и предположим, что утверждение справедливо в случае выполнения одного из двух следующих условий:

$$\dim A \leqslant m, \quad \dim B \leqslant n - 1; \quad (1)$$

$$\dim A \leqslant m - 1, \quad \dim B \leqslant n. \quad (2)$$

Пусть p — точка множества $A \cup B$. Предположим, что $p \in A$.

Пусть U — окрестность точки p в X. В силу A) существует открытое множество V такое, что

$$p \in V \subset U$$

и

$$\dim (W \cap A) \leqslant m - 1,$$

где W есть граница множества V. Но $W \cap B$ — подмножество множества B, и потому

$$\dim (W \cap B) \leqslant n.$$

По индуктивному предположению (1) и (2)

$$\dim [W \cap (A \cup B)] \leqslant m + n.$$

В силу A), этим доказано, что

$$\dim (A \cup B) \leqslant m + n + 1,$$

и индукция оказывается проведенной.

C) Сумма $(n+1)$-го подпространства размерности ≤ 0 имеет размерность 0.

Пример 6. Пусть $0 \leqslant m \leqslant n$. Обозначим через $\mathcal{M}_n^m$ множество точек пространства E_n, имеющих *не более* m рациональных координат, и через $\mathcal{L}_n^m$ —множество точек пространства E_n, имеющих *не менее* m рациональных координат.

Тогда

$$\dim \mathcal{M}_n^m \leqslant m,$$

(В действительности,

$$\dim \mathcal{M}_n^m = m,$$

$$\dim \mathcal{L}_n^m = n - m$$

Кроме того, мы позднее докажем, что каждое n-мерное пространство может быть топологически включено в $\mathcal{L}^n_{2n+1}$)

и $\dim \mathcal{L}^m_n \leqslant n - m$.

Так как очевидно, что

$$\mathcal{M}^m_n = \mathfrak{R}^0_n \cup \mathfrak{R}^1_n \cup \ldots \cup \mathfrak{R}^m_n$$

$$\mathcal{L}^m_n = \mathfrak{R}^m_n \cup \mathfrak{R}^{m+1}_n \cup \ldots \cup \mathfrak{R}^n_n,$$

то утверждение следует из С) и того факта, что каждое слагаемое здесь нульмерно.

Пример 7. Пусть $0 \leq m$. Обозначим через $\mathcal{M}^m_\omega$ множество точек гильбертова параллелепипеда I_ω, имеющих не более m рациональных координат. Тогда

$\dim \mathcal{M}^m_\omega \leqslant m$;

(В действительности, $\dim \mathcal{M}^m_\omega = m$)

так как

$$\mathcal{M}^m_\omega = \mathfrak{R}^0_\omega \cup \mathfrak{R}^1_\omega \cup \ldots \cup \mathfrak{R}^m_\omega,$$

то утверждение следует из С) и того факта, что каждое слагаемое нульмерно.

1.3.3. Теорема сложения для *n*-мерных пространств и теорема о разложении *n*-мерного пространства в сумму нульмерных пространств

Теперь мы докажем наиболее важные теоремы абстрактной части теории размерности.

Теорема 2. Теорема сложения для размерности *n*. *Пространство, являющееся суммой счетного числа замкнутых множеств размерности* $\leq n$, *имеет размерность* $\leq n$.

Доказательство. (По индукции). Обозначим теорему сложения для размерности n через Σ_n. Очевидно, Σ_0, эквивалентна утверждению, что любое пространство, являющееся суммой счетного числа F_σ, множеств размерности $\leq n$, имеет размерность $\leq n$.

Σ_{-1} тривиальна. Мы выведем теперь Σ_n из Σ_{n-1}, использовав Σ_0, которая уже была доказана независимою. Сначала мы докажем, что из Σ_{n-1} вытекает следующее предложение:

Δ_n. Любое пространство размерности $\leq n$ является суммой подпространства размерности $\leq n$ -1 и подпространства размерности ≤ 0.

Доказательство предложения Δ_n . Пусть X — пространство размерности $\leq n$. По предложению 1 В) существует базис открытых множеств пространства X, состоящий из множеств, границы которых имеют размерность $\leq n$ — 1. Так как X—пространство со счетным базисом, то существует счетный базис $\{U_i\}$, $i = 1,2,\ldots$, состоящий из множеств, границы $\{B_i\}$ которых имеют размерность $\leq n$ — 1. Из Σ_{n-1}, следует, что

$$B = \bigcup_{i=1}^{\infty} B_i$$

имеет размерность $\leq n$ — 1. Мы утверждаем, что

$$\dim (X \setminus B) \leqslant 0. \quad (1)$$

В самом деле, границы множеств U_i, очевидно, не пересекаются с $X\setminus B$, и, следовательно, условие предложения 2 А) (с $n = 0$ и $X\setminus B$ вместо X') удовлетворяется. Поэтому Δ_n следует из равенства $X = B \cup (X \setminus B)$. Воспользуемся теперь Σ_{n-1} и Δ_n для доказательства Σ_n. Предположим, что

$$X = C_1 \cup \ldots \cup C_i \cup \ldots,$$
$$\dim C_i \leqslant n,$$

где каждое C_i замкнуто. Мы хотим доказать, что $\dim X \leqslant n$. Пусть

$$K_1 = C_1,$$
$$K_i = C_i \setminus \bigcup_{j=1}^{i-1} C_j = C_i \cap (X \setminus \bigcup_{j=1}^{i-1} C_j), \quad i = 2,3, \ldots$$

Тогда

$$X = \bigcup_{i=1}^{\infty} K_i, \quad (2)$$

$$K_i \cap K_j = 0, \text{ если } i \neq j, \quad (3)$$

$$K_i \text{ есть } F_\sigma \text{ в } X, \quad (4)$$

$$\dim K_i \leqslant n. \quad (5)$$

Соотношения (2) и (3) очевидны. Чтобы доказать (4), заметим, что сумма

$$\bigcup_{j=1}^{i-1} C_j$$

является замкнутым множеством. Поэтому множество

$$X \setminus \bigcup_{j=1}^{i-1} C_j$$

открыто и, следовательно, как открытое множество в метрическом пространстве есть F_σ; множество K_i, как пересечение этого F_σ с замкнутым множеством C_i, также есть F_σ. (5) выполнено в силу того,

что K_i является подмножеством C_i. Неравенство (5) позволяет применить Δ_n к каждому K_i, и мы получаем:

$$K_i = M_i \cup N_i,$$

$$\dim M_i \leqslant n-1, \quad \dim N_i \leqslant 0.$$

Обозначим $\bigcup M_i$ через M и $\bigcup N_i$ через N. Из (2) следует, что

$$X = M \cup N$$

каждое M_i есть F_σ в M. Действительно,

$$M_i = M_i \cap K_i = (M_1 \cup \ldots \cup M_i \cup \ldots) \cap K_i = M \cap K_i,$$

так как $M_i \subset K_i$ и $K_i \cap K_j = 0$ для $i \neq j$ в силу (3). Поэтому M_i, как пересечение M с K_i, являющимся, в силу (4), F_σ -множеством, само есть F_σ в M. Следовательно, применяя Σ_{n-1}, заключаем, что $\dim M \leqslant n-1$. Подобным же образом мжно показать, что каждое N_1 есть F_σ в N, и, следовательно, $\dim X \leqslant n$ в силу $\sum_0$.

Таким образом, $X = M \cup N$, причем $\dim M \leqslant n-1$ и dim N≤0. Из 2 В) следует, что $\dim X \leqslant n$; что и требовалось доказать.

Следствие 1. *Сумма двух подпространств, каждое из которых имеет размерность ≤n и одно из которых замкнуто, имеет размерность ≤n.*

Следствие 2. *Размерность непустого пространства не возрастает при прибавлении к нему одной точки.*

Доказательство. Следствие 2, очевидно, вытекает из следствия 1.

Следствие 3. *Если пространство X′ размерности ≤n содержится в некотором пространстве X, то каждая точка пространства X имеет произвольно малые окрестности (в X), пересечения границ которых с X′ имеют размерность ≤n — 1* [сравнивая с 2 А), обратим внимание на то, что 2 А) налагает ограничение только на окрестности точек $p \in X'$].

Следствие 4. *Каждое пространство размерности ≤n является суммой подпространства размерности ≤n — 1 и подпространства размерности ≤0.*

Доказательство. Следствие 4 есть Δ_n, которое, как было показано при доказательстве теоремы 2, является следствием $\sum_{n-1}$.

Теорема 3. Теорема о разложении *n*-мерного пространства в сумму нульмерных пространств. *Пространство имеет размерность ≤n, n— конечное, в том* и *только в том случае, если оно является суммой* (*n*+1)-го *подпространства размерности* ≤ 0.

Доказательство. Теорема 3 следует из повторного применения следствия 4, доказанного выше, и 2 С).

Следствие. Если $\dim X = n$, а p и q — два целых числа ≥—1 таких, что $p+q+1=n$, то X является суммой двух подмножеств P и Q, размерность которых, соответственно, равна p и q.

Доказательство. Непосредственно следует из теоремы 3.

Пример 8. Мы уже привели разложение пространства E_n на $(n+1)$ пространств $\mathfrak{R}_n^0, \ldots, \mathfrak{R}_n^n$ размерности 0.

1.3.4. Размерность топологического произведения

Теорема 4. *Обозначим через $A \times B$ топологическое произведение двух пространств A и B, по крайней мере одно из которых непусто. Тогда*

$$\dim (A \times B) \leqslant \dim A + \dim B.$$

Доказательство. (По индукции). Предложение очевидно, если

$$\dim A = -1 \text{ или } \dim B = -1.$$

Пусть теперь $\dim A = m$, $\dim B = n$, и предположим, что предложение справедливо для случаев

$$\dim A \leqslant m, \quad \dim B \leqslant n - 1, \quad (1)$$

$$\dim A \leqslant m - 1, \quad \dim B \leqslant n. \quad (2)$$

Каждая точка $p = (a, b)$ пространства $A \times B$ обладает произвольно малыми окрестностями вида $U \times V$, где U—окрестность точки a в A и V—окрестность точки b в B, причем можно предположить, что

$$\dim Fr U \leqslant m - 1, \quad \dim Fr V \leqslant n - 1.$$

Мы имеем

$$Fr(U \times V) = (\bar{U} \times Fr V) \cup (\bar{V} \times Fr U).$$

Здесь каждое слагаемое замкнуто и по индуктивному предположению (1), (2) имеет размерность $\leqslant m + n - 1$. Следовательно, по теореме сложения

$$\dim Fr(U \times V) \leqslant m + n - 1;$$

этим доказано, что

$$\dim (A \times B) \leqslant m + n.$$

Следствие. *Если B нульмерно, то*

$$\dim (A \times B) = \dim A + \dim B. \quad (3)$$

Доказательство. Так как B непусто:

$$A \times B \supset A.$$

Следовательно, $\dim (A \times B) \geqslant \dim A = \dim A + \dim B$, что в соединении с теоремой 4 доказывает следствие.

Замечание. Можно было бы предположить, что логарифмический закон (3) будет справедлив и в общем случае. К сожалению, это не так, ибо ясно, что $\mathfrak{R}_\omega$ (множество точек гильбертова пространства, все координаты которых рациональны) гомеоморфно $\mathfrak{R}_\omega \times \mathfrak{R}_\omega$, в то время как пример 5 показывает, что $\dim \mathfrak{R}_\omega = 1$. (3) не имеет, вообще говоря, места даже в том случае, когда и A, и B являются компактами. Это показано понтрягинским примером двумерных компактов,

произведение которых трехмерно. Можно показать, что (3) имеет место для случая, когда B одномерно, а A является компактом. Проблема установления характеристики пространств B, для которых при произвольном пространстве A имеет место (3), остается открытой.

1.3.5. Отделение множеств в n-мерных пространствах

А) Если пространство X имеет размерность $\leq n$, то любые два непересекающиеся замкнутые множества могут быть отделены в X замкнутым множеством размерности $\leq n - 1$.

Предложение А) содержится (при $X=A$) в более общем предложении:

В) Пусть C_1 и C_2 — два непересекающиеся замкнутые множества пространства X (произвольной размерности), и A — подмножество пространства X размерности $\leq n$. Тогда существует замкнутое множество B, отделяющее C_1 от C_2 и такое, что

$$\dim A \cap B \leqslant n - 1.$$

Доказательство. Если $n = 0$, то или $\dim A = -1$, и В) очевидно, или $\dim A = 0$. Но для этого случая В) было уже доказано.

Допустим теперь, что $n>0$. Применяя к множеству A следствие 2, получаем: $A = D \cup E$, причем $\dim D \leqslant n-1$, $\dim E \leqslant 0$. В силу предложения В) для $n = 0$, существует множество В, отделяющее C_1 от C_2 и не пересекающееся с E. Следовательно,

$$A \cap B \subset D.$$

Но $\dim D \leqslant n-1$; поэтому $\dim A \cap B \leqslant n-1$; что и требовалось доказать.

Замечание. Утверждение, обратное к А), очевидно; так что А) устанавливает эквивалентность между локальным свойством отделения точки от замкнутого множества и свойством «в целом»— отделения двух замкнутых множеств. Это—та связь локальных свойств и свойств «в целом», которая, в значительной степени, является причиной плодотворности понятия размерности.

Следующее предложение будет играть важную роль дальше:

С) Пусть X — пространство размерности $\leq n - 1$, и пусть $C_i, C_i', i = 1, \ldots, n$ пар замкнутых подмножеств пространства X, для которых

$$C_i \cap C_i' = 0.$$

Тогда существуют n замкнутых множеств B_i таких, что B_i отделяет C_i от C'_i и

$$B_1 \cap B_2 \cap \ldots \cap B_n = 0.$$

Доказательство. В силу А) существует замкнутое множество B_1, отделяющее C_1 от C_2, такое, что

$$\dim B_1 \leqslant n-2.$$

Пользуясь предложением В), находим замкнутое множество B_2, отделяющее C_2 от C'_2, такое, что

$$\dim B_1 \cap B_2 \leqslant n-3.$$

Повторением этого применения предложения В), получим n множеств B_i таких, что B_i отделяет C_i от C'_i и

$$\dim (B_1 \cap B_2 \cap \ldots \cap B_k) \leqslant n-k-1, \quad k=1,\ldots,n.$$

При $k=n$ заключаем, что $B_1 \cap B_2 \cap \ldots \cap B_n = 0$.

Замечание. Свойство предложения С) является *характеристическим* для пространств размерности, меньшей, чем n .

1.3.6. Компакты

Рассмотрим следующие свойства пространства *X:* (1) Любые две различных точки могут быть отделены замкнутым множеством размерности $\leq n$ —1.

(2) Любая точка может быть отделена от несодержащего ее замкнутого множества замкнутым множеством размерности $\leq n$ —1, т.е. $\dim X \leqslant n$ (см. определение 1').

(3) Любые два непересекающиеся замкнутые множества могут быть отделены замкнутым множество» размерности $\leq n$ —1.

Ясно, что из (3) следует (2), а из (2) следует (1). Мы уже показали (5А), что из (2) для произвольных (метрических со счетный базисом) пространств следует (3), а ранее мы исследовали отношения между этими свойствами для случая $n = 0$. Теперь мы докажем, что

А) Для компактов свойства (1), (2) и (3) эквивалентны.

Доказательство. Осталось показать, что из (1) следует (2). Мы покажем это, доказав следующее предложение.

В) Пусть X — компакт, C—замкнутое подмножество пространства X, и p—точка из X. Тогда, если p может быть отделена замкнутым множеством размерности $\leq n$ — 1 от каждой точки множества С, то p может быть отделена замкнутым множеством размерности $\leq n$ — 1 от C.

Доказательство. Для каждой точки $q \in C$ существует открытое множество U(q), для которого

$$q \in U(q), \quad p \notin \overline{U}(q),$$

$$\dim Fr\,U(q) \leqslant n-1.$$

Множество С, как замкнутое подмножество компакта, само является компактом. Следовательно, существует конечное число $q_1, q_2, \ldots, q_k$ точек q таких, что

$C \subset U = U(q_1) \cup \ldots \cup U(q_k).$

Пусть B —граница множества U, $B = \overline{U} \setminus U$. Тогда

$B \subset FrU(q_1) \cup \ldots \cup FrU(q_k),$

и, следовательно, по теореме сложения для размерности $n = 1$

$\dim B \leqslant n - 1.$

Кроме того $p \notin \overline{U}$. Следовательно, p отделено от C замкнутым множеством B размерности $\leq n$ — 1, что и требовалось доказать.

C) Компакт имеет размерность $\leq n$ в том и только в том случае, если любые две различные точки в нем могут быть отделены замкнутым множеством размерности $\leq n$ — 1.

Замечание. Предложение C) остается справедливым для локально-компактных пространств,

1.4. РАЗМЕРНОСТЬ ЭВКЛИДОВЫХ ПРОСТРАНСТВ

В этом разделе введенное нами понятие размерности будет оправдано доказательством того, что размерность n-мерного эвклидова пространства в точности равна n.

Следует иметь в виду, что до сих пор не было даже показано существование пространств размерности > 1.

1.4.1. Некоторые топологические свойства пространства E_n

Мы уже знаем , что $\dim E_n \leqslant n$, и остается доказать, что $\dim E_n \geqslant n$. Для того чтобы доказать это, нам нужны некоторые общие теоремы о сферах и эвклидовых пространствах.

Интуитивно ясно, по крайней мере для $n = 1$ и $n = 2$, что n-мерная сфера не может быть стянута в лежащую на ней точку. (Конечно, является тривиальным тот факт, что двумерную сферу можно стянуть в точку в трехмерном эвклидовом пространстве. С другой стороны, резиновую пленку, натянутую на твердый шар, нельзя стянуть в точку, не разорвав ее, и это есть в точности то, что интуитивно понимается под нестягиваемостью двумерной сферы.)

Под этим мы подразумеваем, что точки n-мерной сферы S_n нельзя за конечный промежуток времени переместить вдоль путей, лежащих в S_n и имеющих общую конечную точку так, что положение движущейся точки непрерывно зависит от времени и исходного положения точки. Точнее:

A) Пусть S_n — n-мерная сфера, т. е. множество всех точек $\underline{x} \in E_{n+1}$, находящихся на расстоянии 1 от начала координат. Тогда не существует никакой непрерывной функции $f_t(x)$ двух переменных:

$x \in S_n$ и $0 \leqslant t \leqslant 1$, принимающей значения из S_n и удовлетворяющей граничным условиям

$$f_0(x) = x \quad \text{и} \quad f_t(x) = \text{const} \quad (1)$$

(f_1 является «постоянным» отображением: вообще, отображение f пространства X в пространстве Y называется *постоянным* отображением, если f отображает все X одну точку пространства Y.)

Для читателя, знакомого с понятием степени отображения (n-мерной сферы в себя) и с теоремой о том, что степень отображения остается инвариантной при непрерывных деформациях, мы можем доказать предложение А) очень быстро: тождественное отображение f_0 имеет степень 1, постоянное отображение f_1 имеет степень 0; следовательно, невозможно, чтобы f_0 и f_1 принадлежали к одному и тому же непрерывному семейству f_t.

Теперь мы приведем доказательство.

Доказательство. Мы воспользуемся простым геометрическим фактом, состоящим в том, что если $p_0, ..., p_n$ — точки единичной сферы S_n, попарные расстояния между которыми меньше 1, то $p_0, ..., p_n$ определяют единственный (возможно, вырождающийся) сферический n-мерный симплекс

(*Сферический n-мерный симплекс,* определенный $p_0,.., p_n$, есть проекция на S_n из начала координат наименьшего выпуклого множества в E_{n+1}, содержащего $p_0,.., p_n$. Этот n-мерный симплекс *вырождается,* если он лежит в n-мерной гиперплоскости, содержащей $p_0,.., p_n$ и начало координат. Мы не касаемся вопросов ориентации.)

(i) Пусть T—триангуляция сферы S_n, элементами которой являются сферические симплексы. Предположим, что каждой вершине a_i триангуляции T мы поставили в соответствие некоторую точку $\varphi(a_i) \in S_n$, причем так, что для любого n-мерного симплекса Δ, принадлежащего Т, точки, соответствующие его вершинам, образуют множество диаметра меньшего 1 и, следовательно, определяют новый n-мерный симплекс Δ^φ (который может быть вырождающимся). Таким образом, φ ставит в соответствие триангуляции T сферы S_n некоторую совокупность T^φ n-мерных (сферических) симплексов на S_n; симплексы из T^φ будут, вообще говоря, перекрываться. Для любой точки $p \in S_n$, не лежащей на границе никакого из симплексов Δ^φ, принадлежащих T^φ, обозначим через

$$n(p, \varphi, T)$$

число n-мерных симплексов Δ^φ, содержащих p. Мы утверждаем, что если q— другая точка S_n, не принадлежащая границе никакого Δ^φ из T^φ, то

$$n(p, \varphi, T) = n(q, \varphi, T) \pmod 2. \quad (2)$$

При доказательстве соотношения (2) можно предполагать, что *точки*

$$\varphi(a_i) \text{ находятся в общем положении,} \qquad (3)$$

т. е., если $m \leq n$, то никакие $m+1$ из точек $\varphi(a_i)$ не лежат в m-мерной плоскости, проходящей через начало координат.

(Отсюда следует, что никакой из симплексов Δ^φ не вырождается).

В самом деле, для фиксированных p и q мы можем точки $\varphi(a_i)$ заменить точками $\varphi'(a_i)$, удовлетворяющими условию (3) и такими, что $p \in \Delta^\varphi_i$ тогда и только тогда, когда $p \in \Delta^{\varphi'}_i$ и $q \in \Delta^\varphi_i$ тогда и только тогда, когда $q \in \Delta^{\varphi'}_i$, т. е.

$$n(p, \varphi, T) = n(p, \varphi', T) \text{ и } n(q, \varphi, T) = n(q, \varphi', T).$$

Проведем на S_n дугу, соединяющую p и q, пересекающую каждую (n— 1)-мерную грань любого симплекса из T^φ не более, чем в конечном числе точек, и не содержащую никаких точек, принадлежащих более, чем одной (n—1)-мерной грани. Это возможно в силу (3). Когда точка x пробегает вдоль этой дуги от p к q, целое число $n(x, \varphi, T)$ изменяется только в момент, когда x пересекает некоторую (n— 1)-мерную грань какого-либо симплекса Δ^φ из T^φ. Пусть такая грань определена точками $\varphi(a_1), \ldots, \varphi(a_n)$. Вершины $a_1, \ldots .a_n$ определяют некоторый (n— 1)-мерный симплекс исходной триангуляции T, и этот (n—1)-мерный симплекс является общей гранью в точности двух n-мерных симплексов Δ и Γ триангуляции T.

Рассмотрим симплексы Δ^φ и Γ^φ. Симплекс Δ^φ либо перекрывается симплексом Γ^φ либо нет. В первом случае $n(x, \varphi, T)$ изменяется на 2, когда x пересекает рассматриваемую (n— 1)-мерную грань, а во втором случае $n(x, \varphi, T)$ остается неизменным; в обоих случаях *четность* числа $n(x, \varphi, T)$ остается неизменной. Таким образом, соотношение (2) доказано.

Обозначим теперь через $n(\varphi, T)$ общее значение чисел $n(p, \varphi, T) \pmod 2$.

(ii) Пусть f—отображение сферы S_n в себя. (Начиная с этого момента, под словом «отображение» всегда будет пониматься непрерывное отображение, если специально не оговорено противное.)

Тогда существует триангуляция T сферы S_n, симплексы которой так малы, что образ каждого из них при отображении f имеет диаметр, меньший, чем 1. Рассматривая в (i) $f(a_i)$ вместо $\varphi(a_i)$, определим $n(p, f, T)$ и $n(f, T)$. Кроме того, в действительности $n(f, T)$ зависит только от f:

$$n(f, T) = n(f, T') \qquad (4)$$

для всяких двух триангуляции T и T', для которых эти символы имеют смысл. Так как любые две триангуляции T и T' имеют общее подразделение T'', достаточно, следовательно, доказать (4) для случая,

когда T' является подразделением триангуляции T. Достаточно ограничиться даже случаем, когда T' получается из Т подразделением только одного симплекса Δ_0 триангуляции Т, с сохранением всех остальных симплексов неизменными. Чтобы дать доказательство в этом частном случае, выберем точку p вне Δ_0^f. Равенство (4) следует тогда из. тривиального равенства

$$n(p, f, T) = n(p, f, T').$$

Обозначим через $n(f)$ общее, значение $n(f, T)$. Это целое число, определенное только по модулю 2, называется *степенью* отображения f по модулю 2. Ясно, что если f—тождественное отображение, то $n(f)=1$, а если f—постоянное отображение, то $n(f) = 0$.

(iii) Возвратимся теперь к доказательству предложения А). Допустим, что непрерывная функция $f_t(x)$ описанного в А) типа действительно существует. Из компактности сферы S_n и отрезка $0 \leqslant t \leqslant 1$ следует, что $f_t(x)$ является равномерно непрерывной функцией пары (x, t). Следовательно, существует триангуляция T_0 сферы S_n, обладающая следующим свойством: для каждого t образ любого симплекса триангуляции T_0 при отображении f_t имеет диаметр меньше 1. Очевидно, что если p—произвольная точка, для которой определено $n(p, f_t, T_0)$, то существует $\delta>0$ такое, что $n(p, f_{t'}, T_0)$ также определено и

$$n(p, f_t, T_0) = n(p, f_{t'}, T_0)$$

для каждого t', удовлетворяющего неравенству $|t-t'|<\delta$. Следовательно, $n(f_t) = n(f_{t'})$ при $|t-t'|<\delta$; этим доказано, что $n(f_t)$ постоянно. Прэтому невозможно, чтобы $f_t(x)$ удовлетворяла граничным условиям (1), потому что из них следуют равенства $n(f_0) = 1$ и $n(f_1) = 0$.

Предложение А) можно также сформулировать следующим образом:

В) Пусть K_n—замкнутый шар в E_n, т. е. множество точек пространства E_n, находящихся на расстоянии, меньшем или равном 1 от начала координат, и пусть S_{n-1}—$(n-1)$-мерная сфера, являющаяся границей K_n Тогда не существует никакого отображения F шара K_n *в* S_{n-1} оставляющего неподвижными все точки сферы S_{n-1}.

Доказательство. Действительно, допустим, что F— такое отображение. Тогда функция

$$f_t(x) = F[(1-t)x], \quad x \in S_{n-1}$$

удовлетворяет (1) (мы здесь рассматриваем x, как единичный вектор с началом в начале координат; $(1-t)x$ является тогда вектором, компоненты которого равны компонентам x, умноженным на $(1-t)$. В качестве следствия из В) получаем одну из наиболее известных теорем топологии:

C) **Теорема Брауэра о неподвижной точке.** *При отображении замкнутого шара K_n пространства E_n в себя всегда существует неподвижная точка, т. е. точка, совпадающая со своим образом.*
Доказательство. Допустим, что существует непрерывная функция *g(x),* определенная на K_n, для которой $g(x) \neq x$ для каждой точки *x*. Пусть S_{n-1} — $(n-1)$-мерная сфера, ограничивающая K_n. Для каждой точки $x \in K_n$ рассмотрим луч, соединяющий *x* с *g(x),* направленный от *g(x)* к *x*. Пусть *f(x)* — пересечение этого луча с S_{n-1}. Тогда соответствие

$$x \longrightarrow f(x)$$

является, очевидно, отображением K_n в S_n, оставляющим неподвижными все точки $x \in S_{n-1}$. Такое отображение, однако, существовать не может, так как его существование противоречит предложению B).
D) Пусть I_n — куб в E_n, например множество точек, каждая из координат $x_1, \ldots, x_n$ которых удовлетворяет неравенству

$$|x_i| \leqslant 1.$$

Пусть C_i — грань I_n , определенная уравнением

$$x_i = 1,$$

и C_i'—-противоположная грань. Пусть B_i — замкнутое множество, отделяющее C_i от C_i'. Тогда

$$B_1 \cap \ldots \cap B_n \neq 0.$$

Доказательство. B_i отделяет C_i от C_i' в I_n т. е.

$$I_n \setminus B_i = U_i \cup U_i',$$

$$C_i \subset U_i, \quad C_i' \subset U_i',$$

$$U_i \cap U_i' = 0,$$

причем U_i и U_i' открыты в $I_n \setminus B$, а, следовательно, и в I_n . Для каждой точки $x \in I_n$ пусть *V(x)* есть вектор, *i*-я компонента которого имеет значение

$$\pm \rho(x, B_i)$$

($\rho(x, B_i)$, как обычно, обозначает расстояние от *x* до B_i), где берется знак + , если $x \in U_i'$, и — , если $x \in U_i$. Поместим начальную точку вектора *V(x)* в точку *x* и поставим точке *x* в соответствие в качестве ее образа *f(x)* конечную точку этого вектора. Правило знаков обеспечивает нам, что в любом случае $f(x) \in I_n$. Соответствие *f*(*x*), как легко видеть, непрерывно: *f*(*x*) является непрерывным отображением куба I_n в себя. Из теоремы Брауэра о неподвижной точке (C), которую можно применить

к I_n , так как I_n гомеоморфно K_n, следует, что существует точка x^0 такая, что

$$f(x^0) = x^0.$$

Это означает, что

$$p(x^0, B_i) = 0$$

для каждого i; другими словами, $x^0 \in B_i$. Следовательно,

$$B_1 \cap \ldots \cap B_n \neq 0,$$

что и требовалось доказать.

Замечание. Предложения A), B), C) и D) тесно связаны между собой, и любое из них может быть выведено из другого с помощью весьмапростых рассуждений.

1.4.2. Размерность пространства E_n

Теперь мы докажем наиболее важные результаты этого раздела.

A) $\dim I_n \geqslant n.$

Доказательство. Предположим, что $\dim I_n \leqslant n-1$. Тогда, как известно, существует n таких замкнутых подмножеств B_i куба I_n , отделяющих различные пары противоположных граней, что $B_1 \cap B_2 \cap \ldots \cap B_n = 0$. Но это противоречит 1 D).

B) $\dim E_n \geqslant n.$

Доказательство. Предложение следует из того, что I_n является подмножеством пространства E_n.

Теорема 1. *n-мерное эвклидово пространство имеет размерность* n.

Доказательство. Надо сопоставить B) с примером 4 предыдущего раздела .

Следствие. *n-мерный эвклидов куб имеет размерность n.*

Доказательство. Надо сопоставить A) с тем фактом, что I_n является подмножеством пространства E_n.

Пример 1. $\dim \mathcal{M}_n^m = m$ и $\dim \mathcal{L}_n^m = n - m$. В самом деле, очевидно (см. пример 6 предыдущего раздела),

$$E_n = \mathcal{M}_n^m \cup \mathcal{L}_n^{m+1},$$

Причем $\dim \mathcal{M}_n^m \leqslant m$ и $\dim \mathcal{L}_n^{m+1} \leqslant n - m - 1$.

Следовательно, если хотя бы одно из равенств $\dim \mathcal{M}_n^m = m$ и $\dim \mathcal{L}_n^{m+1} = n - m - 1$ было бы неверным, то оказалось бы, что

$$\dim E_n < n.$$

Пример 2. $\dim \mathcal{M}_\omega^m = m$. В самом деле, рассмотрим множество M точек x гильбертова параллелепипеда,

$x = (x_1, x_2, \ldots, x_i, \ldots),$

первые m координат которых произвольны, а для каждого индекса $k > m$, x_k принимает иррациональное значение $\frac{1}{k\pi}$. M гомеоморфно I_m. Следовательно, $\dim M = m$.

Но M является подмножеством множества $\mathscr{R}^m_\omega$, так что $\dim \mathscr{R}^m_\omega \geqslant m$. Сопоставляя это с примером 7 предыдущего раздела, заключаем, что $\dim \mathscr{R}^m_\omega = m$.

1.4.3. Теорема Лебега о покрытиях

Теорема 2. Теорема Лебега о покрытиях. *Пусть n-мерный куб есть сумма конечного числа замкнутых множеств, ни одно из которых не пересекается ни с какими двумя противоположными гранями. Тогда по крайней мере n+* 1 *из этих замкнутых множеств имеют общую точку.*

Доказательство существенно опирается на предложение 1 D).

Сначала мы докажем два вспомогательных предложения:

А) Пусть A — замкнутое множество пространства X, C и C' — пара непересекающихся замкнутых подмножеств пространства X, и K— замкнутое подмножество множества A, отделяющее $A \cap C$ от $A \cap C'$ в A. Тогда существует замкнутое множество B, отделяющее C от C′ в X, такое, что $A \cap B \subset K$.

Доказательство. Утверждение, что K отделяет в A множество $A \cap C$ от множества $A \cap C'$, означает, что

$$A \setminus K = D \cup D',$$

$$A \cap C \subset D, \quad A \cap C' \subset D',$$

$$D \cap D' = \emptyset,$$

причем D и D' замкнуты в $A \setminus K$. Следовательно, ни одно из непересекающихся множеств $C \cup D$ и $C' \cup D'$ не содержит предельных точек другого. Так как X вполне нормально, существует открытое в X множество W такое, что $C \cup D \subset W$ и $\overline{W} \cap (C' \cup D') = \emptyset$. Отсюда следует, что $B = Fr W$ отделяет C от C′ в X и $B \cap (D \cup D') = \emptyset$; следовательно, $A \cap B \subset K$.

В) Пусть C_i и C_i', (i = 1,..., n) — пары противоположных граней куба I_n. Пусть

$$K_1 \supset K_2 \supset \ldots \supset K_n \quad (1)$$

— убывающая последовательность замкнутых множеств куба I_n таких, что K_1 отделяет C_1 от C'_1 в I_n; K_2 отделяет $K_1 \cap C_2$ от $K_1 \cap C_2'$ в K_1; ...; K_n отделяет $K_{n-1} \cap C_n$ от $K_{n-1} \cap C_n'$ в K_{n-1}.

Тогда K_n не пусто.

Доказательство. Пусть $B_1 = K_1$. Пользуясь предложением A) и принимая K_{i-1}, за A, а C_i и C'_i — за C и C′, расширим каждое K_i ($i = 2,..., n$) до множества, отделяющего C_i от C'_i в I_n. В результате получим систему $B_1.....B_n$ замкнутых множеств, обладающих следующими свойствами:

$$B_i \text{ отделяет } C_i \text{ от } C'_i \text{ в } I_n. \quad (2)$$

$$B_1 = K_1,\ B_i \cap K_{i-1} \subset K_i,\ i = 2, \ldots,\ n \quad (3)$$

Из (1) и (3) следует, что

$$B_1 \cap \ldots \cap B_n \subset K_n; \quad (4)$$

из (2) и предложения 1 D) вытекает, что

$$B_1 \cap \ldots \cap B_n \neq 0; \quad (5)$$

следовательно, K_n не пусто.

Доказательство теоремы 2. Обозначим через L_1 сумму множеств заданного разложения куба I_n, пересекающихся с C_1 через L_2 — сумму множеств, не вошедших L_1 и пересекающихся с C_2, через L_3 — сумму множеств, не принадлежащих $L_1 \cup L_2$ и пересекающихся с C_3,..., наконец, через L_{n+1} — сумму множеств, не пересекающихся ни с одним из C_i. Пусть

$$K_1 = L_1 \cap (L_2 \cup \ldots \cup L_{n+1}),$$

$$K_2 = L_1 \cap L_2 \cap (L_3 \cup \ldots \cup L_{n+1})$$

. . . . . .

$$K_{n-1} = L_1 \cap L_2 \cap \ldots \cap L_{n-1} \cap (L_n \cup L_{n+1}),$$

$$K_n = L_1 \cap L_2 \cap \ldots \cap L_n \cap L_{n+1}.$$

Из условий теоремы, как легко видеть, следует, что множества K_i удовлетворяют условиям предложения B). Следовательно, $K_n \neq 0$, т. е. $L_1 \cap L_2 \cap \ldots \cap L_{n+1} \neq 0$.

Справедливость самой теоремы вытекает из того обстоятельства, что каждое из исходных замкнутых множеств заданного разложения содержится лишь в одном из L_i.

1.4.4. Подмножества пространства E_n

Пусть N—подмножество пространства E_n, содержащее непустое открытое множество. Очевидно, $\dim N = n$. Действительно, существует точка $p \in N$ и положительное действительное число ρ такие, что сферическая окрестность $S(p,$ ρ) точки p радиуса ρ целиком содержится в N, а S $(p,$ ρ), очевидно, гомеоморфна E_n.

Докажем теперь обратное утверждение.

Теорема 3. *Для того чтобы подмножество N пространства E_n было n-мерным, необходимо и достаточно, чтобы N содержало непустое открытое в E_n множество.*

Доказательство. Нам нужно показать, что если dim $N=n$, то N содержит непустое открытое множество. Обозначим через M дополнение к N. Тогда высказанное выше утверждение эквивалентно утверждению, что если M — плотное в E_n множество, то дополнение множества M имеет размерность $\leq n-1$. Не уменьшая общности, можно предположить, чго M счетно. Действительно, M, как подмножество пространства со счетным базисом, содержит счетное плотное в нем подмножество A, а из того, что $\dim (E_n \setminus A) \leqslant n-1$ конечно, вытекает, что $\dim (E_n \setminus M) \leqslant n-1$

Утверждение верно , если в качестве M взять счетное множество $\mathfrak{R}_n$ точек пространства E_n, все координаты которых рациональны. Для того чтобы вывести отсюда справедливость нашего утверждения для произвольного плотного в E_n счетного множества M, достаточно доказать свойство однородности эвклидовых пространств, выражаемое следующим предложением.

А) *Каковы бы ни были два всюду плотных счетных подмножества A и B пространства E_n, существует гомеоморфизм пространства E_n на себя, переводящий A в B.*

Прежде чем доказать предложение А), сделаем несколько предварительных замечаний.

Пусть (x^1, x^2) и (y^1, y^2)—две упорядоченные пары точек пространства E_n. Если оси координат находятся в общем положении, т. е. никакая гиперплоскость, параллельная координатной гиперплоскости, не содержит более, чем одну точку x^1 или x^2, и более, чем одну точку y^1 или y^2, то мы скажем, что (x^1, x^2) и (y^1, y^2) подобно расположены, если векторы x^1-x^2 и y^1-y^2 содержатся в одном и том же «квадранте» E_n, т. е., если для каждого $i = 1,..., n$ действительные числа $x_i^1-x_i^2$ и $y_i^1-y_i^2$ имеют один и тот же знак, где x_i и yi— координаты точек x и y.

Пусть $X=x^1, x^2,\ldots, x^n,\ldots,$ и $Y=y^1, y^2,\ldots, y^n,\ldots$—две счетные последовательности точек или же два конечных множества точек одинаковой мощности. Всегда возможно таким образом выбрать систему координат, чтобы оси координат находились в общем положении (т. е. чтобы никакая гиперплоскость, параллельная координатной гиперплоскости, не содержала более, чем одну точку x, и более, чем одну точку y), так как для этого нужно лишь, чтобы оси не были направлены по не более, чем счетному числу определенных направлений. Предполагая это условие выполненным, мы скажем, что X и Y подобно расположены, если две упорядоченные пары точек

(x^{μ_1}, x^{μ_2}) и (y^{μ_1}, y^{μ_2}) подобно расположены для каждой пары индексов μ_1 и μ_2.

В) Пусть *A* и *B* — два счетных плотных в E_n множества, и пусть координатные оси находятся в общем положении по отношению к *A* и *B*. Тогда *A* и *B* можно переупорядочить в подобно расположенные последовательности.

Доказательство В). Пусть *A* и *B* произвольным образом упорядочены: $A = a^1, \ldots, a^i \ldots$ и $B = b^1, \ldots, b^i \ldots$. Мы определим подобно расположенные последовательности C= c^1,..., c^i,..., и D=d^1,…., d^i,..., совпадающие соответственно с *A* и *B,* но только упорядоченные по-иному. Эта конструкция строится индуктивно, причем каждый шаг индукции состоит из выбора: (1) элемента множества *C* из A, (2) элемента множества *D* из *B,* (3) другого элемента множества *D* из *B,* и, наконец, (4) другого элемента множества C из *A*. Начинаем, полагая $c^1 = a^1$ и $d^1 = b^1$. Затем полагаем $d^2 = b^\sigma$ и $c^2 = a^\sigma$, где σ — наименьший номер такой, что (c^1, a^σ) и (d^1, d^σ) подобно расположены, c^2 существует, в силу того, что A плотно в E^n. Допустим, что $c^1, \ldots, c^{2j}$ и $d^1, \ldots, d^{2j}$ уже выбраны так, что системы $c^1, \ldots, c^{2j}$ и $d^1, \ldots, d^{2j}$ подобно расположены. Обозначим через c^{2j+1} первый элемент *a*, не содержащийся среди элементов $c^1, \ldots, c^{2j}$, и через d^{2j+1}" —первый элемент *b* такой, что $c^1, \ldots, c^{2j}, c^{2j+1}$ и $d^1, \ldots, d^{2j}, d^{2j+1}$ подобно расположены.

Такой элемент d^{2j+1} существует, так как *B* плотно в E_n.

Далее, обозначим через d^{2j+2} первый элемент *b,* не содержащийся среди $d^1, \ldots, d^{2j}, d^{2j+1}$, и через c^{2j+2} — первый элемент *a* такой, что

$$c^1, \ldots, c^{2j}, c^{2j+1}, c^{2j+2} \text{ и } d^1, \ldots, d^{2j}, d^{2j+1}, d^{2j+2}$$

подобно расположены. Элемент c^{2j+2} существует, так как *A* плотно в E^n.

Это завершает индукцию. Ясно, что *C* и *D* подобно расположены и что в *C* входит каждый элемент *A*, а в *D*—каждый элемент B, т. е. *C* и *D* совпадают, соответственно с *A* и B, но только упорядочены по-иному. Предложение В) доказано.

Возвратимся теперь к доказательству предложения А).

Доказательство предложения А). Предложение В) позволяет нам рассматривать счетные плотные в пространстве E_n последовательности *A* и *B* как подобно расположенные. Предложение А) будет доказано тем, что взаимно однозначное соответствие

$$f: a^i \rightleftarrows b^i$$

между *A* и *B* будет продолжено в гомеоморфное отображение пространства E_n на себя.

Пусть $x = (x_1, \ldots, x_n)$ — произвольная точка пространства

E_n, не принадлежащая A. Мы определим $y = f(x)$, задавая для каждого k, $k = 1, \ldots, n$, k-ую координату y_k точки y. Множество A разобьем на два непересекающиеся класса: класс, состоящий из всех точек множества А, k-я координата которых не больше x_k, и класс, состоящий из точек, k-я координата которых больше x_k. С этим разбиением множества A, в силу взаимно однозначного характера соответствия $f(A) = B$, связано разбиение множества B на два непересекающиеся класса. Разбиение множества B, в свою очередь, порождает разбиение на два непересекающиеся класса множества K всех k-ых координат элементов множества B. Так как A и B подобно расположены, то это разбиение множества K обладает тем свойством, что каждый элемент одного класса меньше, чем каждый элемент другого класса. Так как K плотно на числовой, прямой, то полученное сечение определяет действительное число, которое мы и берем в качестве y_k.

Таким образом, А), а следовательно, и теорема 3 доказаны.

Следствие 1. *Для того чтобы подмножество N n-мерного многообразия (связного пространства, каждая точка которого имеет окрестность, гомеоморфную E_n) было n-мерно, необходимо и достаточно, чтобы N содержало непустое открытое подмножество многообразия.*

Следствие 2. *Пусть U—непустое и не всюду плотное открытое множество пространства E_n, и пусть B—граница множества U* (т. е. и U, и его дополнение содержат непустое открытое множество). *Тогда* $\dim B = n - 1$.

Доказательство. Прежде всего, из теоремы 3 следует, что $\dim B \leq n - 1$, так как B не содержит никакого непустого открытого множества. Теперь мы покажем, что невозможно, чтобы B имело размерность $\leq n - 2$.

Предположим сначала, что U ограничено. Пусть p — некоторая точка U. Если бы B имело размерность $\leq n - 2$, то простым процессом сжатия мы могли бы получить произвольно малые окрестности точки p, гомеоморфные U, границы которых гомеоморфны B и, следовательно, имеют размерность $\leq n - 2$. Таким образом, E_n имело бы размерность $\leq n - 1$ в точке p. В силу однородности пространства E_n отсюда следовало бы, что E_n имеет размерность $\leq n - 1$, в противоречии с теоремой 1. Таким образом, $\dim B = n - 1$.

Теперь предположим, что U не ограничено. В этом случае обозначим через p внутреннюю точку дополнения к U. Пусть ρ—столь малое действительное число, что сферическая окрестность $S(p,\rho)$ точки p радиуса ρ содержится в дополнении к множеству U. Произведем инверсию пространства E_n с центром p и радиусом инверсии ρ. Инверсия, конечно, является гомеоморфным отображением $E_n \setminus p$ на

себя и переводит U в непустое ограниченное открытое множество U', целиком содержащееся в $S(p, \rho)$. Если через B' мы обозначим границу множества U', то окажется, что $B'\backslash p$ гомеоморфно B. В силу только что доказанного, $\dim B' = n-1$. Из того факта, что размерность непустого множества не может быть повышена прибавлением одной точки, вытекает, что $\dim B = n-1$.

Замечание. Пусть U— непустое и не всюду плотное открытое множество некоторого n-мерного многообразия, и B — граница множества U. Тогда $\dim B = n-1$. Это вытекает из следствия 1 теоремы 4.

Пример 3. Пусть $X\backslash a$—вполне несвязное множество Кнастера и Куратовского, становящееся связным после прибавления точки a. Тогда

$$\dim (X \setminus a) = \dim X = 1,$$

ибо мы уже знаем, что $X\backslash a$ и X имеют размерность ≥ 1, но ни $X\backslash a$, ни X, являющиеся подмножествами E_2, не содержат открытого множества пространства E_2.

1.4.5. Разбивающие множества в E_n

Определение 1. Подмножество D пространства X *разбивает* X, если $X\backslash D$ несвязно.

А) Следующие три утверждения о пространстве X эквивалентны:

(1) X может быть разбито подмножеством D размерности $\leq m$;

(2) X содержит открытое множество U, непустое и не всюду плотное, граница которого имеет размерность $\leq m$;

(3) $X = C_1 \cup C_2$, где C_1, C_2 —замкнутые истинные подмножества пространства X и $\dim C_1 \cap C_2 \leqslant m$.

Доказательство. (1) → (2). Так как D разбивает пространство X, то

$$X \setminus D = U_1 \cup U_2, \quad U_1 \neq 0, \quad U_2 \neq 0,$$

и

$$(U_1 \cap \overline{U}_2) \cup (\overline{U}_1 \cap U_2) = 0. \qquad (4)$$

Мы можем предположить, что

$$X = \overline{X \setminus D} = \overline{U}_1 \cup \overline{U}_2,$$

так как в противном случае D содержало бы непустое открытое множество V такое, что $\overline{V} \subset D$. Граница множества V, как подмножество множества D, имела бы размерность $\leq m$ и, следовательно, удовлетворяла бы условиям (2).

Пусть $U = X \setminus \overline{U}_1$. Из (4) и (5) заключаем, что $U_2 \subset U \subset \overline{U}_2$.

Этим показано, что U непусто. Кроме того, U не всюду плотно, так как, в силу (4), $\overline{U}_2 \neq X$. Наконец, граница множества V содержится в $\overline{U}_2 \setminus U_2$, следовательно, и в D, а поэтому имеет размерность $\leq m$. Таким образом, X удовлетворяет условию (2).

(2)→(3). Если U—множество, удовлетворяющее условию (2), то $C_1 = \overline{U}$, $C_2 = X \setminus U$—множества, удовлетворяющие условию (3).

(3)→(1). Если C_1 и C_2—множества, удовлетворяющие условию (3), то $D = C_1 \cap C_2$—множество, удовлетворяющее условию (1).

Теорема 4. *E_n не может быть разбито подмножеством размерности $\leq n - 2$.*

Доказательство. Действительно, если бы теорема была не верна, то E_n, в силу А), содержало бы некоторое открытое множество, непустое и не всюду плотное, граница которого имеет размерность $\leq n - 2$, что противоречило бы следствию 2 теоремы 3.

Следствие 1. *S_n, а так же любое n-мерное многообразие, не может быть разбито подмножеством размерности $\leq n - 2$.*

Доказательство. Пусть R — n-мерное многообразие, и пусть D разбивает R, т. е. $R \setminus D = A_1 \cup A_2$, где множества A_1 и A_2 замкнуты в $R \setminus D$ и $A_1 \cap A_2 = 0$. Допустим, что dim $D \leq n - 2$. Тогда R\D плотно в R. Следовательно, каждая точка $p \in D$ является предельной либо для множества A_1, либо для множества A_2. Найдется точка $p \in D$, предельная для обоих множеств A_1 и A_2, ибо, в противном случае, замкнутые множества $\overline{A_1}$ и $\overline{A_2}$ дают в сумме R и $\overline{A_1} \cap \overline{A_2} = 0$, что невозможно, так как R связно. Рассмотрим окрестность U точки p, гомеоморфную пространству E_n. Легко видеть, что U разбивается множеством $D \cap U$, что противоречит теореме 4.

Следствие 2. *Куб I_n не может быть разбит подмножеством размерности $\leq n - 2$.*

Доказательство. Пусть D — подмножество куба I_n, имеющее размерность $\leq n - 2$, а I'_n — внутренность куба I_n. I'_n гомеоморфно пространству E_n. Поэтому $I'_n \setminus D$ связно. Но все точки множества $I_n \setminus D$ являются предельными точками множества $I'_n \setminus D$, следовательно, образуют также связное множество (множество остается связным после прибавления к нему произвольного множества его предельных точек).

1.4.6. Бесконечномерные пространства

Обозначим через J_ω подмножество гильбертова пространства, состоящее из точек, все координаты которых, за исключением конечного числа из них, равны нулю. J_ω содержит топологический образ n-мерного эвклидова пространства для каждого n, следовательно,

$\dim J_\omega = \infty$. J_ω является примером бесконечномерного пространства, которое представимо в виде суммы счетного числа конечномерных пространств; именно $E_1 \cup E_2 \cup \ldots$, где E_n — n-мерное эвклидово пространство, порождаемое первыми n координатными осями.
J_ω некомпактно, но также легко построить компактное подмножество K_ω гильбертова параллелепипеда, являющееся суммой счетного числа конечномерных пространств. Пусть M_n, $n = 1, 2, ...,$ — n-мерный «куб», состоящий из точек $x = (x_1, x_2, \ldots)$, удовлетворяющих условиям:

$$|x_i| \leqslant \frac{1}{n} \quad \text{для} \quad i = 1, \ldots n,$$

$$|x_i| = 0 \quad \text{для} \quad i > n,$$

и пусть $K_\omega = \bigcup_{n=1}^{\infty} M_n$.

Не всякое бесконечномерное пространство является суммой счетного числа конечномерных пространств. В самом деле:
А) Гильбертов параллелепипед не является суммой счетного числа конечномерных пространств.
Доказательство. На основании теоремы о разложении n-мерного пространства в сумму нульмерных пространств доказываемое предложение эквивалентно тому, что гильбертов параллелепипед не является суммой счетного числа нульмерных пространств.
Допустим противное:

$$I_\omega = A_1 \cup A_2 \cup \ldots, \quad \textbf{(1)}$$

где каждое A_i — нульмерно. Пусть C_i — грань куба I_ω, определенная уравнением $x_i = \frac{1}{i}$, а C'_i — противоположная ей грань. Существует замкнутое множество B_i, отделяющее C_i от C'_i и такое, что

$$A_i \cap B_i = 0. \quad \textbf{(2)}$$

Обозначим через B^n_i пересечение множества B_i с n-мерным эвклидовым пространством, порождаемым первыми n координатами гильбертова пространства. Тогда для данного n очевидно, что множества $B^n_1, B^n_2, \ldots, B^n_n$ суть замкнутые множества куба n-мерного эвклидова пространства, отделяющие различные пары его противоположных граней. Следовательно, в силу 1 D), эти n множеств имеют непустое пересечение, и тем более $B_1 \cap \ldots \cap B_n \neq 0$. Из компактности I_ω следует, что

$$\bigcap_{i=1}^{\infty} B_i \neq 0. \quad (3)$$

С другой стороны, из (1) и (2) следует, что

$$\bigcap_{i=1}^{\infty} B_i = 0,$$

чего быть не может.

Следствие. Никакое из пространств L_p и никакое из пространств l_p не является суммой счетного числа нульмерных пространств.
Доказательство. Утверждение следует из того факта, что все пространства L_p и l_p гомеоморфны , а l_2 есть гильбертово пространство.
Возможность разложения в сумму счетного числа конечномерных пространств тесно связана с так называемой «трансфинитной размерностью». Продолжая определение с помощью трансфинйтной индукции, мы скажем, что $\dim X \leqslant \alpha$ (α — порядковое число), если каждая точка пространства *X* обладает произвольно малыми окрестностями, границы которых имеют размерность, меньшую, чем α[3]($\dim B < \alpha$, конечно, означает, что $\dim B \leqslant \beta$ для некоторого $\beta < \alpha$.)

Мы скажем, что $\dim X = \alpha$, если верно, что $\dim X \leqslant \alpha$, но неверно, что $\dim X < \alpha$.
Не всякое пространство имеет трансфинитную размерность, но
B) Если *X* имеет трансфинитную размерность α, то α— число первого или второго порядкового класса.
Доказательство. Нам нужно показать, что α меньше, чем $\Omega = \omega_1$.
Допустим противное. Пусть β — наименьшее порядковое число $\Omega \leqslant \beta \leqslant \alpha$, для которого существует пространство *B* размерности β. По определению трансфинитной размерности, *B* обладает базисом, составленным из открытых множеств, границы которых имеют размерность, меньшую чем β и, следовательно, в силу минимального характера β, меньшую чем Ω. Так как пространство *B* обладает счетным базисом, то можно предположить, что и этот базис счетный. Но счетное множество порядковых чисел первого или второго класса имеет верхнюю границу, принадлежащую второму классу. Это означает, что существует порядковое число
$\gamma < \Omega$
такое, что граница каждого элемента нашего базиса имеет размерность, меньшую, чем γ. Но тогда
$\dim B \leqslant \gamma$
в противоречии с тем, что
$\dim B = \beta \geqslant \Omega$.
C) Если *X* имеет трансфинитную размерность, то *X* является суммой счетного числа конечномерных подпространств.
Доказательство. Пусть $\dim X = \alpha$. Предложение
C) очевидно, если α = 0. Предположим, что предложение справедливо для всех α, меньших β, и покажем, что оно справедливо и для α, равного β. Пусть *X*—пространство размерности β. Рассмотрим счетный

базис пространства *X,* составленный из множеств, границы которых имеют размерность меньше β.
По индуктивному предположению каждая из этих границ равна сумме счетного числа нульмерных пространств. Следовательно, сумма *B* этих границ является суммой счетного числа нульмерных множеств. Доказательство вытекает поэтому из замечания, что *X\B* нульмерно.
Следствие. *Гильбертово пространство и гильбертов параллелепипед не имеют никакой трансфинитной размерности.*
Утверждение, обратное предложению С), неверно:J_ω— множество тех точек из E_ω, у которых все координаты, кроме конечного числа из них, равны нулю, является суммой счетного числа конечномерных пространств, но можно доказать, что J_ω не имеет никакой трансфинитной размерности.
Однако частичное обращение предположения С) имеет место:
D) Если *X*—полное пространство (и тем более, если *X*— компакт), являющееся суммой счетного числа нульмерных пространств, то *X* имеет трансфинитную размерность.
Доказательство. Пусть

$$X = A_1 \cup A_2 \cup \ldots, \quad (5)$$

где каждое A_i нульмерно. Допустим, что *X* не имеет трансфинитной размерности. Тогда существует точка $p \in X$ и окрестность *U* точки *p,* обладающая следующий свойством: если *V* — произвольная окрестность точки *p,* содержащаяся в *U,* то граница *B* окрестности *V* не имеет трансфинитной размерности. Так как A_1 нульмерно, то существует окрестность V_1 точки *p,* содержащаяся в *U,* граница B_1 которой не имеет трансфинитной размерности (следовательно, не пуста) и не пересекается с A_2. Мы можем также потребовать, чтобы диаметр множества B_1 был меньше 1.
Теперь заменим *X* множеством B_1 и построим в B_1 замкнутое множество B_2 диаметра меньше 1/2 , не имеющее . трансфинитной размерности и не пересекающееся с A_2. По индукции получим счетную убывающую последовательность замкнутых непустых множеств B_i, диаметры которых стремятся к нулю и таких, что

$$A_i \cap B_i = 0. \quad (6)$$

Предположение, что *X*—полное пространство, влечет за собой существование точки, общей для всех B_i. С другой стороны, соотношения (5) и (6) делают это невозможным.

2. Покрытия, включения, отображения

2.1. ТЕОРЕМЫ О ПОКРЫТИЯХ И О ВКЛЮЧЕНИИ

Любое подмножество эвклидова пространства является метрическим конечномерным пространством со счетным базисом. Справедливо ли обратное?
Мы докажем, что обратное действительно справедливо— точнее: каждое (метрическое со счетным базисом) пространство размерности $\leq n$ может быть топологически включено в E_{2n+1}(теорема 3). Это значит, что класс конечномерных пространств топологически тождествен классу подмножеств эвклидовых пространств.
(Теорему 3 можно также выразить следующим образом: если dim $X<n$, то среди совокупности всех непрерывных действительных функций, определенных на X, существует множество, состоящее из $(2n+1)$ функций $f_1(x), \ldots, f_{2n+1}(x)$, (координатные функции), образующих базис в том смысле, что любая непрерывная действительная функция f(x), определенная на X, может быть выражена в форме $\varphi(f_1(x), \ldots, f_{2n+1}(x))$, где φ — непрерывная функция $(2n+1)$ действительных переменных.)

Пример А. Флореса (Пример 3) показывает, что невозможно понизить число $2n+$, т. е. произвольное n-мерное пространство включить в эвклидово пространство размерности $2n$.
Другим фундаментальным результатом этой главы является доказательство эквивалентности определения размерности, принятого в этой книге, с определением, опирающимся на свойства покрытий, по-существу, принадлежащим Лебегу (теорема 8). Этот результат приводит к теореме Александрова об аппроксимации n-мериых пространств n-мерными полиэдрами и связывает понятие размерности с комбинаторными свойствами пространств.

2.1.1. Теоремы о покрытиях

Определение 1. Под *покрытием* пространства X мы понимаем *конечную* систему $U_1, \ldots, U_r$ *открытых* множеств пространства X, сумма которых есть X. *Порядком* покрытия называется наибольшее целое число n такое, что существует $(n+1)$ элемент $U_{i_1}, \ldots, U_{i_{n+1}}$ покрытия с непустым пересечением. Если X

ограничено, то *диаметром* покрытия называется наибольший из диаметров множеств U_i.

Пусть M— подмножество пространства X. Мы говорим, что конечная система открытых множеств пространства покрывает M, если сумма множеств этой системы содержит M.

Пример 1. Если покрыть квадрат рядами «кирпичей» таким образом, чтобы стыки кирпичей каждого ряда приходились на середины кирпичей соседних рядов, и затем каждый из кирпичей немного увеличить, то эти кирпичи, рассматриваемые как открытые множества (без границы), образуют покрытие квадрата порядка 2. Ясно, что существуют покрытия этого вида произвольно малого диаметра.

Пример 2. Аналогично, I_n допускает покрытия произвольно малого диаметра, порядок которых равен n.

Определение 2. Покрытие β называется покрытием, *вписанным* в покрытие α, если каждый элемент покрытия β содержится в некотором элементе покрытия α.

А) Пусть X—пространство, и M— его подмножество размерности ≤ 0. Если U_1 и U_2 — два открытые множества пространства X, покрывающие M, то существуют два открытые множества V_1 и V_2, покрывающие M и обладающие тем свойством, что

$$V_1 \subset U_1, \quad V_2 \subset U_2, \qquad V_1 \cap V_2 = 0.$$

Доказательство. А) Очевидно, если $\dim M = -1$, т. е. если M пусто. Допустим теперь, что M нульмерно. Мы можем предположить, что

$$X = U_1 \cup U_2,$$

так как в противном случае мы могли бы заменить X на $U_1 \cup U_2$. Тогда

$$C_1 = X \setminus U_2 \text{ и } C_2 = X \setminus U_1 \quad \textbf{(1)}$$

суть непересекающиеся замкнутые множества. Так как dim $M=0$, можно получить непересекающееся с M замкнутое множество B, отделяющее C_1 от C_2. Отсюда вытекает существование двух открытых множеств V_1 и V_2, обладающих следующими свойствами:

$$V_1 \supset C_1, \quad V_2 \supset C_2, \quad (2)$$

$$V_1 \cap V_2 = 0, \quad (3)$$

$$X \setminus B = V_1 \cup V_2. \quad (4)$$

Из (I), (2) и (3) получаем

$$V_1 \subset U_1, \quad V_2 \subset U_2$$

а из (4) и из того, что $M \cap B = 0$,

$$V_1 \cup V_2 \supset M.$$

Предложение А) доказано.

В) Пусть X—пространство, M—его подмножество размерности ≤ 0, а $U_1, U_2, \ldots, U_r$— открытые множества, покрывающие M. Тогда существуют открытые множества $V_1, V_2, \ldots, V_r$, покрывающие M, такие, что

$$V_i \subset U_i, \quad i = 1, 2, \ldots, r; \quad V_i \cap V_j = 0, \quad i \neq j.$$

Доказательство. Доказательство проведем индукцией по числу множеств U_i. Предложение В) очевидно, если имеется лишь одно U_i. Предположим теперь, что число множеств U_i равно r и что предложение В) справедливо, когда число множеств U_i равно r —1. Положим

$$U'_{r-1} = U_{r-1} \cup U_r.$$

Рассмотрим покрытие множества M, состоящее из открытых множеств

$$U_1, \ldots, U_{r-2}, U'_{r-1},$$

число которых равно r — 1.

По индуктивному предположению существует покрытие

$$V_1, \ldots, V_{r-2}, V'_{r-1}$$

множества M такое, что

$$V_1 \subset U_1 \ldots, \quad V_{r-2} \subset U_{r-2} \quad V'_{r-1} \subset U'_{r-1},$$

причем $V_1, \ldots, V_{r-2}$ и V'_{r-1} попарно не пересекаются.

Множество $V'_{r-1} \cap M$ имеет размерность ≤ 0; множества $U_{r-1} \cap V'_{r-1}$ и $U_r \cap V'_{r-1}$ покрывают $V'_{r-1} \cap M$. Следовательно, в силу А), существуют открытые множества V_{r-1} и V_r, покрывающие $V'_{r-1} \cap M$ и удовлетворяющие условиям:

$$V_{r-1} \subset U_{r-1} \cap V'_{r-1}, \quad V_r \subset U_r \cap V'_{r-1},$$
$$V_{r-1} \cap V_r = 0.$$

Тогда множества

$$V_1, \ldots, V_{r-2}, V_{r-1}, V_r$$

удовлетворяют условию предложения В).

Из В) будет выведена следующая важная :) теорема:

Теорема 1. *Пусть X—пространство размерности $\leq n$, и α— покрытие пространства X. Тогда существует вписанное в α покрытие β порядка $\leq n$.*

Доказательство. Теорема очевидна, если $n = \infty$. Допустим теперь, что n конечно. По теореме о разложении n-мерного пространства в сумму нульмерных :

$$X = A_1 \cup \ldots \cup A_{n+1},$$

где каждое слагаемое A_i имеет размерность ≤ 0. Так как α покрывает каждое множество A_i, то, применив В), можно получить для каждого i конечную систему β^i открытых множеств:

$$\beta^i = (V_1^i, \ldots, V_{r(i)}^i) \quad i = 1, \ldots, n+1,$$

такую, что

$$\beta^i \text{ покрывает } A_i, \text{ и } V_j^i \cap V_k^i = 0, \text{ если } j \neq k. \quad (5)$$

Пусть β — покрытие пространства X, состоящее из всех множеств V_j^i, $i = 1, \ldots, n+1$; $j = 1, \ldots, r(i)$. Тогда мы утверждаем, что βимеет порядок $\leq n$. Действительно, в силу обычного рассуждения о $(n+2)$-х предметах, лежащих в $(n+1)$-м ящике, среди $(n+2)$-х элементов покрытия β всегда имеется два элемента, принадлежащих одному покрытию β^i. Условие (5) тогда показывает, что их пересечение пусто.

Следствие. *Пусть X—компакт размерности ≤ n. Тогда X обладает покрытиями произвольно малого диаметра^ порядок которых ≤ n .*

Доказательство. Рассмотрим для произвольного положительного в систему всех сферических окрестностей в X, имеющих радиус $0{,}5\ r$. Так как X компактно, существует покрытие пространства X, состоящее из элементов этой системы, т. е. покрытие, диаметр которого $\leq \varepsilon$. Применяя теорему к этому покрытию, получаем доказываемое следствие.

2.1.2. Пространство отображений

Определение 3. Пусть X—произвольное пространство, а Y—компакт. Обозначим множество всех отображений пространства X в Y через Y^X, и введем метрику в *пространстве отображений* Y^X , полагая

$$\rho(f, g) = \sup_{x \in X} \rho(f(x), g(x)).$$

Из компактности пространства Y следует, что ρ (f,g) конечно. Непосредственно ясно, чго $\rho(f, g) = 0$ тогда и только тогда, когда $f(x) = g(x)$, и что аксиома треугольника в Y^X следует из аксиомы треугольника в Y.

А) Y^X —полное) пространство.

Доказательство. Пусть f_n— фундаментальная последовательность элементов пространства Y^X . Тогда для каждой точки x пространства X последовательность точек f_n (x) является фундаментальной последовательностью в Y. Так как Y—компактное, следовательно и подавно, полное пространство, то $f_n(x)$ сходится к некоторой точке, которую мы обозначим через $f(x)$. Мы утверждаем, что $f(x)$ непрерывно, т. е. что $f \in Y^X$. Действительно,

$$\rho(f(x), f(x')) \leqslant \rho(f(x), f_n(x)) + \rho(f_n(x), f_n(x')) + \\ + \rho(f_n(x'), f(x')).$$

Мы знаем, что f_n равномерно сходятся к f; следовательно, если x' стремится к x, то $f(x')$ стремится к $f(x)$.

В) Предложение А) позволяет применять к пространству отображений Y^X теорему Бэра: пересечение счетного числа плотных в Y^X G_δ-множеств плотно в Y^X ; в частности, такое пересечение не пусто. (Под G_δ в некотором пространстве мы понимаем пересечение счетного числа открытых множеств).

2.1.3. Включение *n*-мерного компакта в I^{2n+1}

Сначала рассмотрим случаи компактов. Это облегчит понимание главной идеи общего доказательства, излагаемого далее, ибо в случае компактов доказательство проще и не затемняется техническими деталями.

Теорема 2. *Пусть X—компакт,* и dim X≤*n, где n — конечное. Тогда X гомеоморфно подмножеству куба* I^{2n+1}

Более того, множество гомеоморфных отображений пространства X в I^{2n+1} *является плотным* G_δ *в пространстве отображений* I^X_{2n+1}.

(Напоминаем, что гомеоморфизм «в» означает гомеоморфизм «на часть». Утверждение, что гомеоыорфвые отображения образуют всюду плотное множество, интуитивно означает, что каждое отображение пространства X в I^{2n+1} произвольно малым изменением может быть сделано гомеоморфизмом.)

Сначала мы рассмотрим отображения, которые ведут се5я «приблизительно» так же, как гомеоморфизмы.

Определение 4. Пусть X—компакт, ε — положительное число, а g — отображение компакта X в пространство Y. Мы скажем, что g есть ε-отображение, если полный прообраз каждой точки множества $g(X)$ имеет диаметр меньше ε.

А) Если отображение g компакта X в компакт Y есть $\frac{1}{i}$-отображение для каждого натурального числа i, то g есть гомеоморфизм компакта X в Y, и обратно.

Доказательство. Бели g есть $\frac{1}{i}$-отображение при каждом i, то g должно быть взаимно однозначным, а взаимно однозначное отображение компакта является гомеоморфизмом. Обратно, каждый гомеоморфизм является $\frac{1}{i}$-отображением при каждом i.

В) Пусть X—компакт. Тогда для каждого ε > 0 множество G_ε, всех ε-отображений открыто в Y^X.

Доказательство. Пусть g есть ε-отображение. Положим,

$$\eta = \inf \rho(g(x), g(x')),$$

где нижняя грань берется по парам (x, x'), для которых $\rho(x, x') \geqslant a$. В силу компактности X, η равно $\rho(g(x), g(x'))$ для некоторой пары точек x, x' такой, что $\rho(x, x') \geqslant a$; следовательно, η >0, так как в противном случае g не было бы ε-отображением. Пусть теперь f—произвольное отображение, для которого

$$\rho(f, g) < \frac{1}{2}\eta.$$

Пусть точки x и x' таковы, что $f(x) = f(x')$. Тогда расстояние $\rho(g(x), g(x')) < \eta$, и поэтому $\rho(x, x') < a$. Следовательно, f также является ε-отображением.

Доказательство теоремы 2. Рассмотрим пространство отображений I^X_{2n+1}. Пусть $G_{\frac{1}{i}}$ — множество всех $\frac{1}{i}$ -отображений компакта X в куб I_{2n+1}. Положим

$$H = \bigcap_{i=1}^{\infty} G_{\frac{1}{i}}.$$

В силу 3 А), H состоит из гомеоморфных отображений компакта X в I_{2n+1}. В силу 3 В), каждое $G_{\frac{1}{i}}$ открыто и, следовательно, есть G_δ.

Поэтому (см. 2 В)) для доведения до конца доказательства теоремы 2 нам необходимо лишь следующее предложение:

С) Пусть X— компакт, и $\dim X \leqslant n$, где n конечно. Для каждого положительного числа ε обозначим через G_ε , множество ε-отображений компакта X в куб I_{2n+1}. Тогда G_ε, плотно в пространстве отображений I^X_{2n+1}.

Доказательство. Пусть f— произвольный элемент пространства I^X_{2n+1}, и η— положительное число. Мы построим g такое, что

$$\rho(f, g) < \eta, \quad (1)$$

$$g \in G_a. \quad (2)$$

В силу равномерной непрерывности (непрерывная функция на компакте является равномерно непрерывной) отображения f, существует положительное число $\delta < \varepsilon$ такое, что

$$\rho(f(x), f(x')) < \frac{1}{2}\eta,$$

как только

$$\rho(x, x') < \delta.$$

На основании следствия теоремы 1 существует покрытие

$\beta : U_1, \ldots, U_r$

компакта X, обладающее свойствами:

порядок покрытия $\beta \leqslant n$ (3)

и

$\delta(U_i) < \delta, \quad i = 1, \ldots, r.$ (4)

($\delta(M)$ обозначает днаметр множества *M).*

Следовательно,

$\delta(f(U_i)) < \frac{1}{2}\eta, \quad i = 1, \ldots, r.$ (5)

Выберем в I_{2n+1} точки $p_1 \ldots, p_r$ такие, что

$\rho(p_i, f(U_i)) < \frac{1}{2}\eta_i \quad i = 1, \ldots, r,$ (6)

p_i находятся в общем положении в E_{2n+1}, (7)

т. е. никакие $(m + 2)$ из точек p_i $(m = 0, 1, \ldots, 2n)$ не лежат ни в каком *m*-мерном линейном подпространстве пространства E_{2n+1}.

Для каждой точки *x* пространства *X* положим:

$w_i(x) = \rho(x, X_i \setminus U_i), \quad i = 1, \ldots, r.$

(Если $U_i = X$, то принимаем, что $w_i(x) = 1$.)

Очевидно, что

$w_i(x) > 0$, если $x \in U_i$, и $w_i(x) = 0$, если $x \notin U_i$.

Для каждой точки *x*, по крайней мере, одно $w_i(x)$ положительно, так как U_i покрывают *X.* Каждой точке p_i, выбранной выше, поставим в соответствие вес $w_i(x)$, и обозначим через $g(x)$ центр тяжести системы точек p_i, взятых с этими весами. Ясно, что *g* является непрерывным отображением компакта *X* в I_{2n+1}. Покажем теперь, что *g* удовлетворяет условиям (1) и (2).

Доказательство (1): Пусть *x* — произвольная точка компакта *X.* Предположим, чго U_i занумерованы таким образом, что U_1 ,..., U_s есть множество всех U_i, содержащих *x.* Тогда $w_i(x) > 0$ для $i \leq s$, и $w_i(x) = 0$ для $i > s$; поэтому при определении $g(x)$ нам нужно рассматривать лишь $p_1, \ldots, p_s$

Из того, что $x \in U_i$ при $i \leq s$, и из (5) и (6) получаем, что

$\rho(p_i, f(x)) < \eta_i, \qquad i \leqslant s.$

Тем более центр тяжести $g(x)$ вершин p_i удовлетворяет условию

$\rho(g(x), f(x)) < \eta_i$ **(1)**

Доказательство (2). Пусть

$U_{i_1}, \ldots, U_{i_s}$

суть все элементы покрытия β, содержащие данную точку *x* пространства *X*. Из (3) следует, что s≤*n*+1. Рассмотрим линейное (*s* — 1)-мерное пространство *L(x)* в E_{2n+1}, содержащее точки

$$p_{i_1}, \ldots, p_{i_s}.$$

Очевидно, что *g(x)* лежит в *L(x).* Пусть *x′*—некоторая другая точка компакта *X.* Мы утверждаем: *если L(x) и L(x′) пересекаются, то они содержат общую точку* p_i; *следовательно, x и x′ содержатся в некотором общем элементе* U_i *покрытия* β. Действительно, пусть *L(x′)* проходит через точки

$$p_{j_1}, \ldots, p_{j_t}.$$

Снова, в силу (3), $t \leqslant n+1$, и $L(x')$ есть (*t* — 1)-мерное пространство. Если $L(x)$ и $L(x')$ пересекаются, то линейное пространство, наименьшей размерности, содержащее все p_i и p_j, имеет размерность $\leqslant s+t-2 \leqslant 2n$. Отсюда и из (7) следует что, по крайней мере, одна вершина p_i встречается также среди вершин p_j. Итак, справедливость утверждения установлена.

Теперь допустим, что $g(x)=g(x')$. Тогда $L(x)$ и $L(x')$ пересекаются. В силу доказанного выше утверждения, *x* и *x′* принадлежат общему элементу покрытия β. Поэтому из (4) вытекает, что $\rho(x, x')<\delta<\varepsilon$ и, следовательно, *g* является ε- отображением. Этим завершается доказательство предложения С) и, следовательно, теоремы 2.

2.1.4. Включение *n*-мерного пространства в I_{2n+1}

Пользуясь доказательетвом теоремы 2 как моделью, мы окажем теперь следующую теорему:

Теорема 3. *Пусть X — произвольное пространство и* dim X≤*n, где n конечно. Тогда X гомеоморфно подмножеству куба* I_{2n+1}.

Более того, множество гомеоморфных отображений пространства X в I_{2n+1} *содержит плотное* G_δ *в пространстве отображений* I_{2n+1}^X.

ε-отображения, которыми мы пользовались в параграфе 3, не годятся для нашей теперешней цели, так как отображение произвольного пространства, являющееся ε-отображениеи при каждом ε>0, может не быть топологическим, как например, отображение

$$\theta = x$$

полуинтервала $0 \leqslant x < 2\pi$ прямой на окружность $0 \leqslant \theta \leqslant 2\pi$.

Вместе с тем доказано, что в общем вопросе перехода от компактов к произвольным пространствам большое значение имеет метод, состоящий в том, чтобы заменять выражение: «для каждого положительного ε существует множество диаметра меньше ε»,

выражением: «для каждого покрытия α существует покрытие, вписанное в α». Пользуясь этим методом, приходим к следующему видоизменению определения 4.

Определение 5. Пусть α—покрытие пространства X, и *g*— отображение пространства *X* в пространство *Y*. Мы скажем, что *g* есть α-отображение, если каждая точка пространства Y обладает окрестностью в Y, полный прообраз которой целиком содержится в некотором элементе покрытия α.

Определение 6. Пусть α— покрытие пространства X. Обозначим через *S(x)* открытое множество, являющееся суммой элементов покрытия α, содержащих данную точку *x*. Счетная система $\alpha^1, \alpha^2, \ldots$ покрытий называется *базисной последовательностью покрытий,* если для произвольной точки *x* и ее произвольной окрестности *U*, по крайней мере, одно из открытых множеств

$$S_{\alpha^1}(x),\ S_{\alpha^2}(x), \ldots,$$

содержится в *U*.

А) Для каждого пространства *X* существует базисная последовательность покрытий.

Доказательство. Путь $U_1, U_2, \ldots$ —счетный базис пространства *X*. Рассмотрим пары U_n, U_m непустых открытых множеств этого базиса, для которых

$$\overline{U}_n \subset U_m.$$

Обозначим через $\alpha^{n,m}$ покрытие пространства *X*, состоящее из двух элементов: $X \setminus \overline{U}_n$ и U_m. Совокупность покрытий $\alpha^{n,m}$, конечно, счетна. Кроме того, из $x \in U_n$ следует $S_{\alpha^{n,m}}(x) = U_m$. Следовательно, система множеств $\{S_{\alpha^{n,m}}(x)\}$ для данного *x* включает совокупность всех U_m, содержащих *x*. Таким образом, $\{\alpha^{n,m}\}$ является базисным семейством покрытий.

В) Пусть $\alpha^1, \alpha^2, \ldots$ —базисная последовательность покрытий пространства *X*. Тогда, если есть α^i-отображение пространства *X* в пространство *Y* для каждого *i*, то *g* есть гомеоморфизм.

Обратное неверно. В самом деле, пусть *X*—полупрямая $x \geq 0$, *Y*—отрезок $0 \leqslant y \leqslant 1$ и *h*—«сжатие» *X* в *Y*, задаваемое формулой $y = \frac{x}{1+x}$. Пусть α°— покрытие пространства *X*, состоящее из двух открытых множеств: дополнения до множества всех четных и дополнения до множества всех нечетных целых чисел. Хотя *h* — гомеоморфизм, *h* не является α°-отображением, так как никакая окрестность точки $y = 1$ не обладает полным прообразом, содержащимся целиком в каком-либо элементе покрытия α°.

Доказательство. Мы покажем, что если *x*—произвольная точка пространства *X* и *U* — окрестность точки *x,* то существует окрестность *V* точки *g(x)* в *Y,* полный прообраз которой содержится в *U.* Отсюда вытекает взаимная однозначность отображения *g* и непрерывность отображения g^{-1}.

По определению базисной последовательности окрестностей существует покрытие α^i, для которого

$$S_{\alpha^i}(x) \subset U. \quad (1)$$

Так как *g* есть α^i -отображение, то существует окрестность *V* точки *x* и элемент U^i_0 покрытия α^i, для которых

$$g^{-1}(V) \subset U^i_0. \quad (2)$$

Но

$$x \in g^{-1}(V) \subset U^i_0.$$

Отсюда

$$U^i_0 \subset S_{\alpha^i}(x). \quad (3)$$

(2), (3) и (1) доказывают предложение.

C) Пусть α — покрытие компакта *X.* Тогда существует положительное число η, обладающее тем свойством, что каждое подмножество компакта *X,* имеющее диаметр меньше η, целиком содержится в некотором элементе покрытия α.

Доказательство. В противном случае существовала бы последовательность множеств $X_1, X_2, \ldots$, не содержащихся ни в одном элементе покрытия α и таких, что диаметры $\delta(X_i) \to 0$. Пусть $x_i \in X_i$. Так как *X*—компакт, $\{x_i\}$ имеет предельную точку *x,* которая содержится в некотором элементе U_0 покрытия α. Но U_0 открыто, так что $\rho(x, X \setminus U_0) = d > 0$. Тогда всякое X_i находящееся от *x* на расстоянии, меньшем, чем $\frac{d}{2}$, и имеющее диаметр, меньший, чем $\frac{d}{2}$, целиком содержится в U_0, что противоречит предположению.

D) Пусть *X*—произвольное пространство, и Y— компакт. Для каждого покрытия α множество G^α, всех α-отображений в *Y* открыто в Y^x.

Доказательство. Пусть *g* есть α-отображение. Это значит, что каждая точка компакта *Y* имеет окрестность, полный прообраз которой целиком содержится в некотором элементе покрытия α. Так как *Y*— компакт, то существует конечная подсистема системы таких окрестностей, образующая покрытие σ компакта *Y.* В силу C), существует положительное число η, обладающее тем свойством, что произвольное множество компакта *Y,* имеющее диаметр меньше, чем η, содержится в некотором элементе покрытия σ, и, следовательно, его полный прообраз при отображении *g* целиком содержится в некотором

элементе покрытия α. Пусть теперь *f*—1 отображение, удовлетворяющее условию

$$\rho(f, g) < \frac{1}{3}\eta.$$

Возьмем сферические окрестности диаметра $\frac{1}{3}\eta$ вокруг каждой точки пространства *Y*. Пусть *A* — полный прообраз при *f* одной из этих окрестностей. Легко видеть, что *g(A)* имеет диаметр < η, так что, в силу определения η, *A* содержится в некотором элементе покрытия α. Таким образом, *f* является α-отображением.

Доказательство теоремы 3. Рассмотрим пространство отображений $I_{2n}^{X}{}_{+1}$. Пусть $\alpha^1, \alpha^2, \ldots$ —базисная последовательность покрытий пространства *X*, G_{α^i} — множество всех α^i-отображений (см. определение 5) пространства *X* в I_{2n+1}, и

$$H = \bigcap_{i=1}^{\infty} G_{\alpha^i}$$

Каждый элемент множества *H* является, в силу 4 B), гомеоморфизмом пространства *X* в I_{2n+i}.На основании 4 D) каждое G_{α^i} открыто и, следовательно, является G_δ-множеством. Поэтому (см. 2 B) для доказательства теоремы 3 нам необходимо только следующее предложение.

E) пусть $\dim X \leqslant n$, где *n* — конечно. Для каждого покрытия *α* пространства *X* обозначим через G_α множество α-отображений пространства *X* в I_{2n+1}.Тогда G_α плотно в пространстве отображений $I_{2n}^{X}{}_{+1}$.

Доказательство. Пусть *f*—произвольный элемент пространства I_{2n+1}^{X}, и η — положительное число. Мы построим отображение *g* такое, что

$$\rho(f, g) < \eta, \quad (4)$$

$$g \in G_{\alpha^r} \quad (5)$$

Как компакт, I_{2n+1} имеет покрытие диаметра, меньшего, чем $\frac{1}{2}\eta$. Пусть τ — покрытие пространства *X*, состоящее из полных прообразов открытых множеств этого покрытия при отображении *f*. В силу теоремы 1, существует покрытие β порядка $\leq n$ вписанное и в α, и в *τ:*

$$\text{порядок покрытия } \beta \leqslant n, \quad (6)$$

$$\beta \text{ вписано в } \alpha, \quad (7)$$

$$\delta(f(U_i)) < \frac{1}{2}\eta, \quad i = 1, \ldots, r. \quad (8)$$

Теперь мы с помощью *r* точек $p_1, \ldots, p_r$, находящихся в I_{2n+1}

в общем положении, строим g точно так же, как это было сделано выше в доказательстве предложения 3 С), как «барицентрическое» отображение. Доказательство неравенства (4) точно такое же, как раньше, а при доказательстве включения (5) нужно только заменить последний абзац следующим.

Так как имеется только конечное число линейных подпространств $L(x)$ то существует число $\eta > 0$ такое, что любые два из этих линейных подпространств $L(x)$ и $L(x')$ или пересекаются, или, в противном случае, находятся на расстоянии $\geq\eta$ друг, от друга. Если $\rho(g(x), g(x')) < \eta$, то расстояние $\rho(L(x), L(x'))$, конечно, $< \eta$, следовательно, $L(x)$ и $L(x')$ пересекаются. Отсюда, как показано выше, вытекает, что обе точки x и x' содержатся в некотором элементе покрытия β . Следовательно, g является α-отображением.

Пример 3. Пусть s_{2n+2} — $(2n+2)$-мерная клетка и P_n — совокупность всех граней клетки s_{2n+2} размерности $\leq n$.

Тогда P_n является n-мерным пространством, которое нельзя включить в E_{2n}. Доказательство показывает, что число $2n+1$ в теореме 3 не может быть понижено.

2.1.5. Включение произвольных пространств в гильбертов параллелепипед

Теорема 4. *Пусть X— произвольное пространство. Тогда X может быть топологически включено в* I_ω. *Более того, множество гомеоморфных отображений пространства X в* I_ω *содержат плотное* G_δ *в* I_ω^X.

Доказательство. Доказательство почти в точности совпадает с доказательством теоремы 3 с тем упрощением, что не нужно интересоваться порядком покрытия β. Точки p_i выбираются так, чтобы любое конечное подмножество этих точек было линейно независимо.

2.1.6. Универсальное *n*-мерное пространство

Определение 7. Некоторое n-мерное пространство является *универсальным n-мерным пространством,* если каждое пространство размерности $\leq n$ может быть в него топологически включено.

Теорема 5. *Множество*

$$X_n = \sqrt{R_{2n+1}^n} \cap I_{2n+1}$$

точек куба I_{2n+1}, имеющих не более n рациональных координат, является универсальным n-мерным пространством.

Доказательство. Нам нужно показать, что любое пространство размерности $\leq n$ может быть топологически включено в $\mathcal{X}_n$. Доказательство является видоизменением доказательства теоремы 3 и использует те же обозначения. Легко видеть, что если M — фиксированное n-мерное линейное подпространство пространства E_{2n+1}, то небольшим сдвигом вершин $p_1,\ldots, p_r$ можно добиться того, чтобы ни одно из $L(x)$ (см. 3 С) и 4 Е)) не пересекало M. Поэтому, исходя из произвольного отображения f пространства X в I_{2n+1} и положительного числа η мы можем построить отображение g, удовлетворяющее неравенству $\rho(f, g) < \eta$ и добавочному условию:

$$\overline{g(X)} \subset I_{2n+1} \setminus M. \quad (1)$$

Отсюда следует, что множество отображений g пространства X в I_{2n+1}, для которых имеет место (1), плотно в I_{2n+1}^X. Это множество также и открыто, потому что из (1) вытекает, что $g(X)$ находится на положительном расстоянии η от M и, следовательно, для каждого отображения f, для которого $\rho(f,g)<\eta$, оказывается, что

$$\overline{f(X)} \subset I_{2n+1} \setminus M.$$

Дополнение множества $\mathcal{X}_n$ состоит из точек куба I_{2n+1}, имеющих, по крайней мере, $n+1$ рациональную координату, т. е. дополнение множества $\mathcal{X}_n$ является суммой гиперплоскостей в I_{2n+1} вида

$$x_{i_1} = r_1, \ldots, x_{i_{n+1}} = r_{n+1},$$

где все r_i рациональны. Каждая из этих гиперплоскостей имеет размерность $2n+1 - (n+1) = n$, причем имеется лишь счетное число таких гиперплоскостей. Назовем их $M_1, M_2, \ldots$. Пусть G_i — открытое плотное в I_{2n+1}^X множество, состоящее из всех отображений g, для которых

$$\overline{g(X)} \subset I_{2n+1} \setminus M_i.$$

В силу теоремы 3, существует плотное в I_{2n+1}^X G_δ-множество H, все элементы которого являются гомеоморфизмами пространства X в I_{2n+1}. Пусть

$$H' = H \cap \bigcap_{i=1}^{\infty} G_i.$$

H', как счетное пересечение всюду плотных G_δ -множеств, само является всюду плотным G_δ -множеством (см. 2 В); в частности, H' не пусто. Пусть $h \in H'$. Ясно, что h есть гомеоморфизм X в I_{2n+1} и

$$h(X) \subset I_{2n+1} \setminus (M_1 \cup M_2 \cup \ldots) = \mathcal{X}_n \quad (2)$$

Следовательно, для каждого пространства X размерности $\leq n$ существует гомеоморфное отображение h пространства X в такое, что I_{2n+1}

$$\overline{h(X)} \subset \mathcal{X}_n \quad (3)$$

Но $\mathcal{X}_n$ имеет размерность n . Таким образом, теорема доказана.

Замечание. Неизвестно, содержит ли $\mathcal{M}_n^k$, $n < 2k+1$, топологический образ каждого k-мерного подмножества пространства E_n, или даже, существует ли при $n < 2k+1$ какое-либо k-мерное подмножество пространства E_n , содержащее топологический образ каждого k-мерного подмножества пространства E_n.

Теорема 6. *Любое пространство может быть топологически включено в компакт той же размерности.*

Доказательство. Пусть X—некоторое пространство. Теорема очевидна, если $\dim X = \infty$, ибо любое пространство может быть топологически включено в I_ω (теорема 4). Если dim X$\leq n$, где n конечно, то теорема следует из приведенной выше при доказательстве теоремы 5 формулы (3), так как $\overline{h(X)}$, как замкнутое подмножествоявляется I_{2n+1}, компактом.

2.1.7. Размерностный тип Фреше

По аналогии с теорией кардинальных чисел, Фреше в 1909 г. ввел понятие размерностного типа, говоря для двух данных пространств A и B, что размерностный тип пространства A не больше размерностного типа пространства B, если A может быть топологически включено в B.

Если при этом также справедливо, что размерностный тип пространства B не больше размерностного типа пространства A, то говорят, что A и B имеют один и тот же размерностный тип.

n-мерное эвклидово пространство, по определению, имеет размерностный тип n.

Очевидно, что два пространства одного и того размерностного типа имеют одну и ту же размерность. С другой стороны, простой пример окружности и дуги показывает, что два пространства одной и той же размерности не обязательно имеют один и тот же размерностный тип.

Теорему 3 можно перефразировать следующим образом: если dim $X \leq n$, то размерностный тип пространства $X \leq 2n+1$.

Теорема 5 устанавливает, что среди размерностных типов пространств размерности $\leq n$ существует наибольший размерностный тип. Теорема 4 утверждает, что размерностный тип каждого

пространства не больше размерностного типа гильбертова параллелепипеда.
Среди бесконечномерных пространств имеются пространства, по крайней мере, двух размерностных типов, а именно сам гильбертов параллелепипед и его подмножество, являющееся суммой счетного числа конечномерных пространств.

2.1.8. Дополнительные теоремы о покрытиях

Теорема 7. *Если в каждое покрытие пространства X можно вписать покрытие порядка* $\leq n$, *то X имеет размерность* $\leq n$.
Доказательство. Теорема следует из того, что в доказательстре теоремы 5 мы, пользуясь только указанным в условии свойством покрытий пространства *X*, вывели, что *X* может быть включено в универсальное *n*-мерное пространство χ_n.
Следствие. *Пусть X— компакт. Если X обладает покрытиями произвольно малого диаметра, имеющими порядок* $\leq n$, *то X имеет размерность* $\leq n$.
Доказательство. Из 4 С) заключаем, что в каждое покрытие пространства *X* можно вписать покрытие порядка $\leq n$. Поэтому следствие непосредственно вытекает из теоремы 7.
Теорема 8. **Теорема о покрытиях.** *Пространство имеет размерность* $\leq n$ *том и только в том случае, если в каждое его покрытие можно вписать покрытие порядка* $\leq n$.
Доказательство. Теорема является объединением теорем 1 и 7.
Следствие. **Теорема о покрытиях для компактов.**
Компакт имеет размерность $\leq n$ *в том и только в том случае, если он обладает покрытиями произвольно малого диаметра, имеющими порядок* $\leq n$.
Доказательство. Следствие есть объединение следствия теоремы 1 со следствием теоремы 7.

2.1.9. Нервы и отображения в полиэдры

Мы будем рассматривать полиэдры в наиболее элементарном смысле, а именно как точечные множества эвклидова пространства, являющиеся объединением конечного числа клеток.
Вершина или *нульмерная клетка* есть точка; *одномерная клетка* — отрезок без его концевых точек; *двумерная клетка* — треугольник без его сторон; *трехмерная клетка* — тетраэдр без его граней и т. д.

k-мерная клетка, $k = 0, 1, 2, .. .,$ определенная с помощью вершин, сторон, граней, ... *p*-мерной клетки, называется *k-мерной гранью* этой *k*-мерной клетки; мы также включаем саму *p*-мерную клетку в число ее граней. *n-мерный полиэдр* есгь точечное множество, расположенное в некотором E_m и составленное определенным способом из конечной системы непересекающихся *p*-мерных клеток, $0 \leqslant p \leqslant n$, по крайней мере, одна из которых является *n*-мерной клеткой, причем каждая грань каждой клетки системы также должна принадлежать системе.

Рассмотрим теперь произвольное множество объектов, которые мы будем называть *(абстрактными) вершинами.* Под *(абстрактным)* *p*-мерным симплексом s^p, p =0, 1, 2, ..., мы понимаем произвольное множество, состоящее из p+1 вершины. *k*-мерный симплекс, вершины которого выбраны из вершин симплекса s^p, называется *k-мерной гранью симплекса* s^p. s_p является своей собственной *p*-мерной гранью. *(Абстрактный) n-мерный комплекс* есть конечная система *p*-мерных симплексов, $0 \leqslant p \leqslant n$, содержащая вместе с симплексом и все его грани и, по крайней мере, один *n*-мерный симплекс.

Связь между полиэдрами и комплексами становится ясной из следующих двух утверждений:

А) Каждому полиэдру P поставим в соответствие комплекс *N,* называемый *комплексом остовов* полиэдра P; вершины комплексов *N* тождественны с вершинами полиэдра P, а симплексами комплекса остовов являютсч те совокупности вершин, которые определяют клетки полиэдра P.

Наоборот;

B) Для данного *n*-мерного комплекса *N* существует *n*-мерный полиэдр P, называемый *геометрической реализацией* комплекса *N,* комплексом остовов которого является *N;* кроме того оказывается, что *P* можно считать подмножеством куба I_{2n+1}.

Доказательство. Пусть $p_1, ..., p_r$ — вершины комплекса N. Легко можно показать, что с I_{2n+1} можно выбрать *r* вершин, которые мы продолжаем обозначать через $p_1, ..., p_r$, в общем положении, т. е, так, что любые $m + 2 (m \leqslant 2n)$ из этих точек линейно независимы. Пусть P — совокупность всех тех клеток в I_{2n+1}, порожденных вершинами $p_{i_0}, \ldots, p_{i_k}$, для которых $(p_{i_0}, \ldots, p_{i_k})$ является симплексом комплекса *N.* Если бы *P* было полиэдром, то его комплексом остовов был бы комплекс *N.* Таким образом, остается показать, что P есть полиэдр. Для этого докажем, что любые две клетки *s* и *t,* принадлежащие *P,* не пересекаются. Пусть $p_1, ..., p_k$ суть все различные вершины клеток s и *t.* Каждая из клеток полиэдра *P* имеет размерность $\leq n,$ следовательно, $k \leqslant 2n + 2$. Так как точки $p_1, ..., p_2$ находятся в общем положении, то точки $p_1, ..., p_k$ линейно независимы и, следовательно, определяют

(k— 1)-мерную клетку u, которая имеет и s, и t среди своих граней. Наше утверждение следует тогда из того факта, что любые две различные грани клетки не пересекаются.

Замечание. Легко можно показать, что два полиэдра с одним и тем же комплексом остовов гомеоморфны. Поэтому геометрическая реализация комплекса определяется топологически однозначно.

Александров ввел следующий процесс, ставящий в соответствие каждому покрытию пространства некоторый комплекс, называемый его нервом. **Понятие нерва имеет огромное значение в топологии, потому что оно связывает между собой непрерывные и комбинаторные методы.** Нервы можно рассматривать как комбинаторные конфигурации, аппроксимирующие пространство, при этом, чем мельче покрытие, тем лучше аппроксимация.

Определение 8. Пусть $\alpha: U_1, \ldots, U_r$— покрытие пространства. Каждому непустому U_i поставим в соответствие значок p_i и следующим образом построим, считая p_i вершинами, комплекс $N(\alpha)$, называемый *нервом* покрытия α:

$(p_{i_1}, \ldots, p_{i_k})$

является симплексом комплекса $N(\alpha)$ в том и только в том случае, если $U_{i_1} \cap \ldots \cap U_{i_k} \neq 0$.

Ясно, что $\dim N(\alpha) =$ порядку покрытия α.

Мы обозначаем через $P(\alpha)$ геометрическую реализацию комплекса $N(\alpha)$.

С понятием нерва тесно связано понятие «барицентрического» отображения.

Определение 9. Пусть X—пространство, и α — его покрытие. Пусть P— полиэдр, вершины которого $p_1 \ldots . p_r$ находятся во взаимно однозначном соответствии с элементами U_i покрытия α. Пусть Z_i — *звезда* вершины p_i, т. е. открытое множество в P, состоящее из всех клеток, имеющих p_i своей вершиной. Отображение g пространства X в P называется *барицентрическим α-отображением*, если

$$g^{-1}(Z_i) = U_i.$$

C) Так как Z_i образуют покрытие полиэдра P, то барицентрическое α-отображение является α-отображением (см. определение 5).

D) Пусть g—барицентрическое α-отображение пространства X в полиэдр P, и P'—полиэдр, содержащийся в P и состоящий из всех граней клеток полиэдра P, содержащих хотя бы одну точку множества $g(X)$. Тогда комплекс остовов полиэдра P' совпадает с $N(\alpha)$, или, что то же, P' является геометрической реализацией нерва $N(\alpha)$.

Доказательство. Сначала мы покажем, что если s' — произвольная клетка полиэдра P′, которую мы обозначим через (p_1, ..., p_m), то

$$\bigcap_{i=1}^{m} U_i \neq 0, \quad (1)$$

т. е. s' является также симплексом комплекса $N(\alpha)$. По определению P' существует клетка s (обозначим ее через $(p', \ldots, p_m, p_{m+1}, \ldots, p_k)$), содержащая точку $y_0 \in g(X)$ и имеющая s' своей гранью. Тогда

$$y_0 \in Z_i, \qquad i = 1, \ldots, k.$$

Отсюда

$$g^{-1}(y_0) \subset g^{-1}(Z_i) = U_i, \qquad i = 1, \ldots, k,$$

и это доказывает (1).

Наоборот, пусть $s(p_1, \ldots, p_m)$ — произвольный симплекс нерва $N(\alpha)$. Тогда $U_1,\ldots..U_m$ имеет общую точку x_0. Далее

$$g(x_0) \in Z_i \qquad i = 1, \ldots, m.$$

и так как $g(x_0)$ является точкой полиэдра P' то отсюда вытекает, что

$$g(x_0) \in Z_i', \qquad i = 1, \ldots, m,$$

где Z'_i обозначает звезду вершины p_i в полиэдре P'. Следовательно, звезды Z_i', $i = 1, \ldots, m$ имеют непустое пересечение, и потому (p_1, ..., p_m) является клеткой полиэдра P'. Доказательство D) закончено.

Интересный частный случай предложения D) получается, когда $X = P$, α есть покрытие полиэдра P, звездами его вершин, а g— тождественное отображение P на себя. В этом случае P' есть само P, так что D) утверждает, что нерв покрытия α есть *комплекс остовов полиэдра P*.

Предложение D) показывает, что, не уменьшая общности, мы можем при изучении барицентрических α-отображений ограничиться случаем, когда P является геометрической реализацией нерва покрытия α.

Название: «барицентрическое» α-отображение оправдывается следующим предложением.

E) Барицентрические α-отображения пространства X в P(α) совпадают с отображениями g, получаемыми следующей конструкцией. Пусть U_1, .. ., U_r — элементы покрытия α. Определим r непрерывных действительных функций $w_i(x)$ таких, что

$$w_i(x) = 0, \quad \text{если} \quad x \notin U_i, \quad (2)$$

$$w_i(x) > 0, \quad \text{если} \quad x \in U_i. \quad (3)$$

$g(x)$ есть тогда отображение, ставящее в соответствие каждой точке $x \in X$ центр тяжести вершин p_i с весами $w_i(x)$.

Доказательство. Пусть g — барицентрическое α-отображение. Для данной точки $x \in X$ мы определяем $w_i(x)$ следующим образом:

Пусть $s = (p_{i_1}, \ldots, p_{i_k})$—клетка полиэдра P(α), содержащая $g(x)$. Пусть

$w_i(x) = 0$, если i отлично от каждого из $i_1, \ldots, i_k$,

$w_i(x)$ = барицентрической координате [2]) точки относительно p_i, если i—одно из $i_1, \ldots, i_k$.

(Барицентрическими координатами точки p клетки $(p_1, \ldots, p_k)$ являются веса $w_1 \ldots w_k$, в сумме равные 1, которые надо приписать соответствующим вершинам, чтобы получить p в качестве центра тяжести.)

$w_i(x) > 0$ тогда и только тогда, когда $p_i \in s$, т. е. $g(x) \in U_i$.

Следовательно, функции $w_i(x)$ удовлетворяют условиям (2) и (3). Обратное предложение доказывается подобным же образом.

F) Всегда могут быть найдены функции $w_i(x)$, удовлетворяющие условиям (2) н (3) предложения E), иными словами, *для каждого покрытия α пространства X существует барицентрическое α-отображение пространства X в* P(α).

Доказательство. Достаточно положить:

$$w_i(x) = \rho(x, X \setminus U_i).$$

(Если $X \setminus U_i$ пусто, полагаем $w_i(x) = 1$.)

Замечание 1. Если мы просмотрим доказательства предложений 3C) и 4E), то мы увидим, что построенное там отображение g обладает тем свойством, что $\overline{g(x)}$ содержится в некотором n-мерном полиэдре, лежащем в I_{2n+1}; в действительности, g является барицентрическим отображением пространства X в геометрическую реализацию нерва покрытия β. Это замечание позволяет высказать следующее усиление предложения 4E):

4E') Пусть $\dim X \leqslant n$, где n конечно. Для каждого покрытия α пространства X обозначим через G'_α множество α-отображений пространства X таких, что $\overline{g(X)}$ содержится в n-мерном полиэдре, расположенном в I_{2n+i}. Тогда G_α плотно в пространстве отображений I_{2n+1}^X.

Замечание 2. Доказательство теоремы о включении использует следующие идеи:

(a) n-мерное пространство обладает произвольно мелкими покрытиями порядка n;

(b) нерв покрытия порядка n может быть геометрически реализован с помощью полиэдра, лежащего в I_{2n+1};

(c) пространство может быть барицентрически отображено на геометрическую реализацию любого из его нервов (предложение B));

(d) теорема Бэра.

Теперь будет доказана

Теорема 9. Аппроксимация полиэдрами. *Пространство X имеет размерность $\leq n$ в том и только в том случае, если, для каждого покрытия α пространства X существует α-отображение пространства X в полиэдр размерности $\leq n$.*

Доказательство. Необходимость. Пусть dim $X \leq n$. По теореме V 1 существует покрытие α' порядка $\leq n$, вписанное в покрытие α. В силу F), существует α'-отображение (в действительности барицентрическое α'-отображение) пространства X в P(α'). Но α'-отображение и подавно является α-отображением, а P(α') имеет размерность $\leq n$.

Доказательство достаточности содержится в следующем предложении.

G) Пусть dim $X \geq m$. Тогда существует покрытие α пространства X, обладающее тем свойством, что для каждого α-отображения g пространства X в компакт Y

$$\dim g(X) \geqslant m.$$

Доказательство. По приведенной ранее теореме существует такое покрытие α пространства X, что каждое вписанное в него покрытие имеет порядок $\geq m$. Пусть g есть α-отображение пространства X в Y. Каждая точка компакта Y имеет окрестность, полный прообраз которой содержится в некотором элементе покрытия α. Так как Y — компакт, конечное число этих окрестностей покрывает Y, и в пересечении с $g(X)$ это конечное число окрестностей образует покрытие β множества $g(X)$. Допустим, что dim $g(X) < m$. Тогда в β можно вписать покрытие β' порядка $< m$. Полные прообразы элементов покрытия β' образуют покрытие порядка $< m$, вписанное в α, в противоречии с предположением. Следовательно, доказано предложение G), а с ним и теорема 9.

Следствие. Теорема Александрова об.аппроксимации компактов полиэдрами. *Пусть X—компакт. Тогда* dim.Y$\geq n$ в *том и только в том, лучае, если для каждого* $\varepsilon > 0$ *существует ε-отображение компакта X в полиэдр размерности* $\geq n$.

Доказательство, Если $\dim X \leqslant n$, то из теоремы 9 следует, что для каждого покрытия α существует α-отображение g_α пространства X в полиэлр размерности $\geq n$. Если дано $\varepsilon > 0$, то пусть α есть покрытие пространства X, имеюшее диаметр $< \varepsilon$. Такое покрытие существует, так как X— компакт. Для этого α-отображение g_α является ε-отображением.

Наоборот, допустим, что для каждого $\varepsilon > 0$ существует ε-отображение g_t пространства X в полиэдр размерности $\geq n$. Если дано произвольное покрытие α пространства X, то пусть ε— положительное число (см. 4 C)), обладающее тем свойством, что любое множество диаметра $< \varepsilon$ содержится в каком-либо элементе покрытия α. Для этого ε отображе-

ние g_α является α-отображением. Отсюда, по теореме 9, заключаем, что dim X≤*n*. Итак, следствие доказано.

Пусть α — покрытие пространства *X*, и Р(α) — геометрическая реализация нерва покрытия α. Отображение *f* пространства *X* в *P* (α) называется *квази-барицентричестм α-отображением,* если (см. определение 9)

$$f^{-1}(Z_i) \subset U_i.$$

*(*Нетрудно показать, что квази-барицентрические α-отображения совпадают с отображениями, определяемыми непрерывными действительными функциями w_i *(x)* такими, что

$$w_i(x) = 0, \text{ если } X \notin U_i, \quad (2)$$

$$w_i(x) \geqslant 0, \text{ если } x \in U_i, \quad (3)$$

$$\Sigma w_i(x) > 0 \text{ для каждой точки } x. \quad (4)$$

(Сопоставьте эти формулы с формулами (2) и (3) Е)). Если дано нормальное пространство *X,* со счетным базисом или без него, и произвольное покрытие α: $U_1, .. ., U_r$ пространства *X,* то всегда существует квази-барицентрическое α-отобрэжение пространства *X* в *P(α).* Это доказывается следующим образом. Так как *X* нормально, то для каждого U_i найдется открытое множество V_i такое, что

$$\overline{V_i} \subset U_i;$$

V_i образуют покрытие *α'* пространства *X.* Нервы покрытия α и α' тождественны. Также, в силу нормальности пространства *X,* найдутся *r* непрерывных действительных функций w_i *(x),* таких, что

$$0 \leqslant w_i(x) \leqslant 1,$$

$$w_i(x) = 0 \text{ для каждой точки } x \notin U_i,$$

$$w_i(x) = 1 \text{ для каждой точки } x \in V.$$

Эти функции w_i *(x)* удовлетворяет указанным выше условиям (2'), (3') и (4').

Квази-барицентрическое α-отображение является α-отображением,

Н) Рассмотрим все подполиэдры *P'* полиэдра *P* (α), обладающие следующим свойством:

(4) Существует квази-барицентрическое α-отображение *f* пространства *X* в *P* (α) такое, что $f(X) \subset P'$.

Пусть P_o — подполиэдр полиэдра *P* (α), неприводимый по отношению к этому свойству, т. е. такой, что P_o удовлетворяет условию (4), в то время как никакой его подполиэдр не удовлетворяет условию (4). Такой полиэдр, очевидно, существует. Пусть

f_0— квази-барицентрическое α-отображение пространства X в P (α), для которого $f_0(x) \subset P_0$. Тогда

$$f_0(X) = P_0.$$

Доказательство. Допустим, что точка $p_0 \in P_0$ не принадлежит f_0(X). Пусть s— клетка полиэдра P_0, содержащая p_0, и пусть S—*звезда клетки* s, т. е. множество всех клеток, имеющих s своей гранью. Обозначим через $B(S)$ («граница» S) множество всех клеток, не принадлежащих S, но являющихся гранями клеток звезды S. Для каждой точки $x \in X$, образ которой принадлежит S, мы заменим $f_0(x)$ ее проекцией $f'_0(x)$ на B (S) из точки p_0. Легко доказать, что $f'(x)$ также является квази-барицентрическим α-отображением пространства X в P_0. Но $f'(X)$ содержится в собственном подполиэдре полиэдра P_0, что противоречит определению полиэдра P_0. Это показывает, что в теореме 9 отображение «в» может быть заменено отображением «на».

Теорема 10. *Пространство X имеет размерность* $\leq n$ *в том и только в том случае, если для каждого покрытия* α *пространства X существует α-отображение пространства X на полиэдр размерности* $\leq n$.

Доказательство. Нуждается в доказательстве только необходимость. Пусть $\dim X \leq n$. Впишем в покрытие α покрытие α' порядка $\leq n$. Пусть f_0 есть α'-отображение пространства X на неприводимый полиэдр P_0, существование которого вытекает из предложения Н). Тогда f_0 есть требуемое отображение.

2.2. Отображения в сферы

В этом разделе мы изучим отображения топологических пространств в сферы и дадим характеристику размерности в терминах таких отображений. Эта характеристика, содержащаяся в теореме 4, является основным результатом раздела. Она находит свои наиболее важные приложения в последующих разделах, где служит базисом для алгебраической и комбинаторной трактовки теории размерности. Аппарат отображений в сферы и чрезвычайно важное понятие гомотопии применяются далее для того, чтобы получить простые доказательства теорем о разбиении эвклидовых пространств (n-мерная теорема Жордзна и исследовать изменения размерности, производимые непрерывными отображениями пространства .

2.2.1. Устойчивые и неустойчивые значения

В этом подразделе мы исследуем следующую проблему: при каких условиях пространство X можно отобразить в n-мерное эвклидово пространство E_n так, чтобы, по крайней мере, одна точка пространства E_n была покрыта «существенно», т. е. не могла стать непокрытой при произвольно малых изменениях отображения? Другкми словами, при каких условиях возможно определить n непрерывных действительных функций

$$f_i(x), \qquad x \in X, \qquad i = 1, \ldots, n$$

таким образом, чтобы для любой системы непрерывных функций $g_i(x)$, аппроксимирующих функции $f_i(x)$ достаточно близко, система уравнений $g_i(x) = 0$

имела решение $x \in X$?

Оказывается (теоремы 1 и 2), что такая система функций существует в том и только в том случае, если $\dim X \geq n$.

Определение 1. Пусть f — отображение пространства X в пространство Y. Точка y множества $f(X)$ называется *неустойчивым значением* отображения f, если для всякого $\delta > 0$ существует отображение g пространства X в Y, удовлетворяющее условиям:

$$\rho(f(x), g(x)) < \delta \text{ для каждой точки } x \text{ в } X, \quad (1)$$

$$g(X) \subset Y \setminus y. \quad (2)$$

Остальные точки множества $f(X)$ называются *устойчивыми* значениями отображения f.

Пример 1. Пусть f — отображение прямой в себя, определенное функцией $y = x^2$. Точка $y = 0$ является неустойчивым значением отображения f, Остальные значения функции f устойчивы:

Пример 2. Пусть f — тождественное отображение n-мерного куба I_n на себя. Каждая граничная точка куба I_n является неустойчивым значением отображения f, так как куб I_n отображением, очень мало отличающимся от тождественного, может быть преобразован в меньший концентрический куб. С другой стороны, каждая внутренняя точка куба I_n является устойчивым значением отображения f. Ввиду однородности достаточно доказать это для начала координат (0, 0,...,0). Мы покажем, что для любого отображения g, удовлетворяющего

$$\delta = \frac{1}{2},$$

неравенству (1) при начало координат является точкой множества $g(I_n)$.

Пользуясь векторными обозначениями, рассмотрим отображение

$$h(x) = x - g(x) = f(x) - g(x).$$

Это отображение преобразует куб $|x| \leqslant \frac{1}{2}$ в себя и, следовательно по теореме Браузра о неподвижной точке, должна найтись точка *x,* для которой $k(x_0) = x_0$, т. е. $g(x_0) = 0$.

Пример 3. Пусть *X*—-произвольное множество пространства E_n, и *f*— тождественное отображение *X* в E_n. Тогда каждая внутренняя точка *p* множества *X* явтяется устойчивым значением отображения *f*. Ибо *p* принадлежит кубу *Q,* содержащемуся в *X*, и из предыдущего примера следует, что *p* является устойчивым значением для частичного отображения (отображения *f*, рассматриваемого только на *Q)* $f|_Q$ и, следовательно, тем более для *f*. Мы предоставляем читателю доказательство того, что граничные точки множества *X* неустойчивы.

Пример 4. Пусть *X*—произвольное пространство, и *f*—его отображение в I_n . Тогда каждая точка границы куба I_n является неустойчивой. Ибо, если дано положительное δ(< 1), функции

$$g_i(x) = (1 - \delta) f_i(x), \qquad i = 1, \ldots, n, \quad \textbf{(3)}$$

определяют отображение, при котором образ пространства *X* не покрывает ни одной граничной точки куба I_n .

Теорема 1. *Пусть X—пространство, размерность которого меньше n, и f—отображение пространства X в* I_n. *Тогда все значения отображения f неустойчивы.*

Доказательство. Из примера 4 мы знаем, что никакая граничная точка куба I_n не может быть устойчивым значением. Следовательно, достаточно доказать, что нет никаких устойчивых точек среди внутренних точек куба I_n , а это равносильно утверждению, что начало координат не является устойчивым значением отображения *f*. Пусть $f_1(x), \ldots, f_n(x)$— координаты точки *f*(*x*), δ — произвольное положительное число, c_i^+—множество точек пространства *X,* для которых

$$f_i(x) \geqslant \delta,$$

и c_i^-—-множество точек пространства *X,* для которых

$$f_i(x) \leqslant -\delta.$$

Для каждого *i* множества C_i^+ и C_i^- замкнуты и не пересекаются. Следовательно, существуют замкнутые множества $B_1,.....B_n$ такие, что B_i отделяет c_i^+ от c_i^-, т. е.

$$X \setminus B_i = U_i^+ \cup U_i^-,$$

где U_i^+ и U_i^-— непересекающиеся открытые множества, содержащие соответственно c_i^+ и c_i^-, и

$$B_1 \cap \ldots \cap B_n = 0. \quad (4)$$

Мы определим новые функции $g_1(x), \ldots, g_n(x)$:

$$g_i(x) = f_i(x), \text{ если } x \in C_i^+ \cup C_i^-,$$

$$g_i(x) = \delta \frac{\rho(x, B_i)}{\rho(x, C_i^+) + \rho(x, B_i)}, \text{ если } x \in U_i^+ \setminus C_i^+,$$

$$g_i(x) = -\delta \frac{\rho(x, B_i)}{\rho(x, C_i^-) + \rho(x, B_i)}, \text{ если } x \in U_i^- \setminus C_i^-,$$

$$g_i(x) = 0, \text{ если } x \in B_i.$$

Легко видеть, что $g_i(x)$ непрерывны и что

$$|g_i(x) - f_i(x)| \leqslant 2\delta. \quad (5)$$

Кроме того, $g_i(x)$ равно нулю только в том случае, если x принадлежит B_i. Следовательно по (4), не существует ни одной точки в пространстве X, в которой все функции $g_i(x)$ одновременно равны нулю. Это значит, что начало координат не является точкой образа пространства X при отображении, определенном функциями g_i, а вместе с (5) это показывает, что начало координат не является устойчивым значением отображения f.

А) Пусть даны отображение f пространства X в I_n и точка

$$y \in I_n.$$

Если $f(X) \subset I_n \setminus y$, то для всякого $\delta > 0$ существует отображение g пространства X в I_n такое, что

$$\rho(f(x), g(x)) < \delta \text{ для каждой точки } x \in X, \quad \textbf{(6)}$$

$$\overline{g(X)} \subset I_n \setminus y. \quad \textbf{(7)}$$

Доказательство. Если y лежит на границе куба I_n, то для нашей цели годится отображение, определенное формулой (3) в примере 3. Пусть теперь y — внутренняя точка куба I_n, и пусть $f'(x)$— проекция точки $f(x)$ из y на границу куба I_n. Если длина отрезка, соединяющего $f(x)$ с $f'(x)$, $\geqslant \frac{1}{2}\delta$, то определяем $g(x)$, как точку этого отрезка, находящуюся на расстоянии $\frac{1}{2}\delta$ от $f(x)$; если длина отрезка, соединяющего $f(x)$ с $f'(x)$, меньше, чем $\frac{1}{2}\delta$, то мы полагаем, что $g(x)$ равно $f'(x)$. Очевидно, для g условия (6) и (7) выполнены. Итак, А) доказано.

Теперь мы докажем теорему, обратную теореме 1, пользуясь аппаратом пространств отображений.

Теорема 2. *Если X— пространство размерности ≥n, то существует отображение пространства X в I_n, имеющее, по крайней мере, одно устойчивое значение.*

Доказательство. Допустим, что никакое отображение пространства X в I_n не имеет устойчивых значений. Из А) и определения неустойчивых значений следует, что для всякой точки y куба I_n каждое отображение f

пространства X в I_n может быть сколь угодно точно аппроксимировано отображениями g, обладающими тем свойством, что

$$\overline{g(X)} \subset I_n \setminus y.$$

Рассмотрим теперь пространство I^X_ω отображений пространства X в гильбертов параллелепипед. Пусть $M = M(i_1, \ldots, i_n; c_1, \ldots, c_n)$ — линейное подпространство гильбертова пространства, определяемое n уравнениями

$$x_{i_1} = c_1, \ \ldots, \ x_{i_n} = c_n. \quad (8)$$

Обозначим через $G(M)$ множество отображений g пространства X в I_n, обладающих тем свойством, что

$$\overline{g(X)} \subset I_\omega \setminus M.$$

Множество $G(M)$ *плотно* в I^X_ω, так как, если дано произвольное отображение

$$f(x) = (f_1(x), \ f_2(x), \ \ldots), \quad |f_i(x)| \leqslant \frac{1}{i}$$

пространства X в I_ω, то функции

$$f_{i_1}(x), \ \ldots, \ f_{i_n}(x) \quad \textbf{(9)}$$

определяют отображение X в I_n и, как замечено выше, можно освободить точку $(c_1, \ldots, c_n)$ от замыкания образа пространства X, произвольно мало изменив отображение (9).

Далее, множество $G(M)$ *открыто* в I^X_ω. Так как, если g принадлежит $G(M)_t$ то расстояние

$$d = \rho(g(X), M)$$

положительно; каждое же отображение f, для которого $\rho(f, g) < d$, принадлежит множеству $G(M)$.

Теперь мы сосредоточим наше внимание на тех отображениях g пространства X в I_ω, для которых

$$\overline{g(X)} \subset \mathcal{M}^{n-1}_\omega, \quad (10)$$

где $\mathcal{M}^{n-1}_\omega$ обозначает множество точек гильбертова параллелепипеда I_ω, имеющих не более n — 1 рациональных координат. Дополнение множества $\mathcal{M}^{n-1}_\omega$ в I_ω является пересечением I_ω с суммой счетного числа линейных пространств $M_1, M_2, \ldots$, типа (8), именно подпространств, соответствующих всевозможным комбинациям n индексов i_j и n рациональных чисел c_j. Следовательно, (10) - эквивалентно условию $\overline{g(X)} \subset I_\omega \setminus M_i$, или

$$g \in G(M_i) \quad \text{для каждого } i = 1, 2, \ldots$$

I^X_ω, в силу теоремы 4, содержит плотное G_δ-множество H, каждый элемент которого является гомеоморфизмом. Множество

$$H' = H \cap \bigcap_i G(M_i),$$

как пересечение счетного числа плотных G_δ-множеств в полном пространстве, само плотно в I_ω^n; в частности, оно не пусто. Следовательно, существует гомеоморфизм h, отображающий пространство X на подмножество множества $\mathscr{M}_\omega^{n-1}$.Но

$$\dim \mathscr{M}_\omega^{n-1} \leqslant n-1$$

— в противоречии с предположением, что

$$\dim X \geqslant n.$$

Это завершает доказательство теоремы 2.

Замечание. Из теоремы 2 вытекает, что в пространстве, размерность которого $\geq n$, всегда можно определить n пар замкнутых множеств $C_i, C_i'; i = 1, \ldots, n; C_j \cap C_i' = 0,$ удовлетворяющих следующему условию: если B_i, $i = 1.....n,$ есть замкнутое множество, отделяющее C_i от C_i', то

$$B_1 \cap \ldots \cap B_n \neq 0.$$

Ибо, в противном случае, мы могли бы, пользуясь рассуждениями доказательства теоремы 1, получить противоречие с теоремой 2.

Следующее предложение показывает, что для отображений в I_n вопрос о том, будет ли некоторая точка устойчивой или неустойчивой, зависит только от поведения отображения в произвольно малой окрестности этой точки.

В) Пусть f—отображение пространства X в I_n. Внутренняя точка y множества $f(X)$ является неустойчивым значением отображения f в том и только в том случае, если для каждой окрестности U точки y существует отображение g пространства X в I_n, удовлетворяющее условиям:

$$g(x) = f(x), \quad \text{если} \quad f(x) \notin U, \tag{11}$$

$$g(x) \in U, \quad \text{если} \quad f(x) \in U, \tag{12}$$

$$y \notin g(X). \tag{13}$$

Доказательство. Нетрудно видеть, что условие достаточно. Ибо из (11) и (12) мы имеем

$$\rho(f(x), g(x)) \leqslant \delta(U) \text{ для каждой точки } x \in X,$$

так что существуют отображения g, аппроксимирующие f произвольно точно и удовлетворяющие соотношению (2).

Перейдем к доказательству необходимости. Пусть y— внутренняя точка куба I_n, и S—действительное положительное число. Не уменьшая общности, можно предположить, что y — начало координат, а U — сферическая окрестность точки y радиуса δ. Так как y — неустойчивое значение отображения f, то существует отображение g' пространства X в I_n, для которого в векторных обозначениях

$$|f(x)-g'(x)| = \rho(f(x), g'(x)) < \frac{1}{2}\delta. \quad (14\text{-}15)$$
$$g'(x) \neq 0,$$

Построим новое отображение *g* следующим образом:

$$g(x) = g'(x), \quad \text{если} \quad |f(x)| \leqslant \frac{1}{2}\delta, \quad (16)$$

$$g(x) = 2\left(1 - \frac{|f(x)|}{\delta}\right)g'(x) - \left(1 - \frac{2|f(x)|}{\delta}\right)f(x),$$

$$\text{если } \frac{1}{2}\delta < |f(x)| < \delta, \quad (17)$$

$$g(x) = f(x), \quad \text{если} \quad |f(x)| \geqslant \delta. \quad (18)$$

Проверим, что *g(x)* является отображением пространства *X* в I_n, обладающим свойствами (11) —(13). (11) равносильна условию (18). Если $\frac{1}{2}\delta < |f(x)| < \delta$, из (14 и (17) с помощью простых выкладок получаем:

$$|f(x)-g(x)| = 2\left(1 - \frac{|f(x)|}{\delta}\right)|f(x)-g'(x)| < \delta - |f(x)|$$

и, следовательно,

$$0 < |g(x)| < \delta. \quad (19)$$

В силу (14) (15) и (16), неравенство (19) имеет силу и в случае, если $|f(x)| \leqslant \frac{1}{2}\delta$. Это доказывает (12) и (13) и завершает доказательство предложения В). Из В) получаем следующее расширение теоремы I:

С) Пусть *X* — пространство, размерность которого меньше *n,* и *f* — его отображение в пространство *B,* содержащее открытое подмножество *U,* гомеоморфное E_n. Тогда все значения отображения *f,* принадлежащие *U,* неустойчивы.

Доказательство. Пусть *U′*— полный прообраз множества *U* при *f*. Частичное отображение $f|U'$ можно рассматривать как отображение множества *U′* в I_n, так как *U* гомеоморфно множеству внутренних точек куба I_n. Отсюда, в силу теоремы 1 и В), для каждой точки $y \in U$ и окрестности *V* точки *y,* замыкание которой содержится в *U,* существует отображение *g* множества *U′* в *U* такое, что

$$g(x) = f(x), \quad \text{если} \quad f(x) \notin V, \quad (20)$$
$$g(x) \in V, \quad \text{если} \quad f(x) \in V, \quad (21)$$
$$y \notin g(X). \quad (22)$$

Положив *g(x)* равным *f(x)* для $x \in X \setminus U'$, получим отображение (которое будем попрежнему обозначать через *g)* всего пространства *X* в *B.* Новое отображение *g* сохраняет свойства (20), (21), (22). Так как *V,* а следовательно, и $\rho(f, g) =$
$= \sup_{x \in X} \rho(f(x), g(x))$

могут быть сделаны сколь угодно малыми,
каждое значение отображения *f* является неустойчивым.
D) Если *f* есть отображение пространства *X* в *n*-мерную сферу и dim *X* < *n*, то все значения отображения *f* неустойчивы.
Доказательство. Предложение D) является следствием предложения C).

2.2.2. Продолжение отображений

Определение 2. Пусть *A*— подмножество пространства *X*, и *f(x)* — отображение множества *A* в пространство Y. Отображение *F(x)* пространства *X* в *Y*, удовлетворяющее условию

$$F(x) = f(x) \quad \text{для} \quad x \in A,$$

называется *продолжением отображения f ни X относительно Y*. В том случае, когда недоразумения невозможны, слова «относительно Y» будут опускаться.
Теорема 3. Теорема Титце о продолжении.*Пусть C — замкнутое подмножество пространства X, и f*(*x*) *— непрерывная действительная функция, определенная на C и ограниченная константой k:*

$$|f(x)| \leqslant k.$$

Тогда существует действительная функция F(x), являющаяся продолжением отображения f на X такая, что

$$|F(x)| \leqslant k.$$

Доказательство. Мы докажем сначала существование непрерывной функции F_1, определенной на *X*, удовлетворяющей неравенствам:

$$|F_1(x)| \leqslant \frac{1}{3}k, \quad x \in X.$$

$$|F_1(x) - f(x)| \leqslant \frac{2}{3}k, \quad x \in C.$$

Пусть C^+ — множество точек $x \in C$, для которых

$$f(x) \geqslant \frac{1}{3}k,$$

и C^- — множество, для точек которого

$$f(x) \leqslant -\frac{1}{3}k.$$

Тогда функция

$$F_1(x) = \frac{1}{3}k\,\frac{\rho(x, C^-) - \rho(x, C^+)}{\rho(x, C^-) + \rho(x, C^+)}$$

обладает требуемыми свойствами, так как

$$F_1(x) = \frac{1}{3}k, \quad \text{если} \quad f(x) \geqslant \frac{1}{k}k,$$

$$-\frac{1}{3}k \leqslant F_1(x) \leqslant \frac{1}{3}k, \quad \text{если} \quad -\frac{1}{3}k \leqslant f(x) \leqslant \frac{1}{3}k,$$

$$F_1(x) = -\frac{1}{3}k, \quad \text{если} \quad f(x) \leqslant -\frac{1}{3}k.$$

Заменяя $f(x)$ функцией $f(x) - F_1(x)$ и k — числом $\frac{2}{3}k$, определим на X функцию $F_2(x)$ такую, что

$$|F_2(x)| \leqslant \frac{2}{3^2}k, \quad x \in X,$$

$$|f(x) - F_1(x) - F_2(x)| \leqslant \frac{2^2}{3^2}k, \quad x \in C.$$

Продолжая этот процесс, получим последовательность $\{F_n(x)\}$ непрерывных функций, определенных на X, удовлетворяющих неравенствам

$$|F_n(x)| \leqslant \frac{2^{n-1}}{3^n}k, \quad x \in X \quad (1)$$

$$|f(x) - \sum_{i=1}^{n} F_i(x)| \leqslant \frac{2^n}{3^n}k, \; x \in C. \quad (2)$$

Неравенство (1) показывает, что ряд

$$\sum_{n=1}^{\infty} F_n(x)$$

сходится равномерно на X и, следовательно, имеет непрерывную сумму $F(x)$, Из (1) и (2) следуют соотношения:

$$|F(x)| \leqslant k$$

и $F(x) = f(x)$ для $x \in C$,

что и требовалось доказать.

Следствие 1. *Пусть C — замкнутое подмножество пространства X, и f—произвольное отображение множества C в I_n (в I_ω). Тогда f может быть продолжено на X (относительно I_n (I_ω)).*

Доказательство. Утверждение следует из применения теоремы VI 3 к каждой из координат точки $f(x)$.

Если заменить I_n n-мерной сферой S_n и рассмотреть отображение f замкнутого подмножества C пространства X в S_n, то, вообще говоря, f не может быть продолжено на все X относительно S_n. Например, пусть X—замкнутый шар

$$\sum_{i=1}^{n+1} x_i^2 \leqslant 1$$

в E_{n+1} ограниченный сферой S_n. Как было показано ранеев, тождественное отображение S_n на S_n не может быть продолжена в отображение пространства X в S_n. Однако:

Следствие 2. *Пусть C — замкнутое подмножество пространства X, и f—отображение C в n-мерную сферу S_n . Тогда существует открытое в X множество, содержащее C, на которое f может быть продолжено (относительно* S_n).

(Следствия 1 и 2 выражают определенные свойства пространств I_n и S_n. Если пространство A обладает свойством, которым, в силу следствия 1, обладает I_n , то оно называется *абсолютным ретрактом;* если A обладает свойством, которым, в силу следствия 2, обладает S_n, то оно называется *абсолютным окрестностным ретрактом.* Можно показать, что каждый полиэдр является абсолютным окрестностным ретрактом. Эти понятия были введены Борсуком.)

Доказательство. Пусть координаты точки $f(x)$ в E_{n+1} суть

$$f_1(x), \ldots, f_{n+1}(x).$$

Если S_n имеет радиус 1, то

$$\sum_{i=1}^{n+1} (f_i(x))^2 = 1,$$

и, следовательно,

$$|f_i(x)| \leqslant 1, \quad i = 1, \ldots, n+1.$$

В силу теоремы Титце о продолжении, существует продолжение $F_i(x)$ функций $f_i(x)$ на X. Пусть U— множество точек пространства X, для которых

$$\sum (F_i(x))^2 > 0.$$

U является, очевидно, открытым множеством, содержащим C. Если положить, что $F(x)$ является точкой сферы S_n, i-я координата которой равна

$$\frac{F_i(x)}{\left[\sum_{i=1}^{n+1} (F_i(x))^2\right]^{1/2}},$$

то отображение F будет искомым продолжением отображения f на U. Теорема Титце о продолжении делает возможной перефразировку теорем 1 и 2 в терминах отображений в сферы.

Теорема 4. *Пространство X имеет размерность* $\leq n$ *в том и только в том случае, если для каждого замкнутого множества C и отображения f множества C в S_n существует продолжение отображения f на X.*

Доказательство. *Условие необходимо.* Пусть даны замкнутое множество *C* и отображение *f* множества C в сферу S_n, которую мы здесь будем рассматривать, как границу куба I_{n+1} является тогда отображением множества *C* в I_{n+1}. По следствию 1 теоремы Титце о продолжении, существует отображение *F'* пространства *X* в I_{n+1}, которое является продолжением отображения *f*. Так как

$$\dim X \leqslant n,$$

то из теоремы 1 следует, что начало координат не является устойчивым значением отображения *F'*. Тогда 1 В) дает нам отображение *F''* пространства *X* в I_{n+1} такое, что начало координат не содержится в *F'' (X),* тогда как $F''(x) = F'(x)$ для всех значений *F' (x),* не являющихся внутренними точками куба I_{n+1}. В частности, для $x \in C$

$$F''(x) = F'(x) = f(x).$$

Пусть *F(x)* — проекция точки *F''(x)* на границу куба I_{n+1} из начала координат. Тогда, очевидно, *F* является нужным нам продолжением отображения *f.*

Условие достаточно. Для того чтобы доказать, что dim X≤ *n,* достаточно, в силу теоремы 2, показать, что отображение *f* пространства *X* в I_{n+1} не может иметь устойчивых значений. Граничная точка куба I_{n+1} никогда не устойчива (*см.* пример 4). Поэтому пусть *y* — внутренняя точка куба I_{n-1}, и U—сферическая окрестность точки *y* радиуса $\frac{1}{2}\delta$. Будем рассматривать S_n как границу множества *U.* Обозначим через *C* полный прообраз сферы S_n при *f*. C замкнуто, потому что *f* непрерывно. По предположению существует отображение *F* всего пространства *X* в S_n такое, что $F(x) = f(x)$ для $x \in C$. Построим теперь новое отображение *g* (*x*) пространства *X* в I_{n+1}, положив

$$g(x) = f(x), \quad \text{если} \quad f(x) \notin U,$$

$$g(x) = F(x), \quad \text{если} \quad f(x) \in U;$$

g(x) является отображением пространства *X* в $I_{n+1} \setminus U$, и $\rho(f, g) < \delta$. Теорема доказана.

Следствие. *Пусть C—замкнутое подмножество пространства X. Тогда, если X\C имеет размерность ≤ n, то каждое отображение множества C в S_n может быть продолжено на X.*

Доказательство. Пусть *f*—отображение множества C в S_n. Следствие 2 теоремы Титце о продолжении показывает, что существует открытое множество *U,* содержащее C, и продолжение *f'* отображения *f* на *U.* Пусть *V* — открытое в *X* множество такое, что

$$C \subset V \subset \overline{V} \subset U;$$

V существует, так так *X* нормально. Рассмотрим частичное отображение $f \mid \overline{V} \cap (X \setminus C)$.

Это есть отображение в S_n замкнутого подмножества пространства $X \setminus C$. Но $X \setminus C$ имеет размерность $\leq n$. По теореме 4 существует поэтому продолжение f'' этого отображения на $X \setminus C$. Если мы положим

$$F(x) = f(x) \quad \text{для} \quad x \in C,$$

$$F(x) = f''(x) \quad \text{для} \quad x \in X \setminus C,$$

то получим искомое продолжение отображения *f* на *X*.
Важным приложением теоремы 4 является простое доказательство теоремы Брауэра об инвариантности области.

2.2.3. Гомотопия

Определение 3. Пусть *X* и *Y*—два пространства. Мы скажем, что отображение *f* пространства *X* в *Y гомотопно* отображению *g* пространства *X* в Y, если можно найти функцию *f(x, t)* двух переменных *x* и *t,* где *x* — точка пространства *X,* а *t* — действительное число, $0 \leqslant t \leqslant 1$, принимающую значения из пространства *Y,* непрерывную по паре *(x, t)* и удовлетворяющую условиям:

$$f(x,\ 0) = f(x),$$

$$f(x,\ 1) = g(x).$$

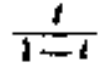
функция *f (x, t)* указанного выше вида является, конечно, ничем иным, как отображением пространства $X \times I$ (произведения пространства *X* на единичный отрезок I) в Y. Интуитивный смысл гомотопии состоит в том, что *g* может быть получено из *f* процессом непрерывной деформации, все фазы которой являются отображениями в *Y.*
Ясно, что соотношение гомотопности симметрично и транзитивно, т. е. если *f* гомотопно *g,* то *g* гомотопно *f*, и если *f* гомотопно *g,* а *g* гомотопно *h*, то *f* гомотопно *h*. Таким образом, возникает разбиение всех отображений пространства *X* в Y на непересекающиеся *гомотопические классы.*
Определение 4. Отображение пространства *X* в пространство *Y* называется *несущественным,* если оно гомотопно постоянному отображению. Отображение, не являющееся несущественным, называется *существенным.*
Пример 5. Пусть и *X,* и *Y*—*n*-мерные сферы S_n, и *f*—тождественное отображение. Тогда, как показывает предложение 1 А), *f* существенно, и этот факт послужил основанием для доказательства того, что E_n

n-мерно. Известно, что существует счетное множество гомотопических классов отображений сферы S_n в себя . Каждый из этих классов характеризуется целым числом, называемым его степенью; грубо говоря, степень — это алгебраическое число навертываний сферы на себя при отображении.

Пример 6. Если пространство Y обладает тем свойством, что каждая, пара точек может быть соединена в нем простой дугой, то для любого пространства *X* все несущественные отображения пространства *X* в *Y* принадлежат одному и тому же гомотопическому классу, так как из указанного свойства, очевидно, следует, что все постоянные отображения гомотопны.

Пример 7. Мы скажем, что *X — стягиваемое* пространство, если тождественное отображение X на X несущественно. Интуитивно это означает, что *X* можно стянуть по себе в точку. E_n и I_n являются примерами стягиваемых пространств; S_n нестягиваемо (см. пример 5). Легко показать, что если хотя бы одно из пространств *X и Y* является стягиваемым, то все отображения пространства *X в Y* несущественны. В частности, всякое отображение куба I_n в произвольное пространство несущественно.

Пример 8. Пусть *f* и *g* — два отображения произвольного пространства *X* в S_n такие, что для любого $x \in X$ точки *f(x)* и *g(x)* находятся друг от друга на расстоянии меньшем, чем диаметр сферы S_n, т. е. никогда не являются диаметрально противоположными. Тогда *f* и *g* гомотопны; ибо всегда существует однозначно определенная меньшая дуга большого круга, соединяющая *f(x)* и *g(x),* и мы можем в качестве



f(x, t) взять точку, делящую эту дугу в отношении

Как следствие получаем: любое отображение *f* пространства *X* в S_n , оставляющее некоторую точку $s \in S_n$ свободной от *f (X),* несущественно, потому что, как только что было установлено, *f* гомотопно постоянному отображению *g* пространства *X* в точку, диаметрально противоположную точке *q.*

Начиная с этого момента, мы будем рассматривать, главным образом, отображения в S_n. Мы возвратимся к вопросу о продолжении отображения, определенного на замкнутом подмножестве пространства, в отображение, определенное на всем пространстве. Оказывается, что существование такого отображения зависит только от гомотопического класса данного отображения,— факт, имеющий много очень важных приложений.

Теорема 5. Теорема Борсука. *Пусть С- замкнутое подмножество пространства X, а f и g — два гомотопных отображения множества*

C в S_n. Тогда, если существует продолжение F отображения f на X, то существует также продолжение G отображения g на X, гомотопное F.

(Заметим, что единственное свойство сферы S_n , используемое в доказательстве, это свойство быть абсолютным окрестностным ретрактом. Поэтому теорема Борсука останется справедливой, если S_n заменить произвольным абсолютным окрестностным ретрактом. Приведенное здесь доказательство принадлежит С. Г. Доукеру).

Предпошлем доказательству предложение относительно открытых множеств в топологическом произведении пространств.

А) Пусть C — подмножество пространства X, а U — открытое множество топологического произведения $X \times I$, содержащее $C \times I$ (I обозначает единичный отрезок [0, 1].

Тогда существует открытое множество V в X такое, что $C \subset V$ и $V \times I \subset U$.

Доказате льство. Сначала мы покажем, что каждая точка $c \in C$ имеет в X окрестность v, для которой $v \times I \subset U$. Рассмотрим отрезок $c \times I$. Он содержится в U. Следовательно, каждая его точка имеет содержащуюся в V окрестность вида $w \times i$, где w—окрестность точки c в X и i — интервал в I. Так как $c \times I$ — компакт, то конечное число этих «прямоугольных» окрестностей покрывает $c \times I$. В качестве окрестности v берем тогда пересечение проекций всех этих окрестностей в X. Предложение А) будет теперь установлено, если в качестве окрестности V взять сумму окрестностей v по всем точкам $c \in C$.

Возвратимся теперь к доказательству теоремы.

Доказательство теоремы Борсука. Гомотопность отображения f и g означает, что существует отображение $f(x, t)$ произведения $C \times I$ в S_n , удовлетворяющее условиям:

$$f(x, 0) = f(x),$$

$$f(x, 1) = g(x),$$

где $x \in C$. По предположению, существует также отображение F пространства X в S_n, совпадающее с f на C. Пусть C' — множество в $C \times I$, состоящее из точек $(x, 0)$ для $x \in X$ и точек (x, t) для $x \in C$ и $0 \leqslant t \leqslant 1$. C является замкнутым подмножеством пространства $X \times I$. Рассмотрим следующее отображение множества C' в S_n:

$$F(x, 0) = F(x) \quad \text{для} \quad x \in X,$$

$$F(x, t) = f(x, t) \quad \text{для} \quad x \in C \quad 0 \leqslant t \leqslant 1.$$

По следствию 2 теоремы 3 существует открытое в $X \times I$ множество $U \supset C'$, на которое $F(x, t)$ может быть продолжено;

продолженное отображение мы будем попрежнему обозначать

через $F(x, t)$. В силу А), существует открытое в X множество V, содержащее С, такое, что $V \times I \subset U$. Заметим, что отображение $F(x, t)$ определено для любого $x \in V$ и $0 \leqslant t \leqslant 1$ и, кроме того, для любого $x \in X$ и $t = 0$.

С и дополнение множества V являются непересекающимися замкнутыми подмножествами пространства X. Следовательно, существует непрерывная действительная функция $p(x)$, определенная[1]) на X, принимающая значения между 0 и 1, равная 1 на С и 0 на $X \setminus V$.

Например, $p(x) = \dfrac{\rho(x, X \setminus V)}{\rho(x, X \setminus V) + \rho(x, C)}$.

Рассмотрим теперь функцию

$$G(x, t) = F(x, tp(x)).$$

$G(x, t)$ определена для всех $x \in X$ и $0 \leqslant t \leqslant 1$ и непрерывна по (x, t). Если мы определим $G(x)$ равенством

$$G(x) = G(x, 1),$$

то ясно, что

$$G(x) = g(x) \quad \text{для} \quad x \in C,$$

так что $G(x)$ является продолжением отображения $g(x)$ на X. Также ясно, что $G(x, 0) = F(x)$; так как $G(x, 1) = G(x)$ по определению, то F и G гомотопны.

Следствие. *Несущественное отображение замкнутого подмножества пространства X в S_n всегда может быть продолжено на X.*

Доказательство. В самом деле, постоянное отображение всегда может быть продолжено.

Теперь мы установим связь между размерностью и гомотопией.

В) Пусть f и g — два отображения пространства X в S_n такие, что точки, в которых $f(x)$ не равно $g(x)$, образуют множества D размерности $\leq n$—1. Тогда f и g гомотопны.

Доказательство. D, очевидно, открыто. Пусть D^* — замкнутое множество произведения $X \times I$, состоящее из точек $(x, 0)$ и $(x, 1)$ для $x \in X$ и точек (x, t) для $x \in X \setminus D$ и $0 \leqslant t \leqslant 1$. Следующим образом определяем отображение $F(x, t)$ множества D^* в S_n:

$$F(x, t) = f(x) = g(x) \text{ для } x \in X \setminus D,$$

$$F(x, 0) = f(x), \; F(x, 1) = g(x).$$

Дополнение множества D^* содержится в $D \times I$, и, в силу известной теоремы , $\dim D \times I \leqslant n$. По следствию известной теоремы , $F(x, t)$ можно продолжить в отображение пространства $X \times I$ в S_n, доказав этим гомотопность a и g.

Теорема 6. *Если X—пространство размерности меньшей чем n, то все отображения X в S_n гомотопны и, следовательно, несущественны.*
Доказательство. Теорема есть непосредственное следствие предложения В).
Соединяя предложение В) с теоремой Борсука, получаем следующий результат, утверждающий, что если отображения, определенные на двух замкнутых частях пространства как бы «приспособлены» друг к другу всюду, за исключением, быть может, множества меньшей размерности, то любое из отображений может быть изменено таким образом, чтобы оно стало «приспособлено» к другому полностью.
С) Допустим, что пространство *X* является суммой двух замкнутых подмножеств C_1 и C_2. Пусть F_1 и F_2 — отображения множеств C_1 и C_2 в S_n. Пусть, далее, точки (в $C_1 \cap C_2$), для которых $F_1(x)$ не равно $F_2(x)$, образуют множество размерности $\leq n$ — 1. Тогда F_1 может быть продолжено на *X*.
Доказательство. Частичные отображения $F_1|C_1 \cap C_2$ и $F_2|C_1 \cap C_2$ отличаются друг от друга только на множестве размерности $\leq n$ —1 и, следовательно в силу В), гомотопны. Так как $F_2|C_1 \cap C_2$ может быть продолжено на C_2, именно в F_2, то из теоремы Борсука следует, что существует продолжение, скажем F'_1, отображения $F_1|C_1 \cap C_2$ на C_2. Полагая

$$F(x) = F_1(x) \text{ для } x \in C_1,$$

$$F(x) = F'_1(x) \text{ для } x \in C_2,$$

получаем искомое продолжение.
Теперь мы установим несколько результатов.
D) Пусть f и g —отображения пространства X в S_n. Допустим, что *X* является суммой двух замкнутых подмножеств C_1 и C_2, пересечение которых имеет размерность $\leq n$ — 2. Если f и g гомотопны на каждом из множеств C_1 и C_2, т. е., если $f|C_1$ гомотопногомотопно $g|C_1$, и $f|C_2\, g|C_2$, то f и g гомотопны.
Доказательство. Рассмотрим топологическое произведение $X \times I$. По предположению, существуют отображения:

$f_1(x, t)$, определенное для $x \in C_1$, $0 \leqslant t \leqslant 1$,

$f_2(x, t)$, определенное для $x \in C_2$, $0 \leqslant t \leqslant 1$,

такие, что

для $x \in C_1$: $f_1(x, 0) = f(x)$, $f_1(x, 1) = g(x)$,

для $x \in C_2$: $f_2(x, 0) = f(x)$, $f_2(x, 1) = g(x)$.

Расширим область определения отображения $f_1(x, t)$, положив,

для $x \in C_2$: $f_1(x, 0) = f(x)$, $f_1(x, 1) = g(x)$.

Каждое из отображений $f_1(x, t)$ и $f_2(x, t)$ определено на замкнутом подмножестве пространства $X \times I$, причем точки, в которых $f_1(x, t)$ и $f_2(x, t)$ определены, но имеют различные значения, принадлежат множеству $(C_1 \cap C_2) \times I$ и, следовательно, образуют множество размерности $\leq n$ —1. Предложение С) показывает тогда, что существует продолжение $F(x, f)$ отображения $f_1(x, f)$ на все $X \times I$. Мы имеем:

$$F(x, 0) = f_1(x, 0) = f(x),$$

$$F(x, 1) = f_1(x, 1) = g(x),$$

откуда следует, что f и g гомотопны.

E) Пусть f—отображение пространства X в S_n. Допустим, что X является суммой двух замкнутых подмножеств C_1 и C_2, пересечение которых имеет размерность $\leq n - 2$. Если f несущественно на каждом из множеств C_1 и C_2, т. е. если $f|C_1$ и $f|C_2$ несущественны, то f несущественно.

Доказательство. Предложение есть частный случай предложения D), когда g —постоянное отображение.

F) Пусть C—замкнутое подмножество пространства X, и $\{V_\lambda\}$ — совокупность открытых множеств, в сумме дающих X, границы которых имеют размерность $\leq n$—1. Если f есть отображение множества C в S_n, допускающее продолжение на каждое из множеств $C \cup \overline{V}_\lambda$, то f можно продолжить на всё пространство X.

Доказательство. Так как X обладает счетным базисом, то можно предположить, что $\{V_\lambda\}$ является счетной совокупностью V_1, V_2..... Определим последовательные продолжения отображения f, взяв в качестве F_1 произвольное продолжение отображения f на $C \cup \overline{V}_1$. Предположим, что мы уже определили F_k как продолжение отображения F_{k-1}, на множество $C \cup \overline{V}_1 \cup \ldots \cup \overline{V}_k$. Разделим $C \cup \overline{V}_1 \cup \ldots \cup \overline{V}_{k+1}$ на две замкнутые части:

$$C_1 = C \cup \overline{V}_1 \cup \ldots \cup \overline{V}_k$$

и

$$C_2 = (C \cup \overline{V}_{k+1}) \setminus (V_1 \cup \ldots \cup V_k).$$

По предположению, f допускает продолжение F^k на C_2. Точки x, для которых $F_k(x) \neq F^k(x)$, принадлежат множеству

$$(C_1 \cap C_2) \setminus C \subset Fr(V_1 \cup \ldots \cup V_k) \subset Fr\, V_1 \cup \ldots \cup Fr\, V_k,$$

которое, в силу теоремы сложения , имеет размерность $\leq n$—1. Следовательно, можно применить С) и получить распространение F_{k+1} отображения F_k на $C \cup \overline{V}_1 \cup \ldots \cup \overline{V}_{k+1}$.

Пусть теперь x — произвольная точка пространства X, и m — первый номер такой, что $x \in V_m$. Полагая $F(x) = F_m(x) =$

$= F_{m+1}(x) = F_{m+2}(x) = \ldots$, получаем отображение F пространства X в S_n, которое, очевидно, является продолжением отображения f.

G) Пусть пространство X является суммой (не обязательно счетной) семейства замкнутых множеств $\{K_\lambda\}$, обладающих следующими двумя свойствами: каждое K_λ имеет размерность $\leq n$, и если даны произвольное K_λ и открытое множество U, содержащее K_λ, то найдется открытое множество V такое, что

$$K_\lambda \subset V \subset U$$

и $\dim Fr\, V \leqslant n-1$.

Тогда X имеет размерность $\leq n$.

Доказательство. В силу теоремы 4, для того чтобы доказать, что X имеет размерность $\leq n$, достаточно показать, что если дано произвольное замкнутое множество $C \subset X$ и произвольное отображение f множества C в S_n, то существует продолжение отображения f на X. Из того обстоятельства, что $\dim K_\lambda \leq n$, вытекает, что для каждого $K \in \{K_\lambda\}$ отображение f можно продолжить на $C \cup K$ (положим в следствии теоремы 4 $X = C \cup K$, тогда $X \setminus (C \cup K)$ и (следствие 2 теоремы 3) что можно даже продолжить f на некоторое открытое множество

$$U \supset C \cup K, \quad (1)\ .$$

Из предположения следует, что найдется открытое множество V такое, что

$$K \subset V \subset \overline{V} \subset U \quad (2)$$

и

$$\dim Fr\, V \leqslant n-1. \quad (3)$$

Так как, по (1) и (2), $C \cup \overline{V} \subset U$, то ясно, что f можно продолжть на $C \cup \overline{V}$. В силу (2), сумма множеств V заполняет X; предложение F) доказывает возможность продолжения f на X, что завершает доказательство.

2.2.4. Отображения, понижающие размерность

Непрерывность однозначного отображения f пространства X в пространство Y можно определить с помощью требования, чтобы обратное отображение f^{-1} переводило замкнутые множества пространства Y *в* замкнутые множества пространства X или, что то же (так как f определено на всем X), открытые множества пространства Y— в открытые множества пространства X. Это наводит нас на мысль говорить, что f^{-1} (многозначное отображение) непрерывно, если само f переводит замкнутые множества пространства X в замкнутые

множества пространства *Y*. Мы могли бы, по этой причине, прийти к решению называть непрерывное отображение *f* в обе стороны непрерывным, если *f* отображает замкнутые множества пространства *X* в замкнутые множества пространства *Y*. Однако против этого имеется серьезное возражение, состоящее в том, чго в такой же степени разумно говорить, что f^{-1} непрерывно, если *f* переводит открытые множества пространства *X* в открытые множества пространства Y*;* к сожалению, два возможных способа определения непрерывности отображения f^{-1} не совпадают , за исключением случая взаимно-однозначных отображений *X* на Y. Следовательно, лучше избегать неясного выражения: «в обе стороны непрерывное» и говорить лишь о замкнутых и открытых отображениях, как они определены ниже.

Определение 5. *Замкнутым (открытым)* отображением пространства *X* в пространство Y называется отображение, переводящее замкнутые (открытые) подмножества пространства *X* в замкнутые (открытые) подмножества пространства Y.

Пример 9. Пусть *X*—интервал $0<x<1$, и *Y* — отрезок $0 \leqslant x \leqslant 1$. Пусть *f*— тождественное отображение $f(x) = x$ пространства *X* в *Y*. Тогда *f* не замкнуто, так как *X,* которое, конечно, замкнуто в *X,* не замкнуто в Y. Легко однако, видеть, что *f* открыто.

Пример 9. 1. Пусть *X* — плоское множество, состоящее из вертикальной и горизонтальной осей, а Y—горизонтальная ось. Пусть *f*—вертикальная проекция *X* в *Y*. Тогда *f* замкнуто, но не открыто.

Замечание. Замкнутое подмножество компакта является компактом, и непрерывный образ компакта также является компактом . Далее, компакт замкнут во всяком содержащем его пространстве; следовательно, каждое отображение компакта замкнуто.

Рассмотрим ортогональную проекцию куба I_{n+k} на I_n ; это отображение понижает размерность на *k* единиц. Полный прообраз каждой точки куба I_n имеет размерность *k.* Теперь мы докажем, что, вообще, непрерывное замкнутое отображение одного пространства в другое не может понижать размерность на *k* единиц без того, чтобы, по крайней мере, одно *k*-мерное множество не сплющивалось в точку.

Теорема *7*. *Пусть f—непрерывное замкнутое отображение пространства X в пространстве Y, и*

$$\dim X \setminus \dim Y = k,$$

$k > 0$. *Тогда существует точка пространства Y, полный прообраз которой имеет размерность* $\geq k$.

Доказательство. Для удобства доказательства представим утверждение теоремы в следующей форме: если для каждого $y \in Y$

$$\dim f^{-1}(y) \leqslant m,$$

то $\dim X \leqslant m + \dim Y.$

Очевидно, можно предположить, что *Y* конечномерно. Мы докажем теорему индукцией по размерности пространства Y, сохраняя *m* фиксированным. Утверждение тривиально, если dim Y= -1, потому что в этом случае *X* также пусто. Предположим теперь, что утверждение справедливо при dim Y ≤ *n* — 1, и докажем его справедливость при dim *Y* = *n.*

Для того чтобы доказать, что $\dim X \leqslant m + n$, достаточно показать, что семейство замкнутых множеств $f^{-1}(y)$ удовлетворяет условиям предложения 3 G), где *n* заменено на *n*+-*m.* По предположению, мы имеем:

$$\dim f^{-1}(y) \leqslant m \leqslant m + n.$$

Далее, если *U*— открытое множество в *X,* содержащее $f^{-1}(y)$, то множество

$$C = f(X \setminus U)$$

замкнуто в Y, как образ замкнутого множества *X\U* при замкнутом отображении *f*. Это множество С не содержит точки *y.* Так как, в силу предположения, dim Y ≤ *n* , то существует окрестность *V* точки *y* в *Y* такая, что

$$C \cap V = 0, \quad (1)$$

$$\dim Fr\, V \leqslant n - 1. \quad (2)$$

Из (1) следует, что

$$f^{-1}(C) \cap f^{-1}(V) = 0,$$

и потому $f^{-1}(V) \subset U.$

Множество $f^{-1}(V)$, содержащее $f^{-1}(y)$, является, как полный прообраз открытого множества *V* при непрерывном отображении, открытым множеством. Пусть *B* — граница этого множества. Легко видеть, что *f*(B) содержится в границе множества *V.* Следовательно в силу (2), $\dim f(B) \leqslant n - 1.$ Применяя индуктивное предположение к отображению множества *B* на *f(B),* заключаем, что $\dim B \leqslant m + n - 1$;это показывает, что условия предложения 3 G) выполнены.

Пример 10. Возвратимся к вполне несвязному пространству Кнастера и Куратовского, которое становится связным после прибавления одной точки *a.* Пользуясь приведенными ранее обозначениями, рассмотрим непрерывное отображение *f* пространства *X\a* на С отображающее каждое L*(*p*) в точку *p* и каждое *L* (q)* в точку *q.* Мы имеем:

$$\dim (X \setminus a) = 1$$

и

$$\dim f(X \setminus a) = \dim \mathcal{C} = 0$$

так что *f* понижает размерность на единицу. Тем не менее, полный прообраз каждой точки множества С нульмерен. Легко видеть, что отображение *f* открыто. Здесь нет противоречия с теоремой 17, ибо *f* не замкнуто.

Замечание. Существует двойственная в некотором отношении теореме 7 теорема о повышающих размерность отображениях: пусть *f*— замкнутое непрерывное отображение пространства *X* на пространство Y, и пусть

$$\dim Y - \dim X = k, \quad k > 0.$$

Тогда существует хотя бы одна точка пространства Y, полный прообраз которой содержит, по крайней мере, *k+1* точку.

2.2.5. Канторовы многообразия

Определение 6. *n*-мерный компакт, при $n \geq 1$, называется *n-мерным канторовым многообразием*, если он не может быть разбит подмножеством размерности $\leq n - 2$.

Пример 11. Из следствий известной теоремы вытекает, что I_n, а также S_n и — обще́е — произвольное *замкнутое* *n*-мерное многообразие являются *n*-мерными канторовыми многообразиями.

А) Легко видеть, что канторово многообразие связно и что *n*-мерное канторово многообразие имеет размерность *n* в каждой своей точке.

В) *n*-мерный компакт С является *n*-мерным канторовым многообразием в том и только в том случае, если невозможно разложение

$$C = C_1 \cup C_2,$$

где C_1 и C_2 — замкнутые подмножества С, а

$$\dim (C_1 \cap C_2) \leqslant n - 2.$$

Основной теоремой о канторовых многообразиях является следующая.

Теорема 8. *Любой n-мерный компакт X содержит n-мерное канторово многообразие.*

Сначала мы установим:

С) Пусть дан компакт *X*, замкнутое множество $\mathcal{C} \subset X$ и отображение *f* множества *C* в S_n, которое не может быть продолжено на *X*. Тогда в *X* существует замкнутое множество *K* такое, что (1) *f* не может быть продолжено на $\mathcal{C} \cup K$, но (2) если *K* — собственное замкнутое подмножество множества *K*, то *f* можно продолжить на $\mathcal{C} \cup K'$.

Доказательство. Рассмотрим семейство $\{K_\lambda\}$ замкнутых множеств, таких, что (3) *f* не может быть продолжено на $\mathcal{C} \cup K_\lambda$. $\{K_\lambda\}$ не пусто, так как

оно содержит *X*. Если *K^0 является пересечением монотонно убывающей последовательности* $\{K_i\}, i = 1, 2, \ldots$ *замкнутых множеств, принадлежащих* $\{K_\lambda\}$, *то* K^0 *также принадлежит* $\{K_\lambda\}$.
Действительно, допустим, что K^0 не принадлежит $\{K_\lambda\}$, т. е. что f можно продолжить на $C \cup K^0$. Тогда, по следствию 2 теоремы 3, существует открытое множество U, содержащее $C \cup K^0$, на которое f можно продолжить. Но, по крайней мере, одно из K_i, скажем K_{i_0}, содержится в U; в противном случае все K_i пересекались бы с дополнением T множества U, что невозможно, так как в этом случае из компактности T следовало бы, что само K^0 пересекается с T. Но из того, что $K_{i_0} \subset U$ вытекает, что f можно продолжить на $C \cup K_{i_0}$, а это противоречит свойству (3).
Мы видим, что семейство $\{K_\lambda\}$ удовлетворяет предположениям теоремы Брауэра. Применяя эту теорему, получаем замкнутое множество K, неприводимое по отношению к свойству (3), т. е. замкнутое множество, удовлетворяющее условиям (1) и (2).
Доказательство теоремы 8. Так как dim X = n, то, в силу теоремы 4, существует замкнутое подмножество C пространства X и отображение f множества C в S_{n-1}, которое не может быть продолжено на X. В силу C), в X существует замкнутое множество K, удовлетворяющее условиям (1) и (2). Мы утверждаем, что K есть n-мерное канторово многообразие. Ибо, в противном случае (см. B)),

$$K = C_1 \cup C_2,$$

где C_1 и C_2 — замкнутые подмножества K, и

$$\dim C_1 \cap C_2 \leqslant n - 2.$$

Из (2) следует, что f можно продолжить в отображения F_1 и F_2, определенные соответственно, на $C \cup C_1$ и $C \cup C_2$. В силу 3 C), F_1 можно продолжить на $C \cup K$ и, следовательно, f можно продолжить на $C \cup K$, в противоречии с условием (1).
Следствие. *Пусть X—n-мерный компакт, и A— его подмножество, состоящее из всех точек, в которых X имеет размерность n. Тогда* dim $A = n$.
Доказательство. Пусть C—n-мерное канторово многообразие, содержащееся в X, Оно существует по теореме 8. Из предложения A) мы знаем, что $C \subset A$. Следовательно,
dim $A = n$.

2.2.6. Инвариантность области в E_n.

Пусть *X*—подмножество произвольного пространства *A*. Допустим, что *h* есть гомеоморфизм, отображающий все *A* в *A*. Если *x*—-внутренняя точка множества *X*, то *h(x)* — внутренняя точка множества *h*(X), и обратно. Предположим, однако, что гомеоморфизм *h* определен только на *X*.

Отнюдь неверно, что и в этом случае *h* переводит внутренние точки во внутренние и граничные точки в граничные.

Пример 12. Пусть *A*— подмножество пространства E_3, состоящее из плоскости (x_1, x_2) и оси x_3, *X*—подмножество» множества *A*, состоящее из оси x_3, и А — гомеоморфизм, отображающий ось x_3 на ось x_1. Тогда точка (0, 0, 1) является внутренней точкой множества *X*, но ее образ не является внутренней точкой множества *h (X)*, так как в *A* нет никакого открытого множества, содержащего этот образ и содержащегося в *h (X)*.

Однако, если *A* — эвклидово пространство, то имеет место

Теорема 9. Теорема Брауэра об инвариантности области. *Пусть X— произвольное подмножество E_n, и h — гомеоморфизм множества X на другое подмножество пространства E_n. Тогда если x — внутренняя точка множества X, то h (x) — внутренняя точка множества h (X), и если x — граничная точка множества X, то n(x) — граничная точка множества n(X). В частности, если A и B — гомеоморфные подмножестза пространства E_n, и A открыто, то и B открыто.*

Доказательство. Мы докажем теорему 9, охарактеризовав внутренние точки множества *X*, или, что то же самое, граничные точки множества *X*, внутренними топологическими свойствами множества *X*, т. е. топологическими свойствами, независящими от пространства, в котором множество *X* расположено.

А) Пусть *X* — произвольное подмножество пространства E_n, и $x \in X$. Тогда *x* является граничной точкой множества *X*, $x \in (\overline{E_n \setminus X}) \cap X$, в том и только в том случае, если *x* обладает произвольно малыми окрестностями *V* в *X*, (т. е. подмножествами множества *X*, открытыми в X и содержащими *x)* обладающими тем свойством, что любое непрерывное отображение множества $X \setminus U$ в S_{n-1} можно продолжить на *X* (относительно S_{n-1}).

Доказательство. *Условие необходимо.* Пусть *x*— граничная точка множества X, *S(x)*—сферическая окрестность точки *x* в E_n, и $U = X \cap S(x)$. Мы покажем, что *U* обладает нужным свойством. Пусть *B* обозначает (*n*— 1)-мерную сферу, являющуюся границей окрестности *S(x)*. По следствию теоремы 4, любое непрерывное отображение *f* множества $X \setminus U$ в S_{n-1} можно продолжить в непрерывное отображение f', определенное на $(X \setminus U) \cup B$. Пусть *q* —

точка S *(x),* не принадлежащая *X*. Для каждой точки $x \in X$ обозначим через *x′* проекцию точки *x* из *q* на *B.* Теперь положим:

$F(x) = f'(x')$ для $x \in U$,

$F(x) = f(x)$ для $x \in X \setminus U$.

F(x) является искомым продолжением.

Условие достаточно. Действительно, пусть *x* — внутренняя точка множества *X*. Пусть *S(x)* — сферическая окрестность точки *x,* замыкание которой содержится в *X*. Мы покажем, что какова бы ни была окрестность U точки *x,* содержащаяся в *S(x),* существует непрерывное отображение множества $X \setminus U$ в S_{n-1}, которое не может быть продолжено на *X*. Отождествим S_{n-1} с границей окрестности *S(x)* и в качестве отображения *f* возьмем проекцию множества $X \setminus U$ из точки *x* на S_{n-1}. Тогда *f* не может быть продолжено на *X,* потому что такое продолжение отображало бы замыкание множества *S(x)* на его границу, оставляя точки границы неподвижными, а это противоречит предложению IV 1 В). Таким образом, доказательство предложения А), а следовательно, и теоремы Брауэра об инвариантности области, закончены.

Следствие. *Теорема* 9 *останется справедливой, если* E_n *заменить произвольным многообразием.*

Доказательство. Следствие вытекает из того, что каждая точка множества *X* имеет в *n*-мерном многообразии окрестность, гомеоморфную E_n.

2.2.7. Разбивающие множества в E_n

Начиная с этого момента мы предположим, что $n \geq 2$, и приступим к изложению теорем о разбиении пространства E_n. Мы будем пользоваться следующими обозначениями. Пусть S_{n-1} $(n-1)$-мерная сфера в E_n радиуса 1 с центром в начале координат 0. Для каждой точки $p \in E_n$ мы обозначаем через π_p отображение множества $E_p \setminus p$ в S_{n-1}, определенное следующим образом: $\pi_p(x)$, $x \in E_p \setminus p$, есть проекция точки $x - p$ (в векторных обозначениях^ из точки 0 на S_{n-1}.

А) Пусть *U*— ограниченное открытое множество в E_n, *p* — точка множества *U,* и С— граница множества *U*. Тогда частичное отображение $\pi_p|C$ нельзя продолжить на $U \cup C = \overline{U}$.

Доказательство. Не уменьшая общности, можно предположить, что *p* — начало координат. Пусть *r* так велико, что $U \cup C$ содержится в шаре S(0, *r)* радиуса *r* с центром в 0. Допустим, что $\pi_0|C$ можно продолжить на $U \cup C$, скажем, в отображение φ. Тогда формулы

$$\psi(x) = \varphi(x) \quad \text{для} \quad x \in U,$$

$$\psi(x) = \pi_0(x) \quad \text{для} \quad x \in S(0, r) \setminus U$$

определяют непрерывное отображение ψ шара S(0,1) на его границу S_{n-1}. Кроме того, для каждого $x \in S_{n-1}$ мы имеем: $\psi(x) = \pi_0(x) = x$, в противоречии с IV 1 B).

Теорема 10. *Пусть C—компактное (т. е. ограниченное замкнутое) множество в E_n. Две точки p, q, ни одна из которых не принадлежит C, отделены множеством C в том и только в том случае, если отображения $\pi_p|C$ и $\pi_q|C$ принадлежат к различным гомотопическим классам.*

Доказательство. Предположив сначала, что *p* и *q* отделены множеством C, мы докажем, что $\pi_p|C$ и $\pi_q|C$ не гомотопны. Нам дано, что

$$E_n \setminus C = U \cup V,$$

где *U, V* — непересекающиеся множества, открытые в $E_n \setminus C$ и, следовательно, в E_n, а $p \in U$, $q \in V$.

Одно из множеств U , *V* ограничено. Действительно, пусть I_n—*n*-мерный куб, настолько большой, что $C \subset I_n$. Тогда, так как $E_n \setminus I_n \subset E_n \setminus C$ и $E_n \setminus I_n$ связно, то оно должно содержаться или в *U,* или в *V*. Предположим, например, что $E_n \setminus I_n \subset V$. Тогда $U \subset I_n$, т. е. *U* ограничено.

Теперь, $\pi_q|C$ можно продолжить на $U \cup C$ (в действительности на $E_n \setminus q \supset U \cup C$). С другой стороны, так как $Fr\,U \subset C$, то, в силу предложения A), $\pi_q|C$ невозможно продолжить на $U \cup C$. Следовательно, $\pi_p|C$ и $\pi_q|C$ не гомотопны, так как их гомотопность противоречила бы теореме Борсука. Предположим теперь, что *p* и q не отделены множеством C. Тогда, по хорошо известному свойству эвклидовых пространств, *p* и *q* можно соединить в $E_n \setminus C$ непрерывной дугой, т. е. можно найти непрерывную функцию *f(t)* действительного параметра *t*, $0 \leqslant t \leqslant 1$, принимающую значения из $E_n \setminus C$, такую, что

$$f(0) = p, \quad f(1) = q.$$

Функция

$$\pi_{f(t)}(x), \quad x \in C$$

является тогда непрерывной функцией по *(x, t),* показывающей, что $\pi_p|C$ и $\pi_q|C$ гомотопны.

Следствие. *Если точки p и q отделены в E_n множеством $C_1 \cup C_2$, где C_1 и C_2 — компакты, пересечение которых имеет размерность $\leq n-3$, то или C_1 или C_2 отделяет p от q.*

Доказательство. В противном случае $\pi_p|C_1$ было бы гомотопно $\pi_q|C_1$, а $\pi_p|C_2$ было бы гомотопно $\pi_q|C_2$. Следовательно по 3 D), отображения $\pi_p|C_1 \cup C_2$ и $\pi_q|C_1 \cup C_2$ были бы гомотопны, так что, в силу теоремы 10, $C_1 \cup C_2$ не отделяло бы *p* от *q*.

Теорема 11. *Пусть C—компактное подмножество пространства E_n, отделяющее p от q, в то время как никакое собственное замкнутое подмножество множества C не отделяет p от q («неприводимое» отделяющее множество). Тогда C есть* (n — 1)*-мерное канторово многообразие.*

Доказательство. Заметим сначала, что C не содержит никакого непустого открытого подмножества. Потому что в противном случае граница множества C была бы его собственным замкнутым подмножеством, отделяющим *p* от *q*.

Следовательно, $\dim C \leq n-1$. Кроме того,по следствию теоремы 10, если

$$C = C_1 \cup C_2,$$

где каждое из C_1 и C_2 является собственным замкнутым подмножеством множества C, то

$$\dim C_1 \cap C_2 \geqslant n-2.$$

Это показывает (5 B)), что C есть (*m*— 1)-мерное канторово многообразие.

Замечание. Теорему 11 можно также сформулировать следующим обраюм: *Если компактное множество C пространства E_n является общей границей двух непересекающихся открытых связных подмножеств A_1 и A_2, то C есть* (*n*—1)*-мерное канторово многообразие.* Действительно, пусть p_1 и p_2 суть, соответственно, точки множеств A_1 и A_2. Очевидно, что p_1 и p_2 отделены множеством C. Однако они не отделены никаким его собственным замкнутым подмножеством *C′*. Действительно, пусть $q \in C \setminus C'$; обозначим через *U* сферическую окрестность точки *q*, такую, что $U \in E_n \setminus C'$. Тогда *U* содержит точки как множества A_1, так и множества A_2, и, следовательно, $A_1 \cup U \cup A_2$ является связным подмножеством множества $E_n \setminus C'$, содержащим p_1 и p_2; таким образом, p_1 и p_2 не отделены множеством C′.

Теорема 12. *Пусть X— компактное (замкнутое ограниченное) подмножество пространства E_n,* и *C — замкнутое подмножество множества X. Для того чтобы существовало непрерывное отображение f множества C в S_{n-1}, которое не может быть продолжено на X, необходимо и достаточно, чтобы существовало непустое открытое подмножество пространства E_n , содержащееся в X\C, в то время как его граница содержится в C.*

Доказательство теоремы 12. Необходимость. Пусть *f*— отображение множества С в S_{n-1}, которое не может быть продолжено на *X*. Тогда по 5 С существует замкнутое яодмножество $K \subset X$, обладающее следующими свойствами:

f не может быть продолжено на $C \cup K$, но (1)

если *K'* — произвольное собственное замкнутое подмножество множества *K,* то *f* может быть продолжено на $C \cup K'$. (2)

Мы утверждаем теперь, что

$K \setminus C \neq 0$, (a)

$K \setminus C \subset X \setminus C$, (b)

$K \setminus C$ открыто, (c)

$Fr(K \setminus C) \subset C$. (d)

Утверждение а) следует из свойства (1), а b) очевидно. Чтобы доказать с), мы рассмотрим произвольную точку s в *K\C* и покажем, что s является внутренней точкой множества *K\C.* В силу предложения б А), достаточно показать, что для каждой окрестности *U* точки s в E_n такой, что $U \cap C = 0$, можно определить непрерывное отображение множества *K\C\U* в S_{n-1}, которое не может быть продолжено на *K\C.* В качестве такого отображения возьмем произвольное продолжение F отображения *f* на $(C \cup K) \setminus U = C \cup (K \setminus U)$; *F* существует в силу свойства (2), так как $K \setminus U \neq K$. $F | K \setminus C \setminus U$ не может быть продолжено на K\ *C* по той причине, что такое продолжение давало бы продолжение самого *f* на $C \cup K$, в противоречии со свойством (1). После того как с) установлено, *d)* следует немедленно. Необходимость доказана.

Достаточность. Пусть *U*—непустое открытое множество; содержащееся в *X\C,* граница *B* которого содержится в *C, V* ограничено как подмножество компактного множества *X*. Пусть $p \in U$. Тогда $\pi_p | B$ не может быть продолжено на $U \cup B$ (предложение 7 А)); так как $B \subset C$ и $U \cup B \subset X$, то тем более верно, что $\pi_q | C$ не может быть продолжено на *X*.

Теорема 13. *Компактное подмножество С пространства E_n разбивает E_n в том и только в том случае, если имеется существенное отображение множества С в S_{n-1}.*

Доказательство. *Необходимость.* Пусть *С* разбивает E_n. Тогда, конечно, существуют две точки *p* и *q,* отделенные друг от друга множеством *С.* По теореме 10 отображения $\pi_p | C$ и $\pi_q | C$ множества *С* в S_{n-1} принадлежат различным гомотопическим классам. Следовательно, по крайней мере, одно из них существенно.

Достаточность. Пусть *f*— существенное отображение множества *С* в S_{n-1}. Пусть I_n—*n*-мерный куб, настолько большой, что $C \subset I_n$.

Продолжить *f* на I_n невозможно, так как иначе *f* было бы несущественным (пример 7). Следовательно, заменяя в теореме 12 пространство *X* кубом I_n , получаем, что существует непустое открытое в E_n подмножество $U \subset I_n \setminus C$, граница которого содержится в *C*. Ясно, что *U* является непустым собственным подмножеством множества E_n\C, одновременно открытым и замкнутым в E_n*C*. Следовательно, С разбивает E_n.

Следствие 1. *Если компакт С разбивает E_n, то каждое подмножество пространства E_n , гомеоморфное С, также разбивает E_n.*

Доказательство. Следствие вытекает из того, что теорема 13 характеризует компактные множества, разбивающие E_n , их внутренними свойствами.

Следствие 2. Теорема Жордана. *Подмножество пространства E_n , гомеоморфное S_{n-1}, разбивает E_n .*

Доказательство. Это — очевидное приложение следствия 1.

Следствие 3. *Теорема* 13 к *следствия 1 и 2 сохраняются, если пространство E_n заменить сферой S_n.*

Доказательство. Легко видеть, что если *C* разбивает S_n, то оно разбивает S_n для любой точки $p \notin C$, » наоборот: но S_n гомеоморфно E_n. Можно дать формулировку теоремы 13, если воспользоваться понятием пространства отображений.

Теорема 14. *Если С — компактное подмножество пространства E_n, то E_n\C связно в том и только в том случае, если связно пространство отображений S^C_{n-1}.*

Доказательство. Достаточно заметить, что компоненты пространства S^C_{n-1} совпадают с гомотопическими классами отображений множества *C* в S_{n-1} так как, с одной стороны, каждый гомотопический класс связен, так как, по самому его определению, любые два элемента в нем можно даже соединить простой дугой, а, с другой стороны, каждый гомотопический класс одновременно открыт (это следует из примера 8) и замкнут.

2.3. Размерность и мера

Этот раздел посвящен выяснению установленной Шпильрайном связи между понятием размерности и понятием меры.

p-мерная мера для каждого неотрицательного действительного числа *p* была определена Хаусдорфон для произвольных метрических пространств. **Эта мера тесно связана с обычной мерой Лебега. Она является метрическим понятием в то время, как размерность**

является понятием чисто *топологическим*. Тем не менее между этими двумя понятиями существует тесная связь, так как оказывается (теорема 2), что пространство размерности *n* должно иметь положительную *n*-мериую меру. Обратное, однако, неверно (см. пример 1). Но если рассмотреть не только само метрическое пространство *X*, но *X* вместе со всеми метриками, которые можно в нем ввести, или, что то же, класс всех метрических пространств, гомеоморфных *X*, тогда, если все эти пространства имеют положительную *n*-мерную меру, то само *X* должно иметь размерность $\geq n$. Основным результатом (доказанным теоремами 2 и 4) является:

Теорема 1. *Для того чтобы пространство X имело размерность* $\leq n$, *необходимо и достаточно, чтобы X было гомеоморфно подмножеству куба* I_{2n+1}, $(n+1)$-*мерная мера которого равна нулю.*

Пример 1. И множество $\mathcal{G}$ иррациональных точек единичного отрезка, и канторово множество $\mathcal{C}$ имеют размерность нуль, хотя $\mathcal{G}$ имеет (линейную) меру единицы, а $\mathcal{C}$—меру нуль. Но так как $\mathcal{G}$ топологически содержится в $\mathcal{C}$, существует топологический образ $\mathcal{G}$, мера которого равна нулю.

(Каждый элемент множества *G* имеет однозначное представление в виде суммы $x = \sum_{n=1}^{\infty} \frac{a_n}{2^n}$ (двоичных дробей). Легко видеть, что $h(x) = \sum_{n=1}^{\infty} \frac{(2a_n)}{3^n}$ есть гомеоморфное отображение множества $\mathcal{C}$ в множество $\mathcal{C}$.)

2.3.1. *p*- мерная мера в общих метрических пространствах

Определение 1. Пусть *X*—пространство, и *p*— произвольное действительное число, $0 \leqslant p < \infty$. Для данного $\varepsilon > 0$ пусть

$$m_p^{\varepsilon} = \inf \sum_{i=1}^{\infty} [\delta(A_i)]^p,$$

где $X = A_1 \cup A_2 \cup \ldots$—произвольное разложение пространства *X* на счетное число подмножеств диаметра, меньшего, чем ε, и *p*— показатель степени.

Положим

$$m_p(X) = \sup_{\varepsilon > 0} m_p^{\varepsilon}(X);$$

$m_{p\varepsilon}(X)$ называется *p*-мерной мерой пространства *X*.

Проверка следующих трех предложений предоставляется читателю.

A) Из δ (A), где A — произвольное подмножество X, обозначает диаметр множества A, условимся считать, что $[\delta(E)]^0 = 0$, если B—пусто и $[\delta(E)]^0 = 1$ в противном случае, следует, что
$m_0(X) = 0$, если X пусто,
$m_0(X) = n$, если X—конечное множество из и точек;
$m_0(X) = \infty$, если X—бесконечное множество.
B) Если $p < q$, то $m_p(X) \geqslant m_q(X)$; в действительности, из $p < q$ и $m_p(X) < \infty$ следует, что $m_q(X) = 0$.
C) n-мерный полиэдр имеет конечную n-мерную меру. Следовательно, его q-мерная мера равна нулю для всех $q > n$.
D) Для того чтобы компакт C имел p-мерную меру, равную нулю, необходимо и достаточно, чтобы для каждого $\varepsilon > 0$ существовало *конечное* разложение компакта C:

$$C = A_1 \cup \ldots \cup A_k,$$

такое, что

$$[\delta(A_1)]^p + \ldots + [\delta(A_k)]^p < \varepsilon. \quad (1)$$

Доказательство. Достаточность условия очевидна. Приступим к доказательству необходимости. Допустим, что $m_p(C) = 0$. По определению 1 существует счетное число подмножеств $A'_1, A'_2, \ldots$ таких, что

$$C = A'_1 \cup A'_2 \cup \ldots$$

и

$$\sum_{i=1}^{\infty} [\delta(A'_i)]^p < \frac{1}{2}\varepsilon. \quad (2)$$

Каждое A'_i можно слегка увеличить до открытого множества A_i так, чтобы

$$[\delta(A_i)]^p < [\delta(A'_i)]^p + \frac{\varepsilon}{2^i + 1}. \quad (3)$$

Так как C — компакт, то конечное число $A_1, \ldots, A_k$ множеств A_i покрывает C. Формула (1) следует тогда из формул (2) и (3).
E) Для подмножеав прямой одномерная мера совпадает с внешней мерой Лебега. Однако n-мерная мера подмножества пространства E_n может количественно отличаться от его внешней меры Лебега. Тем не менее равенство нулю n-мерной меры подмножества пространства E_n эквивалентно равенству нулю его внешней меры Лебега.

2.3.2. *n*-мерное пространство имеет положительную *n*-мерную меру

Теорема 2. *Пусть X — пространство размерности* n, $0 \leqslant n < \infty$. *Тогда* $m_n(X) > 0$.

Эта теорема эквивалентна (при замене n на $n+1$) следующей теореме:

Теорема 3. *Пусть X—пространство такое, что* $m_{n+1}(X) = 0 \quad (0 \leqslant n < \infty)$. *Тогда* $\dim X \leq n$.

Доказательство теоремы 3. Пусть x_0 — произвольная точка пространства X. Для каждого $r > 0$ обозначим через $S'(r)$ границу сферической окрестности точки x радиуса m и через $S(r)$ — множество всех точек $x \in X$, для которых $\rho(x, x_0) = r$. Очевидно, $S'(r) \subset S(r)$.

Выражение «для почти всех n» будет означать: для всех r, исключая множество лебеговой меры нуль. Тогда, если принять во внимание, что множество нульмерной меры нуль пусто (1 А)), то теорема 3 получится по индукции, как только будет доказано следующее предложение:

А) Если X—пространство такое, чго $m_{p+1}(X) = 0$, $0 \leqslant p < \infty$, то для почти всех r множество S (r) имеет p-мерную меру нуль.

Доказательство предложения А). Так как $m_{p+1}(X) = 0$, то существует последовательность разложений пространства X:

$$X = A_1^n \cup A_2^n \cup \ldots,$$

такая, что

$$\lim_{n \to \infty} \sum_{i=1}^{\infty} [\delta(A_i^n)]^{p+1} = 0. \quad (1)$$

Обозначим через $r_i^{(n)}$ и $R_i^{(n)}$ соответственно нижнюю и верхнюю границы чисел $\rho(x, x_0)$ для $x \in A_i^n$. Очевидно, что

$$R_i^{(n)} - r_i^{(n)} \leqslant \delta(A_i^n). \quad (2)$$

Положим

$$d_i^{(n)}(r) = \begin{cases} 0 \text{ для } 0 < r < r_i^{(n)} \text{ и } R_i^{(n)} < r, \\ [\delta(A_i^n)]^p \text{ для } r_i^{(n)} \leqslant r \leqslant R_i^{(n)}, \end{cases}$$

$$d^{(n)}(r) = \sum_{i=1}^{\infty} d_i^{(n)}(r).$$

Из определения $d_i^{(n)}(r)$ и соотношения (2) вытекает:

$$\int_0^{\infty} d_i^{(n)}(r)\, dr \leqslant [\delta(A_i^n)]^{p+1}.$$

(Фактически предложение А) сильнее, чем теорема 3, так как из А), очевидно, следует, что еслито каждая точка $x \in X$ не только имеет произвольно $m_{p+1}(X) = 0$, малые окрестности с границей размерности $\geq n$ -1, но *почти все* сферические окрестности точки x имеют границы размерности $\geq n$—1.)

Каждая функция $d_i^{(n)}(r)$ неотрицательна, поэтому

$$\int_0^\infty d^{(n)}(r)\,dr = \sum_{i=1}^{\infty} \int_0^\infty d_i^{(n)}(r)\,dr \leqslant \sum_{i=1}^{\infty} [\delta(A_i^n)]^{p+1}.$$

Таким образом, в силу (1), последовательность функций $d^{(n)}(r)$ сходится в среднем к нулю. Следовательно, существует подпоследовательность $d^{(n_k)}(r)$, сходящаяся к нулю для почти всех r.

Но мы имеем:

$$[\delta(A_i^{n_k} \cap S(r))]^p \leqslant d_i^{(n)}(r).$$

Поэтому для почти всех r

$$\lim_{k\to\infty} \sum_{i=1}^{\infty} [\delta(A_i^{n_k} \cap S(r))]^p = 0.$$

Следовательно,

$$m_p(S(r)) = 0 \quad \text{для почти всех } r,$$

и предложение А), а значит, и теорема 3 доказаны.

2.3.3. n-мерное пространство гомеоморфно пространству (n+1)-мерной меры нуль

Теорема 4. *Если пространство имеет размерность ≥n, то оно гомеоморфно подмножеству куба I_{2n+1} (n+1)-мерной меры нуль.*

Теорема 4 содержится в следующей теореме:

Теорема 5. *Если* dim $X \geq n$, *то существует гомеоморфное отображение пространства X в куб I_{2n+1} такое, что для любого действительного числа $r > n$*

$$m_r(h(X)) = 0. \quad (1)$$

Более того, пространство I_{2n+1}^X содержит плотное G_δ *множество гомеоморфных отображений, удовлетворяющих условию* (1).

Доказательство теоремы 5. Пусть q> n. Рассмотрим множество K_q всех отображений $f \in I_{2n+1}^X$ таких, что

$$m_q(\overline{f(X)}) = 0. \quad (1)$$

$\overline{f(X)}$ компактно, как замкнутое подмножество куба I_{2n+1}.Следовательно (1 D)), так как $f \in K_q$, для каждого целого числа i = 1,2, ... существует конечное разложение, обозначаемое через *(d):*

$$X = A_1 \cup \ldots \cup A_{k_i}$$

для которого

$$[\delta(f(A_1))]^q + \ldots + [\delta(f(A_k))]^q < \frac{1}{i}. \quad (2)$$

Это можно записать в виде соотношения:

$$K_q = \bigcap_{i=1}^{\infty} \bigcup_{(d)} G_{i,q}^{(d)},$$

где сложение производится по всем конечным разложениям *(d),* а $G_{i,q}^{(d)}$ обозначает множество всех отображений *f*, удовлетворяющих условию (2). Но $G_{i,q}^{(d)}$, очевидно, открыто; следовательно,

K_q есть G_δ в I_{2n+1}^X.

Пусть K^* — множество всех $f \in I_{2n+1}^X$ таких, что $f(X)$ содержится в *n*-мерном полиэдре. K^* плотно в I_{2n+1}^X, и так как q>*n,* то мы имеем (1, C))

$$K^* \subset K_q.$$

Отсюда

K_q есть плотное G_δ в I_{2n+1}^X.

Пусть *H*—плотное в I_{2n+1}^X множество типа G_δ всех гомеоморфных отображений пространства *X* в I_{2n+1}, существующее в силу теоремы V 3. Пусть

$$H^* = H \cap \bigcap_{i=1}^{\infty} K_n + \frac{1}{i}. \quad (3)$$

Тогда H^*, по теореме Бэра, является плотным G_δ в пространстве I_{2n+1}^X и, следовательно, не пусто. Кроме того, если $h \in H^*$, то

a) *h* является гомеоморфным отображением пространства *X* в I_{2n+1},

b) $m_r(\overline{h(X)}) = 0$ для каждого $r > n,$ и теорема доказана.

Замечание 1. Теоремы 5 и 3 снова доказывают, что пространство можно включить в компакт той же размерности.

Замечание 2. Из теоремы 4 и предложения 2 А) получаем такое следствие: пусть *X*— пространство размерности ≤*n*; тогда *X* можно переметризовать таким образом, чтобы пространство *X* в новой метрике обладало тем свойством, что каждая его точка *x* не только имеет произвольно малые окрестности с границами размерности ≤*n* — 1, но *почти все сферические* окрестности точки *x* имеют границы размерности ≤*n* — 1.

2.3.4. Хаусдорфова размерность

Пусть *X*—произвольное метрическое пространство. Назовем хаусдорфовой размерностью пространства *X* верхнюю грань множества всех действительных чисел *p,* для которых $m_p(X) > 0$. Из теоремы 2 следует, что хаусдорфова размерность пространства *X* не меньше dim *X*. Хаусдорфова размерность пространства не обязательно является целым числом. Так, хаусдорфова размерность канторова множества равна

$$\frac{\lg 2}{\lg 3} = 0,63093.$$

Но из теоремы 5 вытекает, что если рассматривать всевозможные пространства, гомеоморфные данному пространствуй, то нижняя грань хаусдорфовой размерности этих пространств равна размерности пространства *X*.

3. Теория гомологии и размерность

Ранее преобладали теоретико-множественные методы. Затем появились комбинаторные методы (в доказательстве того, что *n*-мерное эвклидово пространство имеет размерность *n)* и более полно доказательство теоремы о включении было получено синтезом теоретико-множественных и комбинаторных методов. Причем теоретике-множественные методы нашли свое применение при рассмотрении пространства отображений, а комбинаторные—при использовании некоторых свойств комплексов. В этом разделе мы исследуем теорию размерности с чисто алгебраической, комбинаторной точки зрения. Звеном, соединяющим эти новые методы с нашим предыдущим рассмотрением, является теорема 4 об отображениях в сферы. Главной целью этого раздела является охарактеризование размерности алгебраическими свойствами (теорема 3).

Большая часть раздела посвящена сжатому изложению алгебраической теории связности.

Во всем разделе слово «группа» означает коммутативную группу с групповой операцией записываемой как сложение. Наиболее часто употребляются две группы: группа целых чисел и группа действительных чисел, приведенных по модулю 1. Мы обозначаем их соответственно через F и П. Группа П изоморфна группе вращений

окружности и мультипликативной группе комплексных чисел, по модулю равных 1.

3.1. Комбинаторная теория связности комплекса

Определение 1. Пусть *К*— (конечный) комплекс, и *n* — целое положительное число. Под *ориентированным n*-мерным симплексом s^n комплекса *К* мы понимаем *n*-мерный симплекс комлекса K, вершины которого записаны в некотором определенном порядке:

$$s^n = (p_0, \ldots, p_n).$$

Ориентированные симплексы, имеющие одни и те же вершины, упорядоченные различными способами, рассматриваются как *равные,* если они могут быть получены один из другого с помощью четной перестановки вершин. Следовательно, каждому неориентированному *n*-мерному симплексу соответствуют два различных ориентированных симплекса; если один из них обозначен символом s или + *s,* то другой (полученный из *s* нечетной перестановкой его вершин) будет обозначаться через —s. Например

$$(p_0, p_1, p_3) = (p_1, p_2, p_0) = (p_2, p_0, p_1) = -(p_0, p_2, p_1) = = -(p_2, p_1, p_0) = -(p_1, p_0, p_3).$$

Представляется удобным распространить определения ориентированных *n*-мерных симплексов на случай *n* = 0 и *n* =— 1; причем делается это следующим образом: каждому неориентированному нульмерному симплексу комплекса *К,* т. е. каждой вершине *p* комплекса K, мы ставим в соответствие два символа: +*p* и -*p,* и называем их *ориентированными* нульмерными симплексами, принадлежащими *p.* Далее, условимся называть пустое множество вершин комплекса K неориентированным (— 1)-мерным симплексом комплекса *К (* пустое множество вершин является единственным (—1)-мерным симплексом комплекса K) и ставить ему в соответствие два символа: +Δ и —Δ, как ориентированные (— 1) -мерные симплексы комплекса *К.* Конечно, мы полагаем

$$-(+p) = -p, \quad -(-p) = +p,$$

$$-(+\Delta) = -\Delta, \quad -(-\Delta) = +\Delta.$$

Ориентированный *n*-мерный симплекс $(p_0, p_1, \ldots, p_n)$, $n \geqslant 1$, называется *ориентированной гранью ориентированного (n+1)*- мерного симплекса s^{n+1}, если

$$s^{n+1} = (p_{n+1}, p_0, \ldots, p_n).$$

Для $n \geq 2$ любой ориентированный n-мерный симплекс имеет $n+1$ ориентированные грани. Так, ориентированный двумерный симплекс (p_0, p_1, p_2) имеет три ориентированные грани:

$$(p_1, p_2) = -(p_2, p_1),\ (p_2, p_0) = -(p_0, p_2),\ (p_0, p_1) = -(p_1, p_0).$$

Ориентированные грани одномерного симплекса (p_0, p_1) определяются как два ориентированных нульмерных симплекса: p_1 и $-p_0$.
Условимся, кроме того, считать $+\Delta$ единственной *ориентированной гранью нульмерного симплекса* типа $+p$ и $-\Delta$ единственной *ориентированной гранью нульмерного симплекса* типа $-p$.
В дальнейшем символ s^n будет употребляться для обозначения произвольного ориентированного n-мерного симплекса комплекса K, причем нижние индексы отличают отдельные симплексы. Мы пишем $s^n \prec s^{n+1}$ или $s^{n+1} \succ s^n$; если симплекс s^n является гранью симплекса s^{n+1} и $s^n \prec s^{n+2}$ или $s^{n+2} \succ s^n$, если существует s^{n+1} такой, что $s^n \prec s^{n+1} \prec s^{n+2}$. Легко проверяются следующие простые факты.
А) Соотношение $s^n \prec s^{n+1}$ влечет за собой соотношение $-s^n \prec -s^{n+1}$, но исключает соотношения $s^n \prec -s^{n+1}$ и $-s^n \prec s^{n+1}$. Соотношения:

$$s^n \prec s^{n+2},\ s^n \prec -s^{n+2},\ -s^n \prec s^{n+2},\ -s^n \prec -s^{n+2}$$

эквивалентны и утверждают, что все вершины симплекса s^n являются вершинами симплекса s^{n+2}, т. е. что неориентированный симплекс s^n является гранью неориентированного симплекса s^{n+2}. Если $s^n \prec s^{n+2}$, то существует только один ориентированный симплекс s^{n+1} такой, что $s^n \prec s^{n+1} \prec s^{n+2}$.

Цепи, Δ-циклы, ∇-циклы

Определение 2. Пусть G—абелева группа в аддитивной записи. Пусть K— комплекс, и n—целое число. Под *n-мерной цепью комплекса K по области коэффициентов G* мы понимаем функцию φ, ставящую каждому ориентированному n-мерному симплексу комплекса K в соответствие некоторый элемент группы G и удовлетворяющую условию:

$$\varphi(-s^n) = -\varphi(s^n).$$

Нулевая n-мерная цепь есть функция, ставящая каждому ориентированному n-мерному симплексу комплекса K в соответствие нуль группы G. Для каждого $n > \dim K$ существует одна и только одна n-мерная цепь комплекса K, именно — нулевая n-мерная цепь.
Замечание. (—1)-мерная цепь по области коэффициентов G может быть отождествлена с элементом $\varphi(+\Delta)$ группы G.

Нульмерная цепь—это в сущности функция вершив *p*, принимающая значения из G. При $n \geq 1$ *n*-мерную цепь можно рассматривать как кососимметрическую функцию $(n+1)$ вершин, принимающую значения из G (эта функция определена только на тех множествах, состоящих из $(n+1)$ вершин, которым соответствует некоторый *n*-мерный симплекс комплекса *K).* Если *G*—группа целых чисел, приведенных по модулю 2, то *n*-мерную цепь по области коэффициентов G можно рассматривать как некоторую совокупность неориентированных *n*-мерных симплексов (именно, совокупность симплексов, удовлетворяющих соотношению $\varphi(s^n) = \varphi(-s^n) \neq 0$).

Определение 3. Под суммой $\varphi = \varphi_1 + \varphi_2$ двух *n*-мерных цепей φ_1 и φ_2 комплекса *K* по области коэффициентов G мы понимаем *n*-мерную цепь φ, определенную соотношением:

$$\varphi(s^n) = \varphi_1(s^n) + \varphi_2(s^n).$$

При таком определении сложения *n*-мерные цепи комплекса *K* образуют абелеву группу, называемую *группой n-мерных цепей комплекса K* и обозначаемую через $L^n(K, G)$ или, если никакие недоразумения невозможны, через $L^n(K)$ или даже L^n. Нулем этой группы является нулевая *n*-мерная цепь.

Если $n > \dim K$, то $L^n(K, G)$ сводится к единственному элементу 0.

Если даны симплекс $s_0^n \in K$ и элемент $g_0 \in G$, то символом $g_0 s_0^n$ мы обозначаем *n*-мерную цепь и, определенную следующим образом:

$$\varphi(s^n) = g_0, \quad \text{если } s^n = s^n_0$$
$$\varphi(s^n) = -g_0, \quad \text{если } s^n = -s^n_0$$
$$\varphi(s^n) = 0 \quad \text{в остальных случаях.}$$

Цепи этого типа называются *элементарными* цепями. Каждая *n(*-мерная цепь может быть представлена как сумма $\sum_i g_i s_i^n$ элементарных *n*-мерных цепей , это представление однозначно, если пренебрегать членами с коэффициентом нуль и требовать, чтобы $s_i^n \neq \pm s_{i'}^n$, при $i \neq i'$.

Определение 4. Δ-*граница* (или просто *граница)* Δφ *n*-мерной цепи $\varphi\ (n \geqslant 0)$ есть $(n-1)$-мерная цепь, определенная формулой:

$$\Delta\varphi(s^{n-1}) = \sum_{s^n \succ s^{n-1}} \varphi(s^n);$$

∇ -граница (или *верхняя граница)* $\nabla\varphi$ *n*-мерной цепи φ ($n \geq -1$) есть $(n+1)$-мерная цепь, определенная формулой:

$$\nabla\varphi(s^{n+1}) = \sum_{s^n \prec s^{n+1}} \varphi(s^n).$$

Пример 1. Δ-границей элементарной цепи $g s_0^n\ (n \geqslant 0)$ является цепь

$$\sum_{s^{n-1} \prec s_0^n} g s^{n-1},$$

а ∇ -границей—цепь

$$\sum_{s^{n+1} \succ s_0^n} g s^{n+1}.$$

В) Оператор Δ является, очевидно, гомоморфизмом группы L^n в группу L^{n+1}, а оператор ∇ —гомоморфизмом группы L^n в группу L^{n+1},

Операторы Δ и ∇ обладают важными свойствами:

$$\Delta\Delta\varphi = 0, \quad \nabla\nabla\varphi = 0$$

для любой n-мерной цепи φ (с ограничением $n \geq 1$ для первого соотношения).

(Сумму пустого множества слагаемых мы считаем равной нулю)

Доказательство. В силу определений операторов Δ и ∇ и последней части предложения А), мы имеем

$$\Delta\Delta\varphi(s^{n-2}) = \sum_{s^n \succ s^{n-2}} \varphi(s^n),$$

$$\nabla\nabla\varphi(s^{n+2}) = \sum_{s^n \prec s^{n+2}} \varphi(s^n).$$

Из предложения А) мы знаем также, что из $s^n \prec s^{n+2}$ вытекает, что $s^n \prec - s^{n+2}$ и $-s^n \prec s^{n+2}$. Следовательно, каждый член $\varphi(s^n)$ в сумме погашается членом $\varphi(-s^n) = -\varphi(s^n)$, и это показывает, что $\Delta\Delta\varphi(s^{n-2}) = 0$ для каждого s^{n-2} и $\nabla\nabla\varphi(s^{n+2}) = 0$ для каждого s^{n+2}.

Определение 5. n-мерная цепь φ комплекса K по области коэффициентов G называется Δ-*циклом,* если $\Delta\varphi = 0$, и ∇ -*циклом,* если $\nabla\varphi = 0$. Из В) следует, что любая n-мерная цепь, являющаяся Δ-границей некоторой $(n+1)$-мерной цепи (∇ -границей (n—1)-мерной цепи), оказывается Δ-циклом (∇ -циклом). Цепи этого типа называются *ограничивающими* Δ-*циклами (ограничивающими* ∇ -*циклами).*

И n-мерные Δ-циклы, и n-мерные ∇ -циклы комплекса $K,$ очевидно, образуют подгруппы группы $L^n(K, G).$

(*Ядром* гомоморфизма h группы G в группу G' называется множество элементов $g \in G$ таких, что $h(g) = 0$. Подгруппы Z_Δ^n и Z_∇^n являются, таким образом, ядрами гомоморфизмов Δ и ∇ группы L^n в группу L^{n-1} и группы L^n в группу L^{n+1}.)

Обозначим эти подгруппы соответственно через $Z_{\Delta}^{n}(K, G)$ и $Z_{\nabla}^{n}(K, G)$. Далее, *n*-мерные ограничивающие Δ-циклы образуют, в силу В), подгруппу группы $Z_{\Delta}^{n}(K, G)$, которую мы обозначаем через $H_{\Delta}^{n}(K, G)$.

(H_{Δ}^{n} является образом группы L^{n+1} при гомоморфизме Δ, а H_{∇}^{n} является образом группы L^{n-1} при гомоморфизме ∇.)

Аналогично *n*-мерные ограничивающие ∇-циклы образуютподгруппу группы $Z_{\nabla}^{n}(K, G)$, и мы обозначаем ее $H_{\nabla}^{n}(K, G)$.

Δ- и ∇ - группы комплекса

Определение 6. Фактор-группа $Z_{\Delta}^{n}(K, G) / H_{\Delta}^{n}(K, G)$ группы *n*-мерных Δ-циклов по группе *n*-мерных ограничивающих Δ-циклов называется *n-мерной* Δ*-группой (группой Бетти) комплекса К по области коэффициента G* и обозначается через $\Delta^{n}(K, G)$, или $\Delta^{n}(K)$, или просто Δ^{n}. Элементы группы Δ^{n}, т. е. смежные классы группы Z_{Δ}^{n} по подгруппе H_{Δ}^{n}, называются *n-мерными* Δ*-классами,* а два Δ-цикла, принадлежащие к одному и тому же Δ-классу, называются *гомологичными* между собой. Аналогично, фактор-группа $Z_{\nabla}^{n}(K, G) / H_{\nabla}^{n}(K, G)$ группы *n*-мерных ∇-циклов по группе *n*-мерных ограничивающих ∇-циклов называется *n-мерной* ∇*-группой комплекса К по области коэффициентов G* и обозначается через $\nabla^{n}(K, G)$, или $\nabla^{n}(K)$, или просто ∇^{n}. Элементы группы ∇^{n}, т. е. классы смежности группы Z_{∇}^{n} по подгруппе H_{∇}^{n}, называются *n-мерными* ∇*-классами,* а два ∇-цикла, принадлежащие к одному и тому же ∇-классу, называются *гомологичными* между собой.

Ясно, что два *n*-мерных Δ-цикла ψ и φ гомологичны в том и только в том случае, если существует (*n*+1) -мерная цепь θ такая, что $\Delta\theta = \varphi - \psi$, и два *n*-мерных ∇-цикла ψ и φ гомологичны в том и только в том случае, если существует (*n*— 1)-мерная цепь θ такая, что $\nabla\theta = \varphi - \psi$.

Замечание. Пусть $\mathfrak{Z}$ —группа целых чисел. Тогда, для каждого *n* группа $\Delta^{n}(K, \mathfrak{Z})$ является группой с конечным числом образующих. Следовательно, она может быть разложена в прямую сумму свободной группы и некоторого числа групп конечных порядков. Ранг свободной подгруппы называется *n-мерным числом Бетти комплекса К,* а порядки конечных подгрупп в каноническом разложении называются *n-мерными коэффициентами кручения комплекса К.*

Известно, что числа Бетти и коэффициенты кручения вполне определяют Δ- и ∇ -группы комплекса по произвольной группе О в качестве области коэффициетов. Например, $\nabla^n(K, \mathfrak{Z})$ есть прямая сумма b_n бесконечных циклических групп, где b_n — n-мерное число Бетти комплекса K, и конечных циклических групп, порядки которых равны (n— 1)-мерным коэффициентам кручения комплекса K. Δ^n $(K,$ П) есть прямая сумма b_n групп, изоморфных П, и конечных циклических групп, порядки которых равны $(n$ — 1) -мерным коэффициентам кручения комплекса K.

Пример 2. Если $m > \dim K$, то $\nabla^n(K, G) = \Delta^n(K, G) = 0$.

Пример 3. Комплекс K называется *связным,* если он не может быть разбит на два комплекса без общих вершин, или, что сводится к тому же самому, если любые две вершины p и q комплекса K можно соединить последовательностью одномерных симплексов $(p, p_1), (p_1, p_2), \ldots, (p_n, q)$. Каждый комплекс может быть единственным образом разложен на связные подкомплексы, называемые *компонентами* комплекса K. Можно без труда убедиться в том, что нульмерная цепь комплекса K, т. е. функция $\varphi(p)$ вершин $p \in K$, является ∇ -циклом в том и только в том случае, если на каждой компоненте комплекса K функция φ(p) имеет постоянное значение. Нульмерный ∇ -цикл φ гомологичен нулю в том и только в том случае, если φ имеет постоянное значение на всем комплексе K. Следовательно, группа $\nabla^0(K, G)$ есть прямая сумма m—1 групп, изоморфных G, где m — число компонент комплекса K. Такой же результат легко установить для группы $\Delta^0(K; G)$.

Пример 4. Если n = dim A, то группа $\Delta^n(K, G)$ совпадает с группой $Z_1^n(K, G)$. Кроме того, каждая n-мерная цепь является ∇ -циклом и, следовательно, $\nabla^n(K, G)$ совпадает с группой $L^n(K, G)/H_\nabla^n(K, G)$.

Пример 5. Пусть K—n-мерный комплекс, обладающий следующими свойствами: 1) каждый неориентированный (n — 1)-мерный симплекс комплекса K является общей гранью в точности двух n-мерных симплексов; 2) K не может быть представлен как сумма двух комплексов, не имеющих ни одного общего (n—1)-мерного симплекса; (комплексы такого вида иногда называются *n-мерными псевдомногообразиями).* В качестве области коэффициентов возьмем группу $\mathfrak{Z}$ и определим ∇ - группу $\nabla^n(K) = \nabla^n(K, \mathfrak{Z})$. Каждая n-мерная цепь является ∇ -циклом, и единственный вопрос, на который нам остается ответить, состоит в следующем: когда n-мерная цепь является ограничивающим ∇ -циклом? Следует различать два случая:

а) Допустим, что каждому n-мерному симплексу комплекса K можно придать определенную ориентацию, называемую *положительной* ориентацией,таким образом, чтобы каждый ориентированный $(n-1)$-мерный симплекс являлся ориентированной гранью в точности одного положительно ориентированного n-мерного симплекса (и, следовательно, в точности одного отрицательно ориентированного симплекса). Мы скажем тогда, что комплекс K *ориентируем.* Легко доказать, что каковы бы ни были два произвольных положительно ориентированных симплекса s^n и $\tilde{s}^n$, элементарные цепи $1s^n$ и $1\tilde{s}^n$ гомологичны.

(Это ясно, когда s^n и $\tilde{s}^n$—соседние симплексы; общее утверждение сводится к свойству 2)). Заметим далее, что каждая n-мерная цепь, являющаяся ∇ -границей некоторой $(n-1)$ -мерной цепи, удовлетворяет условию $\sum \varphi(s^n) = 0$, где сумма распространена на все положительно ориентированные симплексы (достаточно проверить это для случая, когда φ является ∇ -границей элементарной цепи). Отсюда можно заключить, что $\nabla^n(K)$ изоморфна группе $\mathfrak{Z}$ и что, если s^n_0 — произвольный ориентированный n-мерный симплекс, элементарные цепи ms^n_0, $m = 0, \pm 1, \pm 2, \ldots$, представляют все ∇ -классы, причем в каждый ∇ -класс входит лишь одна из этих цепей. Аналогично, если G — произвольная группа, группа $\nabla^n(K, G)$ изоморфна G.

b) Допустим, что n-мерные симплексы комплекса K нельзя ориентировать в соответствии с требованиями, высказанными выше. В этом случае K называется *неориентируемым.* Легко показать, что для каждого s^n цепь $2s^n$ гомологична нулю и что группа $\nabla^n(K, \mathfrak{Z})$ является группой порядка 2. Ненулевой элемент группы $\nabla^n(K, \mathfrak{Z})$ представляется каждым из ∇ -циклов $1s^n$ (для того чтобы доказать, что цепь $1s^n$ не гомологична нулю, надо показать, что если $\pm s^n_i$, $i = 1, \ldots, k$, суть все ориентированные n-мерные симплексы комплекса K и φ — n-мерная цепь, являющаяся ∇ -граиицей некоторой (—1)-мерной цепи, то

$$\sum_{i=1}^{k} \varphi(i_i^n) \equiv 0 (\mathrm{mod}\, 2)$$

Пример 6. Пусть комплекс K является суммой двух комплексов K_1 и K_2, не имеющих общих вершин. Тогда, при $n>0$, $\nabla^n(K, G)$ есть прямая сумма групп $\nabla^n(K_1, G)$ и $\nabla^n(K_2, G)$, и $\Delta^n(K, G)$ есть прямая сумма групп $\Delta^n(K_1, G)$ и $\Delta^n(K_2, G)$. При $n = 0$ эти утверждения неверны (см. пример 3).

Мы будем часто рассматривать соответствия между цепями одного и цепями другого комплекса.

с) Пусть K_1 и K_2 — два комплекса, и G — группа. Допустим, что каждой n-мерной цепи φ комплекса K_1 по области коэффициентов G постановлена в соответствие n-мерная цепь h (φ) комплекса K_2 по той же области коэффициентов, n = 0,1, 2,..., и для каждого n оператор h является гомоморфизмом группы L^n (K_1, G) в группу L^n(K_2, G). Допустим, наконец, что h коммутирует с граничным оператором Δ, т. е.

$$h(\Delta(\varphi)) = \Delta(h(\varphi))$$

для каждой n-мерной цепи φ комплекса K_1. Тогда h переводит Δ-циклы в Δ-циклы и ограничивающие Δ-циклы — в ограничивающие Δ-циклы, и, следовательно, гомоморфно отображает группу $\Delta^n(K_1)$ в группу $\Delta^n(K_2)$. Аналогично, если h коммутирует с верхним граничным оператором ∇, т. е. если

$$h(\nabla(\varphi)) = \nabla(h(\varphi)),$$

то h переводит ∇ -циклы в ∇ -циклы и ограничивающие ∇ -циклы в ограничивающие ∇ -циклы, и» следовательно, гомоморфно отображает группу $\nabla^n(K_1)$ в группу $\nabla^n(KC_2)$.

Естественные гомоморфизмы Δ- и ∇ -группы mod *b*

Пусть K— комплекс, L — подкомплекс *) комплекса K.
(Существенно делать различие между цепями комплекса K и цепями подкомплекса L Первые являются функциями, определенными на симплексах комплекса K, а вторые — функциями, определенными на симплексах подкомплекса L. Даже если (для некоторого фиксированного n) множество n-мерных симплексов подкомплекса L совпадает с множеством n-мерных симплексов комплекса K, необходимо отличать n-мерную цепь Φ комплекса K от n-мерной цепи φ подкомплекса L: область определения функции $\nabla \Phi$ состоит из всех (n + 1)-мерных симплексов комплекса K, тогда как область определения функции $\nabla \varphi$ состоит лишь из (n+1)-мерных симплексов комплекса L. Следует быть особенно осторожным, когда цепь записывается как линейная комбинация элементарных цепей, ибо элементарная цепь gs^n_0, где s^n_0 симплекс подкомплекса L, может означать как цепь комплекса L, так и цепь комплекса K.)
Каждой n-мерной цепи Φ комплекса K, естественным образом, соответствует n-мерная цепь $\varphi = h_L \Phi$ комплекса L, полученная ограничением области определения цепи Φ только симплексами подкомплекса L и обратно, каждой цепи φ комплекса L соответствует

цепь $\Phi = h_K\varphi$ комплекса *K,* определенная следующим образом: $\Phi(s^n) = \varphi(s^n)$, если $s^n \in L$, и $\Phi(s^n) = 0$, если $s^n \notin L$. Для каждого *n* мы получаем таким путем гомоморфизм h_L группы $L^n(K, G)$ в группу $L^n(L, G)$ и гомоморфизм h_K группы $L^n(L, G)$ в группу $L^n(K, G)$.

(В действительности h_L является гомоморфнзом группы $L^n(K, G)$ *на* группу $L^n(L, G)$, а h_K — изоморфизмом группы $L^n(L, G)$ в группу $L^n(K, G)$. (Гомоморфизм и изоморфизм группы *G в* группу *H* означает отображение группы *G* на подгруппу группы *H.))*

Легко показать, что оператор h_L коммутирует с верхний граничным оператором ∇, т. е. для любой цепи Ф комплекса *K*

$$h_L \nabla \Phi = \nabla h_L \Phi. \quad \textbf{(1)}$$

Аналогично, оператор h_K коммутирует с граничным оператором Δ, т. е. для любой цепи φ подкомплекса *L:*

$$h_K \Delta\varphi = \Delta h_K \varphi. \quad (1')$$

Из С) заключаем, что h_L порождает гомоморфизм группы $\nabla^n(K, G)$ в группу $\nabla^n(L, G)$ и h_K порождает гомоморфизм группы $\Delta^n(L, G)$ в группу $\Delta^n(K, G)$.

Определение 7. Гомоморфизм H_L группы $\nabla^n(K, G)$ в группу $\nabla^n(L, G)$ и гомоморфизм h_K группы $\Delta^n(L, G)$ в группу $\Delta^n(K, G)$ называются *естественными гомоморфизмами.* Если элемент *e* группы ∇^n(L) является образом элемента $\tilde{e}$ группы ∇^n(K) при h_L, то $\tilde{e}$ называется *продолжением* элемента *e,* а элемент *e — продолжаемым на K.* Элемент группы ∇^n(L) , переходящий при h_K в нуль группы ∇^n(K), называется *ограничивающим* в *K.*

Мы опускаем доказательство следующего предложения:

D) Пусть размерность комплекса *K* равна *n.* Естественный гомоморфизм группы Δ^n(L) в группу Δ^n(K) является изоморфизмом Δ^n(L) в $\Delta^n(K)$, тогда как естественный гомоморфизм группы ∇^n(K) в группу $\nabla^n(L)$ является гомоморфизмом $\nabla^n(K)$ *на* ∇^n (L), т. е. каждый элемент группы ∇^n(L) имеет продолжение (для доказательства нужно воспользоваться примером 4). Пусть *L*—-подкомплекс комплекса *K,* состоящий из всех симплексов комплекса *K* размерности ≤ *m.* Для *m* = *n*— 1 естественный гомоморфизм группы $\Delta^m(L)$ в группу $\Delta^m(K)$ является гомоморфизмом Δ^m (L) *на* $\Delta^m(K)$, а естественный гомоморфизм группы $\nabla^m(K)$ в группу ∇^m (L) является изоморфизмом $\nabla^m(K)$ в ∇^m(L). Если $m \leq n$ — 2, то оба естественных гомоморфизма являются изоморфизмами *на.*

E) Пусть *e* — элемент группы ∇^n(L), и $\tilde{e}$—продолжение

элемента $e, \tilde{e} \in \nabla^n(K)$. Тогда для каждого ∇ -цикла φ подкомплекса *L*, представляющего *e,* существует ∇ -цикл Φ комплекса *K,* представляющий $\tilde{e}$ и являющийся продолжением ∇ -цикла φ.

Доказательство. Предположим сначала, что $\tilde{e} = 0$. Тогда $e = 0$, т. е. существует цепь ψ комплекса *L* такая, что $\nabla\psi = \varphi$. Гомологичный нулю ∇ -цикл $\nabla h_K \psi$ является тогда продолжением φ, ибо $h_L \nabla h_K \psi = \nabla h_L h_K \psi = \nabla\psi = \varphi$. Пусть теперь $\tilde{e}$— произвольный элемент группы ∇^n(K) и пусть Φ_1 — ∇ -цикл комплекса *K,* принадлежащий ∇ -классу $\tilde{e}$. Тогда $\varphi - h_L\Phi_1$ есть ограничивающий ∇ -цикл комплекса *L* и, следовательно, он продолжаем в ограничивающий ∇ -цикл Φ_2 комплекса *K.* ∇ -цикл $\Phi_1+\Phi_2$ является искомым продолжением ∇ -цикла φ.

Рассмотрим теперь функции, принимающие значения из группы G, определенные на ориентированных *n*-мерных симплексах дополнения *K\L. (*Т. е. *n*-мерных симплексах комплекса *K,* не принадлежащих *L. K* рассматривается как множество своих симплексов.)

Если такая цепь удовлетворяет условию: $\varphi(-s^n) = -\varphi(s^n)$, то она называется *n-мерной цепью* mod L, *n*-мерные цепи mod L образуют группу. Δ- и ∇ -граница цепи φ mod L- определяются таким же образом, как для обычных цепей, с тем единственным отличием, что Δφ и ∇ φ рассматриваются как цепи, определенные только для $s^n \in K \setminus L$. Проверяем, что $\Delta\Delta\varphi = 0$ и $\nabla\nabla\varphi = 0$. Как прежде, Δ- и ∇ -циклы mod *L* определяем соотношениями $\Delta\varphi = 0$ и $\nabla\varphi = 0$.

Определяем Δ^n *(K* mod L, G), *n-мерную* ∇ *-группу K* mod *L,* как фактор-группу группы. всех Δ-циклов mod *L* по группе всех ограничивающих Δ-циклов modi, и ∇^n*(K* mod L, G), *n-мерную* ∇ *-группу K* mod *L,* как фактор-группу группы всех ∇ -циклов mod *L* по группе всех ограничивающих ∇ -циклов mod *L.*

Пример 8. Пусть *K*—произвольный комплекс. Образуем новый комплекс *K*,* называемый *конусом* над *K,* прибавляя к вершинам комплекса *K* новую вершину *p* и считая симплексами комплекса *K** все симплексы комплекса *K* и все симплексы вида *(p,* s^n), где симплекс $s^n \in K$. Ясно, что ориентированные *n*-мерные симплексы дополнения *K** \ *K* находятся во взаимно однозначном соответствии с ориентированными (*n* — 1)-мерными симплексами комплекса *K;* следовательно, *n*-мерные цепи *K** mod *K* находятся в изоморфном соответствии с (*n* — 1)-мерными цепями комплекса *K.* Нетрудно доказать, что группы $\nabla^n(K^* \bmod K)$ и $\nabla^{n-1}(K)$ и, точно так же,

группы $\Delta^n(K^* \bmod K)$ и $\Delta^{n-1}(K)$ изоморфны, n=1, 2,...

Пример 9. Если K_1, K_2, L — подкомплексы комплекса K $K = K_1 \cup K_2$, $K_1 \cap K_2 \subset L$, то группа $\nabla^n(K \bmod L, G)$ есть прямая сумма групп $\nabla^n(K_1 \bmod (L \cap K_1), G)$ и $\nabla^n(K_2 \bmod (L \cap K_2), G)$.

Аналогичное утверждение справедливо для Δ-групп (в отличие от утверждений примера 6 эти утверждения имеют место, даже если n = 0).

Если Ф —m-мерная цепь комплекса K, то обозначим через $h_{K \setminus L}\Phi$ ту m-мерную цепь mod L, которая получается из Ф ограничением области определения цепи Ф только симплексами, входящими в $K \setminus L$. Если φ — m-мерная цепь mod L, то обозначим через $h_K\varphi$ m-мерную цепь комплекса K, определенную следующим образом: $h_K\varphi(s^m) = \varphi(s^m)$, если $s^m \in K \setminus L$, и $h_K\varphi(s^m) = 0$, если $s^m \in L$. Легко проверяем, что

$$h_{K \setminus L}\Delta\Phi = \Delta h_{K \setminus L}\Phi \quad (2)$$

для любой m-мерной цепи Ф комплекса K и

$$h_K \nabla \varphi = \nabla h_K \varphi \quad (2')$$

для любой m-мерной цепи φ комплекса K mod L (сопоставьте эти формулы) с формулами (1) и (1') и обратите внимание на то, что операторы Δ и ∇ меняются местами).

(Соотношения (2) и (2′) показывают,что оператор h_K порождает для данных m и G гомоморфизм группы $\nabla^m(K \bmod L, G)$ в группу $\nabla^m(K, G)$ и $h_{K \setminus L}$ порождает гомоморфизм группы $\Delta^m(K, G)$ в группу $\Delta^m(K \bmod L, G)$.)

F) Допустим, что $\nabla^m(K \bmod L, G) = 0$. Пусть Ф —$m$-мерный ∇-цикл комплекса K, и ψ — (m— 1)-меряая цепь подкомплекса L такая, что $\nabla\psi = h_L\Phi$. Тогда существует (m — 1)-мерная цепь ψ комплекса K, обладающая свойствами:

$$h_L\Psi = \psi; \quad \nabla\Psi = \Phi.$$

Доказательство. $\Phi' = \Phi - \nabla h_K\psi$ является m-мерным ∇-циклом комплекса K, обладающим тем свойством, что $h_L\Phi' = 0$, т. е. $\Phi'(s^m) = 0$, если $s^m \in L$, так как, в силу (Г), $h_L\Phi' = h_L\Phi - \nabla\psi = 0$.

Очевидно, $h_{K \setminus L}\Phi'$ является m-мерным ∇-циклом mod L и, так как по предположению все m-мерные ∇-циклы mod L ограничивают, то существует (m — 1)-мерная цепь ψ′ mod L такая, что $\nabla\psi' = h_{K \setminus L}\Phi'$ и, следовательно, в силу (2'), $\nabla h_K\psi' = h_K\nabla\psi' = h_K h_{K \setminus L}\Phi' = \Phi' = \Phi - \nabla h_K\psi$.

Цепь $\Psi = h_K\psi + h_K\psi'$ удовлетворяет условиям $h_L\Psi = \psi$ и $\nabla\Psi = \Phi$.

Из F) получаем:

G) Если $\nabla^m(K \bmod L, G) = 0$, то естественный гомоморфизм группы $\nabla^m(K, G)$ в группу $\nabla^m(L, G)$ является изоморфизмом, а естественный гомоморфизм группы $\nabla^{m-1}(K, G)$ в группу $\nabla^{m-1}(L, G)$ является гомоморфизмом *на* группу $\nabla^{m-1}(L, G)$.

Доказательство. С одной стороны, из F) вытекает, что m-мерный ∇-цикл Ф комплекса K ограничивает, если ограничивает $h_L\Phi$. С другой стороны, применяя F) к m-мерному ∇-циклу Ф = 0 получаем, что каждый *(m*— 1)-мерный ∇-цикл ψ подкомплекса L является образом некоторого ∇-цикла комплекса K при h_L.

Установим теперь предложения, аналогичные F) и G), для граничного оператора Δ. Будем для краткости называть две m-мерные цепи Φ_1 и Φ_2 комплекса K (которые не обязательно являются Δ-циклами) *гомологичными,* если их разность Φ_1 -Φ_2 является ограничивающим Δ-циклом; отсюда, конечно, следует, что Φ_1 и Φ_2 имеют одну и ту же границу.

F') Допустим, что $\Delta^m(K \bmod L, G) = 0$. Пусть φ-*(m*— 1)- мерный Δ-цикл подкомплекса L, и W — m-мерная цепь комплекса K такая, что $\Delta\Psi = h_K\varphi$. Тогда существует m-мерная цепь ψ подкомплекса L такая, что цепь h_L гомологична Ψ. Отсюда следует, что $\Delta\psi = \varphi$ (так как h_K есть взаимно однозначное соответствие и, в силу (1'), $h_K\Delta\psi = \Delta h_K\psi = \Delta\Psi = h_K\varphi$).

Доказательство. $h_{K \setminus L}\Psi$ является m-мерным Δ-циклом mod L (ибо, в силу (2), $\Delta h_{K \setminus L}\Psi = h_{K \setminus L}\Delta\Psi = h_{K \setminus L}h_K\varphi = 0$), и так как по предположению все m-мерные Δ-циклы mod L, ограничивают $h_{K \setminus L}\Psi = \Delta\theta$, где θ — некоторая $(m+1)$-мерная цепь mod L. m-мерная цепь $\Psi' = \Psi - \Delta h_K\theta$ гомологична Ψ, и, в силу (2), мы имеем

$$h_{K \setminus L}\Psi' = h_{K \setminus L}\Psi - \Delta h_{K \setminus L}h_K\theta = h_{K \setminus L}\Psi - \Delta\theta = 0.$$

Это показывает, что $\Psi' = h_K h_L\Psi'$ и что, следовательно, цепь $\psi = h_L\Psi'$ удовлетворяет условиям предложения F').

Из F') получаем:

G') Если $\Delta^m(K \bmod L, G) = 0$, то естественный гомоморфизм группы $\Delta^m(L, G)$ в группу $\Delta^m(K, G)$ является гомоморфизмом *на* гуппу $\Delta^m(K, G)$, а естественный гомоморфизм группы $\Delta^{m-1}(L, G)$ в группу $\Delta^{m-1}(K, G)$ является изоморфизмом.

Доказательство. Из F'), очевидно, следует, что *(m*— 1)-мерный Δ-цикл φ подкомплекса L ограничивает, если $h_K\varphi$ ограничивает. С другой стороны, применяя F') к (*m*—1)-мерному Δ-циклу φ = 0, получаем, что каждый m-мерный Δ-цикл комплекса K является образом некоторого m-мерного Δ-цикла подкомплекса L при h_K.

Мы воспользуемся теперь предложениями G) и G') для определения

Δ- и ∇ -групп простейших комплексов.

Пример 10. Пусть Q^n — комплекс, состоящий из n-мерного симплекса и всех его граней. Тогда для всех m, ∇ - и Δ-группы $\nabla^m(Q^n, G)$ и $\Delta^m(Q^n, G)$ суть нулевые группы. Это доказывается индукцией по n. Утверждение тривиально для Q^0. Допустим, что оно справедливо для Q^{n-1}, Но Q^n есть конус над Q^{n-1}. В силу примера 8, группы $\nabla^m(Q^n \bmod Q^{n-1}, G)$ суть нулевые группы при всех m. В силу G), $\nabla^m(Q^n, G)$ изоморфна $\nabla^m(Q^{n-1}, G)$ и, следовательно, является нулевой группой. Аналогично, $\Delta^m(Q^n, G)$ есть нулевая группа.

Пример 11. Пусть R^n — комплекс, состоящий из всех граней $(n+1)$-мерного симплекса, размерность которых $\leq n$. Назовем этот комплекс *элементарной n-мерной сферой.* Если выбрана некоторая определенная ориентация s^{n+1}: данного$(n+1)$-мерного симплекса, то мы будем говорить об *ориентированной элементарной n-мерной сфере* R^n, причем ориентированные грани симплекса s^{n+1} будем называть *положительно* ориентированными симплексами R^n. *(*R^n, очевидно, является ориентируемым псевдомногообразием). R^n является подкомплексом комплекса Q^{n+1} и, так как, в силу предложения D), для $m \leq n$—1 группы $\nabla^m(R^n, G)$ и $\Delta^m(R^n, G)$ изоморфны соответственно группам $\nabla^m(Q^{n+1}, G)$ и $\Delta^m(Q^{n+1}, G)$, то из примера 10 мы получаем, что $\nabla^m(R^n, G) = \Delta^m(R^n, G) = 0$, $m \leq n$—1. Рассмотрим теперь $\nabla^n(R^n, G)$. Воспользовавшись примером 5, мы могли бы притти к заключению, что $\nabla^n(R^n, G)$ изоморфна группе G. Однако значительно проще следующее рассуждение: n-мерные ∇ -циклы комплекса R^n суть просто все n-мерные цепи, и если n-мерная цепь φ комплекса R^n ограничивает, то n-мерная цепь $\Phi = h_{Q^{n+1}}\varphi$ должна быть ∇ -границей и, следовательно, — ∇ -циклом комплекса Q^{n+1} (так как R^n состоит из всех симплексов комплекса Q^{n+1} размерности $\leq n$); наоборот, каждый n-мерный ∇ -цикл комплекса Q^{n+1} ограничивает (согласно примеру 10). Пусть s_i^n $(i = 0, 1, \ldots, n+1)$ — положительно ориентированные n-мерные симплексы ориентированной элементарной n-мерной сферы. Если φ — некоторая n-мерная цепь комплекса R^n , то, для того чтобы цепь Φ была ∇ -циклом комплекса Q^{n+1}, необходимо и достаточно, чтобы $\sum_{i=0}^{n+1} \varphi(s_i^n) = 0$; в соответствии со сделанным выше замечанием (о том, что n-мерные ограничивающие ∇ -циклы комплекса R^n соответствуют n-мерным ∇ -циклам комплекса Q^{n+1}), это равенство является также необходимым и достаточным условием, для того чтобы ∇ -цикл φ ограничивал в R^n. Отсюда следует, что если φ — произвольная n-мерная цепь комплекса R^n, то элемент

$$\sum_{i=0}^{n+1} \varphi(s_i^n) = \sum_{i=0}^{n+1} \Phi(s_i^n)$$

группы G характеризует ∇ -класс, содержащий φ и что, следовательно, $\nabla^n(R^n, G)$ изоморфна группе G. Различные ∇ -классы представляются цепями $gs^n{}_0$,где *g*— некоторый элемент группы *G.* Заметим, что ориентировать R^n значит выбрать один из двух образующих свободной циклической группы $\nabla^n(R^n, g)$. Рассмотрим теперь группу $\Delta^n(R^n, G)$, которая является ничем иным как группой $Z_\Delta^n(R^n, G)$ *n*-мерных Δ-циклов. *n*-мерная цепь φ является Δ-циклом в том и только в том случае, если $h\varphi^{i+1}\varphi$ ограничивает в Q^{n+1} , *в* силу тех же рассуждений, как и в случае ∇ -гомологии. Далее, элементарные (*n*+1)-мерные цепи комплекса Q^{n+1} суть цепи вида gs^{n+1}*;* границы этих цепей являются функциями, принимающими одно и то же значение *g* на всех симплексах s_i^n и значение — *g* на всех симплексах — s_i^n. Это показывает, что $\Delta^n(R^n, G)$ также изоморфна группе *G.*

Барицентрические подразделения

Пусть *K*— комплекс, и полиэдр *P* — его геометрическая реализация. Пусть *P* — *барицентрическое подразделение* полиэдра *P,* т. е. однозначно определенное подразделение *P* на симплексы, вершинами которых являются вершины и центры тяжести клеток полиэдра *P.* Пусть *K′* — комплекс остовов полиэдра *p′.* Тогда *K′* называется *барицентрическим подразделением* комплекса *K. K′* следующим образом может быть определено абстрактно: каждой паре симплексов s^m и — s^m комплекса *K,* т. е. каждому неориентированному симплексу, ставим в соответствие символ $[s^m]$ *;* эти символы $[s^m]$ суть вершины комплекса *K′.* (Геометрической реализацией символа $[s^m]$ является центр тяжести клетки s^m.*)* Неориентированными симплексами комплекса *K′* являются последовательности

$$([s^{m_0}], [s^{m_1}], \ldots, [s^{m_k}]),$$

где $m_0 > m_1 > \ldots > m_k$ и симплекс s^{m_j} является гранью симплекса $s^{m_{j-1}}$.
Из геометрической картины полиэдра *P* ясно, что кажому ориентированному *m*-мерному симплексу комплекса *K* принадлежит (*m*+ 1)! ориентированных *m*-мерных симплексов комплекса *K′*. Абстрактно симплекс

$$([p_0, p_1, \ldots, p_m], [p_1, \ldots, p_m], \ldots, [p_m]),$$

ориентированный как указано, называется ориентированным *m-мерпыы подсимплексом* ориентированного симплекса $(p_1 ,.. .,p_m)$ комплекса *К,* а ориентированный иначе — ориентированным подсимплексом симплекса — $(p_0 ,.. .,p_m)$.

Если $s_1^m \in K'$ есть ориентированный подсимплекс симплекса $s^m \in K$, то мы пишем: $s_1^m < \cdot s^m$.

Если дан комплекс *К,* то в результате итерации процесса барицентрического подразделения получаем последовательность комплексов, называемых *последовательными барицентрическими подразделениями* комплекса *К* и обозначаемых через $K', K'', \ldots, K^{(n)}, \ldots$.

Мы, естественно, говорим об *n*-мерном симплексе $s^{n(m)} \in K^m$, что он является *ориентированным подсимплексом* симплекса $s^n \in K$, если существует цепочка симплексов $s^{n(i)} \in K^{(i)}$, $i = 1, \ldots, m-1$, для которых $s^{n(m)} < \cdot s^{n(m-1)} < \cdot \ldots < \cdot s^{n'} < \cdot s^n$.

Теперь мы докажем, что процесс барицентрического подразделения не изменяет Δ- и ∇ -групп комплекса.

Каждой цепи φ комплекса *К′* поставим в соответствие цепь Φ=πφ комплекса *К,* определенную формулой

$$\Phi\ (s^m) = \sum_{s^{m'} < \cdot s^m} \varphi\ (s_1^{m'}).$$

Легко видеть, что оператор π коммутирует с верхним граничным оператором: $\nabla\pi\varphi = \pi\nabla\varphi$. Следовательно (предложение С)), π порождает гомоморфизм группы $\nabla^m(K', G)$ в группу $\nabla^m(K, G)$. Более того.

Н) *Гомоморфизм группы* $\nabla^m(K', G)$ *в группу* $\nabla^m(K, G)$, *порожденный оператором π, является изоморфизмом группы* $\nabla^m(K', G)$ *на группу* $\nabla^m(K, G)$.

Рассуждения, которыми мы пользовались при доказательстве G), показывают, что Н) содержится в следующем предложения:

I) Пусть *f* —*m*-мерный ∇ -цикл комплекса *К′,* и Ψ — *(m* — 1)-мерная цепь комплекса *К* такая, что $\pi\varphi = \nabla\Psi$. Тогда существует (*m*—1)-мерная цепь ψ комплекса *К′* такая, что $\pi\psi = \Psi$ и $\nabla\psi = \varphi$, $m = 1, 2, \ldots$.

Доказательство предложения I) индукцией по размерности комплекса *К*. Предложение I) очевидно, если dim K=0, ибо в этом случае *К′* = *К*. Мы докажем I) для dim K=*n*>0, предполагая, что I), а, следовательно, и Н), имеют место для любого комплекса размерности ≤*n* — 1. Обозначим через K''^{n-1} подкомплекс комплекса *К,* состоящий из всех симплексов размерности О — 1, и через K''^{-1} барицентрическое подразделение комплекса K^{n-1} (K^{n-1} является подкомплексом комплекса *К′).* Первый шаг будет состоять в вычислении групп:

$$\nabla^m(K' \bmod K^{n-1'}) = \nabla^m(K' \bmod K^{n-1'}, G),\ m = 0, 1, 2, \ldots n.$$

Пусть $\pm s_1^n, \pm s_2^n, \ldots, \pm s_r^n$ суть все n-мерные симплексы комплекса K; пусть Q_i^n, $i = 1, 2, \ldots, r$,— комплекс, состоящий из симплекса s_i^n и всех его граней, R_i^{n-1} — элементарная $(n-1)$-мерная сфера, состоящая из граней симплекса s_i^{n-1} размерности $\leq n-1$, а $Q_i^{n'}$ и $R_i^{n-1'}$ — барицентрические подразделения комплексов Q_i^n и R_i^{n-1}. В силу примера 9, группа $\nabla^m(K' \bmod K^{n-1'})$ есть прямая сумма групп

$$\nabla^m(Q_i^{n'} \bmod R_i^{n-1'}) \quad i = 1, 2, \ldots, r. \tag{3}$$

Так как $Q_i^{n'}$ является конусом над $R_i^{n-1'}$, то $\nabla^m(Q_i^{n'} \bmod R_i^{n-1'})$ изоморфна (пример 8) группе $\nabla^m(R_i^{n-1'})$, которая, по индуктивному предположению, в свою очередь изоморфна группе $\nabla^{m-1}(R_i^{n-1})$ (при гомоморфизме, порожденном оператором π. Из примера 11 вытекает, что $\nabla^m(Q_i^{n'} \bmod R_i^{n-1'})$ есть нулевая группа при $m < n$ и изоморфна группе G при $m = n$; кроме того, ∇ -класс n-мерного ∇ -цикла φ комплекса $Q_i^{n'} \bmod R_i^{n-1'}$ характеризуется элементом $\sum_{s^{n'} < s_i^n} \varphi(s^{n'})$ группа G. Замечая далее, что для произвольной n-мерной, цепи φ комплекса K' соотношение $\pi\varphi = 0$ эквивалентно соотношению $\sum_{s^{n'} < s_i^n} \varphi(s^{n'}) = 0$ для каждого $i = 1, 2, \ldots, r$, получаем следующее утверждение.

J) Все группы $\nabla^m(K' \bmod K^{n-1'})$ при $m<n$ являются нулевыми группами, n-мерный ∇ -цикл φ комплекса $K' \bmod K^{n-1'}$ гомологичен нулю в том и только в том случае, если $\pi\varphi = 0$.

Продолжим теперь доказательство предложения I). Рассмотрим сначала случай $m = n$. Пусть ψ_1 — $(n-1)$-мерная цепь комплекса K' такая, что $\pi\psi_1 = \Psi$ (такая цепь, конечно, может быть найдена, ибо π является гомоморфизмом группы $L^n(K')$ на группу $L^n(K)$. Тогда мы имеем

$$\pi(\varphi - \nabla\psi_1) = \pi\varphi - \nabla\pi\psi_1 = \pi\varphi - \nabla\Psi = 0.$$

В силу второй части J), это означает, что ∇ -цикл

$$h_{K' \setminus K^{n-1'}}(\varphi - \nabla\psi_1)$$

гомологичен нулю, другими словами

$$\varphi - \nabla\psi_1 = \nabla\psi_2,$$

где ψ_2—некоторая $(n-1)$-мерная цепь комплекса K', удовлетворяющая условию: $\psi_2(s^{n-1'}) = 0$, если $s^{n-1'} \in K^{n-1'}$. Отсюда следует, что $\pi\psi_2 = 0$, а это вместе с (4) показывает, что цепь $\psi = \psi_1 + \psi_2$ удовлетворяет требованиям I).

Пусть теперь $m \leq n-1$. Применяя индуктивное предположение к комплексу K^{n-1}, устанавливаем прежде всего существование $(m-1)$-мерной цепи ψ_1 комплекса $K^{n-1'}$ такой, что $\pi\psi_1 = h_{K^{n-1}}\Psi$ и $\nabla\psi_1 = h_{K^{n-1'}}\varphi$. В силу первой части предложения *J)*, предположения F) выполнены, и,

следовательно, можно построить цепь ψ комплекса K', являющуюся продолжением цепи ψ_1 и имеющую φ своей ∇ -границей. Очевидно, ψ удовлетворяет требованиям предложения I). Таким образом, доказательство I), а следовательно, и H), закончено.

Рассмотрим теперь оператор π', ставящий каждой m-мерной цепи Φ комплекса K в соответствие m-мерную цепь $\pi'\Phi$ комплекса K' определенную следующим образом: $\pi'\Phi(s^{m'}) = \Phi(s^m)$, если $s^{m'} < s^m$, и $\pi'\Phi(s^{m'}) = 0$, если $s^{m'}$ неявляется подсимплексом никакого симплекса s^m. Ясно, что π' устанавливает гомоморфизм группы $L^m(K,G)$ в группу $L^m(K',G)$. Легко проверить, что π' коммутирует с граничным оператором: $\pi'\Delta\Phi = \Delta\pi'\Phi$ для любой m-мерной цепи Φ комплекса K, $m \geq 0$.

Следовательно, π' приводит к гомоморфизму группы $\Delta^m(K,G)$ в группу $\Delta^m(K',G)$. По аналогии с H) имеем:

H') *Гомоморфизм группы* $\Delta^m(K,G)$ *в группу* $\Delta^m(K', G)$, *порожденный оператором* π', *является изоморфизмом группы* $\Delta^m(K,G)$ *на группу* $\Delta^m(K',G)$.

Доказательство предложения H') совершенно аналогично доказательству предложения H). Рассуждения, использованные при доказательстве предложения G'), показывают, что H') содержится в I') Пусть Φ — (m — 1)-мерный Δ-цикл комплекса K, и ψ— m-мерная цепь комплекса K' такая, что $\Delta\psi = \pi'\Phi$. Тогда существует m-мерная цепь Ψ комплекса K такая, что $\pi'\Psi$ гомологично ψ и, следовательно, $\Delta\Psi = \Phi$.

Доказательство предложения I'). I') тривиально, если dim K=0. Пусть $\dim K = n > 0$, и предположим, что I') имеет место для всех комплексов размерности $\leq n$ — 1. Используя это индуктивное предположение, мы установим, в полной аналогии с J), предложение:

J') при $m < n$ группы $\Delta^m(K' \bmod K^{n-1'})$ являются нулевыми группами, n-мерная цепь φ комплекса $K' \bmod K^{n-1'}$ является Δ-циклом mod $K^{n-1'}$ в том и только в том случае, если $h_{K'}\varphi$ является образом некоторой цепи комплекса K при гомоморфизме π'.

Доказательство предложения J') мы опускаем. Докажем теперь I') сначала для случая $m = n$. Так как $\pi'\Phi(s^{n-1'}) = 0$, если $s^{n-1'} \notin K^{n-1'}$, то из соотношения $\Delta\psi = \pi'\Phi$ следует, что $h_{K'\setminus K^{n-1}}\psi$ является Δ-циклом $\bmod K^{n-1'}$. В силу второй части J'), отсюда вытекает, что существует цепь Ψ комплекса K такая, что $\pi'\Psi = \psi$.

Пусть теперь $m \leq n$— 1. В силу F') и J'), существует m-мерная цепь ψ_1 комплекса $K^{n-1'}$ такая, что $h_{K'}\psi_1$ гомологично ψ. По индуктивному предположению, примененному к комплексу $K^{n-1'}$, существует

m-мерная цепь Ψ комплекса K^{n-1} такая, что π'Ψ гомологично ψ_1; это значит, что $\pi' h_K \Psi$ гомологично $h_{K'} \psi_1$ и, следовательно, ψ. Таким образом, цепь $h_K \Psi$ удовлетворяет требованиям предложения I'). **Замечание.** H) и H') представляют собой частные случаи основной теоремы комбинаторной топологии (предложение 4 E)): два комплекса, имеющие гомеоморфные геометрические реализации, имеют изоморфные ∇ - и Δ-группы. Именно эта теорема дает ∇ - и Δ-группам, в отличие от групп Δ-циклов, ∇ -циклов, ограничивающих Δ-циклов и т. д., их топологическое значение.

Пример 11: Пусть R^n, $R^{n'}$, $R^{n''}$, . . .—элементарная *n*-мерная сфера и ее последовательные барицентрические подразделения. Из H) и H') следует, что ∇ - и Δ-группы этих комплексов одни и те же, т. е. (см. пример 11) $\nabla^m (R^{n(k)}, G) = \Delta^m (R^{n(k)}, G) = 0$, если $m < n$, и $\nabla^n (R^{n(k)}, G) = \Delta^n (R^{n(k)}, G) = G$.

Можно расширить понятие барицентрического подразделения, определив *барицентрическое подразделение комплекса К по модулю подкомплекса L*. Мы ограничимся геометрическим определением этого вида подразделения в терминах полиэдров *P и Q*, являющихся геометрическими реализациями комплексов *K* и L, предоставляя читателю абстрактную формулировку. Под барицентрическим подразделением полиэдра *P* mod Q, где *Q*— подполиэдр полиэдра P, мы понимаем однозначно определенное подразделение полиэдра P, вершинами которого являются вершины полиэдра *P* и центры тяжести тех клеток полиэдра P, которые не содержатся в *Q* (так что *Q* не подразделяется). Операторы π и π' определяются так же, как и раньше, и предложения H) и H') остаются справедливыми. Ясно, что можно говорить о *последовательных барицентрических подразделениях комплекса Ктой L.*

3.2. Двойственность

Определение 8. Пусть П—группа действительных чисел, приведенных по модулю 1, и G — произвольная (абелева) группа. Гомоморфизм группы G в группу П называется *характером* группы G. Если χ_1 и χ_2 — два характера группы G, то характер χ, определенный равенством

$$\chi(g) = \chi_1(g) + \chi_2(g), \qquad g \in G,$$

называется *суммой* характеров χ_1 и χ_2 . При таком определении характеры группы *G* образуют группу; эта группа называется *группой характеров* группы G и обозначается через G*. Нулевым характером

группы G является гомоморфизм, отображающий все G в нуль группы П .

Замечание. Если G — счетная дискретная группа, то мы вводим в группу G* топологию, следующим образом определяя сходимость: $\chi_n \to \chi$ в G *, если для каждого $g \in G$ в группе П выполнено соотношение: $\chi_n(g) \to \chi(g)$. Оказывается, что полученное таким образом топологическое пространство компактно и обладает счетным базисом; кроме того, группа *непрерывных* характеров группы G* дискретна и изоморфна группе G

G будет обозначать счетную дискретную группу, а ее группа характеров топологизируется в соответствии с только что сделанным замечанием.

Пример 12. Пусть $G = \mathfrak{Z}$. Тогда G* изоморфна П.

Пример 13. Пусть *G*—группа конечного порядка. Тогда G* изоморфна *G*.

Определение 9. Пусть H—подгруппа группы G. Совокупность характеров группы G , отображающих все элементы подгруппы *H* в нуль группы П, образует замкнутую подгруппу группы G *, называемую *аннулятором* подгруппы H. Обратно, для данной подгруппы *J* группы G * совокупность элементов группы G, отображаемых каждым элементом подгруппы *J* в нуль группы П, образует подгруппу группы G, называемую *аннулятором* подгруппы *J*. Следует обратить внимание на то, что аннулятор подгруппы группы G является подгруппой группы G*, аннулятор подгруппы группы G* является подгруппой группы G, и чем меньше подгруппа, тем больше ее аннулятор.

Наиболее важное свойство аннуляторов содержится в следующем предложении A), доказанном Понтрягиным, которое мы приводим здесь без доказательства.

A) а) Пусть H—подгруппа группы G и *J* — замкнутая подгруппа группы G *. Тогда *J* является аинулятором подгруппы H в том и только в том случае, если H является аннулятором подгруппы *J*.

b) В этом случае факторгруппа G^*/J является группой характеров группы H, и J является группой характеров группы G*/H.*

В качестве простого следствия предложения A) мы докажем:

B) Пусть H—подгруппа группы *G*, *L* — подгруппа группы H, *A (H)* — аннулятор подгруппы H, и *A* (L) — аннулятор подгруппы *L*. Тогда *A (L)/A* (H) есть группа характеров группы *H/L*.

Доказательство. Из предложения A) следует, что *A* (L.) является группой характеров факторгруппы *G/L*. Снова применяя A), на этот раз к группе *G/L* и ее подгруппе H/L, получаем наше утверждение.

Нужно отметить также следующее следствие первой части предложения А):

C) Пусть g фиксировано. Если $\chi(g) = 0$ для произвольного χ, то $g = 0$, т. е. аннулятором группы G* является нулевая подгруппа группы G.

Определение 10. Рассмотрим две группы G_1 и G_2 и гомоморфизм h группы G_1 в группу G_2. Каждому характеру χ группы G_2 поставим в соответствие характер $h^*\chi$ группы G_1, согласно формуле

$$\{h^*\chi\}(g_1) = \chi(h(g_1)),$$

где g_1 — произвольный элемент группы G_1. Таким образом, мы получаем гомоморфизм h^* группы G_2^* в группу G_1^*, и этот гомоморфизм h^* называется *сопряженным* с гомоморфизмом h.

D) Пусть G_1 и G_2—две группы, и h— гомоморфизм группы G_1 в группу G_2. Тогда аннулятор подгруппы h (G_1) группы G_2 является ядром гомоморфизма h^*, сопряженного с h. Далее, аннулятор группы h^* (G^*_1) является ядром гомоморфизма h.

Доказательство. Первая часть, непосредственно следует из определения 10, так как $\chi(h(g_1)) = 0$ означает $\{h^*\chi\}(g_1) = 0$. Чтобы доказать вторую часть заметим, что аннулятор подгруппы $h^*(G_2^*)$ состоит из тех элементов $g_1 \in G_1$, для которых $\{h^*\chi\}(g_1) = \chi(h(g_1)) = 0$ для произвольного $\chi \in G_2^*$; но это означает, в силу C), что $h(g_1) = 0$.

E) Не только гомоморфизм h^* определяется гомоморфизмом h, но и гомоморфизм h определяется гомоморфизмом h^*, т. е. соответствие между гомоморфизмом и сопряженным с ним гомоморфизмом взаимно однозначно.

Доказательство. Пусть h и h' — два гомоморфизма группы G_1 в. группу G_2, и h^* и h'^* — сопряженные с ними гомоморфизмы. Рассмотрим гомоморфизм h — h' группы G_1 в группу G_2, определенный, естественным образом, соотношением

$$\{h - h'\}(g_1) = h(g_1) - h'(g_1).$$

Легко видеть, что гомоморфизмом, сопряженным с h—h', является гомоморфизм h^* — h'^*. Предположим теперь, что $h^* = h'^*$. Тогда (см. определение 10) для каждого характера χ группы G_2 и элемента $g_1 \in G_1$

$$0 = \chi(\{h - h'\}(g_1)).$$

Следовательно, в силу C), $\{h - h'\}(g_1) = 0$, т. е. $h = h'$,

Таким образом, различные гомоморфизмы h и h' не могут иметь один и тот же сопряженный с ними гомоморфизм.

F) Для того чтобы гомоморфизм h был гомоморфизмом группы G_1 *на* группу G_2, необходимо и достаточно, чтобы гомоморфизм h^* группы G_1^* в группу G^*_2 был изоморфизмом.

Доказательство. Вспоминая, что гомоморфизм является изоморфизмом в том и только в том случае, если его ядро состоит

лишь из нуля группы, мы видим из первой части D), что h^* является изоморфизмом в том и только в том случае, если аннулятор группы h (G_1) состоит лишь из нуля группы G^*_2. Но в силу A) а) это эквивалентно утверждению, что h (G_1) является аннулятором нулевой подгруппы группы G^*_2, т. е. $h(G_1) = G_2$.

Двойственность между ∇ - и Δ- группами комплекса

Применим теперь аппарат теории характеров для изучения гомологических свойств комплексов. Пусть K— комплекс, G — счетная группа, $n \geq$ — 1. Пусть L^n — группа всех n-мерных цепей комплекса K по области коэффициентов G. Пусть f—характер группы L^n . Пусть s^n_0 — фиксированный n-мерный симплекс комплекса K. Применяя f к элементарным цепям gs^n_0, получаем характер группы G. Обозначим этот характер через $\chi = \chi(s^n_0)$. Тогда, очевидно,

$$\chi(-s^n_0) = -\chi(s^n_0).$$

Следовательно, χ является n-мерной цепью по области коэффициентов G* (определение 2). Можно легко проверить, что таким образом установлен изоморфизм между группой $L^n(K, G^*)$ и группой характеров группы $L^n(K, G)$. Если φ — n-мерная цепь по области коэффициентов G и ψ — n-мерная цепь по области коэффициентов G*, то элемент ψ (φ) группы П, в который отображается элемент $\varphi \in L^n(K, G)$ характером ψ, может быть вычислен следующим образом: пусть $= s^n_i$, i=1,.. .,k, —ориентированные n-мерные симплексы комплекса K. Тогда

$$\psi(\varphi) = \sum_{i=1}^{k} \{\psi(s^n_i)\}(\varphi(s^n_i)). \quad (1)$$

Теперь мы утверждаем, что если φ* — (n+ 1)-мерная цепь по области коэффициентов G*, и φ — n-мерная цепь по области коэффициентов G, то

$$\{\Delta\varphi^*\}(\varphi) = \{\varphi^*\}(\nabla\varphi). \quad (2)$$

Легко проверить (2), используя (1). Формула (2) показынает, что гомоморфизм Δ группы $L^{n+1}(K, G^*)$ в группу L^n (K, G) является сопряженным (определение 10) с гомоморфизмом ∇ группы $L^n(K,G)$ в группу $L^{n+1}(K,G)$. Используем теперь предложение D). Заметим, что ядром гомоморфизма Δ является группа $Z^{n+1}_\Delta(K, G^*)$ n-мерных Δ-циклов, ядром гомоморфизма ∇ является группа $Z^n_\nabla(K, G)$ n-мерных ∇ -циклов, образом группы $L^{n+1}(K, G^*)$ при гомоморфизме Δ является группа $H^n_\Delta(K,G^*)$ n-мерных ограничивающих Δ-циклов и образом $L^n(K,G)$ при

гомоморфизме ∇ является группа $H^{n+1}_{\nabla}(K, G)$ $(n+1)$-мерных ограничивающих ∇-циклов. Получаем результат (понижая в а) размерность на единицу):

G) а) $Z^n_{\Delta}(K,G^*)$ является аннулятором группы $H^n_{\nabla}(K, G)$,
b) $Z^n_{\nabla}(K,G)$ является аннулятором группы $H^n_{\Delta}(K, G^*)$.

Пользуясь предложением А), заменим b) на b') $H^n_{\Delta}(K,G^*)$ является аннулятором группы $Z^n_{\nabla}(K,G)$.

Из G) получаем следующий результат:

H) $\Delta^n(K, G^*)$ *является группой характеров группы* $\nabla^n(K,G)$. *В частности,* $\Delta^n(K, \Pi)$ *является группой харатеров группы* $\nabla^n(K,\mathfrak{Z})$.

Доказательство. Применяем предложения В) и G), взяв $L^n(K,G)$ в качестве G, $Z^n_{\nabla}(K,G)$ — в качестве H и $H^n_{\nabla}(K,G)$ — в качестве L.

Замечание 1. Пользуясь аналогичными методами, можно доказать, что $\nabla^n(K,G^*)$ является группой характеров группы $\Delta^n(K,G)$.

Замечание 2. $\Delta^n(K,G^*)$ и $\nabla^n(K,G^*)$, как группы характеров счетных групп $\nabla^n(K,G)$ и $\Delta^n(K,G)$, являются компактными группами со счетным базисом. В действительности, топология в этих группах может быть определена непосредственно с помощью топологии группы G*, без использования групп $\nabla^n(K,G)$ и $\Delta^n(K,G)$.

Замечание 3. В теории связности конечных комплексов теория ∇-гомологий не обладает никакими специальными преимуществами по сравнению с теорией Δ-гомологий. Положение, однако, меняется, когда мы переходим к теории связности более общих пространств, где существует преимущество в употреблении ∇-гомологии.

3.3. Симплициальные отображения комплексов

Определение 11. Мы говорим, что f есть *симплициальное* отображение комплекса K_1 в комплекс K_2, если f каждой вершине $p \in K_1$ ставит в соответствие некоторую вершину $f(p) \in K_2$, таким образом, что для любого n-мерного симплекса комплекса K_1 с вершинами $p_0, \ldots, p_n$ точки $f(p_0), \ldots, f(p_n)$ являются вершинами некоторого симплекса комплекса K_2 (размерности, возможно, меньшей, чем n). Пусть $s^n = (p_0, \ldots, p_n)$ — ориентированный n-мерный симплекс комплекса K_1 и f — симплициальное отображение комплекса K_1 в комплекс K_2. Если все вершины $f(p_0), \ldots, f(p_n)$ различны, то через $f(s^n)$ обозначаем ориентированный n-мерный симплекс $(f(p_0), \ldots, f(p_n))$; в противном случае $f(s^n)$ не определено.

А) Пусть f — симплициальное отображение комплекса K_1 в комплекс K_2. Тогда f порождает гомоморфизм hf группы

$\nabla^n(K_2)$ в группу $\nabla^n(K_1)$, определяемый следующим образом:
Если дана n-мерная цепь ψ комплекса K_2, то мы ставим ей в соответствие n-мерную цепь $\varphi = hf\psi$ комплекса K_1 полагая

$$\varphi(s^n) = \psi(f(s^n)), \text{ если } f(s^n) \text{ определено,}$$

$$\varphi(s^n) = 0 \text{ в противном случае.}$$

Очевидно, hf является гомоморфизмом группы $L^n(K_2)$ в группу $L^n(K_1)$. Легко проверить, что hf коммутирует с ∇. Следовательно (предложение 1С)), hf приводит к гомоморфизму группы $\nabla^n(K_2)$ в группу $\nabla^n(K_1)$, который мы продолжаем обозначать символом hf.

В) Аналогично, симплициальное отображение n комплекса K_1 в комплекс K_2 порождает гомоморфизм hf группы $\Delta^n(K_1)$ в группу $\Delta^n(K_2)$, определяемый следующим образом: пусть z^n —произвольный n-мерный симплекс комплекса K_2. Тогда произвольной n-мерной цепи φ комплекса K_1 мы ставим в соответствие n-мерную цепь $\psi = h_f\varphi$ комплекса K_2, полагая

$$\psi(z^n) = \sum_{f(s^n)=z^n} \varphi(s^n).$$

Очевидно, h_f является гомоморфизмом группы $L^n(K_1)$ в группу $L^n(K_2)$. Легко видеть, что h_f коммутирует с Δ. Следовательно (предложение 1 С)), h_f приводит к гомоморфизму группы $\Delta^n(K_1)$ в группу $\Delta^n(K_2)$, который мы продолжаем обозначать через h_f.

Легко доказать, что если f — симплициальное отображение комплекса K_1 в комплекс K_2 и g—синплициальное отображение комплекса K_2 в комплекс K_3, то гомоморфизм h_{gf} группы $\Delta^n(K_1)$ в $\Delta^n(K_3)$, порожденный отображением gf, равен произведению $h_g h_f$. Аналогичное утверждение о транзитивности имеет место для гомоморфизмов ∇-групп, порожденных симплнциальными отображениями.

С) Пусть G^* — группа характеров группы G, f—симплициальное отображение комплекса K_1 в комплекс K_2, h_f — гомоморфизм группы $L^n(K_1, G^*)$ в группу $L^n(K_2, G)$, порожденный отображением f, и h^f— гомоморфизм группы $L^n(K_2,G)$ в группу $L^n(K_2,G)$, порожденный отображением f. Мы утверждаем, что h_f является сопряженным с h^f гомоморфизмом (определение 10), т. е. если ψ—n-мерная цепь комплекса K_1 по области коэффициентов G^* и φ — n-мерная цепь комплекса K_2 по области коэффициентов G, то

$$\{h_f\psi\}(\varphi) = \psi(h^f(\varphi)).$$

Простое доказательство этого факта предоставляется читателю.

D) Пусть G — счетная группа, и G^* — ее группа характеров. Вспомним, что группы $\Delta^n(K_1, G^*)$ и $\Delta^n(K_2, G^*)$ являются группами характеров групп $\nabla^n(K_1, G)$ и $\nabla^n(K_2, G)$. Если в соотношении (1) ψ

будет Δ-циклом, а φ— ∇ -циклом, то мы видим, что гомоморфизм h_f группы $\Delta^n(K_1, G^*)$ в группу $\Delta^n(K_2, G^*)$ является сопряженным с гомоморфизмом h^f группы $\nabla^n(K_2, G)$ в группу $\nabla^n(K_1, G)$. Используя аналогичные» методы, можно доказать, что f порождает сопряженные друг с другом гомоморфизмы: группы $\nabla^n(K_2, G^*)$ в группу $\nabla^n(K_1, G^*)$ и группы $\Delta^n(K_1, G)$ в группу $\Delta^n(K_2, G)$.

Пример 14. Пусть L — подкомплекс комплекса K, и f—-тождественное отображение L в K. Тогда f есть симплициальное отображение. Гомоморфизмы ∇ -групп комплекса K в ∇ -группы комплекса L и Δ-групп комплекса L *в* Δ-группы комплекса K, порожденные n, являются, конечно, естественными гомоморфизмами этих групп (определение 7).

Пример 15. Пусть K—комплекс, и K' — его барицентрическое подразделение. Ранее мы определили гомоморфизм π групп n-мерных цепей комплекса K в группу n-мерных цепей комплекса K, и мы знаем (предложение 1 Н)), что этот гомоморфизм порождает изоморфизм группы $\nabla^n(K', G)$ на группу $\nabla^n(K, G)$.

Поставим теперь каждой вершине p комплекса K' в соответствие определенную, хотя и произвольно выбранную, вершину того симплекса комплекса K, центром тяжести которого является p. Тогда мы получим симплициальное отображение f комплекса K' в K и, следовательно, гомоморфизм h^f (см. А)) группы $L^n(K, G)$ в $L^n(K',G)$, Легко видеть, что πh^f есть тождественное отображение группы $L^n(K$, G) на себя. Следовательно, πh^f в применении к группе $\nabla^n(K, G)$ также является тождественным отображением, и так как, что было замечено выше, π является изоморфизмом группы $\nabla^n(K', G)$ на $\nabla^n(K, G)$, то также должно быть верным, что h^f является изоморфизмом группы $\nabla^n(K, G)$ на $\nabla^n(K', G)$. Таким образом, получаем результат, состоящий в том, что гомоморфизм группы $\nabla^n(K, G)$ в группу ∇^n (K, G), порожденный отображением n, является, в действительности, изоморфизмом группы $\nabla^n(K, G)$ на группу $\nabla^n(K', G)$, обратным гомоморфизму π.

Подобным образом можно показать, что гомоморфизм группы $\Delta^n(K', G)$ в группу , $\Delta^n(K, G)$, порожденный отображением f, является, в действительности, изоморфизмом группы $\Delta^n(K', G)$ на $\Delta^n(K$, G), обратным изоморфизму π'. Кроме того, если G — счетная группа, то изоморфизм π' группы $\nabla^n(K, G^*)$ на $\nabla^n(K', G^*)$ является сопряженным гомоморфизмом с изоморфизмом π группы $\nabla^n(K', G)$ на $\nabla^n(K, G)$.

Пример 16. Пусть K—комплекс,и R — элементарная n-мерная сфера (см. пример 11) или одно из ее последовательных барицентрических подразделений. Пусть f—симплициальное отображение K в R, и $\mathfrak{Z}$— группа целых чисел. Тогда мы знаем, что $\nabla^n(R, \mathfrak{Z})$ есть свободная

циклическая группа, образующий элемент которой определяется ориентацией сферы. Следовательно, гомоморфизм h^f группы $\nabla^n(R, \mathfrak{Z})$ в группу $\nabla^n(K, \mathfrak{Z})$ вполне определяется элементом $h^f(\alpha)$, соответствующим образующему элементу α группы $\nabla^n(R, \mathfrak{Z})$; этот элемент называется *степенью* отображения *f*. Степень отображения *f* легко вычисляется. Пусть z_0^n — произвольный положительно ориентированный *n*-мерный симплекс, т. е. ориентированный подсимплекс одного из ориентированных симплексов элементарной *n*-мерной сферы, *n*-мерной сферы *R*.

В силу примера 11, элементарная цепь $1 \cdot z^n_0$ представляет ∇-класс, являющийся образующим элементом группы $\nabla^n(R, \mathfrak{Z})$. Тогда степень отображения *f* есть ∇-класс следующей цепи φ комплекса *K:* для каждого *n*-мерного симплекса s^n комплекса *K:*

$$\varphi(s^n) = 1, \quad \text{если } f(s^n) = z_0^n,$$
$$\varphi(s^n) = -1, \quad \text{если } f(s^n) = -z_0^n,$$
$$\varphi(s^n) = 0 \quad \text{в иных случаях.}$$

Если *K* также является элементарной *n*-мерной сферой или одним из ее барицентрических подразделений, то степень отображения *f* можно рассматривать как целое число, именно $\sum \varphi(s^n)$, где сумма берется по всем положительно ориентированным симплексам s^n; интуитивно степень показывает, сколько раз симплекс z^n_0 положительно накрывается при отображении *f*, причем для каждого целого числа *m* легко построить отображение *f* степени *m*. .

Пусть *K*—снова произвольный комплекс; рассмотрим группу характеров П группы $\mathfrak{Z}$. Симплициальное отображение *f* порождает гомоморфизм h_f группы $\Delta^n(K, \Pi)$ в группу $\Delta^n(R, \Pi)$. Мы знаем, что $\Delta^n(R, \Pi)$ изоморфна П, так что h_f является характером группы $\Delta^n(K, \Pi)$. Но h_f является сопряженным с h^f гомоморфизмом, и, следовательно, каждый из этих гомоморфизмов определяется другим. Таким образом, степень отображения *f* можно с таким же успехом определять с помощью этого характера группы $\Delta^n(K, \Pi)$.

Замечание. Точно таким же образом можно определить степень отображения для более общего случая симплициального отображения *n*-мерного комплекса в произвольное *n*-мерное ориентируемое псевдомногообразие (см. пример 5). Если комплекс *K* и сам является ориентируемым псевдомногообразием, то степень отображения можно рассматривать как целое число.

Е) Пусть *f* и *g*— два симплициальные отображения комплекса K_1 в комплекс K_2, обладающие следующим свойством: для каждого симплекса *s* комплекса K_1 симплексы *f(s)* и *g(s)* являются гранями

некоторого симплекса комплекса K_2. Тогда f и g порождают одни и те же гомоморфизмы Δ-групп комплекса K_1 в Δ-группы комплекса K_2 и ∇-групп комплекса K_2 в ∇-группы комплекса K_1.

Доказательство. Достаточно доказать Е) для ∇-групп, так как при заданной счетной группе G гомоморфизм группы $\nabla^n(K_2, G^*)$ в группу $\nabla^n(K_1, G^*)$, порожденный отображением f, является сопряженным (предложение D)) с гомоморфизмом группы $\Delta^n(K_1, G)$ в $\Delta^n(K_2, G)$, порожденным n, и, следовательно (предложение 2 Е)), второй определяется первым. Случай произвольной группы G, взятой в качестве области коэффициентов, сводится к случаю счетной группы с помощью следующего простого замечания. Пусть φ — n-мерный Δ-цикл комплекса K_1 по области коэффициентов G, и пусть G_1— счетная подгруппа группы G, порожденная элементом $\varphi(s^n)$. Тогда $h_f\varphi$ и $h_g\varphi$ можно рассматривать как n-мерные Δ-циклы по области коэффициентов G_1, и если они гомологичны по отношению к области коэффициентов G_1, то тем более они будут гомологичны по отношению к области коэффициентов G.

Можно поедположить, что g и f отличаются друг от друга лишь на одной вершине комплекса K_1, так как g можно получить из f с помощью конечного числа изменений, каждая из которых состоит в перемене образа лишь одной вершины. В соответствии с этим пусть p— единственная вершина, для которой $f(p)$ отлично от $g(p)$. Поставим в соответствие каждой n-мерной цепи φ комплекса K_2 (n— 1)-мерную цепь φ^* комплекса K_1 следующим образом: пусть

$s^{n-1} = (p_0,\ldots,p_{n-1})$ - (n — 1)-мерный симплекс комплекса K_1 имеющий p своей вершиной, и такой, что симплекс $f(s^{n-1})$ определен и не имеет $g(p)$ своей вершиной. Пусть z^n — n-мерный симплекс комплекса K_2, вершинами которого являются вершины симплекса $f(s^{n-1})$ и вершина $g(p)$, а ориентация определяется следующим образом:

$$z^n = (g(p), f(p_0), \ldots, f(p_{n-1})).$$

Тогда мы полагаем: $\varphi^*(s^{n-1}) = \varphi(z^n)$. Для всех ($n$—1)-мерных симплексов s^{n-1} комплекса K_2, не обладающих указанными свойствами, полагаем: $\varphi^*(s^{n-1}) = 0$. Обозначим через h^f и h^g гомоморфизмы группы $L^n(K_2, G)$ в группу $L^n(K_1, G)$, порожденные соответственно отображениями f и g. Тогда простые вычисления показывают, что

$$\nabla\varphi^* = -(\nabla\varphi)^* + h^f\varphi - h^g\varphi.$$

Если φ— ∇-цикл, то первый член в правой части равенства равен нулю, так что эта формула устанавливает ∇-гомологию между $h^f\varphi$ и $h^g\varphi$ и доказывает предложение.

3.4. Δ- и ∇ -группы компактов

Одним из достижений топологии было перенесение теории гомологии с полиэдров на произвольные компакты. Это —один из наиболее существенных шагов в установлении связи между комбинаторными и теоретико-множественными методами в топологии.

ПРЯМЫЕ И ОБРАТНЫЕ СПЕКТРЫ

Пусть $\sum$ — *частично упорядоченное множество,* т. е. множество, в котором существует транзитивное соотношение < , определенное для некоторых (не обязательно всех) пар элементов. Подмножество $\sum'$ множества $\sum$ называется *конфинальным* множеству $\sum$, если для любого элемента $\sigma \in \Sigma$ существует элемент $\sigma' \in \Sigma'$, следующий за σ: $\sigma < \sigma'$. $\sum$ называется *направленным* множеством, если для любой пары элементов $\sigma_1, \sigma_2 \in \Sigma$ существует элемент $\rho \in \Sigma$, следующий, как за σ, так и за

$$\tau : \sigma < \rho \text{ и } \tau < \rho.$$

Определение 12. Пусть $\sum$ — направленное частично упорядоченное множество. Пусть каждому элементу $\sigma \in \Sigma$ поставлена в соответствие группа G_σ, а каждой паре элементов $\sigma < \tau$ соответствует гомоморфизм $h^{\sigma\tau}$ группы G_σ в группу G_τ,причем предполагается, что если $\rho < \sigma < \tau$, то всегда

$$h^{\rho\tau} = h^{\sigma\tau}(h^{\rho\sigma}).$$

Такая система групп называется *прямым спектром.* Для данного прямого спектра мы определим теперь новую группу, называемую *предельной группой G прямого спектра.* Элемент g_σ группы G_σ называется *эквивалентным* элементу g_τ группы G_τ, если существует такое $\rho \in \Sigma$, что $\sigma < \rho, \tau < \rho$ и в группе G_ρ: $h^{\sigma\rho} g_\sigma = h^{\tau\rho} g_\tau$. Класс, состоящий из всех элементов, эквивалентных некоторому элементу g_σ группы G_σ , считается, по определению, *элементом* группы G, а g_σ называется *представителем* в группе G_σ этого элемента предельной группы. *Сложение* в G определяется следующим образом. Пусть g_σ, и g_τ — представители двух элементов и g' группы G, соответственно в группах G_σ и G_τ, и пусть ρ — такой элемент множества $\sum$, что $\sigma < \rho$ и $\tau < \rho$. Тогда под *суммой* элементов g и g' понимаем элемент группы G, определенный элементом $h^{\sigma\rho} g_\sigma + h^{\tau\rho} g_\tau$ группы G_ρ.
(Нулем группы G является класс, содержащий нули каждой группы G_σ.)

Пример 17. Пусть G_1, G_2, ...—последовательность групп, причем G_n является подгруппой группы G_{n+1} при каждом n, и пусть h^{mn}, $m \leqslant n$,— тождественное отображение G_m в G_n. Тогда, очевидно, $\{G_n\}$ является прямым спектром, а предельная группа изоморфна теоретико-множественной сумме групп G_n.

Пример 18. Пусть $\{G_\sigma\}$ —прямой спектр такой, что $h^{\sigma\tau}$ является изоморфизмолг группы G_σ на группу G_τ при любых $\sigma < \tau$. Тогда предельная группа спектра $\{G_\sigma\}$ изоморфна каждой группе G_σ.

А) Пусть $\sum$ — направленное частично упорядоченное множество, $\sum'$ — его конфинальное подмножество, и $\{G_\sigma\}, \sigma \in \Sigma$, прямой спектр. Тогда подсистема системы $\{G_\sigma\}$, соответствующая элементам $\sigma \in \Sigma'$, авляется прямым спектром, и легко видеть, что предельная группа этого спектра изоморфна предельной группе всего спектра.

В) Если в прямом спектре $\{G_\sigma\}$, $\sigma \in \Sigma$, каждая группа G_σ счетна и если $\sum$ имеет счетное конфинальное подмножество, то предельная группа прямого спектра счетна.

Определение 13. Пусть $\sum$ — направленное частично упорядоченое множество. Пусть каждому элементу $\sigma \in \Sigma$ поставлена в соответствие группа G_σ, а каждой паре элементов $\sigma < \tau$ соответствует гомоморфизм $h_{\tau\sigma}$ группы G_τ в группу G_σ, причем предполагается, что если $\rho < \sigma < \tau$, то всегда

$$h_{\tau\rho} = h_{\sigma\rho} \cdot (h_{\tau\sigma}).$$

Такая система групп называется *обратным спектром.* Определим теперь *предельную группу G обратного спектра. (* Если $\{G_\sigma\}$ — обратный спектр, в котором группы G_σ —-топологические и гомоморфизмы $h_{\tau\sigma}$ являются непрерывными, то предельную группу спектра $\{G_\sigma\}$ можно рассматривать как топологическую группу; но если $\{G_\sigma\}$ —прямой спектр, то в предельной группе спектра нельзя ввести топологию, исходя из топологии групп G_σ.)

Элементом группы G, по определению, считается система элементов $\{g_\sigma\}$, содержащая в точности по одному элементу g_σ из каждой группы G_σ, и обладающая тем свойством, что если $\sigma < \tau$, то $h_{\tau\sigma} g_\tau = g_\sigma; g_\sigma$ называется *представителем* этого элемента предельной группы в группе G_σ. (Таким образом, элемент группы G имеет в точности по одному представителю в каждой группе G_σ, но различные элементы группы G могут иметь одного и того же представителя в некоторой группе G_σ *).Сложение* в G определяется следующим образом: пусть g и g' — два элемента группы G, а g_σ и g'_σ — их представители в группе G_σ. Тогда под суммой элементов g и g' понимаем систему элементов $\{g_\sigma + g'_\sigma\}$; (нулем предельной группы является система элементов, состоящая из нулей всех групп G_σ).

Пример 19. Пусть $G_1, G_2, \ldots$—последовательность групп, причем G_{n+i} является подгруппой группы G_n при каждом n, и пусть $h_{n,m}, m \leqslant n$,— тождественный гомоморфизм G_n в G_m. Тогда последовательность $\{G_n\}$ является, очевидно, обратным спектром, а предельная группа изоморфна пересечению групп G_n.

Пример 20. Пусть $\{G_\sigma\}$ — обратный спектр такой, что $h_{\tau\sigma}$ является изоморфизмом группы G_τ на группу G_σ при любых $\sigma < \tau$. Тогда предельная группа спектра $\{G_\sigma\}$ изоморфна каждой группе G_σ.

C) Пусть $\sum$—направленное частично упорядоченное множество, и $\sum'$ — его конфинальное подмножество. Пусть $\{G_\sigma\}$, $\sigma \in \Sigma$, - обратный спектр. Тогда подсистема системы $\{G_\sigma\}$, соответствующая элементам $\sigma \in \Sigma'$, является, очевидно, обратным спектром, и легко видеть, что предельная группа этого спектра изоморфна предельной группе всего спектра.

(Было бы ошибочно сделать из C) заключение, по аналогии с B), что если каждая группа G_σ счетна и $\sum$ имеет счетное конфинальное подмножество, то предельная группа обратного спектра счетна.)

D) Пусть $\{G_\sigma\}$ — прямой спектр с гомоморфизмами $h^{\sigma\tau}, \sigma < \tau$. Пусть G_σ * —группа характеров группы G_σ и $h^*_{\sigma\tau}$— гомоморфизм группы G_τ* в группу G_σ*, сопряженный с гомоморфизмом $h^{\sigma\tau}$ (см. определение 10). Тогда: a) группы $\{G_\sigma\}$ с гомоморфизмами $h^*_{\sigma\tau}, \sigma < \tau$ образуют обратный спектр; b) если G— предельная группа спектра $\{G_\sigma\}$, то ее группа характеров G* является предельной группой спектра $\{G^*_\sigma\}$.

Доказательство. Предоставив доказательство утверждения a) читателю, приступим к доказательству утверждения b). Согласно определению предельной группы прямого спектра, каждый элемент $g_\sigma \in G_\sigma$ определяет некоторый элемент группы G, и это соответствие, очевидно, является гомоморфизмом h^σ группы G_σ в группу G. Далее, если $\sigma < \tau$,

$$h^\sigma = h^\tau h^{\sigma\tau}. \quad (1)$$

Пусть теперь χ — характер группы G. При фиксированном σ χh^σ является характером группы G_σ, т. е. элементом g^*_σ группы G^*_σ. Далее, из (1) следует, что если $\sigma < \tau$, то

$$h^*_{\sigma\tau} g^*_\tau = g^*_\sigma, \quad \textbf{(2)}$$

и, следовательно, система элементов $\{g^*_\sigma\}$ является элементом предельной группы обратного спектра $\{G^*_\sigma\}$. Таким образом, каждому характеру группы G соответствует некоторый элемент предельной группы обратного спектра $\{G^*_\sigma\}$. Нетрудно довести это доказательство до конца, показав, что это соответствие является изоморфизмом группы G* на предельную группу спектра $\{G^*_\sigma\}$.

Пример 20. 1. Пусть каждая группа G_i, i=1, 2, ... есть группа $\mathfrak{Z}$ целых чисел, и пусть h^{ik}, $i \leqslant k$,—гомоморфизм группы G_i в группу G_k, задаваемый формулой:

$$h^{ik}(m) = m \cdot 2^{k-i}, \quad m \in \mathfrak{Z}. \quad \textbf{(3)}$$

Тогда группы $\{G_i\}$ с этими гомоморфизмами образуют прямой спектр, предельная группа G которого, как легко видеть, изоморфна группе двоично-рациональных чисел, т. е. группе дробей вида $\frac{m}{2^n}$.В силу D), группа характеров G* группы G является предельной группой обратного спектра $\{G_i\}$ - где группы G^*_i изоморфны группе П, а гомоморфизмы h^*_{ik}, $i \leqslant k$, задаются формулой:

$$h^*_{ik}(r) = r \cdot 2^{k-i}, \quad r \in \Pi.$$

G* называется (ван Данциг) *диадическим соленоидом*; как группа характеров счетной группы G она компактна и по своей топологической структуре является неразложимым одномерным континуумом.

Если в качестве группы G_i, i= 1, 2, ..., взять группу целых чисел, приведенных по модулю 2, а гомоморфизмы снова определить по формуле (3), то G будет группой двоично-рациональных чисел, приведенных *по* модулю 1, а ее группа характеров G* — диаднческой нульмерной группой.

Δ- и ∇ -группы компактов

Рассмотрим теперь произвольный компакт X. Пусть $\sum$ — множество всех покрытий пространства X, частично упорядоченное следующим образом: σ предшествует τ, $\sigma < \tau$, если покрытие τ вписано в покрытие σ. $\sum$ направлено, так как для любых двух покрытий σ и τ существует покрытие, вписанное и в σ и в τ (состоящее из попарных пересечений элементов покрытия σ с элементами покрытия τ). Для каждого покрытая σ рассмотрим его нерв N (σ) . Пусть $\sigma < \tau$, т. е. покрытие τ вписано в покрытие σ. Для каждого элемента покрытия τ выберем некоторый, содержащий его элемент покрытия σ. Получаем симплициальное отображение комплекса N(τ) в комплекс $N(\sigma)$, называемое проекцией $N(\tau)$ в $N(\sigma)$. Вообще говоря, существует много проекций $N(\tau)$ в $N(\sigma)$, соответствующих различному выбору элемента покрытия σ, содержащего каждый данный элемент покрытия τ : но любые две проекции p и p' удовлетворяют условию предложения 3 Е) и, следовательно, порождают один и тот же гомоморфизм $h^{\sigma\tau}$ группы $\nabla^n(N(\sigma), G)$ в группу $\nabla^n(N(\tau), G)$ и один и тот же гомоморфизм $h_{\sigma\tau}$

группы $\Delta^n(N(\tau), G)$ в группу $\Delta^n(N(\sigma), G)$. Если $\rho < \sigma < \tau$ и p является проекцией $N(\tau)$ в $N(\sigma)$, а q — проекцией $N(\sigma)$ в $N(\rho)$, то qp является проекцией $N(\tau)$ в $N(\rho)$, следовательно, $h^{\rho\tau} = h^{\sigma\tau}h^{\rho\sigma}$, и $h_{\tau\rho} = h_{\sigma\rho}h_{\tau\sigma}$. Таким образом, при любом выборе целого числа n и области коэффициентов G, группы $\{\nabla^n(N(\sigma), G)\}$ с гомоморфизмами $h^{\sigma\tau}$ образуют прямой спектр, а группы $\{\Delta^n(N(\sigma), G)\}$ с гомоморфизмами $h_{\sigma\tau}$ образуют обратный спектр.

Определение 14. Пусть X—компакт, G — группа, n — целое число ≥ 0, и $\{\sigma\}$ — совокупность покрытий пространства X *n-меркой* Δ *группой* $\Delta^n(X,G)$ *по области коэффициентов G* (если невозможны недоразумения, мы пишем просто $\Delta^n(X)$) называется предельная группа обратного спектра $\{\Delta^n(N(\sigma), G)\}$, и *n-мерной* ∇ *-группой по области коэффициентов G* (если невозможны недоразумения, мы пишем $\nabla^n(X)$) называется предельная группа прямого спектра $\{\nabla^n(N(\sigma), G)\}$.

Замечание 1. Ясно, что определение 14 является топологически инвариантный, т. е. гомеоморфные пространства имеют одни и те же Δ- и ∇ -группы.

Замечание 2. При изучении комплексов мы замечали полную симметрию между Δ- и ∇ -группами. Но здесь, в теории гомологии компактов, эта симметрия исчезает, так как Δ-группы являются предельными группами *обратных* спектров, а ∇ -группы являются предельными группами *прямых* спектров.

Топологическая инвариантность комбинаторных Δ- и ∇ -групп

Е) Пусть X—полиэдр, K—его комплекс остовов $\nabla^n(X)$ и $\Delta^n(X)$ — (топологически инвариантные) ∇ - и Δ-группы пространства X, а $\nabla^n(K)$ и $\Delta^n(K)$— (комбинаторные) ∇ - и Δ-группы комплекса K. Тогда *группа* $\nabla^n(X)$ *изоморфна группе* $\nabla(K)$, *а группа* $\Delta^n(X)$ *изоморфна группе* $\Delta^n(K)$. *Следовательно, если* K_1 *и* K_2 *— комплексы остовов гомеоморфных полиэдров, то они имеют изоморфные* ∇ *и* Δ*-группы.*

Доказательство. Пусть $K=K_{(0)}$, и обозначим через $K_{(1)}, K_{(2)},\ldots$. последовательные барицентрические подразделения комплекса K. Каждому комплексу $K_{(i)}$ соответствует подразделение X_i полиэдра X на симплексы. Пусть σ_i — покрытие пространства X, состоящее из звезд вершин подразделения X_i. Заметим, что нервом $N(\sigma_i)$ покрытия σ_i является комплекс $K_{(i)}$. Последовательность $\{\sigma_i\}$ конфинальна

множеству всех покрытий, так как диаметр покрытия σ_i стремится к нулю. Следовательно, в силу *A)* и C), группы $\nabla^n(X)$ и $\Delta^n(X)$ являются предельными группами, соответственно, прямого и обратного спектров $\{\nabla^n(N(\sigma_i))\}$ и $\{\Delta^n(N(\sigma_i))\}$. Но проекция f комплекса $N(\sigma_{i+1})$ в комплекс $N(\sigma_i)$ является симплициальным отображением $K_{(i+1)}$ в $K_{(i)}$ типа, рассмотренного в примере 15, и, следовательно, порождает изоморфизмы группы $\nabla^n(N(\sigma_i))$ на группу $\nabla^n(N(\sigma_{i+1}))$ и группы $\Delta^n(N(\sigma_{i+1}))$ на группу $\Delta^n(N(\sigma_i))$.

Из примеров 18 и 20 поэтому следует, что предельные группы, т. е. $\nabla^n(X)$ и $\Delta^n(X)$, изоморфны группам $\nabla^n(\mathrm{K})$ и $\Delta^n(\mathrm{K})$.

Замечание. Пусть φ — ∇-цикл комплекса K, и φ^* — ∇-цикл одного из последовательных барицентрических подразделений $K_{(m)}$ комплекса K (или, более общо, одного из последовательных барицентрических подразделений комплекса K по модулю некоторого подкомплекса) и пусть φ и φ^* связаны между собой следующим образом: если s^n — ориентированный симплекс комплекса K, то

$$\varphi(s^n) = \sum \varphi^*(s^n_{(m)}), \quad (5)$$

где сумма распространена по всем ориентированным подсимплексам $s^n_{(m)}$ симплекса s^n. Это означает, что φ^* получается из φ m-кратным применением оператора π. Тогда из предшествовавших рассмотрений следует, что φ и φ^* представляют один и тот же элемент группы $\nabla^n(\mathrm{P})$, где P обозначает геометрическую реализацию комплекса K.

Пример 21. Пусть I_n — n-мерный куб. Так как I_n гомеоморфен n-мерному симплексу, то $\nabla^m(I_n, G) = \Delta_m(I_n, G) = 0$ для всех m, в силу примера 10.

Пример 22. Пусть S_n—n-мерная сфера. Так как сфера S_n гомеоморфна полиэдру, состоящему из всех собственных граней $(n+1)$-мерного симплекса, то из примера 11 легко следует, что если $m < n$, то и $\nabla^m(S_n, G)$ и $\Delta^m(S_n, G)$ суть нулевые группы, тогда как и группа $\nabla^n(S_n, G)$, и группа $\Delta^n(S_n, G)$ изоморфны группе G.

Пример 22. 1. Пусть компакт X является суммой замкнутых непересекающихся множеств X_1 и X_2. Если $n > 0$, то $\nabla^n(X)$ есть прямая сумма групп $\nabla^n(X_1)$ и $\nabla^n(X_2)$, и $\Delta^n(X)$ есть прямая сумма групп $\Delta^n(X_1)$ и $\Delta^n(X_2)$. Это легко следует из примера 6.

F) ПустьX—компакт. Тогда, если $n > \dim X$, то $\Delta^n(X) = \nabla^n(X) = 0$.

Доказательство. В силу приведенной ранее теоремы, множество покрытий пространства X, имеющих порядок $\leq n$, составляет конфинальиое подмножество множества всех покрытий пространства X. Сопоставление этого обстоятельства с предложениями A) и C) и примером 2 доказывает предложение F).

Двойственность между Δ- и ∇ -группами компакта

Пусть теперь *G* — счетная группа. Заметим, что множество *{σ)* всех покрытии компакта *X* содержит счетное конфинальное подмножество, ибо, если σ_i — некоторое покрытие диаметра $\leq \frac{1}{i}$, $i = 1, 2, \ldots$, то в каждое покрытие компакта *X* вписано хотя бы одно покрытие σ_i.
Из предложения B) мы видим, что ∇ -группа $\nabla^n(X, G)$ счетна.
G) *Пусть G — счетная группа, и G*—ее группа характеров. Тогда группа* $\Delta^n(X, G^*)$ *является группой характеров группы* $\nabla^n(X, G)$.
Доказательство. Мы уже видели (2H)), что для каждого покрытия σ пространства *X* группа $\Delta^n(N(\sigma), G^*)$ является группой характеров группы $\nabla^n(N(\sigma), G)$. Наше утверждение следует поэтому из предложения D).
Из предложения G) и того обстоятельства, что группа $\nabla^n(X, G)$ счетна, вытекает следствие, состоящее в том, что к паре групп $\nabla^n(X, G)$ и $\Delta^n(X, G^*)$ можно применить понтрягинскую теорию двойственности, изложенную в § 2.
Замечание 1. Пусть X— произвольный компакт и G - счетная группа. Как группа характеров счетной дискретной группы, группа $\Delta^n(X, G^*)$ допускает компактную топологию, но не существует никакого естественного способа определения топологии в группе $\Delta^n(X, G^*)$.
Замечание 2. Пусть *X* — компакт, и *G* — *произвольная* группа. Известно, что Δ- и ∇ -группы компакта *X* по области коэффициентов G полностью определяются как Δ-группами компакта *X* по области коэффициентов П, так и ∇ -группами компакта *X* по области коэффициентов $\mathfrak{Z}$.
Замечание 3. Формально определения Δ- и ∇ - групп, данные выше, могут быть приложены к некомпактным пространствам, однако полученная таким путем теория гомологии совершенно неудовлетворительна. Ибо, например, одномерная Δ-группа Бетти прямой даже при наиболее простых группах *G* оказывается чрезвычайно сложной. Кроме того, не существует никакой простой двойственности между Δ- и ∇ -гомологиями.
Пример 22.2. Определим компакт *X* следующим образом: точками пространства *X* являются последовательности комплексных чисел (включая число ∞):

$$(z_1, z_2, \ldots, z_i, \ldots),$$

удовлетворяющих условию:

$z^i_{i+1} = z_i, \qquad i = 1, 2, \ldots .$

Сходимость в *X* означает почленную сходимость этих последовательностей. Очевидно, *X*—компактное пространство со счетным базисом. Пусть *i* — фиксированное целое число >0. Ставя каждой точке $z = (z_1, z_2, \ldots)$ в соответствие ее *i*-ю координату z_i, получаем отображение f_i компакта *X* в комплексную числовую сферу *S*. Каждому покрытию σ сферы *S* ставим в соответствие покрытие σ' компакта *X*, элементами которого являются полные прообразы элементов покрытия σ при отображении f_i. Нерв $N(\sigma^i)$, очевидно, тождественен нерву $N(\sigma)$.

Легко можно построить последовательность $\{\sigma^i\}$ покрытий сферы S, диаметр которых стремится к нулю, таких, что покрытие σ^{i+1}_i вписано в покрытие σ^i_i, $i = 1, 2, \ldots$. Нетрудно видеть, что проекция нерва $N(\sigma^{i+1}_{i+1}) = N(\sigma_{i+1})$ в нерв $N(\sigma^i_i) = N(\sigma_i)$ является симплициальным отображением степени 2 и что, следовательно, порожденные им гомоморфизмы группы $\nabla^q(N(\sigma_i), \mathfrak{Z})$ в группу $\nabla^q(N(\sigma^k_k), \mathfrak{Z})$, $(k \geqslant i)$, суть гомоморфизмы (3) примера 20.1. Группа $\nabla^q(X, \mathfrak{Z})$ является предельной группой прямого спектра $\{\nabla^q(N(\sigma^i_i), \mathfrak{Z})\}$, и поэтому, в силу примера 20.1, она изоморфна группе двоично-рациональных чисел. Ее группа характеров $\Delta^q(X, \Pi)$ есть диадический соленоид.

Пусть *Y*—замкнутое подмножество компакта *X*, состоящее из точек $(z_1, z_2, \ldots)$ таких, что $|z_i| \leqslant 1$, и пусть *Z*—пространство, полученное из Y отождествлением каждой пары точек $\{z_i\}, \{z'_i\}$, удовлетворяющих условию:

$$|z_1| = |z'_1| = 1, \quad \left(\frac{z_i}{z'_i}\right)^{2^i} = 1, \qquad i = 1, 2, 3, \ldots .$$

Методами, аналогичными методам, использованным выше, можно показать, что группа $\nabla^q(Z, \mathfrak{Z})$ является предельной группой прямого спектра $\{G_i\}$, где G_i — циклическая группа порядка 2 и гомоморфизмы h^{ik}, $i \leqslant k$, определяются формулой (3) примера 20.1. Следовательно, группа $\nabla^q(Z, \mathfrak{Z})$ является группой двоично-рациональных чисел, приведенных по модулю 1, а ее группа характеров $\Delta^q(Z, \Pi)$ есть нульмерная диадическая группа.

3.5. Отображения компактов

В этом параграфе мы рассмотрим отображение одного компакта в другой с точки зрения теории гомологии.

А) Пусть *X* и *Y*— два компакта, и *f*—отображение *X* в *Y*. Пусть G — группа, и *n* — целое число. Тогда *f* порождает гомоморфизм h^f группы $\nabla^n(Y, G) = \nabla^n(Y)$ в группу $\nabla^n(X, G) = \nabla^n(X)$, определяемый

следующим образом: пусть σ—покрытие пространства *Y*, σ_f— покрытие пространства *X*, элементами которого являются полные прообразы элементов покрытия σ.

Вершины нерва $N(\sigma_f)$ находятся во взаимно однозначном соответствии с вершинами нерва $N(\sigma)$, и это соответствие, очевидно, является симплициальным отображением комплекса $N(\sigma_f)$ в $N(\sigma)$. В силу предложения 3 А), это симплициальное отображение порождает гомоморфизм группы $\nabla^n(N(\sigma))$ в группу $\nabla^n(N(\sigma_f))$. Кроме того, если τ — другое покрытие пространства *Y*, а $g_\sigma \in \nabla^n(N(\sigma))$ и $g_\tau \in \nabla^n(N(\tau))$— эквивалентные элементы (см. определение 12), то образы элементов g_σ и g_τ в группах $\nabla^n(N(\sigma_f))$ и $\nabla^n(N(\tau_f))$ также будут эквивалентными. Таким образом, для каждого элемента группы $\nabla^n(Y)$ мы получаем некоторый элемент группы $\nabla^n(X)$. Именно это отображение группы $\nabla^n(Y)$ в группу $\nabla^n(X)$ (являющееся, как легко видеть, гомоморфизмом) мы обозначаем через h^f.

В) Аналогично, *f* порождает гомоморфизм h_f группы $\Delta^n(X)$ в группу $\Delta^n(Y)$. В силу предложения 3 В), симплициальное отображение нерва $N(\sigma_f)$ в нерв $N(\sigma)$ порождает гомоморфизм группы $\Delta^n(N(\sigma_f))$ в группу $\Delta^n(N(\sigma))$. Гомоморфизм h_f группы $\Delta^n(X)$ в группу $\Delta^n(Y)$ получается следующим путем: каждому элементу *e* группы $\Delta^n(X)$ ставим в соответствие тот элемент группы $\Delta^n(Y)$, представитель которого в группе $\Delta^n(N(\sigma))$ является образом представителя элемента *e* в группе $\Delta^n(N(\sigma_f))$.

Пусть *f*—отображение компакта *X* в компакт Y, и *g*— отображение компакта *Y* в компакт *Z*. Тогда гомоморфизм группы $\Delta^n(X)$ в группу $\Delta^n(Z)$, порожденный отображением *gf*, совпадает с произведением гомоморфизмов $h_g h_f$. Аналогичное утверждение о транзитивности имеет место для гомоморфизмов ∇-групп, порожденных отображениями *f*, *g* и *gf*.

Пример 23. Пусть *P* и Q —два полиэдра в данных симплициальных подразделениях. Отображение *f* полиэдра *P* в полиэдр Q называется *симплициальным отображением,* если оно обладает следующим свойством: если $p_0, \ldots, p_k$— вершины некоторой клетки полиэдра *P,* то $f(p_0), \ldots, f(p_k)$ являются вершинами некоторой клетки полиэдра Q , и *f* отображает клетку с вершинами $p_0, \ldots, p_k$ в клетку с вершинами $f(p_0), \ldots, f(p_k)$ *линейно.* Симплициальное отображение полиэдра *P* в полиэдр Q, конечно, порождает симплициальное отображение комплекса остовов полиэдра *P* в комплекс остовов полиэдра Q; наоборот, симплициальное отображение комплекса остовов полиэдра *P* в комплекс остовов полиэдра *Q* порождает симплициальное отображение полиэдра *P* в *Q*. Мы предоставим читателю доказать, что

если f— симплициальное отображение полиэдра P в полиэдр Q, то порожденные отображением f гомоморфизмы группы $\Delta^n(P)$ в группу $\Delta^n(Q)$ и группы $\nabla^n(Q)$ в группу $\nabla^n(P)$ совпадают с ранее определенными гомоморфизмами, порожденными симплициальным отображением комплексов остовов (см. § 3). (Рассматриваем покрытие σ полиэдра Q и покрытие τ полиэдра P, порожденные звездами вершин данных симплициальных подразделений, и используем тот факт, что τ вписано в покрытие, образованное полными прообразами элементов покрытия σ.)

Пример 24. Пусть τ — покрытие компакта X, $P(\tau)$ — геометрическая реализация нерва $N(\tau)$, и b — барицентрическое τ-отображение пространства X в полиэдр $P(\tau)$. Тогда гомоморфизм h^b группы $\nabla^n(P(\tau))$ в группу $\nabla^n(X)$, порожденный отображением b, является просто гомоморфизмом, ставящим каждому элементу e_τ группы $\nabla^n(N(\tau))$ в соответствие элемент e группы $\nabla^n(X)$, имеющий e_τ своим представителем. Это непосредственно следует из определения гомоморфизма h^b.

Определение 15. Пусть C — замкнутое подмножество компакта X, и f — тождественное отображение C в X. Гомоморфизмы группы $\Delta^n(C)$ в группу $\Delta^n(X)$ и группы $\nabla^n(X)$ в группу $\nabla^n(C)$, порожденные отображением f, называются *естественными гомоморфизмами.* Элемент группы $\Delta^n(C)$, который отображается этим гомоморфизмом в нуль группы $\Delta^n(X)$, называется *ограничивающим в X.* Элемент группы $\nabla^n(C)$, являющийся образом при этом гомоморфизме некоторого элемента группы $\nabla^n(X)$, называется *продолжаемым* на X.

Замечание. Если X—полиэдр и C его подполиэдр, то это определение согласуется с определением (определение 7) естественных гомоморфизмов комплексов остовов полиэдров X и C; это следует из примеров 14 и 23.

С) Пусть f—отображение компакта X в компакт Y, и h^f — гомоморфизм группы $\nabla^n(Y)$ в группу $\nabla^n(X)$, порожденный отображением f. Тогда для каждого элемента e группы $\nabla^n(Y)$ существует положительное число δ, обладающее тем свойством, что если g—отображение компакта X в Y такое, что

$$\rho(f,g) < \delta, \text{ то } h^g(e) = h^f(e).$$

Доказательство. Пусть дан элемент e группы $\nabla^n(Y)$, и пусть $\nabla^n(N(\sigma))$— группа, в которой e имеет представителя. Допустим, что этот представитель есть e_σ. Пусть, далее, покрытие σ пространства Y состоит из открытых множеств $V_1, \ldots, V_k$, а τ — покрытие пространства X, элементы $U_1, \ldots, U_k$ которого удовлетворяют условию

$$\overline{U}_i \subset f^{-1}(V_i)$$

(существование покрытия τ вытекает из нормальности пространства *X).* Тогда взаимно однозначное соответствие между U_i и V_i дает нам симплициальное отображение *m* нерва *N(τ)* в нерв N(σ), являющееся произведением симплиииального отображения нерва $N(\sigma_f)$ в нерв N (σ), определенного в А), и проекции нерва $N(\tau)$ в $N(\sigma_f)$. Следовательно, отображение *m* порождает гомоморфизм группы $\nabla^n(N(\sigma))$ в группу $\nabla^n(N(\tau))$, причем ясно, что образ элемента e_σ при этом гомоморфизме является элементом группы $\nabla^n(N(\tau))$, представляющим элемент $h^f(e)$. Положим:

$$\delta = \min \rho(f(\overline{U}_i),\ Y \setminus V_i).$$

Тогда, если *g*— отображение компакта *X* в *Y,* для которого $\rho(f, g) < \delta$, $\overline{U}_i \subset g^{-1}(V_i)$, и, следовательно, *g* также порождает то же самое симплициальное отображение нерва N(τ) в нерв N(σ). Следовательно, $h^g(e) = h^f(e)$.

D) Пусть *f*— отображение компакта *X* в компакт *Y,* и h_f— гомоморфизм группы $\Delta^n(X)$ в группу $\Delta^n(Y)$, порожденный отображением *f*. Тогда для произвольно заданного элемента *e* группы $\Delta^n(X)$ и покрытия σ пространства *Y* существует положительное число δ, обладающее тем свойством, что если *g* — отображение компакта *X* в *Y* такое, что $\rho(f, g) < \delta$, то $h_f(e)$ и $h_g(e)$ имеют одного и того же представителя в группе $\Delta^n(N(\sigma))$. Это может быть доказано методами, аналогичными методам, использованным при доказательстве предложения С).

E) Из С) и D) вытекает, что *гомотопные отображения одного компакта в другой порождают один и тот же гомоморфизм* Δ*-групп и один и тот же гомоморфизм* Δ*-zpynn.*

Пример 25. Пусть *f*— отображение компакта *X* в *n*-мерную сферу S_n. Пусть $\sigma = \mathfrak{Z}$. В силу примера 22,группа $\nabla^n(S_n, \mathfrak{Z})$ является свободной циклической группой. Мы говорим, чго S_n *ориентирована,* если выбран один из двух образующих элементов этой группы.

Предполагая, что S_n ориентирована, определим *степень* отображения *f,* в точности как в примере 16, как элемент группы $\nabla^n(X, \mathfrak{Z})$, соответствующий при гомоморфизме h^f, образующему группы $\nabla^n(S_n, \mathfrak{Z})$. Снова, как в примере 16, степень отображения *f* может быть определена как характер h^f группы $\Delta^n(X, \Pi)$.

Предложение Б) показывает, что степень отображения компакта в ориентированную *n*-мерную сферу является гомотопическим инвариантом.

F) Пусть X и Y — компакты, G — счетная группа, G* — ее группа характеров, и n —целое число. Вспомним (4 G)), что группы $\Delta^n(X, G^*)$ и $\Delta^n(Y, G^*)$ являются группами характеров групп $\nabla^n(X, G)$ и $\nabla^n(Y, G)$. Пусть теперь f—отображение компакта X в Y, h^f— гомоморфизм группы $\nabla^n(Y, G)$ в группу $\nabla^n(X, G)$, порожденный отображением f, и h^*_f— гомоморфизм группы $\Delta^n(X, G^*)$ в группу $\Delta^n(Y, G^*)$, также порожденный отображением f. Тогда h^*_f является гомоморфизмом, сопряженным с гомоморфизмом h_f.

Доказательство. Пусть σ — покрытие компакта Y, и σ_f— покрытие компакта X, элементами которого являются полные прообразы элементов покрытия σ. Группы $\Delta^n(N(\sigma_f), G^*)$ и $\Delta^n(N(\sigma), G^*)$ являются группами характеров групп $\nabla^n(N(\sigma_f), G)$ и $\nabla^n(N(\sigma), G)$. Предложение F) является поэтому простым следствием замечания, состоящего в том, что гомоморфизм группы $\Delta^n(N(\sigma_f), G^*)$ в группу $\Delta^n(N_\sigma, G^*)$ является сопряженным с гомоморфизмом группы $\nabla^n(N(\sigma), G)$ в группу $\nabla^n(N(\sigma_f), G)$ (см. 3 C))

3.6. Теорема Хопфа о продолжении отображения

В этом параграфе областью коэффициентов для ∇ - групп будет всегда группа $\mathfrak{Z}$ целых чисел, а областью коэффициентов для Δ-групп будет всегда группа П — группа характеров группы $\mathfrak{Z}$.

Пусть C — замкнутое подмножество компакта X, и f—отображение C в компакт Y. При каких условиях f можно продолжить на X. Мы уже изучали эту проблему, используя методы теоретико-множественной топологии. Мы получим более точные результаты, исследуя вновь эту проблему, на этот раз с точки зрения алгебраической топологии. Сначала мы установим необходимое условие для возможности такого продолжения.

A) Пусть n — целое число, и h^f — гомоморфизм группы $\nabla^n(Y)$ в группу ∇^n (C), порожденный отображением f. Для того чтоэы f было продолжаемо на X, необходимо, чтобы каждый элемент группы ∇^n (C), являющийся образом некоторого элемента группы $\nabla^n(Y)$ при гомоморфизме h^f, был продолжаем на X (см. определение 15).

Доказательство. Предположим, что f можно продолжить на X до отображения F компакта X в Y. Пусть h — естественный гомоморфизм группы ∇^n (X) в группу $\nabla^n(C)$; вспомним, что h есть гомоморфизм, порожденный тождественным отображением множества C в X. Отображение f является теперь результатом последовательного

применения двух отображений: тождественного отображения множества С в X и отображения *F;* формула $h^f = hh^F$ тогда показывает, чго каждый элемент группы ∇^n (С), являющийся образом при гомоморфизме h_f, является также образом при гомоморфизме *h.*

В) Заметим, что необходимое условие предложения *А)* эквивалентно следующему условию в терминах Δ-гомологий: каждый элемент группы Δ^n (С), ограничивающий в *X,* отображается гомоморфизмом h_f в нуль группы Δ^n (Y).

Доказательство. Обозначим через h_f гомоморфизм группы Δ^n (С) в группу Δ^n (Y), порожденный отображением *f*, и через *h** — естественный гомоморфизм группы Δ^n (С) в группу Δ^n *(X).* Вспомним, что h_f и *h** являются гомоморфизмами, сопряженными с гомоморфизмами h^f и *h*. Образы $h^f\nabla^n(Y)$ и $h\nabla^n(X)$ групп $\nabla^n(Y)$ и $\nabla^n(X)$, соответственно при гомоморфизмах h^f и *f*, являются подгруппами группы ∇^n (С). Условие предложения А) требует, чтобы первая из этих подгрупп содержалась во второй. Это эквивалентно утверждению, что аннулятор первой подгруппы содержит аннулятор второй. Но в силу предложения 2 D), аннулятором подгруппы $h^f\nabla^n(Y)$ является ядро гомоморфизма h_f, т. е. множество элементов группы Δ^n (С), отображающихся гомоморфизмом h_f в нуль группы Δ^n (Y). Аналогично, аннулятором подгруппы $h\nabla^n(X)$ является ядро гомоморфизма *h**, т. е. множество элементов группы $\Delta^n(C)$, ограничивающих в *X.* Это показывает эквивалентность условий предложений А) и В).

В общем случае условие предложения А) не является достаточным.

Пример 26. Пусть *Y* — плоское множество, состоящее из двух окружностей C_1 и C_2, касающихся друг друга в точке *P; X*—замкнутый круг, и окружность С—его граница, *f* — отображение множества С в Y, определенное следующим образом. Разобьем С на четыре части последовательно расположенными точками P_1, P_2, P_3, P_4. Каждая из точек P_1, P_2, P_3, P_4 переводится отображением *f* в точку Р. Открытая дуга P_1P_2 топологически отображается на положительно ориентированную дугу $C_1\backslash P$, открытая луга P_2P_3 топологически отображается на положительно ориентированную дугу $C_2\backslash P$, открытая дуга P_3P_4 топологически отображается на отрицательно ориентированную дугу $C_1\backslash P$, и открытая дуга P_4P_1 топологически отображается на отрицательно ориентированную дугу $C_2\backslash$ Р. Тогда можно показать, что *f* не гомотопно постоянному отображению, т. е. не может быть продолжено на *X,* несмотря на то, что гомоморфизм h^f отображает каждый одномерный ∇ -цикл пространства *Y* в ограничивающий ∇ -цикл.

Пример 26. 1. Пусть X—замкнутая область $x_1^2+x_2^2+x_3^2+x_4^2\leqslant 1$ в пространстве E_4, ограниченная сферой S_3. Г. Хопф построил отображение сферы S_3 в сферу S_2, не гомотопное постоянному отображению, т. е. не продолжаемое на X, относительно S_2, Тем не менее, так как $\nabla^n(S_3)=0$ для $n\leqslant 2$ и $\nabla^n(S_3)=0$ для $n>2$, каждый гомоморфизм группы $\nabla^n(S_2)$ в группу $\nabla^n(S_3)$ является нулевым гомоморфизмом. Теорема Хопфа о продолжении отображения утверждает, что условие предложения А) не только необходимо, но и достаточно в случае, когда X имеет размерность $\leq n+1$, а Y есть ni-мерная сфера. Это значит, что топологическая проблема продолжения отображений сводится в этом случае к чисто алгебраической проблеме.

ПРОДОЛЖЕНИЕ СИМПЛИЦИАЛЬНЫХ ОТОБРАЖЕНИЙ В S_n

Прежде чем приступить к доказательству теоремы Хопфа о продолжении отображения, рассмотрим сначала симплициальные отображения полиэдров и докажем одновременной индукцией следующие два предложения (начиная с этого момента, n есть целое число ≥ 1).

C_n) Пусть P — полиэдр размерности $\leq n$, а R — или элементарная $nл$-мерная сфера R_n, или одно из ее последовательных барицентрических подразделений.

(«Элементарная сфера» употребляется в этом параграфе для обозначения, как комплекса, определенного в примере 11, так и его геометрической реализации. Далее, мы пишем симплекс полиэдра P, ∇ -цикл полиэдра P и т. п., понимая под этим симплекс, ∇ -цикл и т. п. комплекса остовов полиэдра P.)

Будем считать, что сфера R ориентирована. Пусть f— симплициальное отображение полиэдра P в R. Если степень отображения f есть нуль, то f гомотопно постоянному отображению.

D_n) Пусть P — полиэдр размерности $\leq n+1$, Q — его подполиэдр и f— симплициальние отображение полиэдра Q в ориентированную элементарную n-мерную сферу R_n. Пусть элемент e группы $\nabla^n(Q)$ является степенью отображения f, а элемент $\tilde{e}\in\nabla^n(P)$ — продолжением элемента e. Тогда f можно продолжить в отображение F полиэдра P в R_n, имеющее $\tilde{e}$ своей степенью.

Сначала мы докажем C_1), затем покажем, чго из C_2) следует D_n), $n\geq 1$, и, наконец, докажем, что из D_n) вытекает C_{n+1}, $n\geq 1$. Тогда C_n) и D_n) будут установлены для всех $n\geq 1$.

Доказательство предложения C_1). Пусть $z^1_0 = (r_0, r_1)$ — положительно ориентированная одномерная клетка полигона R. Пусть φ— одномерный ∇-цикл полиэдра P, определенный следующим образом:

$$\varphi(s^1) = \pm 1, \quad \text{если} \quad f(s^1) = \pm z^1_0,$$

$$\varphi(s^1) = 0 \quad \text{в иных случаях.}$$

Тогда φ представляет степень отображения f и, следовательно, по предположению является ∇-границей некоторого нульмерного ∇-цикла ψ. ψ есть целочисленная функция вершин полигона P, обладающая следующим свойством: пусть (p_0, p_1)—одномерный симплекс полигона P. Если $f(p_0, p_1)$ есть одномерный симплекс (r_0, r_1)), то $\psi(p_1) - \psi(p_0) = 1$; если $f(p_0, p_1)$ не является ни симплексом (r_0, r_1), ни симплексом (r_1, r_0), то

$$\psi(p_1) - \psi(p_0) = 0.$$

Отождествим теперь точки полигона R с элементами групп к. действительных чисел, приведенных по модулю 1. Очевидно, можно предположить, что отрезок z^1_0, направленный от r_0 к r_1, соответствует отрезку $\left(0, \frac{1}{2}\right)$.

Если p — действительное число, то обозначим через $\{p\}$класс действительных чисел, сравнимых с p по модулю 1. Определим теперь на P действительную функцию $F(x)$, накладывая на нее два условия:

1° $\{F(x)\} = f(x)$,

2°, если x принадлежит замкнутому симплексу (p_0, p_1), то

$$-\frac{1}{2} + \frac{\psi(p_0) + \psi(p_1)}{2} \leqslant F(x) < \frac{\psi(p_0) + \psi(p_1)}{2} + \frac{1}{2}.$$

Простые вычисления показывают, что $F(x)$ — однозначная непрерывная функция. Полагая для $0 \leqslant t \leqslant 1$

$$f(x, t) = \{tF(x)\},$$

видим, что отображение f гомотопно постоянному отображению.

Доказательство того, что из C_n) следует D_n). Обозначим через $r_0, r_1, \ldots, r_n, r_{n+1}$, вершины элементарной n-мерной сферы R_n в порядке, соответствующем ориентации, и через z^n_0 n-мерный симплекс $(r_1, r_2, \ldots, r_{n+1})$. z^n_0 положительно ориентирован. ∇-класс e представляется ∇-циклом φ полиэдра Q, определенным следующим образом: $\varphi(s^n) = 1$, если $f(s^n) = z^n_0$; $\varphi(s^n) = -1$, если $f(s^n) = -z^n_0$; $\varphi(s^n) = 0$ в иных случаях, а ∇-класс $\tilde{e}$ в силу 1E) может быть представлин ∇-циклом Ф полиэдра P, являющимся продолжением ∇-цикла φ.

Пусть теперь P_n и Q_n — полиэдры, состоящие из всех клеток размерности $\leq n$ полиэдров P и Q. Пусть $P^{(m)}{}_n$ — m-кратное барицентрическое подразделение полиэдра $P_n \bmod Q_n$, где m настолько велико, что можно осуществить следующую конструкцию: обозначим через $s_1^n, s_2^n, \ldots, s_q^n$ все те n-мерные симплексы полиэдра P_n, которые удовлетворяют условию $\Phi(s_i^n) > 0$, $i = 1, \ldots, q$, и не принадлежат Q. Для каждого $s^n{}_i$ выберем $\Phi(s^n{}_i)$ n-мерных симплексов в $P^{(m)}{}_n$:

$$s_{ik}^n = (p^{ik}_1, p^{ik}_2, \ldots, p^{ik}_{n+1}), \qquad k = 1, \ldots, \Phi(s_i^n),$$

причем каждый симплекс $s^n{}_{ik}$ является ориентированным подсимплексом симплекса $s^n{}_i$, не имеющим ни одной общей вершины как с симплексом $s^n{}_i$, так и с любым отличным от него симплексом $s^n{}_{ik}$.

Продолжим теперь симплициальное отображение f в симплициальное отображение F_1 полиэдра $P^{(m)}_n \cup Q$ в R_n, полагая:

$$F_1(p^{ik}_l) = r_i, \quad i = 1, 2, \ldots; r; \quad k = 1, 2, \ldots, \Phi(s_i^n);$$
$$l = 1, 2, \ldots, (n+1),$$

$F_1(p) = r_0$ для остальных вершин p, принадлежащих $P^{(m)}_n \setminus Q$.

Помимо симплексов полиэдра Q симплексами, отображающимися при F_1 на $z^n{}_0$, являются только симплексы $s^n{}_{ik}$, и степень отображения F_1 представляется цепью Φ*, которая, очевидно, связана с Φ формулой (5), и, следовательно, представляет тот же самый ∇-класс полиэдра P, рассматриваемого как топологическое пространство. Таким образом, имеем:

а) степень отображения F_1 есть элемент группы $\nabla^n(P'^{(m)}_n) = \nabla^n(P_n)$, соответствующий элементу $\tilde{e}$ при естественном гомоморфизме группы $\nabla^n(P)$ в группу $\nabla^n(P_n)$.

Нам нужно теперь показать, что отображение F_1 может быть продолжено на P. Пусть, U — некоторая $(n+1)$-мерная клетка, принадлежащая $P \setminus Q$. Отображение $F_1 | \bar{U} \setminus U$ является симплициальным отображением n-мерного полиэдра в R_n. Пусть $e^* \in \nabla^n(\bar{U} \setminus U)$ — степень отображения $F_1 | \bar{U} \setminus U$.

Тогда степень отображения F_1 является продолжением элемента e^* на P_n и, следовательно, в силу а), e^* продолжаемо даже на P. Тем более, e^* продолжаемо на $\bar{U}$. Но $\nabla^n(\bar{U}) = 0$ (в силу примера 10), и, следовательно, $e^* = 0$. Пользуясь теперь предложением C_n), получаем, что отображение $F_1 | \bar{U} \setminus U$ гомотопно постоянному отображению, или, что в точности то же самое, отображение $F_1 | \bar{U} \setminus U$ продолжаемо на $\bar{U}$. Ясно, что, применяя эти рассуждения к каждой *(n+1)-мерной* клетке, принадлежащей $P \setminus Q$, получим продолжение F отображения F_1 на P. Степенью отображения F является $\tilde{e}$, так как естественный гомоморфизм группы $\nabla^n(P)$ в группу $\nabla^n(P_n)$ является изомор-

физмом, а образ степени отображения F при этом изоморфизме является степенью отображения F_1, которая, в силу а), является также образом элемента $\tilde{z}$.

Доказательство того, что из D_n) следует C_{n+1}).

P обозначает теперь полиэдр размерности $\leq n+1$, а R— или элементарную $(n+1)$-мерную сферу, или одно из ее последовательных барицентрических подразделений. Пусть $z^{n+1}{}_0$ —положительно ориентированный симплекс R, и V—полиэдр, определенный $(n+1)$-мерными симплексами полиэдра P, переходящими в $z^{n+1}{}_0$ при отображении f. Можно предположить, что никакие два из этих симплексов не имеют общей вершины, ибо, в противном случае, можно было бы заменить P и R их двукратными барицентрическими подразделениями P'' и R'' и взять в качестве z_0^{n+1} $(n+1)$-мерный симплекс полиэдра R'', ни одна из вершин которого не является вершиной R (такой симплекс, очевидно, существует и удовлетворяет нашему требованию). Гомоморфизм h^f группы $\mathfrak{L}^n(R,\mathfrak{Z})$ в группу $\mathfrak{L}^n(P,\mathfrak{Z})$ отображает элементарную цепь $1\cdot z_0^{n+1}$ в $(n+1)$ -мерную цепь φ полиэдра P, представляющую степень стображения f и, следовательно, в силу предположения, что степень отсбражения f есть нуль, существует n-мерная цепь ψ такая, что

$$\varphi = \nabla\psi. \quad \textbf{(1)}$$

Пусть $\tilde{\varphi}$ и $\tilde{\psi}$ —цепи $h_V\varphi$ и $h_V\psi$ полиэдра V, соответствующие цепям φ и ψ при естественном гомоморфизме. Из равенства (1) получаем

$$\tilde{\varphi} = \nabla\tilde{\psi}. \quad \textbf{(2)}$$

Пусть $z^n{}_0$ — ориентированная грань симплекса z_0^{n+1}, пусть n-мерная цепь ψ_1 полиэдра V есть образ элементарной цепи $1\cdot z_0^n$ при гомоморфизме, порожденном отображением $f|V$. Так как $1\cdot z_0^{n+1}$ является Δ-границей цепи $1\cdot z_0^n$ в комплексе, состоящем из неориентированного симплекса $z^{n+1}{}_0$ и его граней, а $\tilde{\varphi}$ — образом цепи $1\cdot z_0^{n+1}$, при гомоморфизме, порожденном отображением $f|V$, то мы имеем

$$\tilde{\varphi} = \nabla\psi_1,$$

и это вместе с (2) показывает, что цепь $\psi_1 - \tilde{\psi}$ полиэдра V является ∇-циклом в V и, следовательно (см. примеры 6 и 10), гомологична нулю в ∇. Пусть теперь T—полиэдр, состоящий из всех клеток полиэдра V размерности $\leq n$. Рассмотрим отображение $f|T$ как отображение полиэдра T в элементарную n-мерную сферу R_n, образованную гранями симплекса z_0^{n+1}. ψ_1 можно рассматрирать как

∇-цикл полиэдра *T,* представляющий степень отображения $f\backslash T$. Так как ∇-цикл $\psi_1 - \tilde{\psi}$, как показано выше, ограничивает, то степень отображения $f\backslash T$ представляется также цепью $\tilde{\psi}$. Заметим, что, в силу (1), $\nabla\psi(s^{n+1}) = 0$ для любого $(n+1)$-мерного симплекса, не принадлежащего *V*. Это означает, что цепь $h_{(P\setminus V)\cup T}\psi$ * является n-мерным ∇-циклом полиэдра $(P\setminus V)\cup T$. Так как этот ∇-цикл является продолжением ∇-цикла $\tilde{\psi}$, то, используя D_n), получаем, что отображение $f\backslash T$ может быть продолжено в отображение *F* полиэдра $(P\setminus V)\cup T$ в R_n.

Положим $g(x) = f(x)$, если $x \in V$, и $g(x) = F(x)$ в противном случае; *g* есть отображение полиэдра *P* в *R*. Введем, далее, в *R* сферическую метрику таким образом, чтобы клетка, определенная симплексом z^{n+1}_0 совпадала с полусферой. Ясно, что в этой метрике *f(x)* и *g(x)* никогда не бывают диаметрально противоположными точками *n*, следовательно, *f* и *g* гомотопны. Так как образ полиэдра *P* при отображении *g* является собственной частью $(n+1)$-мернсй сферы, именно, замкнутой клеткой, определенной симплексом $z^{n+1}{}_0$, то *g* гомотопно постоянному отображению , а значит, это верно и для *f*.

Таким образом, доказательство того, чго из D_n) следует C_{n+1}), закончено. Следовательно, предложения C_n) и D_n) полностью доказаны.

ОТОБРАЖЕНИЯ КОМПАКТОВ В S_n

Теорема 1. *Пусть X— компакт размерности* $\leq n+1$, *C - его замкнутое подмножество, и f — отображение множества C в (ориентированную) n- мерную сферу* S_n. *Пусть* $e \in \nabla^n(C)$ *есть степень отображения f. Тогда для того чтобы f было продолжаемо на X, необходимо и достаточно, чтобы e было продолжаемо на X. Кроме того, если* $\tilde{e} \in \nabla^n(X)$ *есть продолжение e на X, то существует продолжение F отображения f на X, имеющее* $\tilde{e}$ *своей степенью.*

Следующая теорема двойственна предыдущей.

Теорема 1′. Теорема Хопфа о продолжении отображения. *Пусть X — компакт размерности* $\leq n+1$, *C — его замкнутое подмножество, и f— отображение множества C в* S_n. *Для того чтобы f было продолжаемо на X, необходимо и достаточно, чтобы каждый элемент группы* Δ^n (C), *ограничивающий в X, отображался в нуль группы* $\Delta^n(S_n)$ *гомоморфизмом* h_f *группы* Δ^n (C) *в группу* $\Delta^n(S_n)$.

(Напомним, что областью коэффициентов ∇ -групп является группа $\mathfrak{Z}$ целых чисел, а областью коэффициентов Δ-групп — группа П действительных чисел, приведенных по модулю 1.)
Эквивалентность теорем VIII 1 и 1' была установлена в предложении В). Таким образом, достаточно доказать лишь первую из них.
Доказательство теоремы 1. Необходимость условия уже была проверена (предложение А)).
Для того чтобы доказать достаточность, отождествим сначала S_n с элементарной n-мерной сферой R_n (см. предложение D_n)). Мы утверждаем, что существует покрытие τ компакта X, обладающее следующими свойствами:
(a) $\tilde{e}$ имеет представителя, $\tilde{e}_\tau$, в группе $\nabla^n(N(\tau))$;
(b) если $U_1,\ldots, U_k$ — элементы покрытия τ, то образ при отображении f множества $U_i \cap C$ содержится в звезде некоторой вершины комплекса R_n ;
(c) Порядок покрытия τ не больше $n+1$.
Чтобы показать это, найдем сначала покрытие τ_1 компакта X, удовлетворяющее условию (а) и покрытие τ_2 (существование которого вытекает из компактности X), удовлетворяющее условию (Ь). Тогда в качестве τ можно взять любое покрытие порядка $n+1$, вписанное и в τ_1, и в τ_2; существование такого покрытия следует из теоремы V 1.
Обозначим через τ|C покрытие множества C, состоящее из пересечений элементов покрытия τ с множеством C. $\dim N(\tau) \leq n+1$, и $N(\tau|C)$ является подкомплексом нерва N(τ). Рассмотрим *самплициалъное отображение g_τ комплекса $N(\tau|C)$ в R_n, полученное тем, что каждому элементу $U_i \cap C$ покрытая $\tau \mid C$ ставится в соответствие некоторая вершина комплекса R_n, звезда которой содержит $f(U_i \cap C)$*. Тогда из определения гомоморфизмов ∇ -групп, порожденных отображением, легко следует, что степенью e_τ отображения g_τ является представитель элемента e в группе $\nabla^n(N(\tau|C))$. Рассмотрим геометрические реализации $P(\tau)$ и $P(\tau|C)$ нервов $N(\tau)$ и $N(\tau|C)$, и будем рассматривать g_τ как симплициальное отображение полиэдра $P(\tau|C)$, а не только его вершин, в R_n.
Можно предположить, что $\tilde{e}_\tau$ является продолжением элемента e_τ.
(Если e'_τ есть элемент группы $\Delta^n(N(\tau|C))$, продолжением которого является $\tilde{e}_\tau$, то e'_τ и e_τ представляют один и тот же элемент группы ∇^n (C), т. е. они имеют одну и ту же проекцию в группу $\nabla^n(N(\tau_1|C))$, где τ_1—подходящим образом выбранное, вписанное в τ покрытие. Заменяя τ покрытием τ_1и g_τ произведением g_τ к проекции комплекса $N(\tau_1|C)$ в комплекс $N(\tau|C)$, можем предположить, что $e'_\tau = e_\tau$.)

В силу D_n существует продолжение G_τ отображения g_τ на $P(\tau)$, имеющее $\tilde{e}$ своей степенью. Возьмем теперь барицентрическое τ-отображение b_τ гфостранства X в полиэдр $P(\tau)$. Частичное отображение $b_\tau \mid C$ является, очевидно, отображением множества С в полиэдр $P(\tau|C)$. Рассмотрин отображение

$$g(x) = g_\tau(b_\tau(x)), \qquad x \in C$$

множества С в R_n и его продолжение

$$G(x) = G_\tau(b_\tau(x)), \qquad x \in X,$$

являющееся отображением пространства X в R_n. Степень отображения g есть образ элемента e_τ в группе $\nabla^n(X)$ при гомоморфизме, порожденном отображением b_τ и, в силу примера 24, это будет элемент e. По тем же соображениям, степенью отображения G является элемент $\tilde{e}$.

Отображение g обладает тем свойством, что если $U_{i_0}, \ldots, U_{i_m}$ — все элементы покрытия τ, содержащие данную точку $x \in C$, то $g(x)$ содержится в клетке полиэдра R_n, определенной симплексом $(p_{i_0}, \ldots, p_{i_m})$, першины которого соответствуют при отображении g_τ множествам $U_{i_0}, \ldots, U_{i_m}$. Из определения отображения g_τ вытекает, что $f(x)$ содержится в звезде каждой вершины $p_{i_0}, \ldots, p_{i_m}$, т. е. $f(x)$ содержится в клетке полиэдра R_n, имеющей симплекс $(p_{i_0}, \ldots, p_{i_m})$ своей гранью; следовательно, $g(x)$ и f(x) содержатся в одной и той же замкнутой клетке полиэдра R_n . Следовательно, f и g гомотопны друг другу, ибо можно точки $g(x)$ и $f(x)$ соединить прямолинейным отрезком и передвигать $f(x)$ к $g(x)$ по этому отрезку.

Теперь g допускает продолжение на X, а именно в G. Следовательно (теорема Борсука), f также допускает некоторое продолжение F на X, гомотопное отображение G. Но G имеет степень $\tilde{e}$; следовательно (предложение 5 В)), F также имеет степень $\tilde{e}$. Теорема 1 доказана.

Следствие 1. *Если X — компакт размерности* $\leq n + 1$, *то каждому элементу e группы* $\nabla^n(X)$ *соответствует отображение компакта X в ориентированную n-мерную сферу* S_n *степени e.*

Следствие 2. *ПустьX — компактразм рности* $\leq$^n+1, *и C — его замкнутое подмножество. Для того чтобы каждое отображение множества C в* S_n *было продолжаемо на X, необходимо и достаточно, чтобы каждый элемент группы* ∇^n(C) *был продолжаем, другими словами, чтобы естественный гомоморфизм группы* ∇^n(X) *в группу* ∇^n(C) *был гомоморфизмом группы* $\nabla^n(X)$ *на группу* ∇^n(C).

Доказательство. *Необходимость:* пусть e — непродолжаемый элемент группы ∇^n(C). В силу следствия 1, существует отображение f множества C в сферу S_n степени e. В силу предложения А), f не может

быть продолжено на *X*. Достаточность следует непосредственно из теоремы 1.

Следствие 3. *Пусть X— компакт размерности ≤n+1 и C — его замкнутое подмножество. Для того чтобы каждое отображение множества C в* S_n *было продолжаемо на X, необходимо и достаточно, чтобы только нулевой элемент группы* Δ^n *(C) ограничивал в X, другими словами, чтобы естественный гомоморфизм группы* Δ^n (C) *в группу* Δ^n *(X) был изоморфизмом группы* Δ^n (C) *в группу* Δ^n *(X).*

Доказательство. Следствие 3 вытекает из следствия 2 и предложения 2 F).

Теорема 2. *Пусть X— компакт размерности ≤n. Два отображения компакта X в ориентированную n-мерную сферу* S_n *гомотопны в том и только в том случае, если они имеют одну и ту же степень. Кроме того, гомотопические классы отображений компакта X в ориентированную сферу* S_n *находятся во взаимно однозначном соответствии с элементами группы* ∇^n *(X).*

Доказательство. Мы уже доказали, что два гомотопные отображения имеют одну и ту же степень (предложение 5 E)). Докажем теперь обратное. Рассмотрим топологическое произведение $X \times I$ компакта X на отрезок I. В силу теоремы III 4, $\dim X \times I \leqslant n+1$. Пусть p — отображение произведения $X \times I$ в X, задаваемое формулой: $p(x, t) = x$, $x \in X$, $0 \leqslant t \leqslant 1$, и q — отображение пространства X в произведение $X \times I$ задаваемое формулой: $q(x) = (x, 0)$. Отображение p порождает гомоморфизм h^p группы $\nabla^n(X)$ в группу $\nabla^n(X \times I)$, а отображение q порождает гомоморфизм h^q группы $\nabla^n(X \times I)$ в группу $\nabla^n(X)$. Так как pq есть тождественное отображение компакта X на себя, то гомоморфизм $h^q h^p$, порождаемый этим отображением, является тождественным автоморфизмом группы $\nabla^n(X)$: для любого элемента $e \in \nabla^n(X)$

$$h^q h^p(e) = e. \quad (3)$$

Пусть X_0 — множество точек вида $(x, 0)$, и X_1 — множество точек вида $(x,1)$. Если $n>0$, то, в силу примера 22.1, группа $\nabla^n(X_0 \cup X_1)$ является прямой суммой групп $\nabla^n(X_0)$ и $\nabla^n(X_1)$. Следовательно, элементы группы $\nabla^n(X_0 \cup X_1)$ могут быть представлены парами (e, e'), где e и e' — произвольные элементы группы $\nabla^n(X_0)$. Из соотношения (3) легко следует, что для любого элемента $e \in \nabla^n(X)$ элемент (e, e) группы $\nabla^n(X_0 \cup X_1)$ является образом элемента $h^p(e)$ при естественном гомоморфизме группы $\nabla^n(X \times I)$ в группу $\nabla^n(X_0 \cup X_1)$, ибо гомоморфизм h^q, очевидно, можно интерпретировать как естественный гомоморфизм группы $\nabla^n(X \times I)$ в группу $\nabla^n(X_0)$ (или в $\nabla^n(X_1)$). Это показывает,

что каждый элемент группы $\nabla^n(X_0 \cup X_1)$ вида (e, e) продолжаем на $X \times I$.

Пусть f_0 и f_1 — отображения компакта X в S_n, имеющие одну и ту же степень e. Рассмотрим отображение множества $X_0 \cup X_1$ в S_n, равное отображению f_0 на X_0 и отображению f_1 на X_1. Его степень есть (e, e), а элемент (e, e) продолжаем на $X \times I$. Следовательно, по теореме Хопфа о продолжении, отображение множества $X_0 \cup X_1$, определенное выше, может быть продолжено на $X \times I$; это означает, что отображение f_0 и f_1 гомотопны.

Доказательство того, что гомотопические классы находятся во взаимно однозначном соответствии с ∇ -классами, вытекает теперь непосредственно из следствия 1 теоремы 1.

Следствие. *Пусть X— компакт размерности $\leq n$. Тогда X допускает существенные отображения в S_n в том и только в том случае, если $\nabla^n(X) \neq 0$, или (двойственно) в том и только в том случае, если* $\Delta^n(X) \neq 0$.

Замечание. Если X — компактное подмножество $(n+1)$-мерного эвклидова пространства, то предположение $\dim X \leq n$ в следствии может быть опущено, ибо, в силу следствия 1 теоремы 1, из того, что $\nabla^n(X) \neq 0$, вытекает существование существенного отображения пространства X в S_n. С другой стороны, из того, что $\nabla^n(X) = 0$, следует (см. теорему 1), что каждое отображение пространства X в S_n может быть продолжено на содержащий пространство X $(n+1)$-мерный куб и, следовательно, является несущественным. Это позволяет дать алгебраическую интерпретацию теоремы VI 13: *компакт $X \subset E_{n-1}$ разбивает E_{n+1} в том и только в том случае, если $\nabla^n(X) \neq 0$, или двойственно, если $\Delta^n(X) \neq 0$.*

3.7. Теория гомологии и размерность

Теперь мы переходим к главному результату главы

Теорема 3. *Пусть X—компакт конечной размерности. Для того чтобы размерность X была не больше n, необходимо и достаточно, чтобы для любого замкнутого подмножества $C \subset X$ каждый элемент группы $\nabla^n(C)$ был продолжаем на X. Это значит, что естественный гомоморфизм группы $\nabla^n(X)$ в группу $\nabla^n(C)$ должен быть гомоморфизмом группы $\nabla^n(X)$ на всю группу $\nabla^n(C)$.*

В двойственной формулировке:

Теорема 3'. *Пусть X—компакт конечной размерности. Для того чтобы размерность X была не больше n, необходимо и достаточно,*

чтобы для любого замкнутого подмножества $C \subset X$ только нулевой элемент группы Δ^n (C) ограничивал в X. Это значит, что естественный гомоморфизм группы Δ^n (C) в группу Δ^n (X), должен быть изоморфизмом группы Δ^n (C) в группу Δ^n (X).

Так как естественный гомоморфизм группы Δ^n(C) в группу Δ^n (X) является сопряженным гомоморфизмом с естественным гомоморфизмом группы ∇^n (X) в группу ∇^n (C) (предложение 5 F) и определение 15)), то условия обеих теорем эквивалентны (предложение 2 F)). Следозателъно, достаточно дать доказательство лишь одной из них.

Доказательство теоремы 3. *Необходимость* можно было бы немедленно получить из теоремы VI 4 и следствия 2 теоремы 1. Можко, однако, дать значителььо более элементарное доказательство. Заметим сначала, что если *K*— комплекс размерности $\leq n$ и *L* — подкомплекс комплекса *K,* то каждый элемент группы ∇^n (L) продолжаем на *K,* ибо в комплексе *K* нет никакой разницы между *n*-мерными цепями и *n*-мерными ∇ -циклами, а каждая *n*-мерная цепь подкомплекса L, конечно, может быть продолжена в *n*-мерную цепь комплекса *K.*

Пусть *e* — некоторый элемент группы ∇^n (C), и σ—покрытие множества C такое, что *e* имеет представителя в группе $\nabla^n(N(\sigma))$. Рассмотрим теперь покрытие τ пространства *X,* обладающее тем свойством, что элементы покрытия τ являются пересечениями множества *C* с элементами покрытия τ. По теореме V 1 существует покрытие τ' порядка $\leq n$, вписанное в τ. Пусть σ' -покрытие множества C, состоящее из пересечений множества *C* с элементами покрытия τ'. Так как σ' вписано в σ, то *e* имеет представителя $e_{\sigma'}$ в группе $\nabla^n(N(\tau'))$. Так как $\dim N(\tau') \leq n$, то, как замечено выше, существует элемент $e_{\tau'}$ группы $\nabla^n(N(\tau'))$, являющийся продолжением элемента $e_{\sigma'}$. Элемент группы ∇^n (*X*), определяемый этим элементом $e_{\tau'}$, является продолжением элемента *e.*

Достаточность. Пусть $\dim X > n$. Тогда, в силу предложения III I D), *X* содержит замкнутое множество X_1 размерности $n+1$. По теорзме VI 4 существуют замкнутое подмножество $C \subset X_1$ и отображение *f* множества C в S_n, которое не может быть продолжено на X_1. Следовательно, согласно теореме 1, существует элемент группы ∇^n (C), не продолжаемый на X_1, а значит, и подавно не продолжаемый на *X.*

ОТНОСИТЕЛЬНЫЕ ГОМОЛОГИИ

В заключение мы кратко рассмотрим так называемые «относительные» Δ- и ∇ -гомологии, с помощью которых связь между размерностью и гомологией может быть выражена в более ясной форме.

Если дан компакт X и его замкнутое подмножество C, то можно определить Δ- и ∇ -группы компакта X mod C. Для каждого покрытия τ компакта X пусть $\tau|C$—покрытие множества C, состоящее из пересечений C с элементами покрытия τ. Тогда система групп $\{\nabla^n(N(\tau) \bmod N(\tau|C), G)\}$ с гомоморфизмами, порождаемыми симплициальными отображениями нервов, является прямым спектром. Аналогично, $\{\Delta^n(N(\tau) \bmod N(\tau|C), G)\}$ является обратным спектром. Предельные группы этих спектров называются, соответственно, n-мерными ∇ - и Δ-группами компакта X mod C по области коэффициентов G. Предложения 1G) и 2H), доказанные для комплексов, могут быть без всяких затруднений перенесены на общие компакты.

(a) Если $\nabla^{n+1}(X \bmod C, G)$— нулевая группа, то естественный гомоморфизм отображает группу $\nabla^n(X, G)$ на всю группу $\nabla^n(C, G)$, т. е. каждый элемент группы $\nabla^n(C, G)$ продолжаем.

(b) Кслн G — счетная дискретная группа и G* — ее группа характеров, то $\Delta^n(X \bmod C, G^*)$ является группой характеров группы $\nabla^n(X \bmod C, G)$.

Возьмем теперь снова в качестве области коэффициентов ∇ -групп группу $\mathfrak{Z}$, а в качестве области коэффициентов Δ-групп группу Π.

Теорема 4. *Пусть X— компакт конечной размерности. Для того чтобы размерность X была не больше n, необходимо и достаточно, чтобы для любого замкнутого подмножества C компакта X группа* $\nabla^{n+1}(X \bmod C)$, *или, что то же, группа* $\Delta^{n+1}(X \bmod C)$, *была нулевой группой.*

Доказательство. Необходимоегь вытекает из рассуждений предложения 4 F), а достаточность — из теоремы 3 и (a).

Можно показать, что группы $\nabla^m(X \bmod C)$ и $\Delta^m(X \bmod C)$ являются топологическими инвариантами открытого множества $X \setminus C$. Таким образом, теорема 4 утверждает, что $\dim X \geq n$ в том и только в том случае, если X содержит открытое множество, несущее на себе существенные n-мерные гомологии.

Часть III. Основы теории измерений

Предметом теории измерений является проблема измерения в широком смысле. При этом измерение рассматривается как основополагающая познавательная процедура, позволяющая получать экспериментальные данные о свойствах объектов, а также устанавливать и проверять правильность научных теорий и законов. Теория измерений как самостоятельная дисциплина оформилась сравнительно недавно. Ее появление обусловлено двумя обстоятельствами: с одной стороны, необходимостью систематизации и обобщения обширных разрозненных знаний по теории и технике измерений, накопленных в естественных и технических науках; с другой -в связи со значительным усложнением измерительных задач и возрастанием требований к точности и достоверности измерений в различных областях научной и практической деятельности.

Теория измерений изучает закономерности хранения, воспроизведения, передачи, получения, обработки, использования, а также оценки качества (точности и достоверности) измерительной информации.

В теории измерений различают два подхода. Первый — классическая, или репрезентационная теория измерений (от англ. *represent* - представлять) изучает **представление свойств объектов числами**. Ее основы были заложены в работах английского ученого Кэмпбела в начале XX в. и позднее развиты в трудах специалистов по математической психологии (Стивене, Супес, Зиннес, Уилкоксон и др.). **Понятие измерения в ней определяется как "представление свойств посредством номеров и чисел"** (отсюда происходит и название теории).

Второе направление - это так называемая алгоритмическая теория измерений , в которой измерение рассматривается с позиций его технической реализации, как **процесс преобразования входного сигнала (измеряемой величины) в выходной (результат измерения) с помощью специальных алгоритмических и аппаратных средств.** Она охватывает построение алгоритмов обнаружения эмпирических закономерностей, а также анализ и систематизацию процедур формирования экспериментальных данных.

В настоящей работе излагаются оба подхода.

1. Основные положения теории измерений

1.1 Взаимосвязь понятий измерения и числа

Одно из основополагающих математических понятий — "число" своим возникновением обязано практической потребности в счете и измерении. По мере развития знаний об окружающем мире понятие числа также развивалось на протяжении нескольких тысячелетий: положительные целые числа *(N),* целые числа (Q), рациональные числа *(Ra),* действительные числа *(Re),* комплексные числа (C), так что каждая последующая система чисел включает предыдущую, являясь ее обобщением. Параллельно развивалось понятие измерения. При этом каждый переход к новой системе чисел сопровождался появлением новых возможностей изучения свойств объектов окружающего мира и установления зависимостей между ними.

Подчеркивая взаимосвязь числа и измерения, греческий мыслитель Филолаос Кратонский (V в. до н. э.) говорил: "Все, что можно узнать, имеет число, без него ничего нельзя понять или осмыслить". Теория чисел Пифагора была положена им в основу модели мироздания.

Неразрывная связь измерения с понятием числа следует из определения, приведенного выше (**измерение — представление свойств посредством номеров и чисел**). **Оба эти понятия олицетворяют два фундаментальных свойства окружающего мира: дискретность и непрерывность. Так система целых чисел — дискретна, система действительных чисел — непрерывна. Математическим образом отдельного дискретного объекта является целое число, а математическим образом совокупности дискретных объектов — сумма целых чисел. Математическим образом непрерывности является линия (пространство, множество).** В измерении происходит соединение этих двух противоположных свойств: непрерывное представляется (измеряется) отдельными (дискретными) числами (единицами). Современная практика измерения использует наряду с целыми и действительными числами и другие системы чисел. Например, случайные числа используются при имитационном моделировании и планировании эксперимента, при статистических измерениях.

Появились и другие обобщения понятия числа, например, нечеткие числа, применяемые для представления, так называемых, качественных свойств (высокий, красивый, богатый, большой и т. д.) в языках знаний.

1.2. Физические величины и их единицы

Мы вводим понятия, давая названия свойствам объектов и явлений, чтобы проводить различия между этими свойствами. В ряде случаев понятию удается сопоставить физическую величину; при этом соответствующее свойство должно быть таким, чтобы для него можно было определить единицу и прямо или косвенно измерить. Говорят, что величина *G* измерена, если известно, сколько раз в *G* содержится некоторая единица, что дает числовое значение *{G}* величины *G*. Если обозначить через [G] - единицу величины *G* (например, единица времени 1 секунда, единица силы тока 1 ампер и т. д.), то получим

$$\{G\} = \frac{G}{[G]} \quad (1.1)$$

Числовое значение является просто числом и не содержит какой-либо иной информации. Соотношение (1.1) можно также записать в виде

$$G = \{G\} \cdot [G] \quad (1.2)$$

Условно можно представить измерительную процедуру, задаваемую уравнением (1.2), в виде некоторого "измерительного прибора", где каждому значению измеряемой величины *G* соответствует определенная отметка шкалы прибора в принятых единицах. Указание значения (измеряемого значения) величины *G* влечет за собой, поэтому, необходимость указания соответствующей единицы. Приводящие к неудобству слишком высокие и низкие порядки численных значений (по отношению к 10) сокращенно выражаются с помощью введения новых разрядов единиц, называемых через старые с добавлением приставки (кратной либо дольной). Так получаются новые единицы, например, 1 мм^3 =1·(10^{-3} м)^3 =10^{-9} м . Сама физическая величина при этом не меняется, так как имеет место равенство:

$$G = \{G\} \cdot [G] = \{kG\} \cdot \left[\frac{G}{k}\right] = \{G'\} \cdot [G']. \quad (1.3)$$

Из (1.3) следует, что если единицу уменьшить в *k* раз, то числовое значение увеличится в *k* раз, т. е. имеет место инвариантность физической величины относительно выбора единицы.

Физические величины связаны соотношениями в форме математических уравнений, выражающих законы природы. Физические величины можно разделить на классы, каждый из которых описывает определенный круг явлений (например, механические, электрические, термодинамические и т. п.). Для

того, чтобы систематизировать обширное множество величин и единиц, стремятся ограничить его возможно меньшим числом, так называемых базисных, или основных величин и соответствующих им единиц. **Базисные величины взаимно независимы и не сводятся одна к другой. Тогда все остальные необходимые величины могут быть найдены и определены на основе базисных как производные.** Как правило, построение новых величин происходит путем умножения и деления старых, тем самым исключается, чтобы в качестве базисной величины использовалась, например площадь, так как иначе пришлось бы при образовании величин типа длины прибегать к операции извлечения квадратного корня. Вопросы построения системы физических величин исследованы в работах Флейшмана. Полученные им результаты сводятся к следующему. Обозначим разные типы величин через *A, B, C,* тогда справедливы утверждения.

1. Из *A* и *B* можно построить новый тип величин $C=A\cdot B$ (мультипликативная связь);
2. Существуют неименованные числа, обозначаемые через $(1)=(A^{\circ})$, которые при умножении на *A* не изменяют типа величины: $A\cdot(1)=A$ (единичный элемент);
3. Всякому типу величин соответствует обратный тип величин, A^{-1}, для которого $A\cdot A^{-1}=(1)$;
4. Связи между величинами разных типов подчиняются ассоциативности: $A\cdot(B\cdot C)=(A\cdot B)\cdot C$ и коммутативности: $A\cdot B=B\cdot A$\
5. Для всех $A\neq(1)$ и $m\in N|_0$ справедливо равенство $A^m\neq(1)$;
6. Полное множество, состоящее из бесконечного числа типов величин, обладает конечной производящей системой. Это означает, что имеется конечное число *n* элементов $C_1, C_2,..., C_n$, через которые любой тип величины *X* может быть представлен в виде: $X=C_1^{\alpha_1}\cdot C_2^{\alpha_2}\cdot\ldots\cdot C_n^{\alpha_n}$ при целочисленных α_i. Однозначность такого представления заранее не предполагается.

Утверждения 1-6 образуют полную систему аксиом абелевой группы и справедливы для множества физических величин. Это позволяет воспользоваться теоремой, справедливой для абелевой группы: Среди *n* элементов производящей системы $C_1, C_2,..., C_n$ имеется подмножество $l\leq n$ элементов $B_1, B_2,..., B_l$, обладающее тем свойством, что каждый элемент может быть однозначно представлен в виде

$$X=B_1^{\beta_1}\cdot B_2^{\beta_2}\cdot\ldots\cdot B_l^{\beta_l}, \quad (1.4)$$

где β_i - целые числа. Элементы $B_1, B_2,..., B_l$ называются базисом группы. Здесь B_i - основные типы величин.

Имеет место теорема: группа, удовлетворяющая аксиомам 1-6, обладает по меньшей мере одним базисом $B_1, B_2,..., B_n$, причем в случае, когда $n>2$, существует бесконечное множество равноценных базисов. Величины входящие в базис называются основными, а все остальные величины — производными. Они определяются уравнениями, в которые входят основные физические величины или их комбинации. Как определить число элементов некоторого базиса? Для этого в данной области физики задается k взаимно независимых уравнений для q типов величин $(q>k)$, тогда $n=q—k$ из них остаются неопределенными и не могут быть выведены из других величин, поэтому являются основными.

Так, в механике наиболее известен базис, состоящий из длины *(l),* массы *(m)* и времени *(t).* Для геометрии достаточно только l, для кинематики l и t, для динамики требуется m, l и t. Следует отметить, что площадь, масса и время базиса не образуют, однако импульс p, энергия W и действие S образуют базис.

В 1960 г. было заключено международное соглашение о выборе основных физических величин. Эти величины, а также производные составляют основу Международной системы единиц СИ *(System International).* Система СИ использует в механике базис (l, m, t), учет электромагнетизма добавляет сюда силу электрического тока (I), термодинамика требует включения температуры (Т), для фотометрии нужно добавлять силу света (I_v), наконец, необходимость описывать количественные соотношения в физико-химии привела к добавлению количества вещества (n). Соответствующие единицы обозначают обычно прописными буквами L (метр), M (килограмм), T (секунда), I (ампер), Θ (кельвин), J (кандела), N (моль).

Систему СИ удобно использовать как в теории так и на практике, и во многих странах она имеет силу закона. Основные величины и наиболее важные производные и их единицы имеют собственные имена и краткие обозначения . Существуют точные определения этих величин, реализуемые на практике лишь с конечной точностью, для чего используют разнообразные методы измерений, которые постоянно совершенствуются. Если обратится к истории вопроса, то видно, как с одной стороны, возрастали требования к точности определения единиц основных величин, а с другой, - возникали принципиально новые способы их измерения. Исследователи стремятся связать основные физические величины с физическими константами, которые можно в любое время измерить с хорошей воспроизводимостью. Характерным примером является единица длины. Вначале метр определялся через длину окружности земного шара, затем через длину волны излучения (с 1927 г. - через длину волны красной линии кадмия, с 1960

г. - через излучение изотопа криптона). В 1983 г. на 17й генеральной конференции по мерам и весам было установлено новое определение метра через скорость света, как "длина отрезка, которую свет проходит в вакууме за *1/с* долю секунды". Скорость света в вакууме (*с*) является фундаментальной константой и равна *с*=299792458 м/с.
Важной характеристикой физической величины является размерность, определяемая соотношением (1.4). **Размерность показывает, как данная величина связана с основными величинами**. Размерность как и сама величина не зависит от выбора единиц измерения. Расстояние между двумя точками, длина каната, толщина доски, радиус окружности все это принадлежит к одному роду величин - к величинам типа длины. При этом говорят, что размерность этих величин - длина. Поэтому нет необходимости определять единицу измерения для каждой физической величины: она выражается через произведение основных единиц с целыми показателями степени и численными множителями, равными 1. Размерность произвольной величины выражается в СИ соотношением, аналогичным (1.4), где основные величины заменены на их единицы:

$$\dim G = L^{\beta_1} \cdot M^{\beta_2} \cdot T^{\beta_3} \cdot I^{\beta_4} \cdot \Theta^{\beta_5} \cdot N^{\beta_6} \cdot J^{\beta_7} \quad (1.5)$$

В этом выражении все показатели степени — целые числа. Если все они равны нулю, то величина *G* будет безразмерной. Например размерность потенциальной энергии $E_{пот}$ равна:

$$\dim E_{пот} = \dim(mgh) = ML^2T^{-2}.$$

Величина и ее размерность не одно и тоже. Одинаковую размерность могут иметь совершенно разные величины, например, работа и вращательный момент, сила электрического тока и напряженность магнитного поля. **Размерность не содержит информации о том, является ли данная величина скаляром, вектором или тензором.** Однако **размерность важна для проверки правильности соотношений между величинами**. Величины с одинаковой размерностью можно складывать, вычитать и т. п., что приводит к возможности их сравнения. **Понятие размерности лежит в основе методов теории подобия, позволяющей установить критериальные соотношения между величинами, используемые при моделировании физических явлений в различных областях.**

1.3. Измерительные шкалы

Измеряемые свойства могут иметь различную природу, быть как количественными, так и качественными. Первые увеличиваются при сложении двух объектов (например, вес), вторые не меняются

(например, удельный вес). Результаты измерения твердости материалов или силы ветра выражаются в балах, т. е. с помощью принятых **числовых индексов (номеров),** тогда как результат измерения длины, массы и других физических величин является **именованным числом**. **Для того, чтобы охватить все многообразие свойств с позиций измерительной практики были введены так называемые измерительные шкалы.** По мнению Стивенса, одного из основоположников теории измерений, **существует 4 типа шкал измерений: 1) наименований; 2) порядковая; 3) интервальная; 4) отношений.** Простейшей измерительной процедурой является классификация (установление шкалы наименований). Затем классы располагаются в зависимости от их порядкового номера, где номера служат не только для указания классов, а имеют более важное значение. Для использования порядковой шкалы не требуется равенства, или регулярности размера классов и существования абсолютного нуля. Условием применения интервальной шкалы является регулярность классов интервалов. Шкала отношений используется тогда, когда существует начало координат, которое выбирается произвольным образом. Все величины можно разделить на группы по их принадлежности к той или иной шкале. Шкалы различаются по степени "произвольности" (степени свободы) и возможности (силе шкалы). В табл. 1 приведены характеристики шкал.

Таблица 1

Измерительные шкалы и их характеристики

Шкала	Действие	Математическое соотношение	Допустимое преобразование	Примеры
Наименований	Установление равенства, или эквивалентности номеров	$x = z$ *) $x \neq z$	Замена типа $y = z$	Присвоение номеров для опознавания; классификация и таксономия (схема расчета; нумерация гоночных автомобилей и т.п.)
Порядковая	Построение упорядоченного класса или установление соотношений неравенства между числами	$X < Z$ *) $X > Z$	$y = f(X)$, где f – монотонно возрастающая функция	Определение качества материалов; твердость; установление соотношений предпочтения
Интервальная	Установление равенства интервалов	$(X - V) = (W - Z)$ $(X - V) \neq (W - Z)$	$Y = a + cX$ (две степени свободы)	Температурные шкалы Цельсия и Фаренгейта, энергия, энтропия, потенциал
Отношений	Установление равенства отношений	$(X / V) = (W / Z)$ $(X / V) \neq (W / Z)$	$Y = bX$ (одна степень свободы; существует абсолютный нуль)	Числа, длина, вес, температурная шкала Кельвина и т. п.

*) Строчные буквы обозначают номера, прописные – числа.

Чтобы лучше понять, что такое допустимое преобразование шкалы, перейдем к ее формализованному описанию. Обозначим через *S* — множество свойств некоторой совокупности объектов, на котором задано отношение *R:*

$$S = \left\{ S_i : S_i \; R \; S_j \right\}$$

При измерении каждому свойству ставится в соответствие некоторое число $\hat{S}_i$, а все множество *S* отображается на множество чисел $\hat{S} = \left\{ \hat{S}_i : \hat{S}_i \; R \; \hat{S}_j \right\}$. Тогда тройка: $Ш = \left\langle S, \Psi, \hat{S} \right\rangle$ образует измерительную шкалу, где Ψ— множество гомоморфных

отображений (гомоморфизмов) из *S* в $\hat{S}$, т. е. таких отображений, которые сохраняют отношения между соответственными элементами множеств *S* и $\hat{S}$. Любое преобразование $\varphi \in \Psi$ не меняет типа шкалы и является допустимым преобразованием. Отношение *R* обладает рядом свойств, которые и определяют возможность измерения характеристик реальных объектов (процессов), т.е. делают возможным их упорядочение по степени проявления некоторого свойства:

— *транзитивность:* если *A* находится в некотором отношении к *B,* а *B* к *C,* то *A* находится в том же отношении к *C*: $A\ R\ B \wedge B\ R\ C \Rightarrow A\ R\ C$;

- *симметричность:* если *A* находится в некотором отношении к *B,* то *B* находится в том же отношении к *A*: $A\ R\ B \Rightarrow B\ R\ A$;

- *антисимметричность* (свойство, противоположное предыдущему): если *A* находится в некотором отношении к *B,* то *B* не находится в том же отношении к *A*: $A\ R\ B \Rightarrow B\ \overline{R}\ A$;

- *рефлексивность: A* всегда находится в данном отношении к самому себе: *A R A;*

- *антирефлексивность* (свойство, противоположное предыдущему): *A* никогда не находится в данном отношении к самому себе: $A\ \overline{R}\ A$.

Этих свойств достаточно для установления порядка и размещения объектов в ряд. Например, отношение порядка применимо к свойству твердости ("тверже чем" либо обратное отношение "мягче чем"). Отношение "тверже чем" является транзитивным, так как если *A* тверже *B* (оставляет царапину на *B),* а *B* тверже С, то отсюда следует, что *A* тверже *C.* Это отношение антисимметрично, так как если *A* тверже *B* (оставляет царапину на *B),* то *B* не может быть тверже *A* (не оставляет царапину на *A).* Это отношение является также антирефлексивным *(A* не может быть тверже самого себя).

2. Обработка результатов измерений

2.1. Классификация ошибок

Введение понятия ошибка (погрешность) имеет глубокий гносеологический смысл и тесно связано с аксиомами теории измерений.

(Английскими учеными в последнее время был поднят вопрос, поддержанный рядом международных организаций, о замене понятия погрешность термином неопределенность измерения на том основании, что этот термин лучше учитывает различную природу отклонения измеренного значения от истинного (как статистическую, так и нестатистическую). По нашему мнению, такая замена некорректна, так как эти понятия относятся к разным аспектам информации: погрешность характеризует содержательную часть информации (значение), а неопределенность - истинность, уверенность, т.е. соответствие реальности. Поэтому, не вдаваясь в существо этой дискуссии на страницах учебного издания, отметим, что мы будем использовать термин "ошибка", используемый в математической статистике.)

Общепринятыми являются две аксиомы:

1. Аксиома существования истинного значения измеряемой величины.

В математической статистике ей соответствует аксиома статистической устойчивости, которую можно сформулировать в следующем виде: "**хотя точное значение результата единичного измерения не может быть найдено, функция от результатов нескольких измерений может быть определена гораздо более точно**". **Эта функция называется статистикой**. Построение подходящих статистик является задачей теории ошибок (см. ниже).

2. Аксиома несоответствия постулирует принципиальное несоответствие между измеренным и истинным значением величины.

Эти две аксиомы вытекают из опыта и дают возможность формального определения ошибки (погрешности) измерения. Опыт показывает, что при многократном повторении одного и того же измерения получаются разные численные значения, даже если все делать совершенно одинаково. Перед экспериментатором сразу возникает вопрос об истинном значении измеряемой величины, а также о точности, с которой его можно определить по имеющимся данным (если такое значение действительно существует — для этого и нужна аксиома статистической устойчивости). **Отклонение результата измерения x от истинного значения x_0 (которое обычно неизвестно) называют ошибкой (или погрешностью) измерения** e (первая буква англ. слова *error* - ошибка):

$$e = x - x_0. \quad (2.1)$$

Ошибки измерений величины необходимо проанализировать, попытаться установить их причину и свести их к минимуму. Ошибки измерений принято делить на две группы: **систематические и случайные** (статистические)(другие важные группы ошибок:

методические и инструментальные, а также статические и динамические будут рассмотрены в гл. 3, посвященной измерительным устройствам). Они подчиняются совершенно разным закономерностям, поэтому различаются и способы устранения этих ошибок.
Систематические ошибки устраняются посредством создания специальных условий измерений, применением специальной техники эксперимента и методов измерений. Случайные — проведением многократных измерений и последующей их обработкой с помощью методов математической статистики.
Формально систематическая ошибка определяется выражением:

$$e_{сист} = E[e] = E\left[x - x_0\right] = E[x] - x_0, \quad (2.2)$$

где *E[x]* - математическое ожидание (от англ. *expectation* - ожидание) величины *x;* в (2.2) учтено, что $E(x_0) = x_0$, т. к. x_0 - постоянная величина. Для случайной ошибки имеем по определению:

$$e_{сл} = e - e_{сист} = e - E[e], \quad (2.3)$$

так что систематическая и случайная ошибки в сумме равны полной ошибке измерения *e.*
Рассмотрим основные источники этих ошибок. Систематические ошибки имеют множество причин и их обычно трудно обнаруживать, так как при повторении измерений они, как правило, сохраняют свое значение. Типичными источниками ошибок являются:
- несовершенство используемой измерительной аппаратуры (ошибки линейности, дрейф нулевой точки, градуировочные ошибки);
- несовершенство используемого метода измерений;
- плохая настройка измерительной аппаратуры;
- недостаточное постоянство условий проведения измерений;
- влияние окружающей среды;
- постоянные ошибки экспериментатора (измерителя);
- неучтенные влияния других параметров.

Для обнаружения и исключения систематических ошибок нет общего предписания. Можно изменить условия проведения измерения или проверить все перечисленные источники ошибок. В сомнительных случаях используют радикальное решение, а именно: нужно перейти к совершенно другому способу измерений. Решающее значение при поиске систематических ошибок имеет критическое отношение экспериментатора к проведению измерений и особенно его опыт. Совершенствование экспериментальной техники позволяет во многих случаях избежать систематических ошибок. Например, измерение параметров пучков атомных и молекулярных частиц сильно затруднены их взаимодействием с молекулами остаточных газов.

Проведение измерений в сверхвысоком вакууме позволяет исключить такого рода ошибки.

Случайные ошибки тоже имеют вполне определенные причины, довольно многочисленные, например, малые флюктуации (колебания) параметров измерительной аппаратуры, влияющих величин и т. п. Однако взаимодействие этих причин приводит к такому разбросу измеряемых значений, который зависит уже только от случая. Предсказать значение случайной ошибки для одного измерения в принципе невозможно, поэтому приходится повторять измерения до определенного разумного предела, а полученную совокупность данных обрабатывать с помощью методов теории вероятности и математической статистики. **На этих дисциплинах базируется так называемая теория ошибок.**

Кроме перечисленных ошибок выделяют ошибки третьего типа — так называемые грубые ошибки (выбросы), которые могут быть вызваны ошибками экспериментатора или отказами измерительного оборудования. Их в принципе легко заметить, а дефектные измерения исключить в процессе самого эксперимента. Иногда момент бывает упущен, и тогда при обработке данных применяют критерий грубых ошибок, используя соотношения:

$$\nu = \frac{x_{\max} - \overline{x}_n}{S_n} \text{ и } \nu = \frac{x_{\min} - \overline{x}_n}{S_n}, \quad (2.4)$$

где $x_{\max}, x_{\min}$ — максимальное и минимальное значение из ряда измерений (так как именно они прежде всего подозрительны на грубую ошибку); $\overline{x}_n, S_n$ — вычисляются по формулам (2.12), (2.13). Функции распределения этих величин определяются методами математической статистики. Они затабулированы и по доверительной вероятности *P* или уровню значимости $\alpha = 1-P$ (см. ниже) для данного числа измерений *n* можно найти по таблице $\nu_{крит}$, т. е. такое значение, которое величина *v* еще может принять по случайным причинам. Если оказалось, что $\nu > \nu_{крит}$, то соответствующее значение отбрасывается.

2.2. Основы теории ошибок

2.2.1. Частота, вероятность, среднее значение, дисперсия

Теория ошибок справедлива только для случайных ошибок. Рассмотрим простой случай, когда одна и та же величина измеряется n раз. Если измеряемая величина x изменяется непрерывно, то область полученных n значений разделяют на некоторое количество интервалов (классов) одинаковой ширины Δx и определяют количество измерений, попавших в каждый из интервалов $\left(x_i \pm \Delta x/2\right)$. Такое частотное распределение можно представить с помощью диаграммы, которую называют гистограммой (рис. 1). Она позволяет наглядно показать исход серии измерений. Хотя результат каждого измерения определяется случайными причинами, из рис. 1 видно, что эта случайность подчиняется определенным законам.

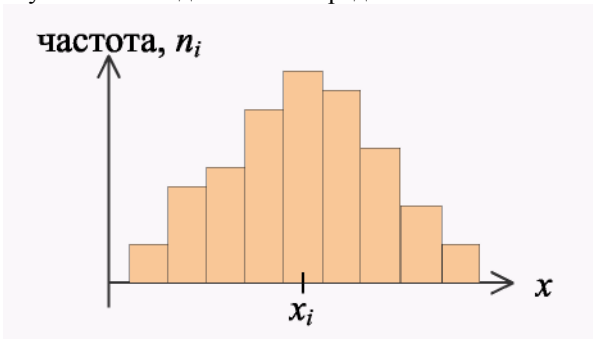


Рис. 1. Гистограмма для серии измерений.

Для описания серий измерений удобно ввести вместо абсолютных частот n_i (где n_i - количество результатов, попавших в класс x_i) *относительные частоты* $h_i=n_i/n$, которые нормированы на единицу: $\sum h_i=1$. При увеличении числа измерений n это распределение стремится *к теоретическому распределению вероятностей,* которое характеризует результаты *бесконечного числа* опытов. Существование теоретического распределения вероятностей является основополагающим предположением теории ошибок, которое, строго говоря, нельзя проверить экспериментально.

Математически предел при $n \to \infty$ для каждого класса x_i выражается в виде:

$$P(x_i) = \lim_{n\to\infty} h_i; \quad \sum_i P(x_i) = 1, \quad (2.4)$$

где *P* - *вероятность* попадания измеряемого значения в интервал (*i*) при одном измерении.

Теоретическое распределение вероятностей переходит при $\Delta x \to 0$ в гладкую кривую. Вероятность попадания исхода одного измерения *x* в интервал Δx равна $p(x) \cdot \Delta x$. Функцию *p(x)* называют *плотностью вероятности.* Вероятность *P* попадания результата измерения в интервал $[x_1, x_2]$ равна:

$$P(x_1 \le x \le x_2) = \int_{x_1}^{x_2} p(x)\, dx\,. \quad (2.5)$$

Справедливо условие нормировки

$$\int_{-\infty}^{\infty} p(x)\, dx = 1 \quad (2.6)$$

Вероятность попадания исхода одного измерения в область от — ∞ до *x* называют в математической статистике функцией распределения *F(x).* Она определяется так:

$$F(x) = \int_{-\infty}^{x} p(z)\, dz, \quad (2.7)$$

где *p(z)* - плотность распределения.

Функция распределения содержит в сжатой форме всю информацию, которую можно получить из опыта, в том числе и истинное значение измеряемой величины x_0. Эту величину для дискретного распределения значений *x* называют *арифметическим средним (E):*

$$x_0 = \overline{x} = E(x) = \sum_i x_i p(x_i), \quad (2.8)$$

а в случае непрерывного распределения - *математическим ожиданием* величины *x,* которое рассчитывается из функции распределения:

$$x_0 = \overline{x} = E(x) = \int_{-\infty}^{\infty} x\,p(x)\,dx = \int_{-\infty}^{\infty} x\,dF(x), \quad (2.9)$$

Очевидно, что если сравнивать результаты нескольких серий измерений одной и той же величины, то наиболее точное значение будет получаться в той серии, в которой кривая распределения самая узкая. Чем она уже, тем меньше ошибка $e = x - \overline{x}$ отдельного измерения, поэтому целесообразно характеризовать распределение вероятностей не только средним значением, но и шириной кривой распределения. Арифметическое значение ошибки *e* для этого не подходит, т.к. оно равно 0.

Поэтому выбирают для этой цели математическое ожидание квадрата ошибки σ^2 , которое называют *дисперсией:*

$$\sigma^2 = E(e^2) = \int_{-\infty}^{\infty} e^2 p(x)\,dx = \int_{-\infty}^{\infty} (x - \overline{x})^2\,p(x)\,dx\,. \quad (2.10)$$

(Для дискретных распределений тоже можно записать соответствующее выражение). $\sqrt{\sigma^2}$ называют *средним квадратичным отклонением* (стандартным отклонением) σ распределения. Оно непосредственно характеризует ширину распределения вероятностей, т.е. разброс измеряемых значений. Решая (2.10) с учетом (2.6), (2.9) получим:

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 p(x)\,dx - \left(\int_{-\infty}^{\infty} x\,p(x)\,dx\right)^2 = \overline{x^2} - (\overline{x})^2 = E(x^2) - [E(x)]^2\,. \quad (2.11)$$

Это выражение справедливо для всех распределений и имеет большое практическое значение.

Совокупность всех возможных исходов измерения в данных условиях называют в математической статистике *генеральной совокупностью.* В нашем случае эта совокупность бесконечно велика, и поэтому теоретическое распределение вероятностей никогда не реализуется. Мы всегда имеем дело с конечным числом *n* измерений, которые называют **выборкой объема (мощности) *n*** . Эти значения представляют собой *случайную выборку* величин из генеральной совокупности. По результатам выборки мы должны как можно точнее узнать характеристики генеральной совокупности. Поэтому нужно определить соответствующие величины выборки, причем следует постоянно помнить, что величины в выборке случайным образом "извлечены" из генеральной совокупности.

Наилучшим приближением истинной величины $\overline{x}$ является так называемое выборочное среднее значение:

$$\overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \quad (2.12)$$

(Этот факт можно обосновать с помощью метода наименьших квадратов (МНК))

По аналогии с (2.10) введем *выборочную дисперсию* S_n, которая определяется как среднее значение квадрата отклонения $(x_i - \overline{x}_n)^2$ (Здесь не идет речь об истинной e_i, т.к. не известно истинное значение величины x).

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 \,. \quad (2.13)$$

(В этой формуле вместо n появился множитель $(n - 1)$, т.к. для расчета разностей $(x_i - \overline{x}_n)$ надо иметь по крайней мере два результата). С математической точки зрения это означает, что только с учетом этого множителя математическое ожидание S_n будет равно дисперсии генеральной совокупности. Величину $\sqrt{S_n^2}$ называют **выборочным стандартным (СТО) или среднеквадратичным (СКО) отклонением** S_n. Оно характеризует разброс отдельных результатов измерений вблизи среднего значения и является наилучшей оценкой среднеквадратичного отклонения σ генеральной совокупности, которую можно получить по выборке объема n . Для практического расчета выборочной дисперсии пользуются формулой, вытекающей из (2.11):

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left[\sum_{i=1}^{n} x_i\right]^2\right\}. \quad (2.14)$$

Кроме среднего значения результатов измерений экспериментатора интересует его точность. Ее можно определить, несколько раз повторяя серии по n измерений. Тогда величины математических ожиданий $\overline{x}_n$ образуют распределение, стандартное отклонение которого $S_{\overline{x}}$ будет характеризовать разброс средних значений $\overline{x}_n$ от выборки к

выборке. Поэтому величину $S_{\bar{x}}$ называют СТО (СКО) выборочного среднего (или его средней ошибкой). Пользуясь законом сложения ошибок (см. п. 2.2.5.) получим:

$$S_{\bar{x}} = \frac{S_n}{\sqrt{n}}. \qquad (2.15)$$

Таким образом, точность измерений достаточно медленно растет с увеличением числа измерений при больших *n.* Поэтому надо стремиться не к увеличению числа измерений, а к улучшению измерительных методов, которые позволяют уменьшить СТО S_n отдельного измерения.

2.2.2. Распределение вероятностей

Обсудим наиболее важные распределения вероятностей (р. в.) для генеральной совокупности, которые часто используются при обработке результатов измерений. На практике могут реализоваться различные распределения вероятностей, т. к. кроме разброса измеряемых значений из-за случайных ошибок существует статистические флюктуации самой измеряемой величины. В качестве примера можно привести радиоактивный распад и спонтанную эмиссию излучения.

2.2.2.1. Гауссово, или нормальное, распределение (н.р.)

Н.р. было найдено К. Ф. Гауссом. Его можно получить *a priori* (до опыта) в рамках теории ошибок. Важная роль гауссова распределения объясняется тем, что оно, с одной стороны, хорошо описывает плотность вероятностей для многих величин, а с другой - распределение численных значений, при самых разных измерениях. Кроме того, многие другие распределения переходят в предельном случае в нормальное, поэтому их можно заменить распределением Гаусса. Плотность вероятности для случайной переменной *x* имеет вид:

$$p(x; x_0, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - x_0)^2}{2\sigma^2} \right), \text{ при } -\infty < x < \infty \qquad (2.16)$$

На рис. 2 показано нормальное распределение со значениями параметра σ =0,5, 1 и 2.

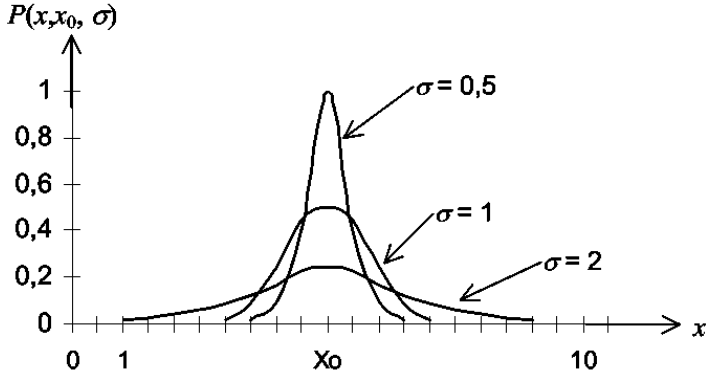


Рис. 2. Плотности вероятностей для нормального распределения при σ =0,5;1;2.

Оно характеризуется следующими особенностями.

1. Распределение симметрично относительно точки $x = x_0$.
2. Математическое ожидание равно:

$$\overline{x} = E(x) = \int_{-\infty}^{\infty} x\, p(x; x_0, \sigma)\, dx = x_0,$$

и ему соответствует максимальная плотность вероятности

$$p(x_0; x_0, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}.$$

3. По обе стороны от максимума величина p падает монотонно и асимптотически стремится к нулю.
4. Дисперсия и среднеквадратичное отклонение (СКО) определяются как

$$D(x) = \int_{-\infty}^{\infty} (x - x_0)^2\, p(x; x_0, \sigma)\, dx = \sigma^2,$$

Среднеквадратичное отклонение (стандартное отклонение) — σ.

5. Из рис. 2 следует, что при увеличении СКО распределение становится шире, а максимальное значение плотности уменьшается.

Вследствие условия нормировки $\int_{-\infty}^{\infty} p\,dx = 1$ площадь под кривой остается постоянной.

Используя величины

$$u = \frac{x - x_0}{\sigma}, \quad (2.17)$$

можно получить нормированное (стандартизованное) нормальное распределение (н. н.р.). Оно имеет вид: $p(x)\,dx = \varphi(u)\,du$, где

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right). \quad (2.18)$$

Функция распределения дается выражением (2.7):

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left(-\frac{(z - x_0)^2}{2\sigma^2}\right) dz. \quad (2.19)$$

Ее нельзя представить в виде элементарных функций, поэтому во многих работах она затабулирована в стандартизованном (нормированном) виде *Ф(u)* :

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-\frac{t^2}{2}\right) dt. \quad (2.20)$$

Часто используют так называемую *функцию ошибок* (интеграл ошибок Гаусса) erf(*u*):

$$\mathrm{erf}(u) = \frac{2}{\sqrt{\pi}} \int_{0}^{u} \exp(-t^2)\,dt = 2\Phi\left(\frac{u}{\sqrt{2}}\right) - 1. \quad (2.21)$$

Можно получить соотношение, которое весьма полезно для практических целей:

$$F(x) = \Phi\left(\frac{x - x_0}{\sigma}\right) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x - x_0}{\sqrt{2}\sigma}\right)\right) \quad (2.22)$$

На рис. 3 приведены нормальное распределение и его функция распределения в стандартизованном виде.

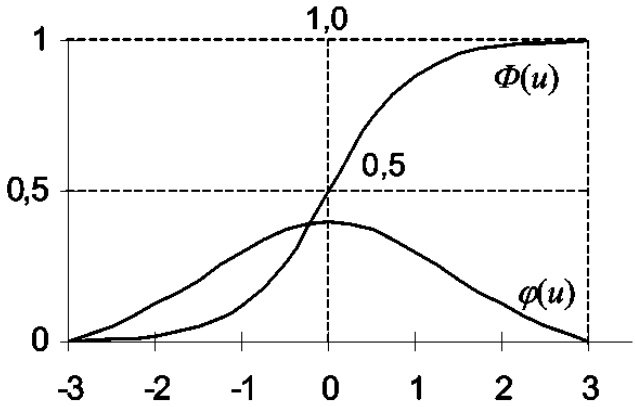


Рис. 3. Стандартизованная форма н.р. *φ(u)* и его функции распределения вероятности *Φ(u)* .

Вероятность того, что случайная переменная *x,* распределенная по нормальному закону, попадает в интервал $[x_1, x_2]$ равна:

$$P(x_1 \le x \le x_2) = F(x_2) - F(x_1) = \Phi\left(\frac{x_2 - x_0}{\sigma}\right) - \Phi\left(\frac{x_1 - x_0}{\sigma}\right) = \quad (2.23)$$

$$= \frac{1}{2}\left\{\mathrm{erf}\left(\frac{x_2 - x_0}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{x_1 - x_0}{\sqrt{2}\sigma}\right)\right\}$$

Величину *P,* выраженную в процентах называют также статистической достоверностью (вероятностью). В табл. 1 приведены ее значения для практически важных интервалов.

Таблица 1

Интервал			*P*, %
$x_0 - \sigma$	$\le x \le$	$x_0 + \sigma$	68,3
$x_0 - 1{,}96\sigma$	$\le x \le$	$x_0 + 1{,}96\sigma$	95
$x_0 - 2\sigma$	$\le x \le$	$x_0 + 2\sigma$	95,5
$x_0 - 2{,}58\sigma$	$\le x \le$	$x_0 + 2{,}58\sigma$	99
$x_0 - 3\sigma$	$\le x \le$	$x_0 + 3\sigma$	99,7

На рис. 4 показаны области $\pm\sigma$ и $\pm 2\sigma$, для нормального распределения.

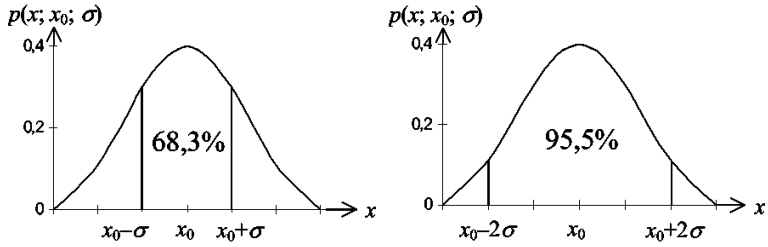


Рис. 4. Интервалы $x_0-\sigma \le x \le x_0+\sigma$ и $x_0-2\sigma \le x \le x_0+2\sigma$.

2.2.2.2. Биномиальное распределение

Это распределение называют иногда *распределением Бернулли.* Оно является наиболее важным дискретным распределением и получило свое название в связи с тем, что его члены представляют собой слагаемые биномиального разложения. Пусть в некотором опыте возможны *только два исхода A* и *B;* причем *p* - вероятность исхода *A*. Повторим наш опыт *n* раз, тогда биномиальное распределение предскажет вероятность того, что исход *A* будет наблюдаться в точности *x* раз:

$$P(x;n,p)=\binom{n}{x}p^x(1-p)^{n-x}, \text{ где } x=0, 1, \dots, n \quad (2.24)$$

Функция распределения имеет вид:

$$F(x)=\sum_{k=0}^{x}\binom{n}{k}p^k(1-p)^{n-k}, \quad (2.25)$$

где $\binom{n}{k}$ - число сочетаний из *n* по *k*.

Разумеется должно выполняться условие нормировки:

$$\sum_x P(x;n,p)=1.$$

Математическое ожидание и дисперсия имеют вид:

$$\overline{x}=E(x)=\sum_{x=0}^{n}xP(x;n,p)=np \quad (2.26)$$

$$\sigma^2 = E\left[\left(x-\bar{x}\right)^2\right] = \sum_{x=0}^{n}\left(x-\bar{x}\right)^2 P(x;n,p) = np(1-p). \quad (2.27)$$

При больших *n* биномиальное распределение приближается к нормальному, что важно для практики, так как с нормальным распределением легче работать.

2.2.2.3. Распределение Пуассона

Если вероятность *p* в биномиальном распределении очень мала, а число возможных исходов *n* велико, то пользоваться распределением в виде (2.24) неудобно. В этом случае полезно перейти к пределу $n \to \infty$ и $p \to 0$ при постоянном значении математического ожидания $\bar{x} = np$. Такое дискретное распределение называют *распределением Пуассона,* а соответствующая функция распределения вероятностей имеет вид:

$$P\left(x;\bar{x}\right) = \frac{\bar{x}^x}{x!}e^{-\bar{x}}, \qquad x=0,1,2,\ldots \quad (2.28)$$

Эта функция однозначно характеризуется *одним* параметром - средним значением $\bar{x}$ числа встречающихся исходов. Математическое ожидание и дисперсия равны:

$$E(x) = \sum_x xP(x;\bar{x}) = \bar{x}$$

$$\sigma^2 = E\left[\left(x-\bar{x}\right)^2\right] = \sum_x\left(x-\bar{x}\right)^2 P(x;\bar{x}) = \bar{x}. \quad (2.29)$$

Функция распределения определяется выражением:

$$F(x) = e^{-\bar{x}}\sum_{k=0}^{x}\frac{\bar{x}^k}{k!}. \quad (2.30)$$

На рис. 5 показаны распределения Пуассона для трех значений параметра $\bar{x}$.

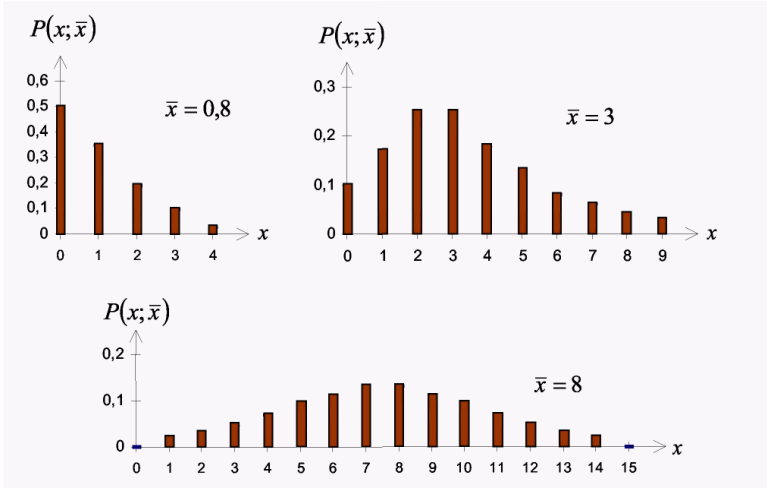


Рис. 5. Распределение Пуассона с различной величиной математического ожидания.

Видно, что с увеличением $\overline{x}$ первоначально асимметричное распределение становится все более симметричным, приближаясь к нормальному распределению с $x_0 = \overline{x}$ и $\sigma = \sqrt{x_0}$. Распределение Пуассона описывает целый ряд явлений, в которых измеряемые величины принимают дискретные целочисленные значения, не зависящие друг от друга. Примером служат измерения в атомной и ядерной физике.

Пусть за секунду фиксируется среднее число частиц $\overline{R}$, а измерение всегда происходит в течение одного интервала времени Δt. Тогда измеренное за это время число частиц x (или скорость счета $R = x/\Delta t$) описывается распределением Пуассона с $\overline{x} = \overline{R}\,\Delta t$ и СКО $\sigma = \sqrt{\overline{R}\,\Delta t}$. Если такое измерение повторить n раз, то стандартное отклонение выборочного среднего в соответствии с (2.15) равно

$$S_{\overline{x}} = \sqrt{\frac{\overline{R}\,\Delta t}{n}} \qquad (2.31)$$

Относительная среднеквадратичная ошибка среднего значения равна

$$\frac{S_{\bar{x}}}{\bar{x}} = \left(n\overline{R}\Delta t\right)^{-\frac{1}{2}} = \frac{1}{\sqrt{N}}. \quad (2.32)$$

Она определяется только числом $N = n\overline{R}\,\Delta t$ всех независимо зафиксированных частиц.

Аналогичные соображения можно применить к электромагнитным волнам. Ограничимся вначале *стабилизированными* колебаниями, под которыми мы будем понимать волновые пакеты бесконечной длины, испускаемые, например, высококачественным генератором или лазером. С помощью соответствующих детекторов с *высоким* временным разрешением можно фиксировать отдельные кванты излучения, причем нужно учитывать статистические свойства самого детектора. Теория показывает, что в этом случае полученное число фотонов тоже описывается распределением Пуассона, в котором x соответствует среднему ожидаемому числу фотонов за фиксированный интервал времени. Фотоны ведут себя в этом случае как классические независимые частицы, а такое состояние фотонов называют *когерентным.* Некогерентное излучение описывается иначе (см. ниже).

2.2.2.4. Другие распределения

При вычислениях вероятностей используется целый ряд других функций распределения. Мы рассмотрим те из них, которые чаще всего используются в науке и технике измерений. Во всех измерениях экспериментатора интересует вероятность $p(t)dt$ того, что после одного события, происшедшего в момент t=0, следующее событие наблюдается в момент t, а точнее в интервале от t до $t+dt$. Если сами события подчиняются распределению Пуассона, то плотность вероятности для интервала t равна:

$$p(t;\overline{R}) = \overline{R}e^{-\overline{R}t} \text{ при } t>0, \quad (2.33)$$

где $\overline{R}$ — средняя скорость счета (количество событий в единицу времени). Таким образом, малые интервалы времени более вероятны, чем большие. Такое распределение называется *экспоненциальным.* Соответствующая функция распределения, математическое ожидание и дисперсия имеют вид:

$$F(t) = 1 - e^{-\overline{R}t}, \quad (2.34) \qquad \bar{t} = E(t) = \frac{1}{\overline{R}}, \quad (2.35)$$

$$\sigma^2 = E\left[\left(t - \bar{t}\right)^2\right] = \frac{1}{\overline{R}^2} = \bar{t}^2 . \quad (2.36)$$

Если при измерениях применяют дискриминатор, который фиксирует только каждое *r*-е событие, то следует пользоваться обобщенным экспоненциальным распределением.

Распределение Коши, больше известное в физике как *распределение Лоренца.* Оно описывает, например, события, которые изучают с помощью метода резонанса. Плотность вероятности функция распределения имеет вид:

$$p(x; x_0, \Gamma) = \frac{1}{\pi} \cdot \frac{\Gamma/2}{(x - x_0)^2 + \left(\Gamma/2\right)^2}, \quad (2.37)$$

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \cdot arctg \frac{x - x_0}{\Gamma/2}. \quad (2.38)$$

Величины математического ожидания и дисперсии нельзя определить, так как интегралы с (2.37) расходятся. Поэтому такое распределение характеризуют медианой x_0 и полушириной Г. Полуширина Г определяется так, чтобы при $x - x_0 = \pm \Gamma/2$ плотность вероятности достигала половины максимального значения. На рис. 6 представлены распределения Гаусса и Лоренца с одинаковой полушириной. Хорошо видно, что распределение Лоренца более широкое, иными словами, плотность вероятности падает медленнее. То же самое справедливо и для вероятностей

$$P\left(x_0 - \Gamma/2 \le x \le x_0 + \Gamma/2\right) = 76\ \% \text{ (для распределения Гаусса) и}$$

$$P\left(x_0 - \Gamma/2 \le x \le x_0 + \Gamma/2\right) = 50\ \% \text{ (для распределения Лоренца).}$$

в интервале $\left[x_0 - \Gamma/2, x_0 + \Gamma/2\right]$:

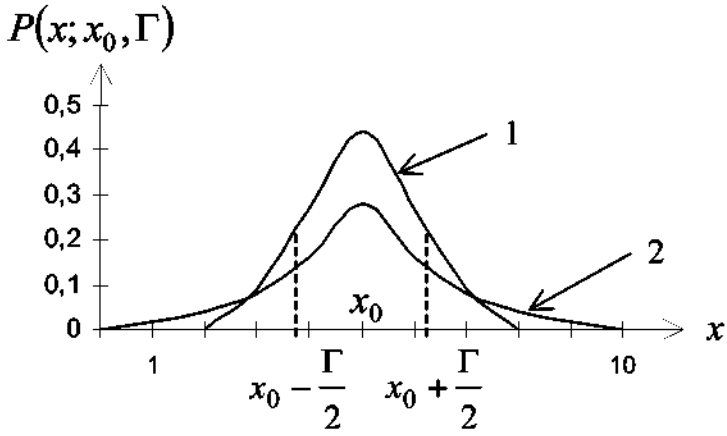


Рис. 6. Распределение Гаусса (1) и Лоренца (2) с одинаковой полушириной Г.

Рассмотрим теперь (в отличии от прежнего случая контролируемых колебаний) фотоны, которые находятся в термическом равновесии со средой (как, например, при описании излучения абсолютно черного тела). Они подчиняются распределению *Бозе-Эйнштейна,* которое можно описать средним числом заполнения $\overline{x}$:

$$P(x;\overline{x}) = \frac{1}{(1+\overline{x})} \cdot \frac{1}{\left(1+\frac{1}{\overline{x}}\right)x}, \text{ при } x=0, 1, 2, \ldots \quad (2.39)$$

Для больших значений среднего числа заполнения это выражение переходит в

$$P(x;\overline{x}) = \frac{1}{\overline{x}} e^{-x/\overline{x}} \quad (2.40)$$

Функция распределения, математическое ожидание и дисперсия равны соответственно:

$$F(x) = \frac{1}{\overline{x}} \sum_{k=0}^{x} \left(1+\frac{1}{\overline{x}}\right)^{-(1+k)}, \quad (2.41)$$

$$E(x) = \overline{x}, \quad (2.42)$$

$$\sigma^2 = \overline{x}^2 + \overline{x}. \quad (2.43)$$

На рис. 7 показаны распределения Пуассона и Бозе-Эйнштейна для средней плотности фотонов $\bar{x} = 10.$ В первом случае СКО σ =3,2 , а во втором σ=10,4 , т. е. величине среднего числа фотонов.

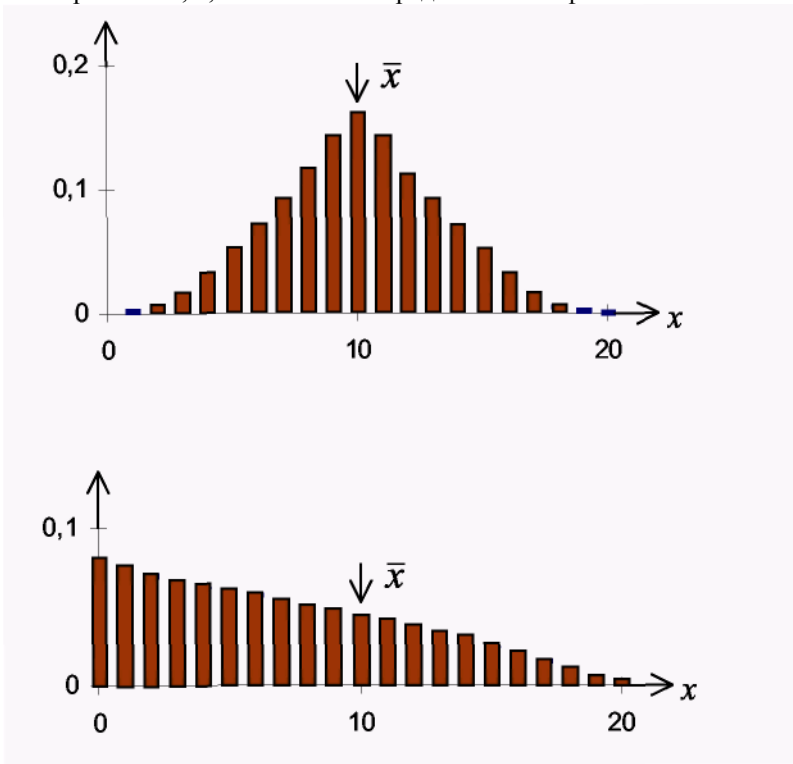
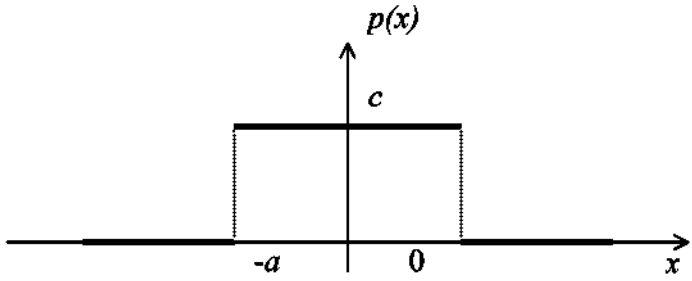


Рис. 7. Распределение Пуассона (а) и Бозе-Эйнштейна (б) для средней плотности фотонов $\bar{x} = 10$.

Отдельную группу составляют так называемые модельные распределения: равномерное (прямоугольное), треугольное (распределение Симпсона) и трапециедальное, которые часто используются для аппроксимации (приближения) неизвестных распределений. Важным свойством этих распределений является то, что дисперсия зависит от значения случайной величины *x*.
Равномерное распределение. Такое распределение имеет например, инструментальная ошибка измерения, задаваемая классом точности прибора. Плотность распределения имеет вид (центр распределения находится в нуле):



$$p(x) = \begin{cases} c & -a \le x \le a \\ 0 & x \notin [-a, a] \end{cases} \quad (2.44)$$

Рис. 8. Плотность равномерного распределения.

Константа *c* находится из условия нормировки $\int_{-\infty}^{\infty} p(x)dx = 1$.

Отсюда, подставляя *p(x)* из (2.44), получим:

$$\int_{-\infty}^{\infty} c\,dx = \int_{-a}^{a} c\,dx = c \cdot 2a = 1 \quad \Rightarrow \quad c = \frac{1}{2a}, \quad (2.45)$$

т. е. константа *c* однозначно связана с шириной интервала. Математическое ожидание и дисперсия равны:

$$E(x) = \bar{x} = \int_{-\infty}^{\infty} x\, p(x)\, dx = \int_{-a}^{a} \frac{1}{2a} \cdot x\, dx = 0; \quad (2.46)$$

$$D(x) = \sigma^2 = \int_{-\infty}^{\infty} x^2\, p(x)\, dx = \int_{-a}^{a} x^2 \frac{1}{2x} dx = \frac{a^2}{3} = \frac{(2a)^2}{12}; \quad (2.47)$$

$$\sigma = \frac{a}{\sqrt{3}} = \frac{(2a)}{2\sqrt{3}} \quad (2.48)$$

Интегральная функция имеет вид:

$$F(x) = \frac{1}{2a}(x + a)$$

Треугольное распределение Симпсона. Такое распределение имеет сумма двух величин *x* и *y,* распределенных по равномерному закону в одном и том же интервале. Плотность распределения имеет вид:

$$p(x)=\begin{cases}0; & x\notin[-2a,2a]\\ \dfrac{1}{2a}+\dfrac{x}{4a^2}; & -2a\le x\le 0\\ \dfrac{1}{2a}-\dfrac{x}{4a^2}; & 0\le x\le 2a\end{cases} \quad (2.49)$$

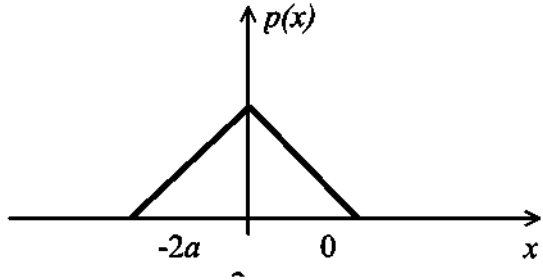


Рис. 9. Плотность распределения Симпсона.

Математическое ожидание и дисперсия равны:

$$E[x]=0; \quad (2.50)$$

$$D[x]=\frac{2}{3}a^2; \quad (2.51)$$

Интегральная функция имеет вид:

$$\left.\begin{aligned}F(x)&=\frac{x^2}{8a^2}+\frac{x}{2a}+\frac{1}{2}; \quad -2a\le x\le 0\\ F(x)&=-\frac{x^2}{8a^2}+\frac{x}{2a}+\frac{1}{2}; \quad 0\le x\le 2a\end{aligned}\right\} \quad (2.52)$$

Трапецеидальное распределение. Такое распределение имеет сумма двух величин *x* и *y,* распределенных по равномерному закону в разных интервалах. Плотность распределения имеет вид:

$$p(x)=\begin{cases}0; & x\notin\left[-(a+b),(a+b)\right]\\ \dfrac{a+b+x}{4ab}; & -(a+b)\le x\le -(a-b)\\ \dfrac{1}{2a}; & -(a-b)\le x\le (a-b)\\ \dfrac{a+b-x}{4ab}; & (a-b)\le x\le (a+b)\end{cases} \quad (2.53)$$

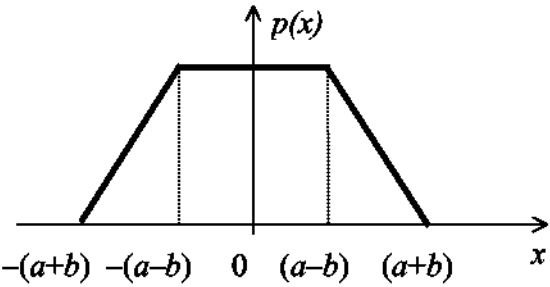


Рис. 10. Плотность трапецеидального распределения.

Математическое ожидание и дисперсия равны:

$$E\left[x\right]=0, \quad (2.54)$$

$$D\left[x\right]=\frac{a^2+b^2}{3}. \quad (2.55)$$

Интегральная функция равна:

$$\left.\begin{aligned} F(x)&=\frac{(a+b+x)^2}{8ab}; && -(a+b)\le x\le -(a-b)\\ F(x)&=\frac{a+x}{2a}; && -(a-b)\le x\le (a-b)\\ F(x)&=1-\frac{(a+b-x)^2}{8ab}; && (a-b)\le x\le (a+b)\end{aligned}\right\}. \quad (2.56)$$

2.2.3. Доверительный интервал

Понятие статистической достоверности мы ввели в п. 2.2.2.1. при обсуждении н.р. и использовали его для определения вероятности того, что измеряемая величина при фиксированной функции распределения окажется в пределах заданных границ. **Эти границы называют доверительными границами, а интервал - доверительным интервалом**. Величина статистической достоверности в каждом конкретном случае зависит от требуемой надежности измерений. Особый интерес представляет доверительный интервал для среднего значения $\overline{x}$, генеральная совокупность которого описывается н.р. с дисперсией σ^2. Относительно просто описывается случай, когда дисперсия известна, так как выборочные средние значения при мощностях выборок n тоже распределены возле x по нормальному закону, а значит их дисперсия равна σ^2/n. По аналогии с (2.17) введем преобразование:

$$u = \frac{(\overline{x}_n - \overline{x})\sqrt{n}}{\sigma} \quad (2.57)$$

и перейдем к стандартному виду $\Phi(u)$ н.р. Для произвольных доверительных границ $\pm u_p$ доверительный интервал составит

$$\left[\frac{-u_p\sigma}{\sqrt{n}}; \frac{u_p\sigma}{\sqrt{n}}\right]$$ с вероятностью (2.23):

$$P\left(-u_p \le u \le u_p\right) = \Phi\left(u_p\right) - \Phi\left(-u_p\right) = P\left(\overline{x} - \frac{u_p\sigma}{\sqrt{n}} \le \overline{x}_n \le \overline{x} + \frac{u_p\sigma}{\sqrt{n}}\right) \quad (2.58)$$

С этой вероятностью истинное значение $\overline{x}$ лежит в интервале

$$\overline{x}_n - \frac{u_p\sigma}{\sqrt{n}} \le \overline{x} \le \overline{x}_n + \frac{u_p\sigma}{\sqrt{n}}, \quad (2.59)$$

который теперь называется доверительным интервалом выборочного среднего. Однако в общем случае дисперсия генеральной совокупности неизвестна, и поэтому кроме выборочного среднего $\overline{x}_n$ нужно также

знать выборочную дисперсию S_n^2. Тогда в отличие от (2.57) вводят переменную *t:*

$$t = \frac{\bar{x}_n - \bar{x}}{\frac{S_n}{\sqrt{n}}}, \qquad (2.60)$$

которая не распределена по нормальному закону. Закон распределения этой величины называют распределением Стьюдента или *t-распределением.* Оно было впервые опубликовано английским ученым У. С. Госсетом под псевдонимом "Студент". Его плотность вероятности равна:

$$p(t;n) = \frac{p_n}{\left[1 + \frac{t^2}{n-1}\right]^{n/2}},$$

где $$p_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)}. \qquad (2.61)$$

Эта величина зависит от объема (мощности) выборки $n \geq 2$. Величину $f=n-1$ называют *числом степеней свободы распределения.* Это распределение симметрично и внешне похоже на колоколообразную кривую н.р., но ее максимум ниже. В то же время на большом расстоянии от $t=0$ плотность распределения Стьюдента совпадает с плотностью н.р. На рис. 8 показаны *t* - распределения для различных *n.* Для $n=2$ оно совпадает с распределением Лоренца, а с увеличением *n* стремится к нормальному распределению с математическим ожиданием 0 и дисперсией 1. При $n>30$ два распределения совпадают настолько хорошо, что можно пользоваться обычным н.р. При малых мощностях выборки ($n<30$) следует использовать *t*-распределение. Функция распределения, как обычно получается интегрированием (2.61):

$$F(t) = p_n \int_{-\infty}^{t} \left[1 + \frac{\vartheta^2}{n-1}\right]^{-n/2} d\vartheta \quad (2.62)$$

Эта функция табулирована, ее математическое ожидание и дисперсия равны:

$$\bar{t} = 0 \text{ для } n \geq 3$$

$$\sigma^2 = \frac{n+1}{n-1} \text{ для } n \geq 4$$

Для $n=2$ и $n=3$ дисперсия не определена, а при $n=2$ не определено и математическое ожидание.

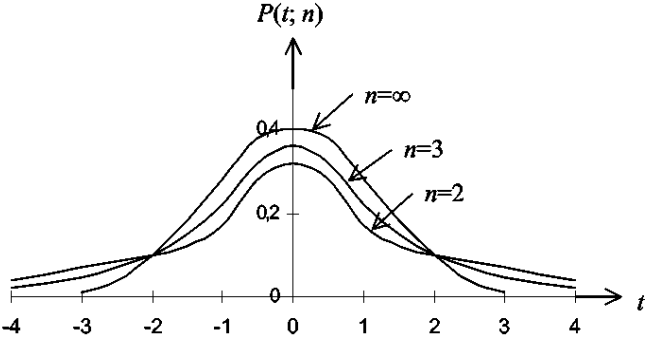


Рис. 11. Распределение Стьюдента при $n=2;$ 3; ∞.

Избранное значение статистической достоверности (доверительной вероятности) $P(\%)$ определяет границы доверительного интервала $\left[-t_p, t_p\right]$, где t_p для $\bar{x}$ определяется по аналогии с (2.59):

$$\bar{x}_n - \frac{t_p S_n}{\sqrt{n}} \leq \bar{x} \leq \bar{x}_n + \frac{t_p S_n}{\sqrt{n}} \quad (2.65)$$

Табулированная функция распределения позволяет легко узнать значение t_p для любых величин статистической достоверности:

$$P = F(t_p) - F(-t_p) = 2F(t_p) - 1, \quad (2.66)$$

где $F(t_p) = \frac{1}{2}(1+P)$.

На рис. 12 показаны эти значения для практически используемых величин статистической достоверности. Так называемая *центральная предельная теорема* математической статистики позволяет показать, что при не слишком малых мощностях выборки распределение выборочных средних, полученное для разных исходных функций распределения, достаточно хорошо описывается н.р. Поэтому в дальнейшем можно пользоваться приведенными выше соотношениями. Точно также можно определить доверительный интервал при фиксированной статистической достоверности для выборочного СТО. При этом используются результаты измерений, распределенные по нормальному закону. Теория позволяет получить для случайной переменной функцию распределения:

$$\chi^2 = (n-1)\frac{S_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2 \quad (2.67)$$

которую называют *хи-квадрат распределением, или распределением Пирсона.* С его помощью можно определить доверительный интервал для σ.

Если распределение результатов измерений не известно, то оценить статистическую достоверность можно, используя неравенство Чебышева. Эта оценка получается из следующих соображений. Запишем выражение для дисперсии (2.10):

$$\sigma^2 = D[x] = \int_{-\infty}^{\infty}\left(x - E(x)\right)^2 p(x)\,dx$$

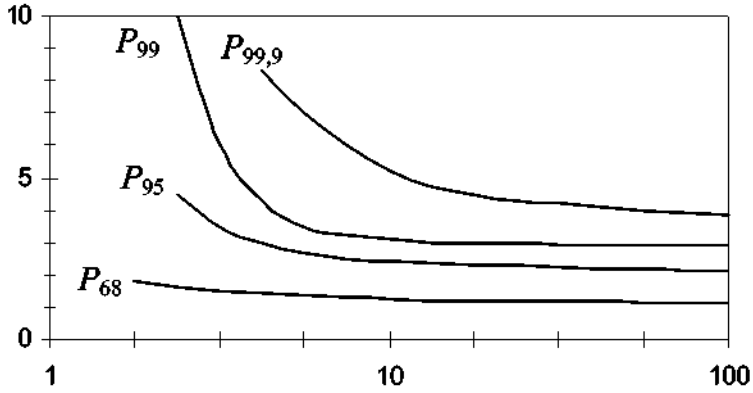


Рис. 12. Значения t_p для различных величин статистической достоверности P в зависимости от мощности выборки n: P_{68}=68%; P_{95}=95%; P_{99}=99% и $P_{99,9}$=99,9%.

Нас интересует оценка доверительной вероятности, при которой истинное значение величины x будет отличаться от $E(x)$ не более чем на ε, т. е. $|x - E(x)| \le \varepsilon$. Заменяя x-$E(x)$ на ε, получим:

$$\sigma^2 \ge \varepsilon^2\left(1 - P\left\{|x - E(x)| \le \varepsilon\right\}\right),$$

откуда и следует неравенство Чебышева:

$$P\left\{|x - E(x)| \le \varepsilon\right\} > 1 - \frac{\sigma^2}{\varepsilon^2}, \quad (2.68а)$$

или в другой форме:

$$P\left\{|x - E(x)| > \varepsilon\right\} \le \frac{\sigma^2}{\varepsilon^2}. \quad (2.68б)$$

Неравенство Чебышева дает слабую оценку, что не удивительно, так как не делается никаких предположений о законе распределения случайной величины x. Например, если $\varepsilon = 3\sigma$, то из (2.68б) найдем вероятность того, что результат измерения отличается от истинного значения на величину, большую 3σ

$$P\left\{|x - E(x)| > 3\sigma\right\} < 11\%.$$

2.2.4. Критерий Пирсона (хи-квадрат)

Рассмотрим распределение вероятности для результатов измерений выборки мощности *n* и попробуем оценить, с какой вероятностью мы можем судить по этому распределению о распределении вероятности для генеральной совокупности. Для этого нужно либо сравнить оба этих распределения, либо оценить их сходство со всеми возможными типами распределения вероятности. Рассмотрим эмпирическое распределение вида, представленного на рис. 1. Разделим затем область полученных значений *x* на *k* независимых классов так, чтобы каждый содержал в среднем 5 отдельных событий. Число классов в данном случае тоже будет близко к 5. Мерой согласия эмпирического и теоретического распределений будет сумма квадратов отклонений эмпирической частоты n_i класса *i* и теоретически рассчитанной частоты nP_i, где P_i - вероятность, предсказанная гипотетическим распределением для данного класса, рассчитываемая по (2.5). Тогда можно определить величину:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - nP_i\right)^2}{nP_i} = \sum_{i=1}^{k} \left(\frac{n_i^2}{nP_i} \right) - n. \quad (2.69)$$

Если $\chi^2 = 0$, то наблюдаемая и ожидаемая частоты в точности совпадают; если $\chi^2 \neq 0$, то нет; причем, чем больше χ^2 тем больше отклонение наблюдаемого распределения от ожидаемого. Поскольку эти отклонения имеют также статистическую природу, то для χ^2 существует свое распределение, которое для выборок большой мощности совпадает с так называемым *хи-квадрат распределением* с *k* -1 степенями свободы. Оно было введено Хелмертом. Плотность распределения имеет вид:

$$P\left(\chi^2; f\right) = P_f \left(\chi^2\right)^{\frac{f-2}{2}} \cdot \exp\left(-\frac{\chi^2}{2} \right),$$

где $P_f^{-1} = 2^{f/2}\,\Gamma\left(f/2\right)$ для $\chi^2 > 0.$ (2.70)

Математическое ожидание и дисперсия равны:

$$\overline{\chi^2} = f, \quad (2.71а)$$

$$\sigma^2 = 2f, \quad (2.71б)$$

где *f=k—1* - число степеней свободы. Если определить по экспериментальным данным еще *r* параметров гипотетического распределения, то отклонения ожидаемой частоты от наблюдаемой налагают еще *r* условий. Тогда число степеней свободы распределения равно:

$$f = k - r - 1. \quad (2.72)$$

На рис. 13 показано χ^2 -распределение для разных *f*. При *f*=1 и *f*=2 кривые монотонно понижаются с увеличением χ^2. при *f>2* наблюдается максимум вблизи значения $\chi^2 = f - 2.$ Функция распределения имеет вид:

$$F\left(\chi^2\right) = P_f \int_0^{\chi^2} \vartheta^{\frac{f-2}{2}} \cdot \exp\left(-\frac{\vartheta}{2}\right) d\vartheta. \quad (2.73)$$

Она табулирована, причем для больших *f* вместо нее можно приближенно использовать н.р.

$$F(\chi^2) \approx \Phi\left(\sqrt{2\chi^2} - \sqrt{2f-1}\right). \quad (2.74)$$

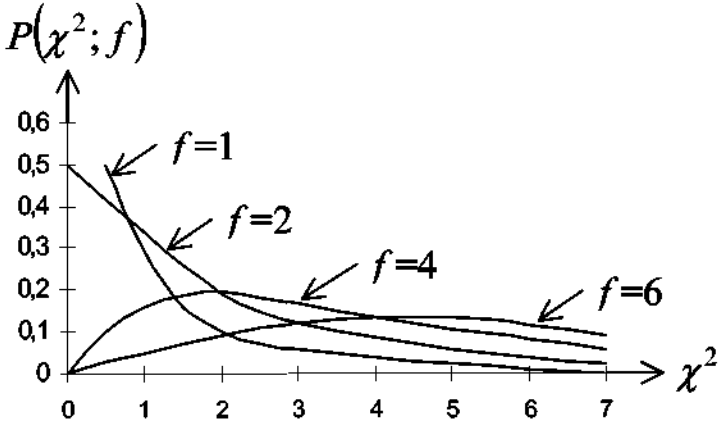


Рис. 13. Плотность вероятности хи-квадрат распределения при числе степеней свободы $f=1; f=2; f=4; f=6$.

Для χ^2 эта величина соответствует верхней границе χ^2_p, ниже которой еще можно считать, что гипотетическое распределение совпадает с истинным распределением генеральной совокупности. При

$\chi^2 > \chi^2_p$ эта гипотеза несправедлива. Величина χ^2_p определяет допустимую вероятность всех возможных отклонений. Величина $\alpha = 1 - P$ определяет вероятность того, что отклонена истинная гипотеза:

$$P\left(0 < \chi^2 \le \chi^2_p\right) = F\left(\chi^2_p\right). \quad (2.75)$$

На практике чаще всего выбирают вероятность P=0,95; 0,99. На рис. 14 показаны некоторые границы χ^2_p. Значение χ^2_p / f представлено как функция числа степеней свободы *f*. Выше каждой из кривых гипотеза о согласии неверна.

В основе χ^2-критерия лежит предположение о гипотетическом распределении для генеральной совокупности. В то же время параметры этого распределения обычно определяются по экспериментальным значениям. Так, например, выборочное среднее является наилучшей оценкой для математического ожидания генеральной совокупности (см. п. 2.2.1.). С помощью ЭВМ можно относительно легко варьировать параметры гипотетического распределения, чтобы достичь *минимальной величины* χ^2 согласно (2.69). Полученное распределение будет наиболее вероятным.

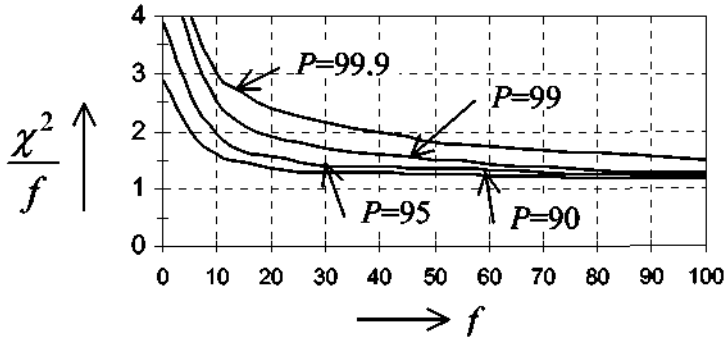


Рис. 14. Верхнее предельное значение χ^2 - распределения в зависимости от числа степеней свободы при разных вероятностях *P*.

2.2.5. Сложение ошибок

Во многих случаях не удается непосредственно измерить интересующую величину, и ее приходится рассчитывать на основе

значений других измеряемых величин (такие измерения называются косвенными). Рассмотрим пример, в котором искомая величина Z является известной функцией независимых друг от друга измеряемых величин *x* и *y:*

$$Z = f(x, y) \quad (2.76)$$

Величины *x* и *y* измеряются соответственно *m* и *n* раз. Выборочные средние равны $\overline{x}_m$ и $\overline{y}_n$, выборочные дисперсии S_m^2, S_n^2. Для каждой пары значений x_i и y_k получим величину Z_{ik}, а выборочное среднее определим по выборке мощностью $m \cdot n$:

$$\overline{Z}_{mn} = \frac{1}{mn}\sum Z_{ik} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n} f(x_i, y_k). \quad (2.77)$$

Разложим

ве

$$Z_{ik} = f(\overline{x}_m, \overline{y}_n) + \frac{\partial f}{\partial x}(x_i - \overline{x}_m) + \frac{\partial f}{\partial y}(y_k - \overline{y}_n) + \ldots$$

личину Z_{ik} в ряд Тейлора в окрестности значения $f(\overline{x}_m, \overline{y}_n)$:

Здесь вместо $x = \overline{x}_m$ и $y = \overline{y}_n$ следует подставить парциальные отклонения. Тогда выборочное среднее $\overline{Z}_{mn}$ искомой величины будет равно (если пренебречь членами высоких порядков):

$$\overline{Z}_{mn} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\left[f(\overline{x}_m, \overline{y}_n) + \frac{\partial f}{\partial x}(x_i - \overline{x}_m) + \frac{\partial f}{\partial y}(y_k - \overline{y}_n)\right],$$

$$\overline{Z}_{mn} = \frac{1}{mn}\left[mn f(\overline{x}_m, \overline{y}_n) + n\frac{\partial f}{\partial x}\sum(x_i - \overline{x}_m) + m\frac{\partial f}{\partial y}\sum(y_k - \overline{y}_n)\right].$$

так как $\sum(x_i - \overline{x}_m) = 0$ и $\sum(y_k - \overline{y}_n) = 0$, то получим:

$$\overline{Z}_{mn} = f(\overline{x}_m, \overline{y}_n) \quad (2.78)$$

Таким образом, искомое выборочное среднее равно (с точностью до членов 2^го^ порядка) величине Z, рассчитанной по средним значениям $\overline{x}_m$ и $\overline{y}_n$. Дисперсия выборочного среднего в том же приближении равна:

$$S^2_{mn} = \frac{1}{mn-1}\sum_{i=1}^{m}\sum_{k=1}^{n}\left(Z_{ik} - \overline{Z}_{mn}\right)^2 \approx \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\left[\frac{\partial f}{\partial x}\left(x_i - \overline{x}_m\right) + \frac{\partial f}{\partial y}\left(y_k - \overline{y}_n\right)\right]^2.$$

Перекрестный член равен нулю, поэтому получаем:

$$S^2_{mn} = \left(\frac{\partial f}{\partial x}\right)^2 S^2_m + \left(\frac{\partial f}{\partial y}\right)^2 S^2_n. \quad (2.79)$$

Это выражение называется гауссовым *законом сложения ошибок.* Его можно обобщить и на случай многих переменных. Пусть функция Z зависит от *l* величин: $x^{(1)}, x^{(2)}, \ldots, x^{(l)}$, которые измеряются соответственно n_1, n_2,..., n_l раз. Тогда выборочное среднее:

$$\overline{Z}_N = f\left(\overline{x}^{(1)}_{n_1}, \ldots, \overline{x}^{(l)}_{n_l}\right), \quad (2.80)$$

а дисперсия:

$$S^2_{\overline{Z}_N} = \sum_{j=1}^{l}\left(\frac{\partial f}{\partial x^{(j)}}\right)^2\Bigg|_{x^{(j)}=\overline{x}^{(j)}_{n_j}} \cdot S^2_{n_j}, \quad (2.81)$$

где $S^2_{n_j}$ - дисперсия $\overline{x}^{(j)}_{n_j}$.

Соотношение (2.81) справедливо, если

$$f\left(\overline{x}^{(1)}_{n_1}, \ldots, \overline{x}^{(l)}_{n_l}\right) >> \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^{(1)2}}\Bigg|_{x^{(1)}=\overline{x}^{(1)}_{n_1}} + \ldots + \frac{\partial^2 f}{\partial x^{(l)2}}\Bigg|_{x^{(l)}=\overline{x}^{(l)}_{n_l}}\right)$$

Если же оно не выполняется, то

$$\overline{Z}_N = \frac{1}{N}\sum_{i=1}^{N} f\left(x^{(1)}_i, \ldots, x^{(l)}_l\right), \quad (2.82)$$

где $N = n_1 \cdot n_2 \cdot \ldots \cdot n_l$.

Дисперсия равна:

$$S^2_{\overline{Z}_N} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\left[f\left(x^{(1)}_i, \ldots, x^{(l)}_i\right) - \overline{Z}_N\right]^2. \quad (2.83)$$

Доверительный интервал в котором с доверительной вероятностью *P* заключено значение $\overline{Z}_N$ равен:

$$\Delta_P = t_P \cdot S_{\overline{Z}_N}, \quad (2.84)$$

где t_P находится из таблиц распределения Стьюдента по вероятности P и числу степеней свободы:

$$k = \frac{\left(\sum_{j=1}^{l}\left(\frac{\partial f}{\partial x^{(j)}}\right)\cdot S_{n_j}^2\right)^2}{\sum_{j=1}^{l}\frac{1}{n_j-1}\cdot\left(\frac{\partial f}{\partial x^{(j)}}\right)^2\cdot S_{n_j}^4}. \quad (2.85)$$

Рассмотрим как пример случай, когда искомая величина пропорциональна произведению измеряемых величин в некоторой степени:

$$Z = Cx^{\alpha}y^{\beta}, \quad (2.86)$$

$$S_{mn}^2 = \left(\alpha C\overline{x}_m^{\alpha-1}\overline{y}_n^{\beta}\right)^2 S_m^2 + \left(\beta C\overline{x}_m^{\alpha}\overline{y}_n^{\beta-1}\right)^2 S_n^2.$$

Разделив полученную величину на выборочное среднее: $\overline{Z}_{mn}^2 = \left(C\overline{x}_m^{\alpha}\overline{x}_n^{\beta}\right)^2$, получим для квадрата относительной ошибки величины $\overline{Z}_{mn}$:

$$S_{mn}^2 \Big/ \overline{Z}_{mn}^2 = \alpha^2\frac{S_m^2}{\overline{x}_m^2} + \beta^2\frac{S_n^2}{\overline{y}_n^2}. \quad (2.87)$$

2.2.6. Взвешенное среднее значение

На практике часто приходится рассчитывать величины по нескольким выборочным средним, определенным с разной точностью (полученных в разных сериях измерений или с помощью разных методик). Если соответствующие выборочные средние равны $\overline{x}_a, \overline{x}_b, \ldots,$ а выборочные дисперсии соответственно $S_a^2, S_b^2, \ldots,$ то по этим величинам можно определить так называемое *взвешенное среднее,* если

каждое выборочное среднее умножить на множитель *w* (называемый *весом):*

$$\bar{x} = \frac{w_a \cdot \bar{x}_a + w_b \cdot \bar{x}_b + \ldots}{w_a + w_b + \ldots}. \quad (2.88)$$

Вес определяет точность каждого выборочного среднего: чем он выше, тем меньше выборочное СТО. Закон сложения ошибок позволяет получить для $\bar{x}$ дисперсию выборочного среднего. Используя (2.79), найдем:

$$S^2 = \frac{w_a^2 S_a^2 + w_b^2 S_b^2 + \ldots}{\left(w_a + w_b + \ldots\right)^2}. \quad (2.89)$$

Значения весов должны быть выбраны так, чтобы величина S^2 была минимальной. Рассмотрим для простоты два значения w_i. Пусть сумма $w = w_a + w_b$ будет постоянной, тогда можно записать дисперсию в виде:

$$S^2 = \frac{w_a^2 S_a^2 + (w - w_a)^2 S_b^2}{w^2}. \quad (2.90)$$

Из условия $\partial S^2 \Big/ \partial w_a = 0$ следует:

$$\frac{w_a}{w_b} = \frac{S_b^2}{S_a^2}, \quad (2.91)$$

т. е. веса обратно пропорциональны выборочным дисперсиям. Этот результат справедлив и для случая нескольких выборок. Обычно полагают:

$$w_a : w_b : w_c : \ldots = \frac{1}{S_a^2} : \frac{1}{S_b^2} : \frac{1}{S_c^2} : \ldots$$

и (2.89) приводится к виду:

$$S^2 = \frac{1}{\sum_{j=1}^{m} \frac{1}{S_j^2}}, \qquad (2.92)$$

где m - число серий; S_j^2 - дисперсия выборочного среднего в j-й серии. Доверительный интервал равен:

$$\Delta_P = t_P \cdot S, \quad (2.93)$$

где t_P находится из таблиц распределения Стьюдента при доверительной вероятности P и числе степеней свободы

$$k = \frac{m^2}{\sum_{j=1}^{m} \frac{1}{n_j - 1}}, \qquad (2.94)$$

где n_j - число измерений в j-й серии.

2.3. Сглаживание экспериментальных зависимостей. Метод наименьших квадратов

2.3.1. Линейная регрессия

Важной задачей является нахождение функциональных зависимостей между величинами. При этом стараются обычно так сформулировать задачу, чтобы изучать только две величины, в то время как остальные переменные остаются постоянными. В эксперименте получаются пары значений: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, которые образуют выборку мощности n из двумерной генеральной совокупности. В общем случае обе измеряемые величины характеризуются ошибками измерений. Прежде всего изобразим полученные пары значений в прямоугольной системе координат. Тогда через экспериментальные точки, как правило, можно провести гладкую кривую, которая приближенно описывает результаты. На рис. 15 показаны экспериментальные точки, группирующиеся вдоль прямой линии. В случае подобных линейных зависимостей обычно можно достаточно точно провести прямую "на глаз". Однако наилучшая из возможных прямых (*прямая регрессии)* получается, если использовать

объективный метод - так называемый *метод наименьших квадратов Гаусса* (МНК).

Для простоты будем считать величины x независимыми переменными, значения которых измерены с пренебрежимо малой ошибкой. Пусть величина y_i соответствующая значению x_i отклоняется от истинной величины $y(x_i)$ на $y_i - y(x_i) = \varepsilon_i$.

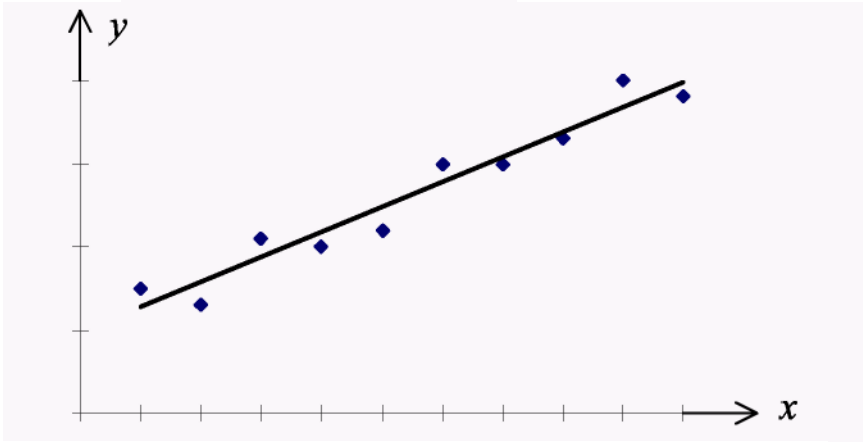


Рис. 15. Экспериментальные значения и линия регрессия.

Наилучшей прямой:

$$y = ax + b \quad (2.95)$$

является такая, на которой достигается минимум суммы квадратов отклонений ε_i.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 = S(a,b,n) \to \min . \quad (2.96)$$

Условием минимума является равенство нулю первых частных производных: $\partial S / \partial a = 0$ и $\partial S / \partial b = 0$. Отсюда получим:

$$\begin{cases} \sum_{i=1}^{n}(x_i y_i - a x_i^2 - b x_i) = 0, \\ \sum_{i=1}^{n}(y_i - a x_i - b) = 0 \end{cases} \quad (2.97)$$

(2.98)

Решая эту систему уравнений для неизвестных a и b, найдем:

$$a = \frac{\sum_1^n x_i \sum_1^n y_i - n\sum_1^n x_i y_i}{\left(\sum_1^n x_i\right)^2 - n\sum_1^n x_i^2},$$

$$a = \frac{\sum_1^n x_i \sum_1^n x_i y_i - \sum_1^n y_i \sum_1^n x_i^2}{\left(\sum_1^n x_i\right)^2 - n\sum_1^n x_i^2}. \quad (2.99)$$

(2.100)

Построим среднее арифметическое всех значений x_i и y_i:

$$\overline{x} = \frac{1}{n}\sum_1^n x_i \text{ и } \overline{y} = \frac{1}{n}\sum_1^n y_i .$$

Тогда уравнение (2.98) будет иметь вид:

$$\overline{y} = a\overline{x} + b. \quad (2.101)$$

Полученная прямая идет через эти средние значения, поэтому можно записать:

$$y - \overline{y} = a(x - \overline{x}). \quad (2.102)$$

Наклон прямой называют коэффициентом регрессии. Мерой разброса значений y_i возле прямой регрессии является дисперсия S_n^2:

$$S_n^2 = \sum_1^n \frac{\left[y_i - y(x_i)\right]^2}{n-2} = \frac{S(a,b,n)\big|_{\min}}{n-2}. \quad (2.103)$$

Число степеней свободы равно здесь *n-2*, так как для определения прямой регрессии необходимо выполнение двух дополнительных условий. Дисперсии величин *a* и *b* равны:

$$S_a^2 = \frac{S_n^2 \cdot n}{n\sum_1^n x_i^2 - \left(\sum_1^n x_i\right)^2}, \quad (2.104)$$

$$S_b^2 = \frac{S_n^2 \sum_1^n x_i^2}{n\sum_1^n x_i^2 - \left(\sum_1^n x_i\right)^2}. \quad (2.105)$$

Для наклона прямой регрессии *a* и отсекаемого ею отрезка *b* в принципе справедливы все соображения изложенные в разделе 2.2.3., если бы было известно распределение вероятности для двумерной генеральной совокупности. Здесь вновь возникает вопрос о доверительном интервале, который показывает, с какой статистической достоверностью эти величины можно определить по данной выборке. Если *x* и *y* распределены нормально, то доверительный интервал определяется с использованием распределения Стьюдента. Если обе переменные равноправны или между ними нет функциональной зависимости, то для обработки результатов измерений используется корреляционный анализ. Задача о *нелинейной регрессии* решается аналогично, причем в качестве кривых регрессии используют полиномы разной степени. Ниже приведены соответствующие результаты.

2.3.2. Нелинейная регрессия

Пусть зависимость между величинами *x* и *y* дана в виде полинома:

$y = a_0 + a_1 x + \ldots + a_n x^n$. Требуется определить неизвестные параметры $a_0, a_1, \ldots, a_n$. Алгоритм решения этой задачи строится следующим образом:

1. Проводится N совместных измерений величин x и y $(N > n+1)$.

2. Составляется система условных уравнений:

$$y_i = a_0 + a_1 x_i + \ldots + a_n x_i^n + \varepsilon_i, \quad i = 1,2,\ldots,N$$

где $\varepsilon_i = y_i - y(x_i)$ - как и выше отклонение измеренного значения y_i от истинного $y(x_i)$.

3. В предположении, что результаты измерений распределены нормально, взаимнонезависимы и ошибкой измерения x_i можно пренебречь, оценки параметров могут быть получены минимизацией суммы квадратов отклонений (невязок):

$$S = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} \left[y_i - (a_0 + a_1 x_i + \ldots + a_n x_i^n \right]^2 \to \min,$$

где параметры a_k рассматриваются как неизвестные. Приравнивая к 0 первые производные от S по каждому параметру получаем систему уравнений:

$$\frac{\partial S}{\partial a_0} = 0; \quad \frac{\partial S}{\partial a_1} = 0; \quad \ldots; \quad \frac{\partial S}{\partial a_n} = 0.$$

После несложных преобразований система нормальных уравнений записывается в виде:

$$\sum_{i=m_{\min}}^{m_{\max}} a_j \left[X^{j+k} \right] = \left[YX^k \right], \qquad (2.106)$$

где $m_{\min}$ - наименьшая степень полинома; $m_{\max}$ - наибольшая степень полинома; $k = m_{\min}, \ldots, m_{\max}$; $[\cdot]$ - скобки Гаусса:

$$\left[X^{j+k} \right] = \sum_{i=1}^{N} x_i^{j+k}; \left[YX^k \right] = \sum_{i=1}^{N} y_i x_i^k.$$

4. Оценки параметров получаются решением системы (2.106):

$$\hat{a}_j = \frac{\Delta_j}{\Delta}, \qquad (2.107)$$

где Δ - главный определитель системы; Δ_j- - определитель, получаемый из главного заменой *j*-го столбца столбцом правых частей.

5. Оценки дисперсий параметров даются выражением:

$$S^2_{\hat{a}_j} = \frac{\Delta_{jj}}{\Delta} \cdot S^2_0, \quad (2.108)$$

где Δ_{jj} - определитель, получаемый из главного вычеркиванием *j*-го столбца и *j*-й строки;

$$S^2_0 = \frac{1}{N-(n+1)} \sum_{i=1}^{N} \left[y_i - (a_0 + a_1 x_i + \ldots + a_n x_i^n \right]^2 . \quad (2.109)$$

6. Оценка дисперсии функции равна:

$$S^2_{y(x)} = S^2_{\hat{a}_0} + x^2 S^2_{\hat{a}_1} + \ldots + x^{2n} S^2_{\hat{a}_n}$$

7. Доверительные интервалы для параметров и функции:

$$\Delta\hat{a}_j = t(P,k) S_{\hat{a}_j}; \Delta y(x) = t(P,k) S_{y(x)},$$

где *t(P, k)* - коэффициент, определяемый из таблиц распределения Стьюдента для доверительной вероятности *P* и *k=n+1*- степеней свободы.

Рассмотрим несколько частных случаев:

1)зависимость имеет вид: $y=a_0$. Составим систему (2.106):

$$m_{\min} = 0; m_{\max} = 0 \Rightarrow j, k = 0 \Rightarrow a_0 \left[x^0\right] = \left[yx^0\right] = [y] \Rightarrow \hat{a}_0 = \frac{[y]}{\left[x^0\right]} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

$$S^2_{\hat{a}_0} = \frac{1}{N} S^2_0; S^2_0 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \hat{a}_0)^2$$

2) зависимость имеет вид: $y = a_0 + a_1 x$ Составляем систему (2.106):

$$m_{\min} = 0; m_{\max} = 1 \Rightarrow j = 0;1; k = 0;1 \Rightarrow \begin{cases} a_0 N + a_1 [x] = [y] \\ a_0 [x] + a_1 \left[x^2\right] = [yx] \end{cases}$$

Из формулы (2.107) найдем:

$$\hat{a}_0 = \frac{\Delta_0}{\Delta}; \Delta = \begin{vmatrix} N & [x] \\ [x] & \left[x^2\right] \end{vmatrix}; \Delta_0 = \begin{vmatrix} [y] & [x] \\ [yx] & \left[x^2\right] \end{vmatrix}; [x] = \sum_{i=1}^{N} x_i; \left[x^2\right] = \sum_{i=1}^{N} x_i^2; [y] = \sum_{i=1}^{N} y_i;$$

$$[yx] = \sum_{i=1}^{N} y_i x_i; \hat{a}_1 = \Delta_1 \Big/ \Delta; \Delta_1 = \begin{vmatrix} N & [y] \\ [x] & [yx] \end{vmatrix}$$

Из (2.108) имеем для дисперсий:

$$S^2_{\hat{a}_0}=\frac{[x^2]}{\Delta}S^2_0;S^2_{\hat{a}_1}=\frac{N}{\Delta}S^2_0;S^2_0=\frac{1}{N-2}\sum_{i=1}^{N}[y_i-(\hat{a}_0+\hat{a}_1x_i)]^2;\ S^2_{\hat{y}(x)}=S^2_{\hat{a}_0}+x^2S^2_{\hat{a}_1}.$$

3) зависимость имеет вид: $y=a_1x$.

Составляемсистему (2.106):

$$m_{\min}=1;m_{\max}=1;\ \Rightarrow\ j=1;k=1\ \Rightarrow\ a_1\left[x^2\right]=[yx]\ \Rightarrow$$

$$\Rightarrow\ \hat{a}_1=\frac{[xy]}{\left[x^2\right]}=\sum_{i=1}^{N}\frac{y_ix_i}{\sum_{i=1}^{N}x_i^2}$$

$$S^2_{\hat{a}_1}=\frac{S^2_0}{\sum_{i=1}^{N}x_i^2};S^2_0=\frac{1}{N-1}\sum_{i=1}^{N}(y_i-a_1x_i)^2;S^2_{\hat{y}(x)}=x^2S^2_{\hat{a}_1}$$

2.4. Методы оценки числа измерений

Для того, чтобы определить ошибку измерения, необходимы многократные измерения. Интуитивно ясно, что выполнять очень много измерений бессмысленно, так как точность результата всегда ограничена точностью метода. Кроме того из (2.15) следует, что СКО выборочного среднего $S_{\bar{x}}$ убывает с увеличением числа измерений n не слишком быстро, всего лишь как $\sqrt{n}$, тогда как затраты на эксперимент растут пропорционально n, т.е. могут быстро превысить потери от неточного знания результата измерения. Для получения объективных оценок необходимого числа измерений используется несколько методов.

2.4.1. Оценка числа измерений, необходимого для получения $\bar{x}$ с требуемой точностью

а) **Точечная оценка**. Потребуем, чтобы СКО среднего не превышало некоторого допустимого значения:

$$S_{\bar{x}}\le S_0,\ (2.112)$$

где S_0 - допустимое значение СКО среднего, либо систематическая ошибка, например, ошибка метода. Тогда из (2.112) следует, что

$$\frac{S_n}{\sqrt{n}} \le S_0,$$

откуда и получаем оценку для *n:*

$$n_{\bar{x}} \ge \frac{S_n^2}{S_0^2} \qquad (2.113)$$

где S_n^2 - выборочная дисперсия единичного измерения.

Вводя относительные ошибки: $\varepsilon_n = S_n / \bar{x}$ **и** $\varepsilon_0 = S_0 / \bar{x}$,

можно (2.113) представить в виде:

$$n_x \ge \frac{\varepsilon_n^2}{\varepsilon_0^2} \qquad (2.114)$$

б) Более сильная точечная оценка может быть получена из критерия ничтожных ошибок. Пусть ошибка измерения является суммой нескольких составляющих $S_\Sigma = \sqrt{\sum_k S_{(k)}^2}$, в частности

$$S_\Sigma^2 = S_{сл.}^2 + S_{сист.}^2,$$

где $S^2_{сл}$ - ошибка выборочного среднего, $S^2_{сист}$ - например, методическая ошибка. Из критерия ничтожных ошибок следует, что частной ошибкой можно пренебречь, если $S_k \le 0{,}3 S_\Sigma$, т. е. случайной ошибкой можно

пренебречь, если $S_{ñèñò} / S_{c'.} \ge 3$, что дает:

$$n \ge 11\left(\frac{S_n}{S_{NENO}}\right)^2 \qquad (2.115)$$

в) **Интервальная оценка:**

Для получения этой оценки потребуем, чтобы доверительный интервал для среднего не превышал некоторого допустимого значения:

$$t_{\alpha,k} \cdot S_{\bar{x}} \le \Delta_0 \qquad (2.116)$$

где Δ_0 - допустимое значение доверительного интервала. Из (2.116) следует, что

$$n_{\bar{x}} \geq \left(\frac{t_{\alpha,k} \cdot S_n}{\Delta_0} \right)^2 \qquad (2.117)$$

где $t_{\alpha,k}$ - значение определяемое из таблиц распределения Стьюдента при уровне значимости *a* и числе степеней свободы $k = n - 1$. Соотношение (2.117) является функциональным уравнением, так как *n* входит в обе части неравенства. Оно решается приближенно численным методом, подбором *n* с использованием таблиц распределения Стьюдента. Если $n>30$, то для оценки *n* можно использовать нормированное нормальное распределение и (2.117) преобразуется к виду:

$$n_{\bar{x}} \geq \left(\frac{u_\alpha \cdot S_n}{\Delta_0} \right)^2, \qquad (2.118)$$

где u_α - определяется из таблицы н. н. р. и уже не зависит от *n*.

г) **Точечная оценка числа измерений при косвенных измерениях.** Пусть величина *y* связана с непосредственно измеряемыми величинами известной функциональной зависимостью (см. раздел 2.2.5.)

$$y_N = f\left(x_{n_1}^{(1)}, \ldots, x_{n_k}^{(k)}\right).$$

Тогда оценка для *n* может быть получена из условия равных влияний, а именно, потребуем, чтобы парциальные ошибки по всем аргументам были одинаковы (см. раздел 2.2.5.):

$$\left| \frac{\partial f}{\partial x_{n_1}^{(1)}} \right| \cdot S_{\bar{x}_{n_1}^{(1)}} \approx \ldots \approx \left| \frac{\partial f}{\partial x_{n_k}^{(k)}} \right| \cdot S_{\bar{x}_{n_k}^{(k)}} = \frac{S_{\bar{Z}_N}}{\sqrt{k}} \qquad (2.119)$$

где $\bar{Z}_N = f\left(\bar{x}_{n_1}^{(1)}, \ldots, \bar{x}_{n_k}^{(k)}\right).$

Отсюда для числа измерений найдем:

$$n_i = \frac{k\left(\frac{S_{n_i}}{S_{\overline{Z}_N}}\right)^2}{\left.\frac{\partial f}{\partial x^{(i)}}\right|^2_{\overline{x}^{(i)}_{n_i}}} \qquad (2.120)$$

где n_i - число измерений по аргументу $x^{(i)}$; значение частной производной берется при среднем значение аргумента.

2.4.2. Оценка числа измерений, необходимого для получения СКО среднего с требуемой точностью

В разделе 2.2.3. было показано, что доверительный интервал для дисперсии определяется распределением Пирсона, а именно:

$$\chi^2_{k;\frac{\alpha}{2}} \le \frac{(n-1)S_n^2}{\sigma^2} \le \chi^2_{k;1-\frac{\alpha}{2}} \text{ с доверительной вероятностью } P.$$

Отсюда после несложных преобразований получим для минимального числа измерений

$$n_S - 1 = \frac{4\varepsilon^2}{\left(\frac{1}{\chi_{k;1-\frac{\alpha}{2}}} - \frac{1}{\chi_{k;\frac{\alpha}{2}}}\right)^2}, \qquad (2.121)$$

где ε— относительная ошибка измерения СКО; $k = n - 1$.
Соотношение (2.121) является функциональным уравнением, так как n входит в обе части. Оно решается приближенно с использованием таблиц распределения Пирсона.
При $n>30$ число измерений n можно оценить из соотношения:

$$n_S - 1 = 0{,}5 \cdot \left(\frac{u_\alpha}{\varepsilon}\right)^2, \qquad (2.122)$$

где u_α - находится из н. н. р. и не зависит от n.

2.4.3. Оценка числа измерений для определения допустимых границ

Пусть требуется оценить минимальной число измерений, необходимое для того, чтобы с заданной вероятностью *P* некоторая часть генеральной совокупности γ находилась между минимальным и максимальным выборочными значениями. Оценка находится из уравнения Уилкса:

$$n\gamma^{n-1} - (n-1)\gamma^{n} = \alpha \quad (2.123)$$

В табл. 2-6 приведены результаты расчетов числа измерений различными методами.

Таблица 2

Расчет минимального числа измерений $(n_{\bar{x}})$ для получения $\bar{x}$ с требуемой точностью

S_n / S_0	1	2	3	5	10	15	20	25
$n_{\bar{x}}$ по (2.113)	1	4	9	25	100	225	400	625

Таблица 3

Расчет $n_{\bar{x}}$ по (2.115)

$S_n / S_{сист}$	1	2	3	4	5	10	20
$n_{\bar{x}}$ по (2.115)	11	44	99	176	275	1100	4400

Таблица 4

Расчет $n_{\bar{x}}$ по (2.117) и (2.118)

S_n/Δ_0	1	2	3	4	5	10	20
$n_{\bar{x}}$ по (2.117) P=0,95	7	19	36	44	68	165	660
$n_{\bar{x}}$ по (2.117) P=0,99	9	26	49	87	135	233	932
$n_{\bar{x}}$ по (2.118) P=0,95	3	11	25	44	68	165	660
$n_{\bar{x}}$ по (2.118) P=0,99	6	22	49	87	135	233	932

Таблица 5

Расчет n_S по (2.122)*

ε	0,01	0,02	0,05	0,10	0,30	0,50	1
n_S по (2.122) P=0,95	13600	3400	545	137	17	7**	3**
n_S по (2.122) P=0,99	27100	6775	1085	272	32	12**	4**

*) Оценки n_S по (1.121) отличаются от приведенных лишь при больших ε, дляР=0,95: n=6 при ε=0,71; n=11 при ε=0,36; n=4 при ε=1; n=21 при ε=0,28; для Р=0,99: *n=7* при *ε=1;* n=11 при ε=0,66; n=21 при ε=0,41.

**) Оценки даны с округлением.

Таблица б

Расчет *n* по (2.123)

P	γ					
	0,50	0,90	0,95	0,99	0,999	0,9999
0,50	3	17	34	168	1679	16783
0,80	5	29	59	299	2994	29943
0,90	7	38	77	388	3889	38896
0,95	8	46	93	473	4742	47437
0,99	11	64	130	662	6636	66381
0,999	14	89	181	920	9230	92330
0,9999	18	113	230	1171	11751	117559

2.5. Статистическая проверка гипотез

Проверка гипотез, наряду с задачей статистической оценки параметров, рассмотренной в предыдущих параграфах, составляет одну из важнейших процедур принятия статистических решений. Она широко используется в измерительном эксперименте при анализе данных и обработке результатов измерений. Под гипотезой H_0 понимается некоторое предположение о случайной величине x (о виде распределения, параметрах распределения и т.п.). Путем статистической проверки необходимо установить, насколько данные, полученные из выборки $(x_1, x_2, \ldots, x_n)$, согласуется с гипотезой, т. е. можно ли на их основании принять или отвергнуть гипотезу. Абсолютно надежное решение получить нельзя. Необходимо заранее допустить возможность ошибочного решения. Обозначим через α вероятность того, что гипотеза H_0 будет отвергнута, хотя на самом деле она верна. Ее называют также уровнем значимости проверки гипотезы или вероятностью ошибки первого рода. Эта величина или величина $P = 1 - \alpha$, называемая статистической достоверностью или доверительной вероятностью, т. е. вероятностью принять правильную гипотезу, должны быть выбраны экспериментатором. При решении экономических или технических проблем, обычно выбирают $\alpha = 0{,}05$ или $\alpha = 0{,}01;$ в медицинских исследованиях, где цена ошибки очень высока, полагают $\alpha \approx 0{,}001.$

Процедура проверки гипотезы заключается в следующем: выбирается некоторая подходящая выборочная функция (критерий проверки гипотезы) $T(x_1, x_2, \ldots, x_n; H_0)$, определяемая выборкой и выдвинутой гипотезой H_0. Затем устанавливается область K, в которую в случае справедливости гипотезы H_0 значение функции T попадает с вероятностью $P=\alpha$. Область K называется критической областью. Если конкретное значение функции T, найденное по выборке $T(x_1, x_2, \ldots, x_n; H_0)$ попадает в критическую область K, то гипотеза отклоняется, в противном случае принимается. При этом вероятность того, что гипотеза H_0 будет отвергнута в случае, когда на самом деле она верна, оказывается равной заданной величине α. При любом значении α существует множество различных возможностей для выбора критической области. Наиболее часто используется три типа критической области: симметричная,

квазисимметричная, и односторонняя. На рис. 16 приведена функция плотности распределения *f(t)* для критерия *T:*

$$P\{T < t\} = \int_{-\infty}^{t} f(t)\,dt, \quad (2.124)$$

которая симметрична относительно нуля и близка по виду к кривой нормального распределения или распределения Стьюдента.
Критическая область выбрана здесь симметричной относительно нуля, а именно, *K:*

$$|t| \geq \varepsilon.$$

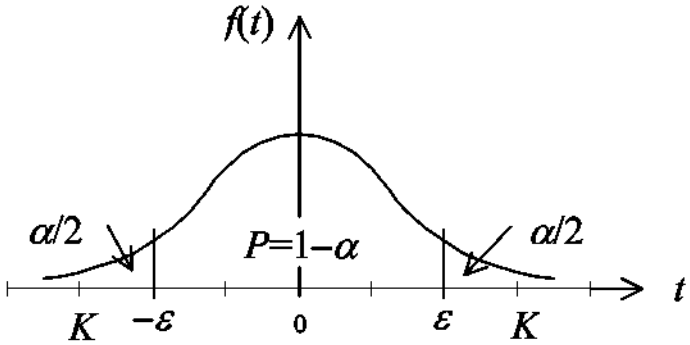


Рис.16. Симметричная критическая область.

На рис. 17, 18 приведены функции плотности распределения, близкие по виду χ^2 и F- распределению Фишера. Критическая область на рис. 18 расположена в диапазоне больших значений критерия, $K: t \geq \varepsilon.$ На рис. 17 показана квазисимметричная критическая область, *K:*
$0 \leq t \leq \varepsilon_1;\ \varepsilon_2 \leq t,$ одна часть которой располагается правее нуля, а другая - в области больших значений критерия.
Заштрихованная часть, расположенная над критической областью, на всех трех рисунках имеет площадь равную *α.*

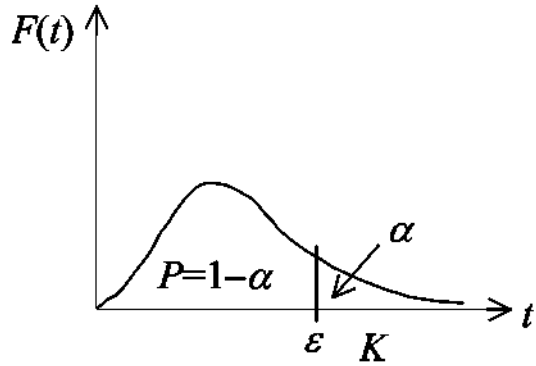


Рис.17. Квазисимметричная критическая область.

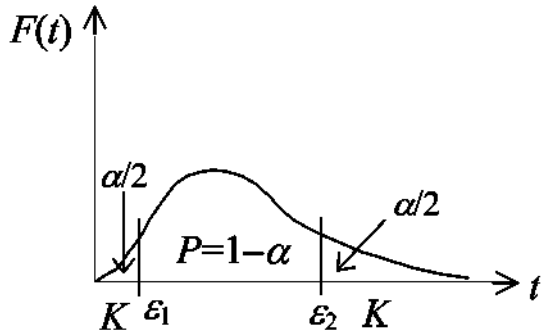


Рис. 18. Односторонняя критическая область.

Выбор критической области определяется из следующих соображений.

Обозначим $\overline{H_0}$ - альтернативную гипотезу по отношению к H_0.

Тогда, $P_{\overline{H_0}}\{T \in K\}$ - вероятность того, что H_0 отвергается, если она не верна, т.е. когда истинной является альтернативная гипотеза. Эта вероятность характеризует избирательность критерия и называется мощностью критерия. Чем больше мощность критерия проверки гипотезы, тем меньше вероятность ошибки второго рода $\beta = 1 - P_{\overline{H_0}}$, характеризующей вероятность принять неправильную гипотезу.

При заданной вероятности ошибки первого рода критическая область выбирается так, чтобы обеспечить максимальную избирательность критерия:

$$P_{\overline{H_0}}(T \in K) = \max . \quad (2.125)$$

Введение двух пороговых вероятностей α и β отражает тот факт, что принятие статистического решения - это всегда компромисс между необходимым и возможным. Возможное (риск исполнителя) характеризуется значением вероятности α, а необходимое (риск заказчика) - значением β. Рассмотрим несколько примеров, иллюстрирующих выбор критической области.

2.5.1. Проверка гипотезы о среднем значении нормально распределенной случайной величины *x* с известной дисперсией

Гипотеза H_0: среднее значение $E(x)=m_0$ (постоянная величина); гипотеза $\overline{H_0}: E(x)\neq m_0$.

В качестве критерия выберем функцию:

$$T = \frac{\overline{x}_n - m_0}{\sigma/\sqrt{n}}, \quad (2.126)$$

где $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ - выборочное среднее.

Отметим, что критерий всегда выбирается безразмерной величиной, поэтому числитель $(\overline{x}_n - m_0)$ делится на СКО выборочного среднего $\sigma/\sqrt{n}$. Критерий T при справедливости гипотезы H_0 имеет нормированное нормальное распределение. Критическую область K, соответствующую уровню значимости α, выберем симметричной, $K: |t| \geq \varepsilon$.

Для определения ε решается уравнение:

$$P\{|T| < \varepsilon\} = \Phi(\varepsilon) - \Phi(-\varepsilon) = P = 1 - \alpha, \quad (2.127)$$

где Φ(ε) - функция нормированного нормального распределения.

Если T попадает в область K, то гипотеза H_0 отвергается. Поясним этот вывод расчетами. Пусть из нормально распределенной совокупности с дисперсией $\sigma^2 = 25$ извлечена выборка объема $n=16$, с помощью которой получена оценка среднего $x_n = 22$. Требуется проверить гипотезу $E(x)=20$. Зададим уровень значимости $\sigma=0{,}05$; по таблице нормального распределения найдем $\varepsilon=1{,}96$. При этом критическая область $K: |t| > 1{,}96.$

Подсчитаем выборочное значение критерия:

$$t = T = \frac{22-20}{\sqrt{25}/\sqrt{16}} = 1{,}6 < \varepsilon,$$

т.е. гипотеза H_0 принимается.

2.5.2. Проверка гипотезы о значении дисперсии нормально распределенной случайной величины *x* при неизвестном среднем

Гипотеза $H_0: \sigma^2 = \sigma_0^2$ (постоянная величина);

$$\overline{H_0}: \sigma^2 \neq \sigma_0^2.$$

В качестве критерия используем функцию:

$$T = \frac{(n-1)S^2}{\sigma_0^2}, \quad (2.128)$$

где S^2 - выборочная дисперсия (оценка дисперсии σ^2).
Если H_0 справедлива, то T подчиняется χ^2 - распределению Пирсона. Критическую область, соответствующую уровню значимости α выберем квазисимметричной:

$$P\{T < \varepsilon_1\} = P\{T > \varepsilon_2\} = \frac{\alpha}{2}. \quad (2.129)$$

Если T попадает в критическую область $K: 0 \le T \le \varepsilon_1;\ \varepsilon_2 \le T$, то гипотезу H_0 следует отвергнуть. Проведем расчеты. Пусть из нормально распределенной совокупности извлечена выборка объема $n=40$, с помощью которой рассчитана оценка дисперсии $S^2=20{,}61$. Требуется проверить гипотезу $\sigma^2 = 20$. Зададим уровень значимости $\alpha=0{,}05$ и по таблице распределения Пирсона найдем

$\varepsilon_1=24,4;\quad \varepsilon_2=59,3.$ При этом критическая область $K:\ 0 \le \chi^2 \le \varepsilon_1;\ \ \chi^2 \ge \varepsilon_2.$ Подсчитаем выборочное значение критерия: $\chi^2 = T = \frac{39 \cdot 20,61}{20} = 40,2.$ Так как T не попадает в критическую область, то гипотеза H_0 принимается.

В рассмотренных примерах критическая область выбрана наилучшим образом в смысле обеспечения максимальной избирательности критерия. Положение критической области существенно зависит также от выбора альтернативной гипотезы $\overline{H_0}$. Мы выбираем $\overline{H_0}: E(x) \neq m_0$ (в первом примере) либо $\sigma^2 \neq \sigma_0^2$ (во втором примере). Если использовать другую гипотезу, например, $\overline{H_0}:\ \sigma^2 > \sigma_0^2,$ то в качестве критической следует выбрать одностороннюю критическую область на рис.18.

Использованные критерии зависят от вида распределения случайной величины *x;* такие критерии называют параметрическими. Существует и другая группа критериев, применение которых не связано с предположениями о законе распределения. Они называются непараметрическими. В табл. 7 приведены наиболее распространенные виды критериев и области их применения.

Таблица 7

Статистическиекритерии проверки гипотез

Параметрические критерии		Непараметрические критерии	
Критерий	Область применения	Критерий	Область применения
Аббе Стьюдента	Проверка гипотез об однородности, независимости, стационарности данных; проверка	Серий Знаков Уилкоксона или ранговых сумм (одно- и	Проверка гипотез об однородности, независимости, стационарности

	гипотез о средних	двухвыборочный).	и сдвиге.
Пирсона Фишера Кокрена Бартлета	Проверка гипотез о дисперсиях; проверка гипотез о равнорассеянности данных	Ансари-Бредли	Проверка гипотез о дисперсиях
Хотеллинга Шеффе Уилкса	Проверка гипотез об однородности многомерных совокупностей		
Пирсона Крамера-Мизеса-Смирнова Колмогорова	Проверка гипотез о согласии (соответствии) выбранной модели распределения с исходными данными.		

Рассмотрим несколько часто применяемых непараметрических критериев, свободных от вида распределения.

2.5.3. Проверка гипотез о независимости и стационарности данных

Пусть имеется последовательность, состоящая из *m* элементов *а* и *n* элементов *в (а* - знак «плюс», *в* - знак «минус»). Серией называется часть последовательности, состоящая из элементов одного вида. Обозначим *k* -общее число серий в данной последовательности. Гипотеза H_0*:* элементы *а* и *в* расположены случайно; гипотеза $\overline{H_0}$**:** в расположении *а* и *в* наблюдается закономерность. Для проверки гипотезы используется так называемый критерий серий, который имеет вид (при больших *m, n* и отношении *m/n):*

$$T = \frac{k - E(k) + 0{,}5}{\sqrt{D(k)}}, \quad (2.130)$$

где $E(k) = 1 + \frac{2mn}{m+n}$ (математическое ожидание величины *k),(2.*131)

$$D(k) = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)} \text{ (дисперсия } k\text{)}, \quad (2.132)$$

0,5 - поправка на непрерывность.

Если $|T| \geq u_{1-\alpha/2}$, где $u_{1-\alpha/2}$ определяется по таблицам нормированного нормального распределения, то гипотеза H_0 отвергается, в противном случае она принимается.

2.5.4. Проверка гипотез о положении (сдвиге), симметрии распределения, однородности данных

Пусть имеется две выборки x_i и y_i (i=1,2,...,n), являющиеся результатами измерений на n объектах, например, до и после воздействия. Предполагается, что элементы обеих выборок взаимно независимы и подчиняются непрерывным распределениям. Гипотеза H_0: значение медианы разностей $z_i = x_i - y_i$ равно нулю (эффект воздействия отсутствует); гипотеза $\overline{H_0}$: значение медианы отлично от нуля (эффект имеется). Для проверки гипотезы о сдвиге (однородности) применяется одновыборочный критерий Уилкоксона. Процедура проверки состоит из следующих шагов:

- вычисляется разность $z_i = x_i - y_i;\ i=1,2,...,n;$
- строится вариационный ряд из абсолютных значений z_i (по возрастанию значений);
- значениям $|z_i|$ присваиваются ранги в общей последовательности, при этом нулевые разности отбрасываются, а для совпадающих значений $|z_i|$ определяются их средние ранги;
- каждому рангу присваивается знак величины z_i в соответствии со знаком разности (см. выше).

Затем вычисляется значение критерия:

$$T = \min(\Sigma_1, \Sigma_2), \quad (2.133)$$

где $\sum_1$ - сумма рангов положительных значений z_i, Σ_2 - сумма рангов отрицательных значений z_i.

Решение принимается следующим образом: если $T \geq W(\alpha_2, k)$ или $T \leq \left(k(k+1)/2 - W(\alpha_1, k)\right)$, то гипотеза H_0 отвергается; если же $k(k+1)/2 - W(\alpha_1, k) < T < W(\alpha_2, k)$,

то гипотеза принимается. Здесь k - число ненулевых разностей z_i; $W(\alpha_1,k)$, $W(\alpha_2,k)$ - табличные значения критерия Уилкоксона; $\alpha=\alpha_1+\alpha_2$ (обычно $\alpha_1=\alpha_2=\alpha/2$).

2.6. Определение вида закона распределения значений измеряемой величины

При обработке экспериментальных данных, а именно при определении выборочно среднего, дисперсии и доверительного интервала используется информация о виде закона распределения вероятности значений измеряемой величины.

Наиболее распространенным является случай нормального распределения, так как в большинстве реальных экспериментальных ситуаций справедлива центральная предельная теорема, из которой следует, что при достаточно больших объемах выборок распределение выборочных средних, полученное из различных исходных функций распределения, достаточно хорошо описывается нормальным распределением. Кроме того, многие практически важные распределения (Пирсона, Стьюдента и др.) уже при n>30 мало отличаются от нормального. В этом случае для обработки результатов используются стандартные (классические) статистические процедуры, рассмотренные в §2.3. Их недостаток состоит в том, что они весьма чувствительны к довольно малым отклонениям от предположений о нормальности, и если истинная функция распределения отличается от нормальной, то классические методы уже нельзя бездумно использовать. Поэтому появились так называемые робастные процедуры, т. е. нечувствительные к малым отклонениям от предположений. Приведем пример, показывающий к каким последствиям приводит отсутствие робастности по распределению в классических процедурах обработки .

Пример. Предположим, что имеется выборка объема n, составленная из большого числа "хороших" и "плохих" случайно перемешанных измерений x_i некоторой величины m. Каждое "хорошее" измерение появляется с вероятностью $1-\varepsilon$, а "плохое" - с вероятностью ε, где ε - малое число. Хорошие измерения x_i имеют нормальное распределение $N\left(m,\sigma^2\right)$, плохие - нормальное распределение $N\left(m,9\sigma^2\right)$.

Иными словами все значения имеют одно и то же среднее *m,* а стандартное отклонение для некоторых из них (плохих) в три раза больше (дисперсия равна $9\sigma^2$), чем у остальных (дисперсия равна σ^2). Приведенная ситуация описывается следующим образом: величины x_i независимы и имеют одно и то же распределение:

$$F(x) = (1-\varepsilon)\Phi\left(\frac{x-m}{\sigma}\right) + \varepsilon\,\Phi\left(\frac{x-m}{3\sigma}\right), \quad (2.134)$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy$ - функция н. н. р.

Рассмотрим две широко известные оценки разброса - среднее абсолютное отклонение:

$$d_n = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \quad (2.135)$$

и среднее квадратичное отклонение (СКО):

$$S_n = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}}. \quad (2.136)$$

Известно, что для нормально распределенных измерений величина S_n примерно на 12% более эффективна, чем d_n, однако на практике применение d_n оказывается зачастую более правильным. Разумеется, величинами S_n и d_n оцениваются разные характеристики распределения. Например, если измеренные значения имеют в точности нормальное распределение, то величина S_n сходится к σ, в то же время d_n стремится к $\sqrt{\frac{2}{\pi}}\sigma \approx 0{,}8\sigma$. Поэтому следует уточнить, как проводить сравнение этих оценок. Обычно используют так называемую асимптотическую относительную эффективность (АОЭ) оценки d_n по оценке S_n, определяемую следующим образом:

$$AOЭ(\varepsilon) = \lim_{n\to\infty} \frac{D(S_n)/(E(S_n))^2}{D(d_n)/(E(d_n))^2} = \frac{\left[3(1+80\varepsilon)/(1+8\varepsilon)^2 - 1\right]/4}{\pi(1+8\varepsilon)/\left(2(1+2\varepsilon)^2\right) - 1},$$

где $D(S_n),\ D(d_n)$ - дисперсии, а $E(S_n),\ E(d_n)$-математические ожидания соответствующих величин.

Значения этого показателя при различных ε приведены ниже:

ε	0	0,001	0,002	0,005	0,01	0,02
$АОЭ(\varepsilon)$	0,876	0,948	1,016	1,198	1,439	1,752
ε	0,05	0,10	0,15	0,25	0,5	1,0
$АОЭ(\varepsilon)$	2,035	1,903	1,689	1,371	1,017	0,876

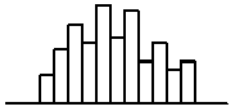
Приведенные результаты показывают, что уже двух плохих наблюдений на тысячу (ε=0,002) достаточно, чтобы свести на нет 12%-ое преимущество среднеквадратичной оценки, причем наибольшее значение, которое показатель *АОЭ(ε)* принимает вблизи ε=0,05 превосходит 2.

Следует отметить, что типичные выборки "хороших" данных довольно точно моделируются законом распределения (2.134); где ε меняется в пределах от 0,01 до 0,1. Приведенный пример убедительно показывает, что удлинение "хвостов" распределения ухудшает классическую оценку S_n и значительно меньше влияет на d_n.

Иногда возникают экспериментальные ситуации, когда распределение неизвестно, т. е. может быть произвольным. В этом случае подход к обработке данных зависит от требований к точности результатов. При этом классические оценки оказываются неэффективными, а иногда и просто не применимы (например, как уже отмечалось в §2.3. для распределения Коши математическое ожидание не определено, так как соответствующий интеграл расходится). Если требования к точности невысоки, то приемлемыми являются оценки, получаемые с использованием непараметрических критериев, например, знакоранговых или неравенства Чебышева. Однако, если требования к точности достаточно высокие, то информация о виде закона распределения является необходимой. Для определения вида закона распределения используются две группы методов: графические и аналитические. При использовании графических методов по имеющийся выборке объема *n* строится гистограмма (см. ниже), по виду гистограммы выдвигается гипотеза о законе распределения, а затем она проверяется по критериям согласия. В реальной ситуации, однако, редко получается "идеальная" форма гистограммы, так как она может маскироваться рядом факторов: интервал группирования данных, наличие ошибок и "плохих" данных и т. д. Поэтому принятие решения о законе распределения по виду гистограммы зачастую оказывается затруднительным или даже невозможным. На рис. 19 даны

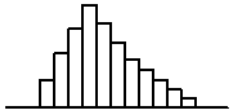
примеры наиболее типичных "неидеальных" гистограмм, встречающихся на практике .

1. Гребенка (мультимодальный тип). Классы через один имеют более низкие частоты. Такая форма встречается, когда число единичных наблюдений, попадающих в класс, колеблется от



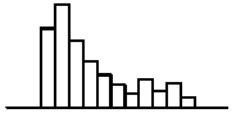
класса к классу или когда действует определенное правило округления данных.

2. Положительно (отрицательно) скошенное распределение. Среднее значение гистограммы локализовано слева (справа) от центра размаха. Частоты довольно резко спадают при движении



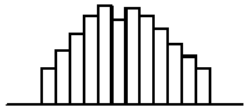
влево (вправо) и, наоборот, медленно вправо (влево). Форма асимметрична. Такая форма встречается, когда нижняя (верхняя) граница регулируется либо теоретически, либо по значению допуска или когда левое (правое) значение недостижимо.

3. Распределение с обрывом слева (справа). Среднее арифметическое гистограммы локализуется далеко слева (справа) от центра размаха. Частоты резко спадают при движении влево



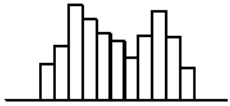
(вправо), и, наоборот, медленно вправо (влево). Форма асимметрична. Эта форма часто встречается при 100% просеивании изделий из-за плохой воспроизводимости процесса, а также, когда проявляется резко выраженная положительная (отрицательная) асимметрия.

4. Плато (равномерное и прямоугольное распределение). Частоты в разных классах образуют плато, так как все классы имеют более или менее одинаковые ожидаемые частоты. Эта



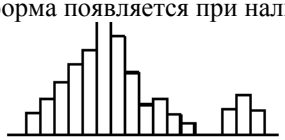
форма встречается в смеси нескольких распределений, имеющих разные средние.

5. Двухпиковый тип (бимодальный). В окрестностях центра диапазона данных частота низкая, зато есть по пику с каждой стороны. Эта форма встречается, когда смешиваются два



распределения с далеко отстоящими средними значениями.

6. Распределение с изолированным пиком. Наряду с распределением обычного типа появляется изолированный пик. Эта форма появляется при наличии малых включений данных из



другого распределения. Например, как в случае нарушения нормального процесса, появления ошибки измерения или простого включения данных другого процесса.

Рис. 19. Типы "неидеальных" гистограмм, встречающихся на практике.

В этом случае чтобы избежать ошибки следует использовать аналитические методы, которые состоят в определении по выборке объема *n* оценок различных показателей формы эмпирического распределения: асимметрии, эксцесса, коэффициента формы распределения, контрэксцесса, энтропийного коэффициента и т. п. Найденные оценки сравниваются с допустимыми значениями показателей для теоретических законов распределения. Если оценки показателей формы эмпирического распределения, найденные по выборке, попадают в интервал значений показателей, допустимый для какого-то теоретического распределения, то этот закон принимается в качестве гипотезы и проверяется его соответствие экспериментальным данным по критериям согласия. В реальном случае из-за ошибок измерения оценки некоторых показателей могут выходить за допустимые пределы для какого-то одного закона распределения, либо оценки разных показателей могут соответствовать разным теоретическим законам распределения. Поэтому приходится выдвигать несколько гипотез и проводить сравнение с несколькими теоретическими распределениями. При использовании ЭВМ это может быть сделано достаточно быстро. Рассмотрим обе группы методов более подробно.

2.6.1. Аналитические методы

К ним относятся:
- метод основанный на определении характеристик формы распределения: коэффициента асимметрии и коэффициента эксцесса;
- метод основанный на определении коэффициента формы распределения;
- метод основанный на определении энтропийного коэффициента и контрэксцесса.

Прежде чем использовать эти методы следует проверить наличие в выборке грубых ошибок (выбросов) и исключить их. Ясно, что признание того или иного измерения выбросом зависит от вида распределения. Для распределений, близких к нормальному, для исключения промахов используется правило "трех сигм" при доверительной вероятности P=0,9973. В случае, когда вид распределения заранее не известен, из выборки исключаются такие значения x_i для которых выполняются неравенства $x_i < x_{r-}$ или $x_i > x_{r+}$; x_{r+}, x_{r-} - границы выбросов, определяемые выражениями:

$$x_{r\mp} = \bar{x}_n \mp S_n\left(1 + A\sqrt{\frac{1}{\aleph_n^2} - 1}\right) \qquad (2.137)$$

где $\bar{x}_n$ - выборочное среднее; S_n - выборочное СКО; $\aleph_n$ - выборочный контрэксцесс: $\aleph_n = \left(S_n^4 / \mu_{4n}\right)^{\frac{1}{2}}$; μ_{4n} - выборочный четвертый центральный момент эмпирического распределения: $\mu_{4n} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^4$;

A - коэффициент, значение которого выбирается в зависимости от доверительной вероятности в диапазоне от 0,85 до 1,30 (рекомендуется выбирать максимальное значение A, соответствующее вероятности $P=0,9913$).

2.6.1.1. Определение выборочного коэффициента асимметрии γ_{an} и коэффициента эксцесса $\gamma_{эп}$

$$\gamma_{an} = \frac{\mu_{3n}}{S_n^3}, \quad (2.138)$$

где μ_{3n} - выборочный третий центральный момент эмпирического распределения: $\mu_{3n} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^3$.

$$\gamma_n = \frac{\mu_{4n}}{S_n^4} - 3, \quad (2.139)$$

2.6.1.2. Определение коэффициента формы распределение α

Коэффициент формы определяется из уравнения:

$$Э_n = \frac{\alpha\,\Gamma\left(1+\frac{1}{\alpha}\right)\cdot\Gamma\left(\frac{5}{\alpha}\right)}{\Gamma^2\frac{3}{\alpha}}, \quad (2.140)$$

где $Э_n$ - выборочный эксцесс: $Э_n = \mu_{4n} / S_n^4$; Г(x) - гамма-функция от *x*, определяемая по таблицам .
Уравнение (2.140) решается приближенно численными методами. Ниже в табл. 8 приведены значения α для различных значений эксцесса, полученные из (2.140).
Таблица 8

Значения α по (2.140)

α	0,5	0,55	0,73	0,77	0,83	0,9	1,0	1,5	2	6
$Э_n$	25	20	10	9	8	7	6	4	3	2

2.6.1.3. Определение энтропийного коэффициента

$$K_n = \frac{\Delta_{эп}}{S_n}; \quad \Delta_{эп} = \frac{d \cdot n}{2 \cdot 10^{-\frac{1}{n}\sum_{j=1}^{m} n_j \lg n_j}}, \tag{2.141}$$

где d - ширина интервала группирования данных; m - число интервалов группирования; n_j — число значений в каждом интервале (см. ниже).

Полученные выборочные значения величин γ_{an}, $\gamma_{эп}$, α_n, K_n и $\aleph_n$ сравниваются с допустимыми значениями этих же величин для различных теоретических распределений (см. табл. 9) и по их согласованию делается предварительное заключение о законе распределения. Для грубого отбора бывает достаточно использовать два значения γ_{an} и $\gamma_{эп}$, однако такой подход требует осторожности, особенно при $n<200$.

При сравнении следует иметь в виду, что из-за ошибок в данных может наблюдаться расхождение выборочных значений критериев и допустимых. Рекомендуется принять, что расхождение незначимо, если выполняется следующее соотношение:

$$\left|Z_n - Z_{доп}\right| \le (2...3) S_Z,$$

где Z_n, $Z_{доп}$ - выборочное и допустимое значение величины Z соответственно; S_z - выборочное СКО величины Z.

В частности, для γ_{an}: $S_{\gamma_{an}} = \left[\frac{6(n-1)}{(n+1)(n+3)}\right]^{\frac{1}{2}}$;

для $\gamma_{эп}$: $S_{\gamma_{эп}} = \left[\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}\right]^{\frac{1}{2}}$;

для α_n: $S_{\alpha_n} = \alpha_n \frac{S_{\gamma_{эп}}}{\gamma_{эп}}$; для $\aleph_n$: $S_{\aleph_n} = \frac{1}{2}\aleph_n \frac{S_{\gamma_{эп}}}{\gamma_{эп}}$;

для K_n: $S_{K_n} = K_n\left(\frac{2}{n-1}\right)^{\frac{1}{2}}$.

Полученное предварительное заключение о виде закона распределения рассматривается как гипотеза (их может быть и несколько). Чтобы принять окончательное решение с определенной статистической достоверностью необходимо гипотезу проверить по одному из критериев согласия (см. ниже).

Таблица 9

Допустимые значения показателей формы для различных распределений

Распределение	Допустимые значения показателей				
	γ_a	$\gamma_э$	α	$\aleph$	K
Нормальное	0	0	2	0,56	2,07
Треугольное	0	–0,6	5	0,65	2,02
Трапецеидальное	0	0…–1,2	2…10	0,58…0,74	1,7…2,07
Равномерное	0	–1,2	10	0,75	1,73
Симметричное экспоненциальное островершинное	0	0,75…22	0,5…1,5	0,2…0,52	1,35…2,02

При использовании аналитических методов наибольшую трудность представляет необходимость определения выборочных значений параметров распределения, принятого в качестве гипотезы, при которых достигается наибольшее соответствие между теоретическим и эмпирическим распределениями. Эта задача решается методом моментов. Он заключается в том, что для произвольного распределения значения параметров определяются приравниванием теоретических значений моментов их выборочным оценкам:

$$\alpha_1(\text{первый момент}) = \int_{-\infty}^{\infty} x p(x)\,dx = \frac{1}{n}\sum_{i=1}^{n} x_i\ ;$$

$$\alpha_2(\text{второй момент}) = \int_{-\infty}^{\infty} x^2 p(x)dx = \frac{1}{n}\sum_{i=1}^{n} x_i^2\ ;$$

$$\alpha_3(\text{третий момент}) = \int_{-\infty}^{\infty} x^3 p(x)dx = \frac{1}{n}\sum_{i=1}^{n} x_i^3 \quad \text{и т. д.}$$

Число уравнений берется равным числу определяемых параметров. Полученная система уравнений решается, и находятся неизвестные значения параметров распределения. Например, если эмпирическое распределение мы хотим описать функцией нормального распределения, то необходимо определить два параметра *m* и *σ*. В этом случае в качестве оценок для *m* и *σ* выбираем соответственно выборочные значения среднего и СКО. Аналогично определяют параметры распределения и в других случаях. Рассмотрим пример. Пусть в качестве гипотезы выбрано прямоугольное (равномерное) распределение. Оно содержит два неизвестных параметра *m* и *a (m* - центр распределения; *a* -полуинтервал). Составляем систему двух уравнений:

$$\begin{cases} \alpha_1 = \int_{-\infty}^{\infty} x p(x)\,dx = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \alpha_2 = \int_{-\infty}^{\infty} x^2 p(x)\,dx = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases}$$

Используя выражение для функции плотности:

$$p(x) = \begin{cases} 1/2a & \text{при} \quad x \in [m-a, m+a] \\ 0 & \text{при} \quad x \notin [m-a, m+a] \end{cases},$$

найдем из первого уравнения:

$$\int_{-\infty}^{\infty} x \frac{1}{2a}\,dx = \int_{m-a}^{m+a} \frac{1}{2a} x\,dx = \frac{1}{2a}\frac{x^2}{2}\Big|_{m-a}^{m+a} = m$$

Отсюда находим параметр m: $m = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$.

Аналогично из второго уравнения найдем *a:*

$$\int_{-\infty}^{\infty} x^2 \frac{1}{2a} dx = \frac{1}{2a}\int_{m-a}^{m+a} x^2 dx = \frac{1}{2a}\frac{x^3}{3}\Bigg|_{m-a}^{m+a} = m^2 + \frac{a^2}{3}$$

Отсюда получаем: $\frac{a^2}{3} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m^2 = \overline{x^2} - \left(\overline{x}\right)^2$.

2.6.2. Графические методы

Графические методы сводятся к построению гистограммы и полигона. Для этого проводят вспомогательные вычисления.
Сначала значения в выборке ранжируют, т. е. располагают в порядке возрастания, и получают вариационный ряд:

$$x_1 \le x_2 \le \ldots \le x_n.$$

Такое представление является удобным, так как наглядно и позволяет сократить время при последующих расчетах.
Затем определяют число интервалов группирования *m:*

$$m = \frac{4}{\aleph}\lg\left(\frac{n}{10}\right).$$

Полученное значение округляется до большего целого числа. Данная рекомендация не является "жесткой", и для одного и того же объема выборки *n* в зависимости от предположений могут быть получены различные оценки для числа интервалов (см. табл. 10). Следует иметь ввиду, что при малом числе интервалов гистограмма будет сильно сглаженной и можно просмотреть ее особенности, а при большом - начинают преобладать ошибки измерений.

Таблица 10
Оценки числа интервалов группирования, получаемые из различных предположений

Объем выборки, *n*	Число интервалов, *m*
50	2, 3, 4, 7, 8, 9, 18
100	3, 4, 5, 8, 9, 10, 14, 15, 21
500	5, 6, 8, 9, 10, 11, 12, 14, 21

После оценки числа интервалов группирования определяются ширина интервала группирования *d,* число значений вариационного ряда n_j, попадающих в каждый интервал (эмпирические частоты), частость в каждом интервале W_j и сумма частостей по всем интервалам *W:*

$$d = \frac{(x_n - x_1)}{m}; \quad W = \sum_{j=1}^{m} W_j; \quad W_j = \frac{n_j}{n}. \tag{2.142}$$

Сумма частостей *W* должна быть равна 1.

На следующем шаге строятся гистограмма и полигон. Для построения гистограммы на оси абсцисс отмечают границы всех интервалов. На каждом интервале как на основании, строят прямоугольник такой высоты, чтобы его площадь была равна частости этого интервала. Высота каждого прямоугольника представляет собой среднюю эмпирическую плотность вероятности того, что истинное значение измеряемой величины находится в соответствующем интервале. Общая площадь между осью абсцисс и ступенчатой кривой должна быть равна единице. Масштаб графика рекомендуется выбирать так, чтобы высота гистограммы относилась к ее основанию как 3 к 5 (это делает график наиболее наглядным).

Полигон распределения значений по интервалам получается соединением середин верхних сторон прямоугольников гистограммы. По внешнему виду гистограммы (полигона) либо наложением на него теоретической кривой функций плотности для различных распределений выдвигают гипотезу о виде закона распределения или более точно о согласовании эмпирического и теоретического распределения. Проверка гипотезы осуществляется по критериям согласия, так же как и для аналитических методов.

2.6.3. Проверка гипотезы о согласовании эмпирического и теоретического распределения по критериям согласия

Обычно используют так называемые критерии согласия Пирсона *(χ^2)* и Крамера-Смирнова *(ω^2).* В ряде случаев используют также критерий Колмогорова, основанный на сравнении интегральных функций распределения, но он менее информативен . Критерий Пирсона можно

применять при объемах выборок $n \geq 50$, а критерий (ω^2) уже при $n \geq 40$. Последний критерий не требует предварительного группирования данных, т. е. свободен от ошибок, связанных со способом их группирования. Наиболее широко применяется критерий Пирсона, так как соответствующее распределение χ^2 затабулировано и общеизвестно; кроме того он позволяет использовать результаты, полученные на предыдущих шагах. Проверка гипотезы по критерию Пирсона выполняется следующим образом.

Для каждого интервала группирования определяют величину $Z_i = \frac{(x_{io} - \bar{x}_n)}{S_n}$, где x_{i0}, - абсцисса, соответствующая середине i-го интервала.

Для вычисленных значений Z_i находят значения плотности вероятности $\varphi(Z_i)$ теоретического распределения, проверяемого в качестве гипотезы, используя статистические таблицы или рассчитывая их самостоятельно по виду распределения с эмпирическими параметрами, определенными ранее методом моментов.

По теоретической кривой плотности распределения $\varphi(Z_i)$ вычисляют теоретические числа значений (выравнивающие частоты) $\hat{n}_i$ в каждом интервале:

$$\hat{n}_i = n\frac{d}{S_n}\varphi(z_i). \quad (2.143)$$

Объединяют соседние интервалы, эмпирическое число значений в которых меньше 5 (для сглаживания ошибок в данных).

Для каждого интервала после объединения вычисляют величину χ^2_i:

$$\chi^2_i = \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}. \quad (2.144)$$ Вычисляют величину χ^2, суммируя χ^2_i по всем интервалам:

$$\chi^2 = \sum_{i=1}^{m_0} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}, \quad (2.145)$$

где m_0 - общее число интервалов после объединения интервалов с малыми частотами.

Определяют число степеней свободы, соответствующее величине χ^2:

$$k=m_0-1-r,$$

где r - число оцениваемых по выборке параметров теоретического распределения.

Например, для нормального распределения по выборке определяют два параметра m и σ, поэтому $r=2$, а $k = m_0 - 3$. Такое выражение для k в случае нормального распределения получается потому, что частоты подчинены трем связям. Действительно, помимо условия, что сумма эмпирических частот (объем выборки n) фиксирована, от теоретического распределения естественно потребовать, чтобы выравнивающие частоты давали среднее значение и СКО, равные соответствующим параметрам, определенным по выборке. Таким образом, имеем три связи и $k = m_0$ - 3.

При подборе другого распределения, например, биномиального: $k = m_o$-2, так как в этом случае имеются две связи: а) сумма эмпирических частот фиксирована и б) выравнивающие частоты должны давать среднее значение, равное соответствующему параметру, определенному по выборке. Аналогично определяется k для других распределений.

По полученному значению k, выбрав уровень значимости α (вероятность ошибки первого рода), определяем по статистическим таблицам распределения Пирсона критические (нижнее и верхнее) значения критерия χ^2_H и χ^2_B для двух значений вероятности:

$$P\left\{\chi^2 \le \chi^2_H\right\} = P\left\{\chi^2 \ge \chi^2_B\right\} = \alpha/2; \quad P\left\{\chi^2_H < \chi^2 < \chi^2_B\right\} = 1 - \alpha/2.$$

Данная рекомендация для определения критических значений критерия соответствует квазисимметричной критической области на кривой плотности критерия χ^2 (см. §2.5.). Можно использовать также одностороннюю критическую область для больших значений критерия, при этом определяется критическое значение χ^2_{kp}, такое, что

$$P\left\{\chi^2 < \chi^2_{kp}\right\} = 1 - \alpha.$$

Гипотеза о согласовании эмпирического и теоретического распределений принимается, если $\chi^2_H < \chi^2 < \chi^2_B$

(для квазисимметричной критической области) или $\chi^2 < \chi^2_{kp}$ (для односторонней критической области).

В. И. Романовский предложил очень простое правило, значительно облегчающее применение критерия согласия Пирсона для оценки расхождения между эмпирическими и выравнивающими частотами. Если $\frac{|\chi^2 - k|}{\sqrt{2k}} \geq 3$, то расхождение можно считать существенным и гипотеза отклоняется, если же $\frac{|\chi^2 - k|}{\sqrt{2k}} < 3$, то расхождение можно считать случайным и гипотеза принимается. Это правило основано на том, что математическое ожидание и СКО величины χ^2 равны: $E(\chi^2) = k;\ \sigma^2_{\chi^2} = 2k$, а также на том, что вероятность значений χ^2, отличающихся от k меньше чем не $3\sigma^2$, т. е. на $3\sqrt{2k}$ в ту или иную сторону, близка к единице. Иногда оказывается, что условия проверки выполняются для нескольких распределений; тогда в качестве искомого принимается то, которое имеет наибольшую статистическую достоверность. Для этого уменьшают последовательно значение α и повторяют проверку оставшихся теоретических распределений до тех пор, пока не останется единственное, согласующееся с эмпирическим, которое и принимается за искомое.

2.6.4. Оценка истинного значения и ошибки измерения

Информация о виде закона распределения позволяет получить точечную и интервальную оценки истинного значения измеряемой величины. За оценку истинного значения принимается оценка центра распределения, положение которого зависит от закона распределения. 1)Для симметричных экспоненциальных распределений с $\aleph \in [0;0,45]$ эффективной оценкой является меридиана, *Me:*

$$Me_n = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) \text{ при четном } n, \quad (2.147)$$

$$Me_n = x_{\frac{(n+1)}{2}} \text{ при нечетном } n, \quad (2.148)$$

2) Для распределений, близких к нормальному с $\aleph \in [0{,}45;0{,}67]$, эффективным иоценками являются среднее *x*~ или усредненное среднее

$$\overline{x}_n(0{,}05),\ \overline{x}_n(0{,}1):$$

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i\ , \quad (2.149)$$

$$\overline{x}_n(\varepsilon) = \frac{1}{n-2l}\sum_{i=l+1}^{n-l} x_i\ , \quad (2.150)$$

где $\varepsilon \cdot n \le l \le \varepsilon_n + 1$ для случая, когда с каждого конца вариационного ряда исключают по *l* значений для получения более устойчивой оценки центра распределения. Обычно используют значения ε=0,05 либо ε=0,1. Данная оценка должна применяться с известной осторожностью, так как необоснованное исключение данных может исказить информацию, содержащуюся в выборке (см. пример в начале этого параграфа).

3)Для распределений, близких к равномерному и арксинусоидальному с $\aleph \in [0{,}67;1]$ целесообразно использовать центр размаха, x_{R1}:

$$x_{R_1} = \frac{x_1 + x_n}{2}. \quad (2.151)$$

4) Для двухмодальных распределений с $\aleph \in [0{,}67;1]$ эффективной оценкой

являетсяцентр срединного размаха, x_{R_2}:

(2.152)

$$x_{R_2} = \frac{1}{2}\left(x_{\frac{n}{4}+1} + x_{\frac{3n}{4}} \right) \text{ при } n, \text{ кратном 4,}$$

$$x_{R_2} = \frac{1}{2}\left(x_{\frac{n+2}{4}} + x_{\frac{3n+2}{4}} \right) \text{ при четном } n, \quad (2.153)$$

$$x_{R_2} = \frac{1}{2}\left(x_{\frac{n-1}{4}+1} + x_{n-\frac{n-1}{4}} \right) \text{ при } (n-1), \text{ кратном 4,} \quad (2.154)$$

$$x_{R_2} = \frac{1}{2}\left(x_{\frac{n+1}{4}+1} + x_{n-\frac{n+1}{4}} \right) \text{ при } (n+1), \text{ кратном 4,} \quad (2.155)$$

Оценка дисперсии получается непосредственным вычислением. Например, для центра размаха, x_{R1} имеем:

$$D(x_{R_1}) = \frac{1}{4}\left[D(x_1) + D(x_n) \right] = \frac{D(x)}{2}, \quad (2.156)$$

т. е. дисперсия равна половине выборочной дисперсии. Значение дисперсии $D(x)$ определяется через параметры распределения (см. §2.2). При выводе (2.156) использовано очевидное равенство: $D(x_1) = D(x_2) = D(x)$, так как x_1 и x_n имеют одинаковую дисперсию, являясь значениями одной и той же величины x.

Оценка доверительного интервала определяется требуемым значением доверительной вероятности. Для модельных распределений (равномерного, треугольного, трапецеидального) доверительный интервал рассчитывается непосредственно по известным параметрам распределения на основе простой связи между ним и доверительной вероятностью. Для нормального и других сложных распределений следует пользоваться статистическими таблицами. Информация об истинном значении измеряемой величины представляется в виде:

$$x = x_u \pm \Delta \quad (2.157)$$

либо

$$x = x_H \ldots x_B \quad (2.158)$$

с указанием доверительной вероятности P и оценки центра распределения $x_ц$, где $x_н$, $x_в$ - нижняя и верхняя границы интервала соответственно:

$x_H = x_ц - \Delta(P)$; $x_B = x_ц + \Delta(P)$; $x_ц$ - оценка центра распределения. В частности, если распределение нормальное, то оценкой центра является выборочное среднее (см. §2.2). Более сложной является совместная обработка количественных и качественных данных.

3. Измерительные устройства

3.1. Основные блоки измерительных устройств

Для проведения измерений используют приборы и передаточные элементы, образующие в совокупности измерительные устройства (ИУ). С позиций системного анализа измерительные устройства в целом так же как и их отдельные блоки функционально одинаковы, т.е. являются системами (подсистемами, модулями): на вход системы подается входной сигнал x (измеряемая величина), а после преобразования на ее выходе появляется выходной сигнал y (результат измерения). Эти величины связаны соотношениями, характеризующими систему, вида:

$$y = Ax, \quad (3.1)$$

где A - оператор, соответствующий алгоритму преобразования (алгоритму измерения).

В сложном измерительном устройстве алгоритм включает цепочку типовых преобразований, так что:

$$A = A_1 \cdot A_2 \cdot A_3, \quad (3.2)$$

где A_1 — соответствует группе аналоговых преобразований (АП),
A_2 - соответствует аналого-цифровому преобразованию (АЦП),
A_3 - соответствует группе цифровых преобразований (ЦП).

При наличии цикла обратной связи к ним добавляется цифро-аналоговое преобразование (ЦАП) и соответствующее соотношение имеет вид:

$$y = Ax - A \cdot By, \quad (3.3)$$

где B - оператор преобразования в цепи обратной связи.

Функциональная схема измерительного устройства представлена на рис. 20, в виде набора типовых модулей:

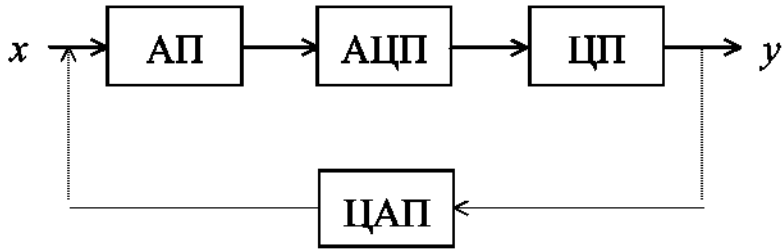


Рис. 20. Функциональная схема измерительного устройства.

Каждый из модулей ИУ в свою очередь является сложной системой и может состоять из ряда элементов. Типовая блок-схема ИУ с линейной структурой дана на рис. 21.

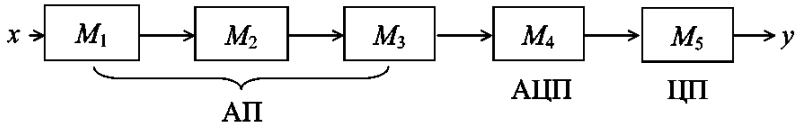


Рис. 21. Блок-схема ИУ с линейной структурой.

Измеряемая величина поступает на вход модуля M_1 (датчик, чувствительный элемент), а на его выходе возникает сигнал, предназначенный для дальнейшей обработки и зависящий от измеряемой величины (например, типичным датчиком является фотодиод). Затем слабые сигналы подаются на модуль M_2, где усиливаются специальным усилителем (например, операционным усилителем) или с помощью преобразователя преобразовываются к более удобному виду (например, в электрический сигнал). Передаточный элемент M_3 (электрическая линия, световод и т.п.) передает сигнал на устройство вывода (модуль M_4), которое либо выдает информацию экспериментатору, либо хранит ее для дальнейшей обработки с помощью процессора (в последнем случае M_4 может включить АЦП). Наконец, модуль M_5 (процессор, ЭВМ) производит обработку данных непосредственно в процессе измерений и представляет результат в числовой (цифровой) форме.
С алгоритмической и аппаратурной реализацией процесса измерения связано появление методической и инструментальной ошибок. Источниками методической ошибки могут быть ограниченная применимость физических законов, используемых для описания изучаемого явления, несоответствие истинного свойства объекта его

модельному представлению, несовершенство методов и методики измерения, т.е. методическая ошибка обусловлена неидеальностью используемых моделей и техники эксперимента. Источником инструментальной ошибки является неидеальность измерительной аппаратуры, что обусловлено несовершенством технологии изготовления средств измерений. Пусть A^0 - алгоритм соответствующий идеальной алгоритмической и аппаратурной реализации процесса измерения, дающий истинные значения измеряемой величины, так что $y_{ист.} = A^0 x$ (этот алгоритм на практике не реализуется); A^u - алгоритм, соответствующий идеальной аппаратурной реализации: $y_{ид} = A^u x$, а A^p - алгоритм, соответствующий неидеальной (реальной) аппаратурной реализации процесса измерения: $y_p = A^p x$. Тогда методическая ошибка равна:

$$e_м = y_{ид} - y_{ист} = \left(A^u - A^0\right)x, \quad (3.4)$$

а инструментальная ошибка:

$$e_{ин} = y_p - y_{ид} = \left(A^p - A^u\right)x. \quad (3.5)$$

Легко видеть, что почленное сложение (3.4) и (3.5) дает полную ошибку:

$$e = e_м + e_{ин}. \quad (3.6)$$

Деление ошибки на методическую и инструментальную конструктивно, так как они имеют разную природу и различные методы устранения.

3.2. Передаточные характеристики

Соотношения вида (3.1) для измерительного устройства называются передаточными характеристиками (ПХ). Многие ПХ имеют общую природу и не зависят от конкретного прибора. Стационарное состояние измерительного устройства (или отдельного элемента) достигается, когда заканчиваются все переходные процессы после подачи на его вход постоянного сигнала *x*. В этом случае ПХ называются статическими. Между выходным *y* и входным *x* сигналами существует функциональная зависимость:

$$y = f(x). \quad (3.7)$$

Обычно требуют, чтобы такая функциональная зависимость для измерительной системы была однозначной, например, она не должна являться гистеризисом, т.е. при возрастании и убывании измеряемой величины зависимость *f(x)* должна оставаться одной и той же. В этом смысле опасны медленные необратимые изменения передаточной характеристики, которые могут быть связаны со старением отдельных элементов. Заметить это изменение можно только с помощью повторного контроля зависимости между входной и выходной величинами. Нужно также учитывать и влияние внешних условий (температуры, давления воздуха, разогрева приборов при длительной работе). Если функциональная зависимость (3.7) представлена графически, то ее называют характеристической кривой. С точки зрения техники измерений удобнее всего работать с линейными зависимостями, т.е. прямыми, которые к тому же проходят через начало отсчета:

$$y = Kx. \quad (3.8)$$

Величину *K* (K=const(*x*)) называют коэффициентом передачи, а ее размерность равна $\left[K\right] = \left[y\right] \cdot \left[x\right]^{-1}$. Если речь идет о сложном приборе (устройстве), то *K* обычно называют чувствительностью *S*. Значение чувствительности показывает, какое изменения Δx входного сигнала необходимо, чтобы выходной сигнал изменился на Δy:

$$S = \frac{\Delta y}{\Delta x} \equiv K. \quad (3.9)$$

Если схема содержит нелинейный элемент, то чувствительность определяется по нелинейной характеристической кривой как производная:

$$S(x, y) = \frac{dy}{dx}. \quad (3.10)$$

В этом случае чувствительность уже не постоянна, а зависит от рабочей точки *(x, y).* Сигнал на выходе измерительного устройства равен:

$$y(x_i) = \int_0^{x_i} S(x, y)\, dx. \quad (3.11)$$

При небольшом изменении измеряемых величин нелинейную характеристическую кривую часто можно приближенно заменить касательной к ней в рабочей точке *(x, y).*

Соотношение (3.8) определяет "идеальную" ПХ. Реальная ПХ отличается от идеальной, так как искажена ошибками. Наиболее характерными являются ошибки трех видов: аддитивная, мультипликативная и ошибка нелинейных искажений.

Аддитивная ошибка постоянна во всем диапазоне измерений системы и не зависит от значения входного сигнала:

$$y = Kx + e_{ад}. \quad (3.12)$$

Она приводит к параллельному смещению прямой по оси ординат.

Мультипликативная ошибка зависит линейно от *x* и проявляется в отличии коэффициента передачи по сравнению с идеальной ПХ:

$$y = K'x = K'x - Kx + Kx = Kx + \Delta Kx, \quad (3.13)$$

где $\Delta K = K' - K$; ΔKx - мультипликативная ошибка.

Она приводит к повороту прямой вокруг начала координат.

Ошибка нелинейных искажений состоит в отклонении зависимости (3.8) от линейной.

3.3 Динамические свойства измерительных устройств

Если входная величина *x(t)* меняется со временем, то выходная величина может содержать искажения, обусловленные инерционными свойствами измерительного устройства. Эти искажения принято называть динамической ошибкой, которая является дополнительной к статической ошибке, характеризующей ошибку измерения в стационарных условиях. Сумма двух ошибок дает полную ошибку. По определению полная ошибка равна:

$$e(t) = y(t) - y_{ист} = y(t) - y_{ист} + y_{ст/t} - y_{ст/t} = \\ = y(t) - y_{ст/t} + \left(y_{ст/t} - y_{ист}\right) = e_{дин} + e_{ст}, \quad (3.14)$$

где $y_{ст/t}$ - выходной сигнал в стационарных условиях, отнесенный к моменту регистрации сигнала *t;* $e_{дин} = y(t) - y_{ст/t}$ - динамическая ошибка; $e_{ст} = y_{ст/t} - y_{ист}$ - статическая ошибка. Передаточные характеристики ИУ в принципе можно рассчитать, зная характеристики всех элементов, однако только непосредственная экспериментальная проверка позволяет учесть все факторы, которые могут исказить входной сигнал. Для этого входная величина изменяется по заданному закону, а выходная регистрируется с достаточно большим разрешением по времени. Зависимости *x(t),* используемые для контроля ПХ прибора, называют контрольными функциями, а результирующие зависимости *y(t)* на выходе -

функциями отклика. Наиболее важными контрольными функциями являются ступенчатая, единичная импульсная *(δ-* функция) и синусоидальная функции.

3.3.1 Передача непериодического сигнала

На рис. 22 показаны характеристики системы 1-го порядка (такая система описывается дифференциальным уравнением 1-го порядка), на вход которой подается сигнал (рис. 22) в форме ступени (скачкообразная функция или функция Хевисайда). Функция отклика на скачкообразный сигнал (рис. 22), описывающая сигнал на выходе, экспоненциально стремится к постоянному значению $y_0=Kx_0$:

$$y(t) = Kx_0\left[1 - \exp\left(-\frac{t - t_0}{T}\right)\right] \quad (3.15)$$

для $t \geq t_0$, где *T*- постоянная времени.
В момент t_0+T выходной сигнал составляет ≈63% нового стационарного значения, а через *5T*- 99% стационарного значения.
Для сравнения различных систем функцию отклика на скачкообразный сигнал на входе делят на величину ступени x_0 входного сигнала:

$$h(t) = \frac{y(t)}{x_0(t)} = K\left[1 - \exp\left(-\frac{t - t_0}{T}\right)\right]. \quad (3.16)$$

Эту нормированную функцию отклика называют переходной функцией (рис.22). Она полностью определяет динамические свойства системы.

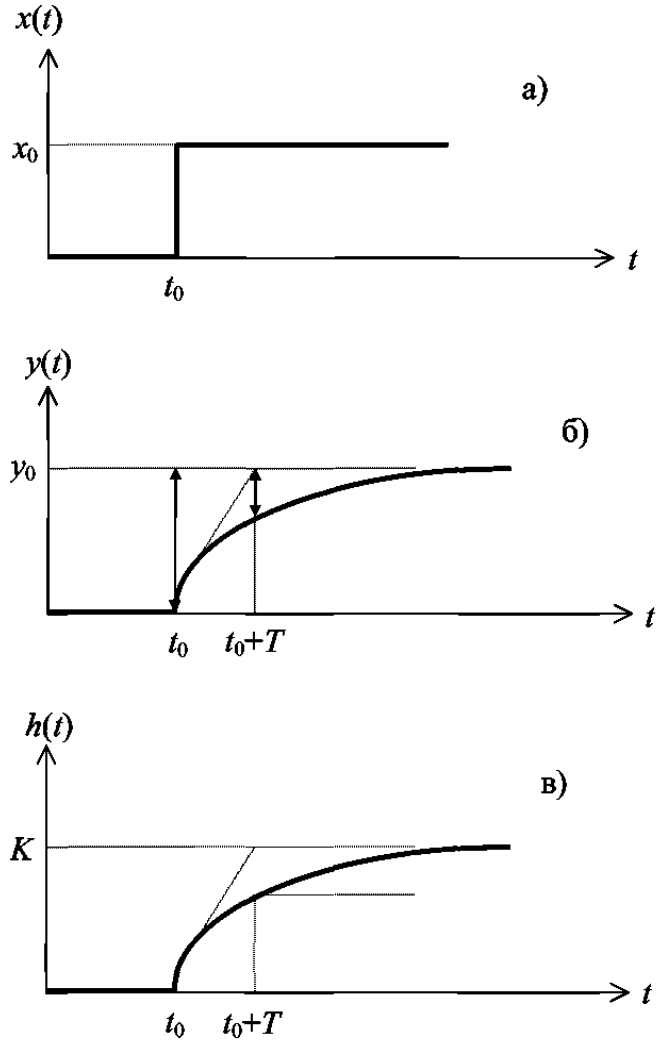


Рис.22. Зависимость входного и выходного сигналов от времени для системы 1-го порядка.

Наряду с этим на практике используют частные характеристики: время установления T_y и время нарастания $T_н$. Временем установления называют промежуток времени, в конце которого выходной сигнал $y(t)$ отличается от стационарного значения Kx_0 на 5%; 1% или 0,1%. Время

нарастания -это время, за которое функция отклика *h(t)* нарастает от 0,1 К до *0,9К.*

К сожалению, большинство систем описывается дифференциальным уравнением 2-го порядка и более высоких порядков (системы 2-го и более высоких порядков). Поэтому переходная функция достигает стационарного значения не экспоненциально, а по более сложному закону. Функция отклика может стремиться к *К* плавно или с затухающими колебаниями возле *К.* Плавное изменение обычно более предпочтительно. Соответствующие примеры показаны на рис.23.

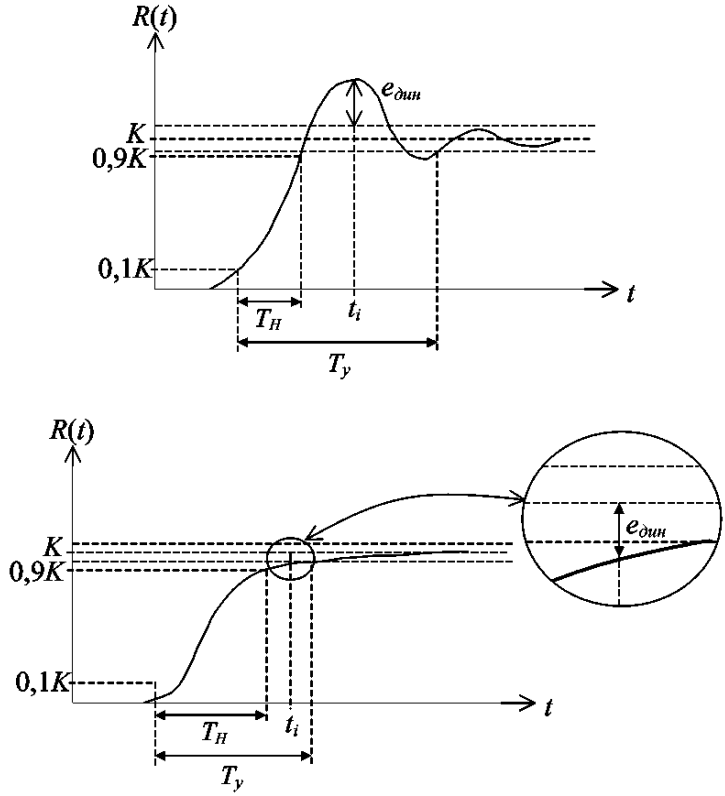


Рис. 23. Переходные функции систем 2-го и более высокого порядка.

В случае колебаний временем установления называют время, после которого функция отклика не выходит за пределы определенной полосы возле *К* (на рис. 23 показана полоса шириной ±5%). В системах с плавным изменением функции отклика время установления *Т*

определяется так же как в системах 1-го порядка. Определим динамическую ошибку. В соответствии с (3.14) для системы 1-го порядка:

$$e_{дин} = y(t) - y_{ст/t} = -Kx_0 \exp\left(-\frac{t-t_0}{T}\right). \quad (3.17)$$

Из (3.17) следует, что она максимальна (по модулю) при $t=t_0$ и стремится к 0 при $t\to\infty$. Динамическая ошибка уменьшается (по модулю) по мере приближения функции отклика к стационарному значению K. Очевидно поэтому, что момент измерения нужно выбирать большим, чем время установления, чтобы динамическая ошибка не превышала 5%; 1% или 0,1%. Этот вывод справедлив для систем любого порядка. На рис. 22 отмечены значения $e_{дин}$ при $t=t_0$ и $t=t_0+T$, а на рис. 23 при $t=t_i$ (они получаются нормированными на величину x_0).

Использование ступенчатой функции в качестве контрольной оправдано, если переходные процессы в ИУ растянуты по времени. В других случаях удобно описывать передаточные характеристики с помощью единичной импульсной функции. Например, в ФЭУ электрон, вылетевший с фотокатода вызывает на аноде импульс тока, ширина и форма которого определяется разбросом времен свободного пробега в ФЭУ. Другой пример - это импульс света, возникающий в сцинтилляторе, когда туда попадает квант излучения или частица. Важно, что во всех этих случаях изучаемые процессы имеют малую длительность.

Единичная импульсная функция $x(t)$ представляет собой короткий импульс прямоугольной формы, продолжительностью Δt с амплитудой x_0. При этом Δt должно быть настолько малым, чтобы за этот промежуток времени не возникало сигнала $y(t)$ на выходе ИУ. В пределе $\Delta t\to 0$ единичная импульсная функция описывается δ-функцией Дирака:

$$x(t) = x_0 \cdot \delta(t) \quad (3.18)$$

причем $$\int_{-\infty}^{\infty} x(t)\,dt = x_0 \quad (3.19)$$

Функцию отклика на единичный импульс $y(t)$ обычно нормируют на x_0, т.е. на площадь под кривой $x(t)$. В этом случае ее называют реакцией на единичный импульс или весовой функцией:

$$g(t) = \frac{y(t)}{x_0}. \quad (3.20)$$

Тогда для любого сигнала $x(\tau) \neq 0$ при $t > 0$ можно представить функцию отклика как интеграл Дюамеля от произведения функции $x(\tau)$ на весовую функцию:

$$y(t) = \int_0^t x(\tau) \cdot g(t-\tau)\, d\tau, \quad (3.21)$$

т. е. , функция отклика равна среднему взвешенному входного сигнала, причем в качестве весов выступают значения весовой функции (отсюда и ее название). Так как δ-функция является производной от функции Хевисайда, то единичная импульсная функция получается при дифференцировании ступенчатой функции. Поэтому отклик системы на единичную импульсную функцию связан с откликом на ступенчатый входной сигнал соотношением:

$$g(t) = \frac{dh(t)}{dt} \quad (3.22)$$

Если задана переходная функция, то (3.21) интегрированием по частям приводится к виду:

$$y(t) = \int_0^t x(\tau) \frac{dh(t-\tau)}{d\tau}\, d\tau = \int_0^t \frac{dx(\tau)}{d\tau} * h(t-\tau)\, d\tau. \quad (3.23)$$

Если требуется восстановить по выходному сигналу форму входного, то это можно сделать, используя интеграл Дюамеля (3.21). Для этого с помощью преобразования Лапласа получают функцию-изображение для входного сигнала. Тогда, используя обратное преобразование Лапласса, имеем:

$$L\{y(t)\} = L\{x(t)\} \cdot L\{g(t)\}, \quad (3.24)$$

где преобразование

Лапласа: $F(S) = L[f(t)] = \int_0^\infty f(t) e^{-st}\, dt.$

Из (3.24) видно, что функция-изображение, полученная с помощью преобразования Лапласа для весовой функции, полностью определяет передаточные свойства системы. Поэтому ее называют передаточной функцией системы. В общем виде она определяется выражением:

$$H(S)=\frac{L\{y(t)\}}{L\{x(t)\}}=L\{g(t)\}, \quad (3.25)$$

где $S=\sigma+i\omega.$ (3.26)

Такой способ описания передаточной функции имеет преимущество, так как позволяет определять ПХ системы по ПХ ее отдельных элементов. В частности, если ИУ состоит из элементов, соединенных последовательно, то общая передаточная функция устройства равна произведению передаточных функций отдельных элементов:

$$H(S)=\prod_{i=1}^{n}H_i(S). \quad (3.27)$$

3.3.2. Передача периодического сигнала

При передаче периодического сигнала в качестве контрольной функции используют синусоидальную функцию. После завершения переходных процессов входной периодический сигнал вида:

$$x(t)=x_0\cdot\exp(i\omega t) \quad (3.28)$$

вызывает на выходе периодический сигнал с такой же угловой частотой ω, но с другой амплитудой y_0 и со сдвигом по фазе φ, которые зависят от ω:

$$y(t)=y_0(\omega)\cdot\exp(i(\omega t+\varphi(\omega))). \quad (3.29)$$

Зависимость между входным и выходным сигналами называют комплексной частотной характеристикой (КЧХ):

$$H(\omega)=\frac{y(t)}{x(t)}=\frac{y_0(\omega)}{x_0}\exp(i\varphi(\omega)). \quad (3.30)$$

В пределе $\omega\to 0$ КЧХ переходит в статический коэффициент передачи K, т.е. $H(\omega)$ имеет ту же размерность, что и K: $[H]=[K]=[y]/[x]$.

Функцию $H(\omega)$ можно представить с помощью годографа на комплексной плоскости, однако на практике обычно используют представление КЧХ с помощью диаграммы Боде. Она представляет собой зависимости фазы и логарифма отношения амплитуд от логарифма частоты. На рис. 24 показаны как пример амплитудная частотная характеристика (АЧХ) и фазовая частотная характеристика (ФЧХ) системы 1-го порядка.

Из рис. 24 видно, что при низких частотах амплитуда не зависит от частоты. Начиная с некоторой частоты ω_g (граничная частота), выходной сигнал становится все слабее, а разность фаз возрастает. При высоких частотах передаточные характеристики системы ухудшаются, поэтому принято определять так называемую граничную частоту ω_g, при которой амплитуда сигнала падает до $1/\sqrt{2}$ (≈71%) исходного значения. При измерениях такие большие искажения не допустимы, поэтому в качестве допустимого отклонения выбирают значения 10%; 5%; 1%, либо наибольшей допустимой частотой считают частоту, которая в 10 раз ниже граничной:

$$\omega_m \le \frac{\omega_g}{10}. \quad (3.31)$$

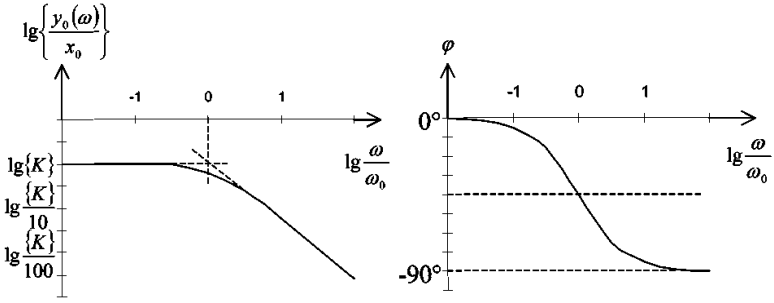


Рис. 24. КЧХ системы 1-го порядка, представленная диаграммой Боде.

На графиках значения частоты обычно нормируют на частоту ω_0, в качестве которой используют граничную частоту ω_g или частоту собственных колебаний в системе. На рис. 24 представлены характеристики системы, называемой фильтром нижних частот (ФНЧ). Простейшим ФНЧ является RC-цепочка (рис. 25), у которой входной и выходной сигнал имеют одинаковую природу. Ее КЧХ можно представить в виде отношения полных сопротивлений, если RC-цепочка подключена как делитель напряжения:

$$H(\omega) = \frac{U_{вых}}{U_{вх}} = 1 \Big/ i\omega c \Big/ \left[R + \frac{1}{i\omega c} \right] = \frac{1}{1 + i\omega RC}. \quad (3.32)$$

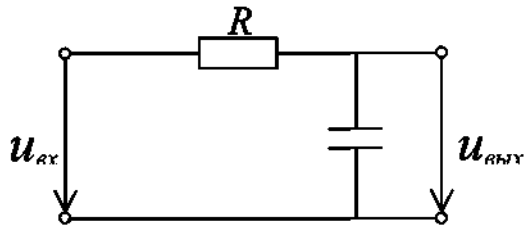


Рис. 25. RC-фильтр нижних частот.

Из (3.32) получаем:

$$|H(\omega)| = \frac{1}{\sqrt{1+\omega^2 R^2 C^2}}; \quad \varphi = -\operatorname{arctg}\omega RC, \qquad (3.33)$$

граничная частота равна:

$$\omega_g = \frac{1}{RC}. \quad (3.34)$$

Подставляя (3.34) в (3.33), найдем после простых преобразований:

$$\lg|H(\omega)| = -\frac{1}{2}\lg\left(1+\frac{\omega^2}{\omega_g^2}\right); \varphi = -\operatorname{arctg}\left(\frac{\omega}{\omega_g}\right). \quad (3.35)$$

Кривые на рис. 24 соответствуют выражениям (3.35) с точностью до постоянного множителя *K* на амплитудной характеристике.
Область частот от 0 до ω_g называют полосой пропускания, а интервал от 0 до $\omega_g/10$ называют полосой пропускания измерительной системы. Решая дифференциальное уравнение, описывающее систему, можно получить постоянную времени для передаточной функции RC-цепочки:

$$\omega_g = \frac{1}{T}. \quad (3.36)$$

Выражение (3.36) справедливо для всех систем 1-го порядка. Учитывая (3.16), можно получить выражение для времени нарастания $T_н$:

$$T_н = T\ln 9 = \frac{\ln 9}{\omega_g} \approx \frac{2{,}2}{\omega_g}. \quad (3.37)$$

Для времени установления (на уровне 1% от стационарного значения) имеем:

$$T_y = 2T \ln 10 = \frac{2 \ln 10}{\omega_g} \approx \frac{4{,}6}{\omega_g}. \quad (3.38)$$

Аналогичные выражения можно получить и для систем более высоких порядков. На практике и в этих случаях часто пользуются соотношениями (3.37) и (3.38).

Системы со свойствами фильтра верхних частот (ФВЧ) можно использовать только для динамических измерений. Такие системы полностью подавляют низкочастотную составляющую вместе с постоянной составляющей сигнала. На рис. 26 показаны типичные АЧХ и ФЧХ фильтра верхних частот, а на рис. 27 - простейший ФВЧ - CR-цепочка (АЧХ на рис.26 описывает свойства CR-цепочки с точностью до постоянного множителя).

Нижняя граничная частота ФВЧ соответствует $1/\sqrt{2}$ максимального значения амплитуды. Аналогично случаю ФНЧ можно записать:

$$H(\omega) = \frac{U_{вых}}{U_{вх}} = \frac{R}{R + \frac{1}{i\omega c}} = \frac{i\omega RC}{1 + i\omega RC}. \quad (3.39)$$

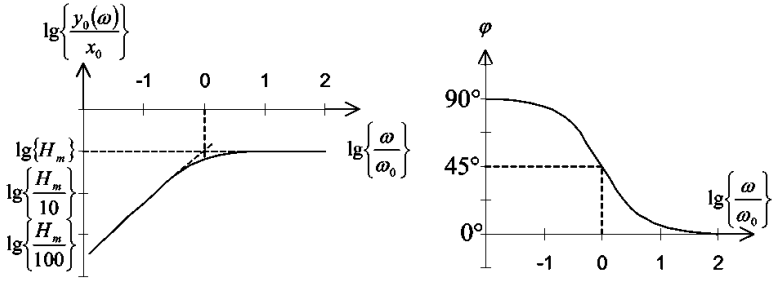


Рис. 26 КЧХ фильтра верхних частот, представленная диаграммой Боде.

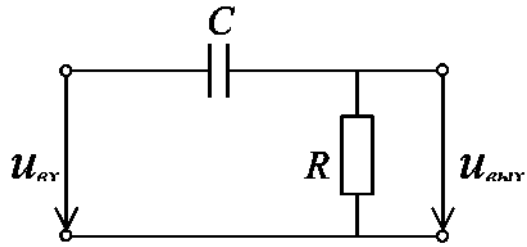


Рис. 27. CR-фильтр (ФВЧ.).

Амплитудная и фазовая характеристики даются выражениями:

$$|H(\omega)| = \frac{\omega RC}{\sqrt{1+\omega^2 R^2 C^2}}; \quad \varphi(\omega) = \operatorname{arctg}\left(\frac{1}{\omega RC}\right). \quad (3.40)$$

Нижняя граничная частота равна:

$$\omega_{gн} = \frac{1}{RC} \quad (3.41)$$

и, следовательно:

$$\lg|H(\omega)| = \lg\frac{\omega}{\omega_{gj}} - \frac{1}{2}\lg\left(1+\frac{\omega^2}{\omega_{gt'}^2}\right), \quad (3.42а)$$

$$\varphi(\omega) = \operatorname{arctg}\frac{\omega_{gн}}{\omega}. \quad (3.42б)$$

Если в системе могут возникать собственные колебания с частотой ω_0, то на АЧХ появляется характерный максимум возле ω_0. На рис. 28 показан пример для ФНЧ. Комбинация ФВЧ и ФНЧ позволяет получить так называемый полосовой фильтр (рис. 29)

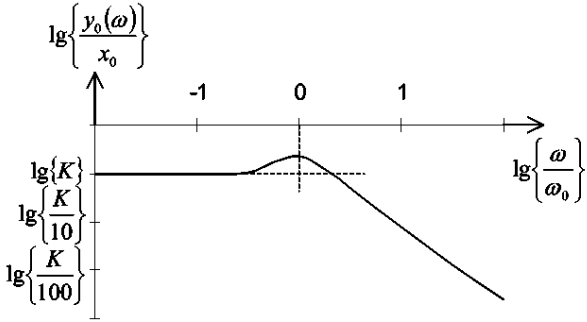


Рис. 28. АЧХ системы, в которой возможны собственные колебания при ω_0.

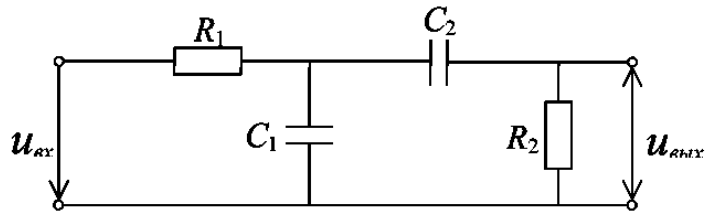


Рис. 29. Полосовой фильтр.

В этом случае, рассуждая аналогично, получим:

$$H(\omega) = \frac{i\omega RC}{(1+i\omega RC)^2}. \quad (3.43)$$

АЧХ и ФЧХ даются выражениями (для простоты сопротивления и емкости в обоих фильтрах приняты одинаковыми, т. е. $R_1=R_2=R$ и $C_1=C_2=C$):

$$|H(\omega)| = \frac{\omega RC}{1+\omega^2R^2C^2}, \quad (3.44)$$

$$\varphi = \operatorname{arctg}\frac{1-\omega^2R^2C^2}{2\omega RC}. \quad (3.45)$$

Вводя граничную частоту

$$\omega_g = \frac{1}{RC}, \quad (3.46)$$

получим:

$$\lg|H(\omega)| = \lg\frac{\omega}{\omega_g} - \lg\left(1+\frac{\omega^2}{\omega_g^2}\right), \quad (3.47а)$$

$$\varphi = \operatorname{arctg}\left(\frac{1}{2}\left(\frac{\omega_g}{\omega} - \frac{\omega}{\omega_g}\right)\right). \quad (3.47б)$$

В общем случае следует ввести две граничные частоты: верхнюю и нижнюю:

$$\omega_{gв} = \frac{1}{R_1 C_1}\ ; \qquad \omega_{gн} = \frac{1}{R_2 C_2} \qquad (3.48)$$

и мы имеем:

$$\lg|H| = \lg\frac{\omega}{\omega_{gн}} - \frac{1}{2}\lg\left(1+\frac{\omega^2}{\omega_{gн}^{\ 2}}\right) - \frac{1}{2}\lg\left(1+\frac{\omega^2}{\omega_{gв}^{\ 2}}\right), \qquad (3.49а)$$

$$\varphi = arctg\left[\frac{1}{\frac{\omega}{\omega_{gв}} + \frac{\omega}{\omega_{gн}}} - \frac{1}{\frac{\omega_{gв}}{\omega} + \frac{\omega_{gн}}{\omega}}\right]. \qquad (3.49б)$$

КЧХ полосового фильтра приведена на рис. 30.

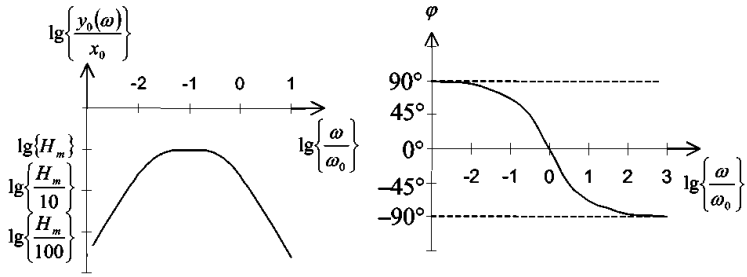


Рис. 30. КЧХ полосового фильтра в виде диаграммы Боде.

КЧХ полосового фильтра и ее графическое представление могут быть получены без утомительных вычислений из (3.27). КЧХ в принципе описывает и специальный случай передаточной функции, когда $S=i\omega$. Поэтому по аналогии с (3.27) функция $H(\omega)$ всего измерительного устройства при последовательном соединении элементов тоже равна произведению частотных характеристик каждого элемента:

$$H(\omega) = \prod_{k=1}^{n} H_K(\omega) \qquad (3.50)$$

или с учетом (3.28):

$$H(\omega)=\prod_{k=1}^{n}\left|H_k(\omega)\right|\exp\left[i\sum_{k=1}^{n}\varphi_k(\omega)\right]. \quad (3.51)$$

Из (3.51) следует, что диаграмму Боде всего ИУ можно получить с помощью простого графического сложения АЧХ и ФЧХ отдельных элементов, так как логарифмы отношений амплитуд складываются по модулю (для нашего примера с полосовым фильтром *n=2).* Обратное тоже верно: если нужно улучшить ИУ и обнаружить ее "слабые" места, то следует рассмотреть по отдельности КЧХ ее элементов.

3.4. Принцип обратной связи

Передаточные характеристики прибора или отдельного элемента ИУ можно изменить, используя обратную связь. На рис. 31 показана принципиальная схема устройства с обратной связью. Выходной сигнал *y(t)* с помощью преобразователя в цепи обратной связи подается на вход и складывается с входным сигналом *x(t).*

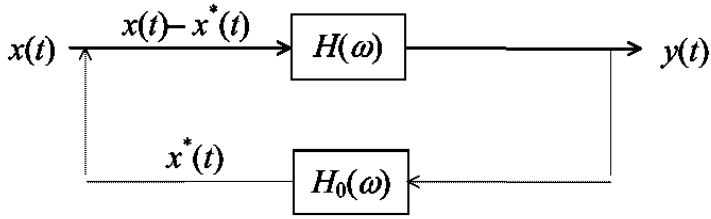


Рис. 31. Система с обратной связью.

Если $H(\omega)$ и $H_0(\omega)$ - КЧХ устройств в прямой и обратной цепочке соответственно, то при синусоидальных собственных колебаниях с частотой ω можно получить синусоидальную функцию отклика. Полагая в (3.3) $A=H(\omega)$; $B=H_0(\omega)$, найдем:

$$y(t)=H(\omega)\left[x(t)-H_0(\omega)y(t)\right] \quad (3.52)$$

что дает:

$$y(t)=\frac{H(\omega)}{1+H(\omega)\cdot H_0(\omega)}x(t). \quad (3.53)$$

КЧХ всей системы имеет вид:

$$H_{эфф}(\omega) = \frac{H(\omega)}{1 + H(\omega)H_0(\omega)}. \quad (3.54)$$

Соотношение (3.54) можно записать иначе:

$$H_{эфф}(\omega) = \frac{1}{H_0} \cdot \frac{1}{1 + \frac{1}{H \cdot H_0}} = \frac{1}{H_0} \cdot F_0, \quad (3.55)$$

т. е. величину $H_{эфф}$ можно представить как произведение КЧХ устройства в цепи обратной связи на поправочный множитель $F_0 = \left(1 + \frac{1}{H \cdot H_0}\right)^{-1}$. Если выполняется условие:

$$|H| >> |H_0|^{-1}, \quad (3.56)$$

то множитель $F_0=1$.

Преимущества такой схемы очевидны. Во многих случаях свойства ИУ в прямой цепи H сильно зависят от влияния окружающей среды. Тогда, подключив параллельно элементу H пассивное звено обратной связи, можно добиться, чтобы возмущения в прямой цепи H практически не влияли на передаточные характеристики всей системы с обратной связью. Выбирая нужную характеристику H_o, можно получить требуемую частотную зависимость свойств системы с обратной связью. При этом следует иметь в виду, что при введении обратной связи уменьшается чувствительность системы в соответствии с (3.55). Схемы с обратной связью чаще всего используются при конструировании усилителей, а также функциональных преобразователей.

Часть IV. Интервалы и операции над ними

Введение

Излагаемый в этом и последующих разделах материал относится к актуальному направлению вычислительной математики, получившему название «интервальный анализ» или даже «интервальная

математика». Интерес к этому направлению обусловлен, в первую очередь, широким применением ЭВМ для всевозможных расчетов. Если эти расчеты проводятся по традиционной схеме, то часто очень трудно, а иногда просто невозможно дать математически строгий ответ на естественный вопрос о соотношении числа, напечатанного машиной, и истинного значения вычисляемой величины. Интервальный анализ дает возможность получить такой ответ ценой увеличения времени счета.

Основная идея интервального анализа чрезвычайно проста. Вещественное число представляется в памяти ЭВМ не одним, а двумя машинными числами — оценкой снизу и оценкой сверху, образующими **интервальное число**. Арифметические операции над этими числами выполняются так, что если $[a_1, a_2] = [b_1, b_2] \circ [c_1, c_2]$, $b \in [b_1, b_2]$, $c \in [c_1, c_2]$, то $b \circ c_1 \in [a_1, a_2]$, где $\circ \in \{+, —, \times, /\}$. Таким образом, **интервальный анализ дает возможность автоматически учитывать погрешности в задании исходных данных и погрешности, вызываемые машинными округлениями.** Это создает основу для аккуратного учета погрешностей, вызываемых используемым приближенным методом вычислений.

Первой монографией по интервальному анализу была вышедшая в 1966 г. книга Р. Е. Мура, которая во многом способствовала становлению этого направления. В настоящее время общее число опубликованных работ в этой области составляет несколько тысяч. Такое обилие публикаций объясняется, в частности, тем, что практическая реализация описанной выше основной идеи интервального анализа сталкивается с большими трудностями. Например, оказалось, что алгоритм Гаусса может стать неприменимым к системе линейных уравнений с интервальными коэффициентами из-за возникающих делений на интервалы, содержащие нуль. В других случаях, когда традиционный численный метод переносился на интервальные числа, в результате вычислений получались интервалы, гарантированно содержащие истинное значение, но столь широкие, что найденные двусторонние оценки были практически бесполезными. Выяснилось, что для успешного применения интервального анализа нужно пересмотреть весь арсенал численных методов.

Такой пересмотр и делается в настоящей работе, которая базируется на книге Г. Алефельда, Ю. Херцберга «Введение в интервальые вычисления». При этом Г. Алефельд, Ю. Херцберг не ограничиваются только описанием различных алгоритмов, но и проводят их сравнение как по достигаемой точности, так и по вычислительной сложности. Большая часть излагаемого материала посвящена задачам линейной алгебры, что неудивительно ввиду той базисной роли, которую играет

линейная алгебра в численных методах. Вместе с тем вне рамок настоящей работы остались многие другие важные для приложений разделы математики, в которых интервальный анализ успешно применяется, например, обыкновенные дифференциальные уравнения.

Интервальный анализ вышел за рамки чисто теоретического исследования и достаточно широко применяется на практике с помощью соответствующего программного обеспечения. За прошедшее время, с одной стороны, появились реализации интервальной арифметики на более мощных языках. С другой стороны, был разработан стандарт ANSI/IEEE на машинное представление чисел и правила выполнения операций над ними. Этот стандарт реализован как программно, так и аппаратно в микропроцессорах. Это в значительной степени облегчает программную реализацию интервальных вычислений, которые становятся ненамного более медленными, чем традиционные.

Следует обратить внимание на терминологию, которая еще полностью не устоялась ни в мире, ни на русском языке, что надо иметь в виду при чтении работ других авторов. На ***интервальные числа*** можно смотреть двояко: ***как на способ задания вещественных чисел, которые мы знаем лишь с некоторой погрешностью, и как на самостоятельные объекты*** Это различие почти нигде не ощущается, пожалуй, его единственное проявление — это определение равенства интервалов. При первом подходе равенство $[a_1, a_2] = [b_1, b_2]$ выполняется тогда и только тогда, когда $a_1 = a_2 = b_1 = b_2$, при втором, принятом в этой книге, оно справедливо тогда и только тогда, когда $a_1 = b_1$ и $a_2 = b_2$.

1. Вещественная интервальная арифметика

1.1. Основные понятия и определения

В этом и последующих разделах поле вещественных чисел будет обозначаться через $\mathbb{R}$, а строчные буквы *a, b, c, ..., x, y, z* будут использоваться для обозначения его элементов. Подмножество A множества $\mathbb{R}$, такое что

$$A = [a_1, a_2] = \{x \mid a_1 \leqslant x \leqslant a_2,\ a_1, a_2 \in \mathbb{R}\},$$

будет называться **замкнутым вещественным интервалом**, или просто **интервалом**, если это не сможет вызвать недоразумения.

Впоследствии в некоторых случаях, чтобы избежать путаницы, мы будем обозначать границы интервала A через

$$i(A) = a_1 \quad \text{и} \quad s(A) = a_2.$$

Множество всех замкнутых вещественных интервалов обозначим через $I(\mathbb{R})$, а прописные буквы $A, B, C, \ldots, X, Y, Z$ зарезервируем для обозначения его элементов. Всякое вещественное число x из $\mathbb{R}$ может считаться особым элементом из $I(\mathbb{R})$, имеющим вид $[x, x]$; чаще всего мы будем называть его точечным интервалом.

Определение 1. Два интервала $A = [a_1, a_2]$ и $B = [b_1, b_2]$ называются равными (записывается: $A = B$), если они равны в теоретико-множественном смысле.

Из этого определения непосредственно следует, что

$$A = B \Leftrightarrow a_1 = b_1, \quad a_2 = b_2.$$

Отношение равенства между двумя элементами из $I(\mathbb{R})$ рефлексивно, симметрично и транзитивно.

Теперь мы можем обобщить арифметику вещественных чисел, введя операции над элементами из $I(\mathbb{R})$.

Определение 2. Пусть $* \in \{+, -, \cdot, :\}$ — бинарная операция на множестве вещественных чисел. Если $A, B \in I(\mathbb{R})$, то

$$A * B = \{z = a * b \mid a \in A, b \in B\} \tag{1}$$

определяет бинарную операцию на $I(\mathbb{R})$.

В определении предполагается, что в случае деления $0 \notin B$, и в дальнейшем это явно указываться не будет. Заметим также, что символы операций на множествах I(R) и R совпадают. Это не должно вызывать затруднений, поскольку из контексга всегда ясно, к чему применяется операция: к вещественным числам или интервалам.

Результат операции над интервалами $A = [a_1, a_2]$ и $B = [b_1, b_2]$ может быть получен явно с помощью формул

$$\begin{cases} A + B = [a_1 + b_1, a_2 + b_2], \\ A - A = [a_1 - b_2, a_2 - b_1] = A + [-1, -1] \cdot B, \\ A \cdot B = [\min\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}, \\ \qquad\qquad \max\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}], \\ A : B = [a_1a_2] \cdot [1/b_2, 1/b_1]. \end{cases} \tag{2}$$

Их обоснованием служит тот факт, что $z = f(x, y) = x * y$, где $* \in \{+, -, \cdot, :\}$ — непрерывная функция на компактном множестве. Следовательно, $f(x, y)$ принимает как наименьшее и наибольшее значения, так и все прочие значения между ними. Таким

образом, $A * B$— также замкнутый вещественный интервал. Теперь понятно, что (2) — это формулы для вычисления наименьшего и наибольшего значений *f(x, y)*. Из сказанного следует замкнутость множества $I(\mathbb{R})$ относительно введенных таким образом операций, а также изоморфизм между вещественными числами *x, y, ...* и интервалами $[x, x]$, $[y, y]$, ... Поэтому всюду далее операция $[x, x] * A$, в которой участвуют точечный интервал $[x, x]$ и произвольный интервал *A,* будет записываться в упрощенной форме $x * A$. Кроме того, мы часто будем опускать знак умножения.

Набор операций вида (1) может быть дополнен другими традиционными, в основном унарными операциями над интервалами.

Определение 3. Если *r(x)*— непрерывная унарная операция на R, то

$$r(X) = [\min_{x \in X} r(x),\ \max_{x \in X} r(x)]$$

определяет соответствующую ей операцию на $I(\mathbb{R})$.

Примерами таких унарных операций могут служить

$$X^k (k \in \mathbb{R}),\quad e^X,\quad \ln X,\quad \sin X,\quad \cos X \text{ и т. д.}$$

Соберем теперь вместе наиболее важные свойства операций на $I(\mathbb{R})$.

Теорема 4. *Пусть* $A, B, C \in I(\mathbb{R})$. *Тогда*

$$A + B = B + A,\quad A \cdot B = B \cdot A \ (\textit{коммутативность});$$

$$(A + B) + C = A + (B + C), \tag{3}$$

$$(A \cdot B) \cdot C = A \cdot (B \cdot C) \ (\textit{ассоциативность}); \tag{4}$$

$X = [0, 0]$ *и* $Y = [1, 1]$ *—единственные нейтральные элементы соответственно сложения и умножения, т. е.*

$$A = X + A = A + X \quad \textit{для всех} \quad A \in I(\mathbb{R}) \Leftrightarrow X = [0, 0],$$
$$A = Y \cdot A = A \cdot Y \quad \textit{для всех} \quad A \in I(\mathbb{R}) \Leftrightarrow Y = [1, 1]; \tag{5}$$

I(R) *не имеет делителей нуля;* (6)

произвольный элемент $A = [a_1, a_2] \in I(\mathbb{R})$, *у которого* $a_1 \neq a_2$, *не имеет обратного ни по сложению, ни по умножению. Тем не менее,*

$$0 \in A - A \quad \textit{и} \quad 1 \in A : A; \tag{7}$$

$$A(B + C) \subseteq AB + AC \ (\textit{субдистрибутивность}),$$

$$a(B + C) = aB + aC, \quad \textit{где} \quad a \in \mathbb{R}, \tag{8}$$

$$A(B + C) = AB + AC, \quad \textit{где} \quad bc \geqslant 0 \quad \textit{для всех} \quad b \in B \quad \textit{и} \quad c \in C.$$

Доказательство. (3): Пусть $* \in \{+, \cdot\}$. Тогда

$$A * B = \{z = a * b \mid a \in A,\ b \in B\}$$
$$= \{z = b * a \mid b \in B,\ a \in A\} = B * A$$

(4): Пусть $* \in \{+, \cdot\}$. Тогда

$$(A * B) * C = \{z = y * c \mid y \in A * B,\ c \in C\}$$
$$= \{z = (a * b) * c \mid a \in A,\ b \in B,\ c \in C\}$$
$$= \{z = a * (b * c) \mid a \in A,\ b \in B,\ c \in C\}$$
$$= \{z = a * x \mid a \in A,\ x \in B * C\} = A * (B * C).$$

(5): Необходимость доказывается тривиально. Если N и $\bar{N}$ -два нейтральных элемента сложения, то

$$N + \bar{N} = \bar{N} \quad \text{и} \quad \bar{N} + N = N.$$

Из свойства коммутативности (3) следует, что $N = \bar{N}$.

Единственность $Y = [1, 1]$, нейтрального элемента умножения, может быть показана подобным же образом.

(6): Пусть $A \cdot B = 0$, т. е.

$$A \cdot B = \{z = a \cdot b \mid a \in A,\ b \in B\} = [0, 0].$$

Из этого следует, что по крайней мере один из интервалов A и B, принадлежащих $I(\mathrm{R})$, должен быть равен [0, 0].

(7): Утверждения, которые нужно доказать, эквивалентны следующим:

$$A - B = [0, 0] \Rightarrow A = [a, a] = B,$$
$$A \cdot B = [1, 1] \Rightarrow A = [a, a], \quad B = [1/a, 1/a].$$

Пусть

$$A - B = \{z = a - b \mid a \in A,\ b \in B\} = [0, 0].$$

Отсюда следует, что $z = a - b = 0$ для всех $a \in A,\ b \in B$. Фиксируя b из B, получаем, что $a = b$ для всех a из A, т. е. $A = [b, b]$. Соответственно можно заключить, что $B = [a, a]$ и, следовательно, $a = b$. Второе утверждение доказывается подобным же образом.

Так как для a из A

$$0 = a - a \in \{z = x - y \mid x \in A,\ y \in A\},$$

то очевидно, что $0 \in A - A$. Аналогично, $1 \in A : A$, где $0 \notin A$.

(8): $$A(B + C) = \{z = a(b + c) \mid a \in A,\ b \in B,\ c \in C\}$$
$$\subseteq \{y = ab + \tilde{a}c \mid a, \tilde{a} \in A,\ b \in B,\ c \in C\}$$
$$= AB + AC.$$

Для того чтобы показать невыполнение равенства в общем случае, приведем один пример:

$$A=[0,\ 1],\quad B=[1,\ 1],\quad C=[-1,\ -1],$$
$$A(B+C)=[0,\ 0]\subset[-1,\ 1]=AB+AC.$$

Далее имеем

$$\begin{aligned}a(B+C)&=\{z=a(b+c)\mid b\in B,\ c\in C\}\\&=\{z=ab+ac\mid b\in B,\ c\in C\}\\&=\{x=ab\mid b\in B\}+\{y=ac\mid c\in C\}\\&=aB+aC.\end{aligned}$$

Доказывая последнее равенство, будем считать b_1 и c_1 неотрицательными, что не приведет к потере общности. Если $a_1 \geq 0$, то

$$A(B+C)=[a_1(b_1+c_1),\ a_2(b_2+c_2)]$$

и

$$AB+AC=[a_1b_1,\ a_2b_2]+[a_1c_1,\ a_2c_2]=[a_1(b_1+c_1),\ a_2(b_2+c_2)],$$

т. е. для этого случая утверждение доказано,

Случай $a_2 \leq 0$ может быть сведен к $a_1 \geq 0$ путем замены A на $-A$. Если $a_1a_2 < 0$, то получаем

$$A(B+C)=[a_1(b_2+c_2),\ a_2(b_2+c_2)],$$

а также

$$AB+AC=[a_1b_2,\ a_2b_2]+[a_1c_2,\ a_2c_2]=[a_1(b_2+c_2),\ a_2(b_2+c_2)],$$

что доказывает утверждение (8) и для этого случая.

Теперь мы хотим остановиться на вопросе разрешимости уравнения

$$AX = B,$$

Где $A\neq[0,\ 0]$ и $X\in I(\mathbb{R})$. Для того чтобы ответить на этот вопрос, введем вспомогательную функцию χ:

$$\chi(A)=\begin{cases}a_1/a_2, & \text{если } |a_1|\leqslant|a_2|,\\ a_2/a_1 & \text{в остальных случаях}\end{cases}$$

(эту функцию предложил Ратшек).

Справедливо следующее утверждение: уравнение $AX=B$ разрешимо относительно X из $I(\mathrm{R})$ тогда и только тогда, когда

$$\chi(A)\geqslant\chi(B).$$

Решение не единственно лишь в случае

$$\chi(A)=\chi(B)\leqslant 0.$$

Проиллюстрируем приведенное утверждение примером. Пусть дано уравнение

$$[1,\ 2]\,X=[-1,\ 3].$$

Равенство выполняется лишь при X = [—1/2, 3/2], поскольку

$$\chi[1, 2] = \frac{1}{2} > \chi[-1, 3] = -\frac{1}{3}.$$

С другой стороны, если рассмотреть множество решений всех уравнений вида

$$ax = b,$$

у которых

$$a \in [1, 2], \quad b \in [-1, 3],$$

то получим

$$\{x = b/a \mid a \in [1, 2], b \in [-1, 3]\} = [-1, 3]/[1, 2] = [-1, 3] \supset X.$$

Это множество решений существенно отличается от интервала X, удовлетворяющего равенству $AX=B$. По этой причине мы не называем X решением уравнения $AX=B$, а предпочитаем говорить об «алгебраическом» решении.

Вообще, можно доказать следующее утверждение. Пусть уравнению $AX = B$, где $0 \notin A$, удовлетворяет некоторое X из $I(\mathrm{R})$. Тогда

$$X \subseteq B : A.$$

Действительно,

$$x \in X \Rightarrow \text{существуют } a \in A, \ b \in B, \text{ для которых}$$
$$ax = b \Rightarrow x = b/a \in B : A.$$

Заметим также, что равенство $AX = B$ может быть выполнено, даже если $B : A$ не определено. Примером служит уравнение

$$\left[-\frac{1}{3}, \ 1\right] X = [-1, 2],$$

для которого единственным решением является X = [—1, 2], причем $\chi\left[-\frac{1}{3}, \ 1\right] > \chi[-1, 2]$.

Основное свойство интервальных вычислений — монотонность включения. Следующая теорема разъясняет это свойство.

Теорема 5. *Пусть*

$$A^{(k)}, \ B^{(k)} \in I(\mathbb{R}), \ k = 1, 2,$$

и предполагается, что

$$A^{(k)} \subseteq B^{(k)}, \quad k = 1, 2.$$

Тогда для операции $*$ *из* $\{+, -, \cdot, /\}$ *имеем*

$$A^{(1)} * A^{(2)} \subseteq B^{(1)} * B^{(2)}. \tag{9}$$

Доказательство. Так как $A^{(k)} \subseteq B^{(k)}, \ k = 1, 2$, то

$$A^{(1)} * A^{(2)} = \{z = x * y \mid x \in A^{(1)}, \ y \in A^{(2)}\}$$
$$\subseteq \{w = u * v \mid u \in B^{(1)}, \ v \in B^{(2)}\} = B^{(1)} * B^{(2)}.$$

Приведем частный случай теоремы 5

Следствие 6. *Пусть* $A, B \in I(\mathbb{R})$ *и* $a \in A$, $b \in B$. *Тогда*

$$a * b \in A * B,$$

где $* \in \{+, -, \cdot, :\}$.

Унарные операции *r(X)* из определения 3 обладают сходными свойствами:

$$X \subseteq Y \Rightarrow r(X) \subseteq r(Y),$$
$$x \in X \Rightarrow r(x) \in r(X). \qquad (10)$$

Непосредственное обобщение этих соотношений на случай интервальных выражений дано в теореме 3 п. 7.3.

Замечания. Это элементарное введение в вещественную интервальную арифметику соответствует описанию, данному Муром. Большинство унарных операций из определения 3 легко задаются в виде функций от левой и правой границ интервального аргумента. К примеру, это можно без труда проделать для монотонных функций x^k и $\sqrt{x}$.

Четыре основные операции (+, —, • и :) на точечных множествах общего вида ввел Янг. Им были получены и некоторые элементарные соотношения, например (3), (4) и (8).

Кулиш исследовал, какие свойства операций, заданных на множестве *M,* переносятся на множество всех его подмножеств *P(M).* Интервальные операции вида (1) получаются при этом как частный случай в числе прочих результатов.

Представление интервалов, которое применил Сунага, соответствует круговым комплексным интервалам, описываемым в последующих микромодулях. При этом способе записи пара чисел *(a, r)* обозначает интервал $[a - r, a + r]$. Данное представление было использовано им для явного описания и последующего применения интервальных операций вида (1).

Ортольф отождествил интервалы $[a_1, a_2]$ с точками (a_1, a_2) из R×R. На этой основе ему удалось построить определение операций над всеми элементами R × R. При $a_1 \le a_2$ его определение сводится к операциям вида (2). Подобным же образом над точками из $R \times R$ вводятся отрицание (аддитивная инверсия) и, если $0 \notin [a_1, a_2]$, обращение (мультипликативная инверсия).

Кахан предложил обобщение интервальных операций вида (2). Наряду с обычными вещественными числами аргументами обобщенных операций могут быть $+\infty$ и $-\infty$. В некоторых случаях результатом операции оказывается «интервал» Ω, включающий все вещественные числа. Кроме Ω допускаются также интервалы

$[a_1, a_2[$, $] a_1. a_2[$, $] a_1, a_2]$, причем разрешены $a_1 = \pm\infty$ и $a_2 = \pm\infty$. Более того, a_2 может быть меньше, чем a_1 (Например, запись [3,2] заменяет выражение $[-\infty, 2] \bigcup [3, +\infty]$ Подобные объекты могут возникать в результате разрешенного в этой арифметике деления на интервал, содержащий нуль). Для интерпретации такого представления интервалов используется ориентированная окружность, на которой располагаются вещественные числа. Введенные подобным образом интервалы могут содержать $\infty \equiv +\infty \equiv -\infty$, быть открытыми и полуоткрытыми. Их арифметика определяется в соответствии с (1).

В общем виде интервальные вычисления в частично упорядоченных пространствах описал Апостолатос**.** И на этот раз $I(\mathrm{R})$ возникает как частный случай.

Клауа разработал трехзначную теорию множеств. Он вводит так называемые частичные множества и частичные кардинальные числа Получающаяся в результате арифметика кардинальных чисел для конечного случая в точности соответствует интервальной. Таким образом, аналогом интервальной арифметики на $I(\mathrm{R})$ служат операции над трехзначными числами. Наряду с отношением = из определения 1 применяется более слабое отношение $=_{\#}$, которое для $A=[a_1, a_2]$ и $B = [b_1, b_2]$ задается следующим образом:

$$A =_{\#} B \Leftrightarrow A \cap B \neq \varnothing \Leftrightarrow \max\{a_1, b_1\} \leqslant \min\{a_2, b_2\}.$$

Отношение $=_{\#}$ рефлексивно и симметрично; кроме того,

$$A = B \Rightarrow A =_{\#} B.$$

Это означает, что если $A \neq_{\#} B$, то для всех a и b, таких что $a \in A$ и $b \in B$, мы всегда имеем $a \neq b$. Соответственно

$$A \neq_{\#} B \Rightarrow A \neq B.$$

Имея в виду отношение $=_{\#}$, можно рассмотреть $I(\mathrm{R})$ как разновидность обобщенного поля и, например, доказать следующие свойства:

$X - X =_{\#} 0$ для $X \in I(\mathbb{R})$;

$AX =_{\#} B \Leftrightarrow X =_{\#} B : A$ для $A, B, X \in I(\mathbb{R})$, причем $A \neq_{\#} 0$;

$X(Y + Z) =_{\#} XY + XZ$ для $X, Y, Z \in I(\mathbb{R})$.

Каухер предложил расширенное множество $\overline{I(\mathbb{R})}$, получив его как результат дополнения $I(\mathrm{R})$ так называемыми нерегулярными интервалами, т. е. интервалами отрицательной ширины. В этом случае точечные интервалы $[a, a]$ больше не являются минимальными элементами в смысле порядка, задаваемого отношением $\subseteq$. Все

структуры $I(\mathbb{R})$ переносятся на $I(\mathbb{R}) \cup \overline{I(\mathbb{R})}$, и с помощью несобственных элементов p и $-p$ достигается замкнутость. Подобным образом можно определить деление на интервал $[a_1, a_2]$, у которого $a_1 \leqslant 0 \leqslant a_2$ и $a_1 \neq a_2$.

1.2. Свойства интервальной арифметики

Введем теперь понятие расстояния на множестве вещественных интервалов.

Определение 1. Расстояние $q(A, B)$ между двумя интервалами A и B, такими что $A = [a_1, a_2]$, $B = [b_1, b_2] \in I(\mathbb{R})$, определяется равенством

$$q(A, B) = \max\{|a_1 - b_1|, |a_2 - b_2|\}.$$

Легко показать, что отображение q задает на $I(\mathbb{R})$ метрику. Действительно, q обладает следующими свойствами:

$$q(A, B) \geqslant 0 \quad \text{и} \quad q(A, B) = 0 \Leftrightarrow A = B,$$

$$q(A, B) \leqslant q(A, C) + q(B, C) \text{ (неравенство треугольника).}$$

Выполнение неравенства треугольника проверяется следующим образом:

$$\begin{aligned} q(A, C) + q(B, C) &= \max\{|a_1 - c_1|, |a_2 - c_2|\} \\ &\quad + \max\{|b_1 - c_1|, |b_2 - c_2|\} \\ &\geqslant \max\{|a_1 - c_1| + |b_1 - c_1|, |a_2 - c_2| + |b_2 - c_2|\} \\ &\geqslant \max\{|a_1 - b_1|, |a_2 - b_2|\} = q(A, B). \end{aligned}$$

Если применить введенное таким способом расстояние к точечным интервалам, то оно сведется к обычному расстоянию между вещественными числами. Иначе говоря,

$$q([a, a], [b, b]) = |a - b|.$$

Предложенная здесь метрика является для $I(\mathbb{R})$ хаусдорфовой. Хаусдорфова метрика обобщает понятие расстояния между двумя точками в метрическом пространстве (у нас таким пространством является R с $q(x, y) = |x - y|$) на случай пространства всех компактных непустых подмножеств данного пространства. Если U и V — непустые компактные множества вещественных чисел, то хаусдорфово расстояние определяется как

$$q(U, V) = \max\{\sup_{v \in V} \inf_{u \in U} q(u, v), \sup_{u \in U} \inf_{v \in V} q(u, v)\}.$$

Существуют другие полезные определения хаусдорфовой метрики. Легко убедиться в том, что для вещественных интервалов A и B хаусдорфова метрика задается выражением из определения 1.

Вводя на множестве $I(\mathrm{R})$ метрику, мы делаем его топологическим пространством. При этом понятия сходимости и непрерывности могут использоваться обычным образом, как и в случае метрического пространства. В этой связи мы получаем, что последовательность интервалов $\{A^{(k)}\}_{k=0}^{\infty}$ сходится к интервалу $A=[a_1, a_2]$ тогда и только тогда, когда последовательность границ отдельных членов последовательности сходится к его соответствующим границам. Следовательно, мы можем записать

$$\lim_{k\to\infty} A^{(k)} = A \Leftrightarrow \left(\lim_{k\to\infty} a_1^{(k)} = a_1 \text{ и } \lim_{k\to\infty} a_2^{(k)} = a_2\right). \quad (1)$$

Доказательство этого утверждения мы опускаем, так как его легко получить непосредственно из определения расстояния между двумя интервалами.

Введенная нами метрика используется в следующей теореме

Теорема 2. *Метрическое пространство* $(I(\mathrm{R}), q)$ *с метрикой из определения* 1 *является замкнутым метрическим пространством.*

(Это означает, что любая интервальная последовательность Коши сходится к интервалу.)

В теореме 3 рассматривается характер сходимости широко используемого класса интервальных последовательностей.

Теорема 3. *Каждая последовательность интервалов* $\{A^{(k)}\}_{k=0}^{\infty}$, *для которой справедливо соотношение*

$$A^{(0)} \supseteq A^{(1)} \supseteq A^{(2)} \supseteq \ldots,$$

сходится к интервалу $A = \bigcap_{k=0}^{\infty} A^{(k)}$.

Доказательство. Пусть имеется последовательность границ, такая что

$$a_1^{(0)} \leqslant a_1^{(1)} \leqslant a_1^{(2)} \leqslant a_1^{(3)} \leqslant \ldots \leqslant a_2^{(3)} \leqslant a_2^{(2)} \leqslant a_2^{(1)} \leqslant a_2^{(0)}.$$

Тогда последовательность нижних границ интервалов из $\{A^{(k)}\}_{k=0}^{\infty}$, является монотонной неубывающей последовательностью вещественных чисел, ограниченной сверху величиной $a_2^{(0)}$. Эта последовательность сходится к вещественному числу a_1. Аналогично, монотонная невозрастающая последовательность вещественных чисел $\{a_2^{(k)}\}_{k=0}^{\infty}$ сходится к вещественному числу a_2, причем $a_1 \leq a_2$. Равенство

$$A = \bigcap_{k=0}^{\infty} A^{(k)}$$

проверяется столь же простым способом.

Как видно из доказательства, каждая последовательность $\{A^{(k)}\}_{k=0}^{\infty}$, для которой

$$A^{(0)} \supseteq A^{(1)} \supseteq A^{(2)} \supseteq A^{(3)} \supseteq \ldots \supseteq B,$$

сходится к такому интервалу A, что $A \supseteq B$.

Для арифметических, а также других определенных выше операций справедлива

Теорема 4. *Введенные ранее операции сложения, вычитания, умножения и деления интервалов непрерывны.*

Доказательство. Мы приводим доказательство только для операции сложения. Пусть $\{A^{(k)}\}_{k=0}^{\infty}$ и $\{B^{(k)}\}_{k=0}^{\infty}$ — две последовательности интервалов, причем $\lim\limits_{k\to\infty} A^{(k)} = A$ и $\lim\limits_{k\to\infty} B^{(k)} = B$. Из (1) вытекает, что последовательность интервальных сумм $\{A^{(k)} + B^{(k)}\}_{k=0}^{\infty}$ имеет предел

$$\begin{aligned}\lim_{k\to\infty}(A^{(k)} + B^{(k)}) &= \lim_{k\to\infty}\left[a_1^{(k)} + b_1^{(k)},\ a_2^{(k)} + b_2^{(k)}\right] \\ &= \left[\lim_{k\to\infty}\left(a_1^{(k)} + b_1^{(k)}\right),\ \lim_{k\to\infty}\left(a_2^{(k)} + b_2^{(k)}\right)\right] \\ &= [a_1 + b_1,\ a_2 + b_2] = A + B.\end{aligned}$$

Доказательство непрерывности остальных операций может быть проведено аналогичным способом.

Обобщением теоремы 4 служит (см. определение 3 п.7.1)

Следствие 5 . *Пусть r — непрерывная функция и*

$$r(X) = \left[\min_{x\in X} r(x),\ \max_{x\in X} r(x)\right].$$

Тогда r(X) — непрерывное интервальное выражение.

Доказательство этого следствия основывается непосредственно на факте непрерывности функции r и поэтому здесь будет опущено. Следствие 5 гарантирует непрерывность выражений, подобных X^k, $\sin X$ и e^x.

Определение 6. Пусть $A = [a_1, a_2] \in I(\mathbb{R})$. Абсолютной величиной этого интервала будем называть величину

$$|A| = q(A, [0, 0]) = \max\{|a_1|, |a_2|\}.$$

Абсолютную величину интервала можно записать и в виде

$$|A| = \max_{a\in A}|a|. \qquad (2)$$

Очевидно, что если $A, B \in I(\mathbb{R})$, то

$$A \subseteq B \Rightarrow |A| \leqslant |B|. \qquad (3)$$

Докажем теперь некоторые свойства, связанные с метрикой на $I(\mathbb{R})$.

Теорема 7. *Пусть*

$$A=[a_1,\ a_2],\ B=[b_1,\ b_2],\ C=[c_1,\ c_2],\ D=[d_1,\ d_2]\in I(\mathbb{R}).$$

Тогда

$$q(A+B,\ A+C)=q(B,\ C), \tag{4}$$

$$q(A+B,\ C+D)\leqslant q(A,\ C)+q(B,\ D), \tag{5}$$

$$q(aB,\ aC)=|a|q(B,\ C),\quad a\in\mathbb{R}, \tag{6}$$

$$q(AB,\ AC)\leqslant|A|q(B,\ C). \tag{7}$$

Доказательство. (4): Из определения метрики q следует, что

$$q(A+B,\ A+C)=\max\{|a_1+b_1-(a_1+c_1)|,\ |a_2+b_2-(a_2+c_2)|\}$$
$$=\max\{|b_1-c_1|,\ |b_2-c_2|\}=q(B,\ C).$$

(5): Из неравенства треугольника, предыдущего свойства (4) и симметричности q вытекает, что

$$q(A+B,\ C+D)\leqslant q(A+B,\ B+C)+q(C+D,\ B+C)$$
$$=q(A,\ C)+q(B,\ D).$$

(6): $q(aB,\ aC)=\max\{|ab_1-ac_1|,\ |ab_2-ac_2|\}=|a|q(B,\ C).$

(7): Пусть $A=[a_1,\ a_2]$. Для краткости будем использовать обозначения $i(A)=a_1$ и $s(A)=a_2$. Тогда утверждение (7) можно записать в виде

$$\max\{|i(AB)-i(AC)|,\ |s(AB)-s(AC)|\}\leqslant|A|q(B,\ C).$$

Докажем, что

$$|i(AB)-i(AC)|\leqslant|A|q(B,\ C).$$

Неравенство

$$|s(AB)-s(AC)|\leqslant|A|q(B,\ C)$$

доказывается аналогично.

Перепишем предыдущее соотношение (6):

$$\max\{|i(aB)-i(aC)|,\ |s(aB)-s(aC)|\}=|a|q(B,\ C).$$

Теперь без потери общности можно предположить, что

$$i(AB)\geqslant i(AC).$$

(Случай $i(AB)<i(AC)$ рассматривается точно так же.)

Поскольку

$$AC=\{ac\,|\,a\in A,\ c\in C\},$$

существует такое a из A, что

$$i(AC)=i(aC).$$

Из свойства монотонности включения следует, что

$$aB \subseteq AB \text{ и } i(aB) \geqslant i(AB),$$

откуда видно, что

$$i(aB) - i(aC) \geqslant i(AB) - i(AC) \geqslant 0.$$

Итак,

$$\begin{aligned} |i(AB) - i(AC)| &= i(AB) - i(AC) \leqslant i(aB) - i(aC) \\ &= |i(aB) - i(aC)| \leqslant |a| q(B, C) \\ &\leqslant |A| q(B, C). \end{aligned}$$

Отождествляя $|A|$ с $q(A, 0)$, получаем следующие легко проверяемые свойства абсолютного значения:

$$\begin{aligned} &|A| \geqslant 0 \text{ и } |A| = 0 \Leftrightarrow A = [0, 0], \\ &|A + B| \leqslant |A| + |B|, \\ &|xA| = |x| |A| \text{ для } x \in \mathbb{R}, \\ &|AB| = |A| |B|. \end{aligned} \tag{8}$$

Вот доказательство последнего равенства:

$$\begin{aligned} |AB| = \max_{c \in AB} |c| &= \max_{a \in A,\ b \in B} |ab| = \max_{a \in A,\ b \in B} (|a| |b|) \\ &= \max_{a \in A} |a| \max_{b \in B} |b| = |A| |B|. \end{aligned}$$

Остальные соотношения доказываются подобным же образом.

Определение 8. Шириной интервала $A = [a_1, a_2]$ будем называть

$$d(A) = a_2 - a_1 \geqslant 0.$$

Множество точечных интервалов можно теперь описать как

$$\{A \in I(\mathbb{R}) \mid d(A) = 0\}.$$

Из определения 8 сразу же получаем свойства

$$A \subseteq B \Rightarrow d(A) \leqslant d(B), \tag{9}$$

$$d(A \pm B) = d(A) + d(B). \tag{10}$$

Утверждение (9) доказывается тривиально — достаточно определение 8 переписать в виде

$$d(A) = \max_{a,\ b \in A} |a - b|. \tag{11}$$

Проверим свойство (10) для операции сложения:

$$\begin{aligned} d(A + B) &= d([a_1 + b_1,\ a_2 + b_2]) \\ &= a_2 + b_2 - (a_1 + b_1) \\ &= a_2 - a_1 + b_2 - b_1 = d(A) + d(B). \end{aligned}$$

Вычитание проверяется точно так же Кроме того, имеет место теорема.

Теорема 9. *Пусть A и B — вещественные интервалы из $I(\mathbb{R})$ Тогда*

$$d(AB) \leqslant d(A)|B| + |A|d(B), \tag{12}$$

$$d(AB) \geqslant \max\{|A|d(B), |B|d(A)\}, \tag{13}$$

$$d(aB) = |a|d(B),\ a \in \mathbb{R}, \tag{14}$$

$$d(A^n) \leqslant n|A|^{n-1} d(A),\ n = 1, 2, \ldots, \tag{15}$$

$$(A^n := A \cdot A \cdot \ldots \cdot A,\ n \text{ раз}),$$

$$d((X-x)^n) \leqslant 2(d(X))^n, \text{ где } x \in X,\ n = 1, 2, \ldots, \tag{16}$$

$$((X-x)^n := (X-x)(X-x) \ldots (X-x),\ n \text{ раз}).$$

Если $C \in I(\mathbb{R})$ *и* $0 \in C$, то

$$|C| \leqslant d(C) \leqslant 2|C|. \tag{17}$$

Доказательство (12): Используя тождество (11), получаем

$$\begin{aligned} d(AB) &= \max_{a\ a' \in A,\ b\ b' \in B} |ab - a'b'| \\ &= \max_{a, a' \in A,\ b, b' \in B} |ab - ab' + ab' - a'b'| \\ &\leqslant \max_{a, a' \in A,\ b, b' \in B} \{|a(b-b')| + |(a-a')b'|\} \\ &\leqslant \max_{a \in A\ b, b' \in B} |a||b-b'| + \max_{a, a' \in A,\ b' \in B} |a-a'||b'| \\ &= \left(\max_{a \in A}|a|\right)\left(\max_{b,\ b' \in B}|b-b'|\right) + \left(\max_{a\ a' \in A}|a-a'|\right)\left(\max_{b' \in B}|b'|\right) \\ &= |A|d(B) + d(A)|B|. \end{aligned}$$

(13): Сначала докажем, что

$$\begin{aligned} d(AB) &= \max_{a, a' \in A\ \ b\ b' \in B} |ab - a'b'| \geqslant \max_{a \in A,\ b, b' \in B} |ab - ab'| \\ &= \max_{a \in A,\ b\ b' \in B} |a||b-b'| = |A|d(B). \end{aligned}$$

Подобным образом можно показать, что

$$d(AB) \geqslant |B|d(A),$$

откуда сразу же вытекает (13).

(14): $$\begin{aligned} d(aB) &= \max_{b,\ b' \in B} |ab - ab'| = \max_{b,\ b' \in B} \{|a||b-b'|\} \\ &= |a| \max_{b, b' \in B} |b-b'| = |a|d(B). \end{aligned}$$

(15): При n=1 имеет место равенство. Если неравенство выполняется для некоторого $n \geq 1$, то, используя (12) и последнее соотношение из (8), имеем

$$d(A^{n+1}) = d(A^n A) \leqslant d(A^n)|A| + |A|^n d(A)$$
$$\leqslant n|A|^{n-1} d(A)|A| + |A|^n d(A)$$
$$= (n+1)|A|^n d(A).$$

(16): Поскольку $x \in X$, из (9) и свойства монотонности включения получаем

$$d((X-x)^n) \leqslant d((X-X)^n) = d([-d(X),\ d(X)]^n)$$
$$= d([-(d(X))^n,\ (d(X))^n]) = 2(d(X))^n.$$

(17): Так как $0 \in C = [c_1, c_2]$, то $c_1 \leqslant 0 \leqslant c_2$, откуда имеем

$$d(C) = c_2 - c_1 = |c_2| + |c_1| \geqslant \max\{|c_1|, |c_2|\} = |C|.$$

Итак,

$$d(C) = |c_1| + |c_2| \leqslant 2\max\{|c_1|, |c_2|\} = 2|C|.$$

Теперь докажем следующую теорему.

Теорема 10. *Пусть* $A,\ B \in I(\mathbb{R})$, *причем* A *— симметричный интервал, т. е.* $A = -A$. *Тогда имеют место следующие свойства:*

$$AB = |B|A, \tag{18}$$

$$d(AB) = |B|d(A). \tag{19}$$

Если $b_1 \geqslant 0$ или $b_2 \leqslant 0$, то второе свойство выполняется и в случае, когда $0 \notin A$.

Доказательство. Предположим, что $A = -A$, или, что то же самое, $a_2 = a = -a_1$. Тогда

$$AB = [\min\{ab_1,\ ab_2,\ -ab_1,\ -ab_2\},\ \max\{ab_1,\ ab_2,\ -ab_1,\ -ab_2\}]$$
$$= [a\min\{b_1,\ -b_1,\ b_2,\ -b_2\},\ a\max\{b_1,\ -b_1,\ b_2,\ -b_2\}]$$
$$= [a(-|B|),\ a|B|] = [-a,\ a]|B| = |B|A.$$

Опираясь на равенство (14), получаем (19) Остальные случаи могут быть доказаны аналогичным образом.

Теорема 11. *Для интервалов* A *и* B *из* $I(\mathrm{R})$ *справедливы следующие свойства:*

$$d(A) = |A - A|, \tag{20}$$

$$A \subseteq B \Rightarrow \frac{1}{2}(d(B) - d(A)) \leqslant q(A,\ B) \leqslant d(B) - d(A). \tag{21}$$

Доказательство.

(20): $$d(A) = a_2 - a_1 = |A - A|.$$

(21): Пусть $A \subseteq B$. Тогда $b_1 \leqslant a_1 \leqslant a_2 \leqslant b_2$ и, следовательно,

$$q(A,\ B) = \max\{|a_1 - b_1|,\ |a_2 - b_2|\} = \max\{a_1 - b_1,\ b_2 - a_2\}$$
$$\leqslant b_2 - a_2 + a_1 - b_1 = b_2 - b_1 - (a_2 - a_1) = d(B) - d(A),$$

откуда

$$q(A,\ B) = \max\{a_1 - b_1,\ a_2 - b_2\} \geqslant \frac{1}{2}(a_1 - b_1 + b_2 - a_2)$$

$$= \frac{1}{2}(d(B) - d(A)).$$

Введем теперь на $I(\mathbb{R})$ еще одну бинарную операцию. Пусть $A,\ B \in I(\mathbb{R})$. Тогда отношение

$$A \cap B = \{c \mid c \in A,\ c \in B\} \tag{22}$$

представляет собой теоретико-множественное пересечение двух интервалов. Результат этой операции принадлежит $I(\mathrm{R})$ тогда и только тогда, когда пересечение не пусто. В этом случае

$$A \cap B = [\max\{a_1,\ b_1\},\ \min\{a_2,\ b_2\}]. \tag{23}$$

В приводимом ниже следствии собраны важные свойства операции пересечения.

Следствие 12. *Пусть* $A,\ B,\ C,\ D \in I(\mathbb{R})$. *Тогда*

$$A \subset C,\quad B \subseteq D \Rightarrow A \cap B \subseteq C \cap D \ (\textit{монотонность включения}). \tag{24}$$

Пока операция пересечения не выводит из $I(\mathrm{R})$, *она непрерывна.*

Доказательство. Монотонность включения (24) вытекает из определения (22). Доказательство непрерывности может быть получено с помощью (23).

Замечания. Мур использовал хаусдорфову метрику на $I(\mathrm{R})$, соответствующую определению 1. Некоторые из правил для вычисления абсолютного значения $|A|$ и ширины $d(A)$ имеются в работах Мура и Кулиша. Важное в приложениях неравенство (7) впервые доказал Майер.

Иногда абсолютное значение вводится следующим способом, основанным на определении 3 п. 7.1:

$$\mathrm{abs}(A) = \left[\min_{a \in A} |a|,\ \max_{a \in A} |a|\right].$$

Поскольку такое определение редко применяется в приложениях, мы также не будем его использовать.

Согласно С. М. Румпу (частное сообщение), в выражении (16) можно обойтись без сомножителя 2, если уточнить соответствующую оценку. Для x, принадлежащего X, $X - x = [a,\ b]$, где $a \leqslant 0,\ b \geqslant 0$. Предположим, что $b \geqslant -a = |a|$ (если это не так, будем иметь дело с $x - X$). Тогда

$$(X - x)(X - x) = [ab,\ b^2],$$

и с помощью полной индукции получаем

$$(X - x)^n = [ab^{n-1},\ b^n].$$

Следовательно,

$$d\left((X-x)^n\right) = b^n - ab^{n-1} = b^{n-1}(b-a).$$

Теперь $b-a = d(X-x) = d(X)$, и, поскольку $a \leqslant 0,\ b \geqslant 0$, имеем $b \leqslant d(X-x) = d(X)$. Итак,

$$d((X-x)^n) \leqslant d(X)^n.$$

1.3. Интервальное оценивание

В этом разделе мы обсуждаем непрерывные вещественные функции. Пусть f относится к их числу. Аналитическое выражение для $f=f(x)$ представляет собой запись вычислительной процедуры, выдающей значение функции f для произвольного аргумента x Примем при этом, что все выражения, с которыми мы будем иметь дело, составлены из операций и операндов, число которых конечно, одновременно мы предполагаем, что если эти выражения вычисляются в интервальной арифметике, то составляющие их операции трактуются в соответствии с определениями 2 п. 1.1 и 3 п.1.1. Выражение, содержащее константы $a^{(0)}$, ..., $a^{(m)}$, будет для наглядности записываться в виде $f(x;\ a^{(0)}, ..., a^{(m)})$. Чтобы упростить изложение, в дальнейшем мы всегда будем предполагать, что каждая из констант $a^{(k)}$, $0 \le k \le m$, встречается в аналитическом выражении функции только один раз Этого всегда можно добиться, вводя новые константы, равные константам, встречающимся неоднократно.

Пример. Двумя аналитическими выражениями функции g являются

$$g^{(1)}(x;\ a) = \frac{ax}{1-x}, \quad x \neq 1, \quad x \neq 0,$$

и

$$g^{(2)}(x;\ a) = \frac{a}{1/x - 1}, \quad x \neq 1, \quad x \neq 0.$$

Запись

$$W\left(f,\ X;\ A^{(0)},\ \ldots,\ A^{(m)}\right) = \left\{f\left(x;\ a^{(0)},\ \ldots,\ a^{(m)}\right) \mid x \in X,\ a^{(k)} \in A^{(k)},\ 0 \leqslant k \leqslant m\right\} = \left[\min_{\substack{x \in X \\ a^{(k)} \in A^{(k)}, \\ 0 \leqslant k \leqslant m}} f(x;\ a^{(0)},\ \ldots,\ a^{(m)}),\ \max_{\substack{x \in X \\ a^{(k)} \in A^{(k)}, \\ 0 \leqslant k \leqslant m}} f(x;\ a^{(0)},\ \ldots,\ a^{(m)})\right]$$

будет в дальнейшем обозначать диапазон изменения функции f, причем предполагается, что x из X и $a^{(k)}$ из $A^{(k)}$, $0 \leqslant k \leqslant m$, не зависят друг от друга. Согласно этому определению, интервал

$W(f, X; A^{(0)}, \ldots, A^{(m)})$ будет одним и тем же при любом аналитическом выражении для *f*.

Пример. Возьмем *g* из предыдущего примера. Для

$$A = [0, 1] \text{ и } X = [2, 3]$$

получаем

$$W(g, [2, 3]; [0, 1]) = \left\{ \frac{ax}{1-x} \,\middle|\, 2 \leqslant x \leqslant 3,\ 0 \leqslant a \leqslant 1 \right\} = [-2, 0].$$

Введем теперь понятие интервального оценивания вещественной функции *f*. Пусть для *f* имеется аналитическое выражение. Заменив в этом выражении все вещественные операнды и операции над ними на интервальные операнды и операции, получим выражение $f(X; A^{(0)}, \ldots, A^{(m)})$. Если все операнды попадают в области, на которых заданы операции из определений 2 п. 1.1 и 3 п.1.1, то $f(X; A^{(0)}, \ldots, A^{(m)})$ называется *интервальной оценивающей функцией,* или, для краткости, *оценкой f,* а получение ее значения— *вычислением,* или *оцениванием, f в интервальной арифметике.*

Для функций, рассматриваемых нами, замена описанного типа возможна всегда. Константы $a^{(0)}, \ldots, a^{(m)}$, как и переменная *x,* превращаются в интервалы. Очевидно, что результат оценивания функции *f* зависит от выбора для нее аналитического выражения. Впоследствии мы будем использовать этот факт. А сейчас приведем простой пример.

Пример. Пусть *g* — функция из предыдущих двух примеров. Для $A = [0, 1]$ и $X = [2, 3]$ получим две различные оценки:

$$g^{(1)}([2, 3]; [0, 1]) = \frac{[0, 1][2, 3]}{1 - [2, 3]} = [-3, 0],$$

$$g^{(2)}([2, 3]; [0, 1]) = \frac{[0, 1]}{1/[2, 3] - 1} = [-2, 0] \neq g^{(1)}([2, 3]; [0, 1]).$$

Введенные выше обозначения можно распространить на функции от нескольких переменных. В этом случае множеством значений $f(x^{(1)}, \ldots, x^{(n)}; a^{(0)}, \ldots, a^{(m)})$ при независимых $x^{(k)}$ из $X^{(k)}$, $1 \leqslant k \leqslant n$, и $a^{(j)}$ из $A^{(j)}$, $0 \leqslant j \leqslant m$, становится $W(f, X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)})$. Подобным же образом обобщается понятиеинтервальной оценки — теперь она обозначается через

$$f(X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)})$$

(Множество $W(f, X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)})$ Мур и другие авторы называют объединенным интервальным расширением, а функцию $f(X^{(1)}, \ldots, X^{(n)}; A^{(0)}, \ldots, A^{(m)})$ — естественным интервальным расширением).

Приведем теперь пример выражения, из которого не удается получить его всюду определенный интервальный аналог путем простой замены операций и операндов. Вещественная функция

$$f(x) = 1\Big/\left(x^2 + \frac{1}{2}\right)$$

определена для всех x из R. Представим f в виде

$$\tilde{f}(x) = 1\Big/\left(x \cdot x + \frac{1}{2}\right).$$

Заменим теперь независимую переменную x на интервал X=[—1, 1], содержащийся в области определения f. Замена всех операций на соответствующие им интервальные приводит к интервальному выражению

$$\tilde{f}([-1,\ 1]) = \frac{1}{[-1,\ 1][-1,\ 1] + \frac{1}{2}} = \frac{1}{[-1,\ 1] + \frac{1}{2}} = \frac{1}{\left[-\frac{1}{2},\ \frac{3}{2}\right]},$$

которое не определено.

Познакомимся с рядом свойств интервального оценивания. Два свойства, используемые в последующих утверждениях, легко выводятся из теоремы 5 п.1.1 и следствия 6 п.1.1.

Теорема 1. *Пусть f — непрерывная вещественная функция, $f(x^{(1)}, \ldots, x^{(n)};\ a^{(0)}, \ldots, a^{(m)})$ — аналитическое выражение для f. Предположим также, что для интервалов $Y^{(1)}, \ldots, Y^{(n)}, B^{(0)}, \ldots, B^{(m)}$ имеется оценка $f(Y^{(1)}, \ldots, Y^{(n)};\ B^{(0)}, \ldots, B^{(m)})$. Тогда*

а) *для всех*

$$X^{(k)} \subseteq Y^{(k)}, \quad A^{(j)} \subseteq B^{(j)}, \quad 1 \leqslant k \leqslant n, \quad 0 \leqslant j \leqslant m$$

справедливо свойство включения

$$W(f,\ X^{(1)}, \ldots, X^{(n)};\ A^{(0)}, \ldots, A^{(m)}) \subseteq f(X^{(1)}, \ldots, X^{(n)};\ A^{(0)}, \ldots, A^{(m)}); \tag{1}$$

б) *для всех*

$$X^{(k)} \subseteq Z^{(k)} \subseteq Y^{(k)}, \quad A^{(j)} \subseteq C^{(j)} \subseteq B^{(j)}, \quad 1 \leqslant k \leqslant n, \quad 0 \leqslant j \leqslant m$$

имеет место монотонность включения

$$f(X^{(1)}, \ldots, X^{(n)};\ A^{(0)}, \ldots, A^{(m)}) \subseteq f(Z^{(1)}, \ldots, Z^{(n)};\ C^{(0)}, \ldots, C^{(m)}). \tag{1'}$$

Пример. *Функция f задана выражением*

$$f(x;\ a) = a - x/(1 + x), \quad x \neq -1.$$

Для

$$X=\left[-\tfrac{1}{2},\ 1\right],\quad Z=\left[-\tfrac{1}{2},\ 2\right],\quad A=C=[2,\ 3]$$

получаем

$$W\left(f,\left[-\tfrac{1}{2},\ 1\right];\ [2,\ 3]\right)=\left[\tfrac{3}{2},\ 4\right]\subset f\left(\left[-\tfrac{1}{2},\ 1\right];\ [2,\ 3]\right)=[0,\ 4],$$

$$f\left(\left[-\tfrac{1}{2},\ 1\right];\ [2,\ 3]\right)=[0,\ 4]\subset f\left(\left[-\tfrac{1}{2},\ 2\right];\ [2,\ 3]\right)=[-2,\ 4].$$

Свойство включения (1) позволяет соотнести множество значений функции с ее интервальной оценкой. Позднее мы дадим формулы для их качественного сравнения.

Можно привести примеры, когда в (1) достигается равенство. Очевидно, что к их числу относится случай однократного вхождения каждой из величин $x^{(1)},\ \ldots,\ x^{(n)};\ a^{(0)},\ \ldots,\ a^{(m)}$ в выражение $f(x^{(1)},\ \ldots,\ x^{(n)};\ a^{(0)},\ \ldots,\ a^{(m)})$.

Теорема 2. *Пусть p — многочлен от вещественной переменной x, определяемый выражением*

$$p(x;\ a^{(0)},\ \ldots,\ a^{(m)})=(\ldots((a^{(m)}x+a^{(m-1)})^{n_{m-1}}+a^{(m-2)})^{n_{m-2}}+\ldots+a^{(1)})^{n_1}+a^{(0)},$$

где $n_\nu\geqslant 2,\ 1\leqslant\nu\leqslant m-1$. Если встречающиеся в этом выражении степени вычисляются по формуле

$$X^k=\left[\min_{x\in X}x^k,\ \max_{x\in X}x^k\right]$$

(см. определение 3 п. 1.1), *то*

$$W(p,\ X;\ a^{(0)},\ \ldots,\ a^{(m)})=p(X;\ a^{(0)},\ \ldots,\ a^{(m)}).$$

Доказательство. Для $m=2$ истинность теоремы очевидна:

$$p(x;\ a^{(0)},\ a^{(1)},\ a^{(2)})=(a^{(2)}x+a^{(1)})^{n_1}+a^{(0)}.$$

Остальную часть доказательства получаем полной индукцией.

Не всякий многочлен можно привести к форме, требуемой теоремой 2. Однако многочлен второй степени

$$p(x;\ b^{(0)},\ b^{(1)})=x^2+b_{(1)}x+b^{(0)}$$

может быть преобразован к виду

$$p(x;\ a^{(0)},\ a^{(1)})=(x+a^{(1)})^2+a^{(0)},$$

где

$$a^{(1)}=b^{(1)}/2,\quad a^{(0)}=b^{(0)}-(b^{(1)})^2/4.$$

Наряду с носящей общий характер теоремой 1 и разобранными выше частными случаями, представляет также интерес качественное утверждение о приближении множества значений функции f с

помощью ее интервальной оценки. В случае функции одной вещественной переменной справедлива

Теорема 3. *Пусть f — вещественная функция от вещественной переменной x. $f(x; a^{(0)}, \ldots, a^{(m)})$ — ее аналитическое выражение. Обозначим через $\tilde{f}(x^{(1)}, \ldots, x^{(n)}; a^{(0)}, \ldots, a^{(m)})$ выражение, полученное заменой в $f(x, a^{(0)}, \ldots, a^{(m)})$ каждого вхождения x на новую переменную $x^{(k)}$, $1 \leq k \leq n$. Пусть определена оценивающая функция $f(Y; A^{(0)}, \ldots, A^{(m)})$, где $Y, A^{(0)}, \ldots, A^{(m)} \in I(\mathbb{R})$.*

Кроме того, предположим, что для каждой переменной $x^{(k)}$, $1 \leq k \leq n$, из интервала Y и произвольно выбранных $x^{(j)}$ из Y,

$$1 \leqslant j \leqslant n, \quad j \neq k, \quad u \quad a^{(j)} \quad из \quad A^{(j)}, \quad 0 \leqslant j \leqslant m,$$

выражение $\tilde{f}(x^{(1)}, \ldots, x^{(n)}; a^{(0)}, \ldots, a^{(m)})$ удовлетворяет условию Липшица. В остальном обозначения имеют тот же смысл, что и в теореме 1. *При этих предположениях для $X \subseteq Y$ имеем*

$$q(W(f, X; A^{(0)}, \ldots, A^{(m)}), f(X; A^{(0)}, \ldots, A^{(m)})) \leqslant \gamma d(X), \quad (2)$$
$$\gamma \geqslant 0.$$

Доказательство. Во-первых, заметим, что

$$\tilde{f}(x, \ldots, x; a^{(0)}, \ldots, a^{(m)}) = f(x; a^{(0)}, \ldots, a^{(m)}), \quad x \in Y.$$

Теперь мы можем получить интервальную оценку для f в виде

$$f(X; A^{(0)}, \ldots, A^{(m)}) = W(\tilde{f}, X, \ldots, X; A^{(0)}, \ldots, A^{(m)}), \quad X \subseteq Y.$$

Остается показать, что

$$q(W(f, X; A^{(0)}, \ldots, A^{(m)}), W(\tilde{f}, X, \ldots, X; A^{(0)}, \ldots, A^{(m)})) \leqslant \gamma d(X), \quad X \subseteq Y.$$

Если теперь для $X \subseteq Y$ записать

$$W(f, X; A^{(0)}, \ldots, A^{(m)}) = [f(u; a^{(0)}, \ldots, a^{(m)}), f(v; b^{(0)}, \ldots, b^{(m)})],$$
$$u, v \in X, \quad a^{(j)}, b^{(j)} \in A^{(j)}, \quad 0 \leqslant j \leqslant m,$$
$$W(\tilde{f}, X, \ldots, X; A^{(0)}, \ldots, A^{(m)})$$
$$= [\tilde{f}(x^{(1)}, \ldots, x^{(n)}; c^{(0)}, \ldots, c^{(m)}),$$
$$\tilde{f}(y^{(1)}, \ldots, y^{(n)}; e^{(0)}, \ldots, e^{(m)})],$$
$$x^{(k)}, y^{(k)} \in X, \quad 1 \leqslant k \leqslant n, \quad c^{(j)}, e^{(j)} \in A^{(j)}, \quad 0 \leqslant j \leqslant m,$$

и принять во внимание соотношение

$$W(f, X; A^{(0)}, \ldots, A^{(m)}) \subseteq W(\tilde{f}, X, \ldots, X; A^{(0)}, \ldots, A^{(m)}),$$

то получим

$$\begin{aligned}
&|f(u;\ a^{(0)},\ \ldots,\ a^{(m)}) - \tilde{f}(x^{(1)},\ \ldots,\ x^{(n)};\ c^{(0)},\ \ldots,\ c^{(m)})| \\
&= f(u;\ a^{(0)},\ \ldots,\ a^{(m)}) - \tilde{f}(x^{(1)},\ \ldots,\ x^{(n)};\ c^{(0)},\ \ldots,\ c^{(m)}) \\
&\leqslant f(u;\ c^{(0)},\ \ldots,\ c^{(m)}) - \tilde{f}(x^{(1)},\ \ldots,\ x^{(n)};\ c^{(0)},\ \ldots,\ c^{(m)}) \\
&= \tilde{f}(u,\ \ldots,\ u;\ c^{(0)},\ \ldots,\ c^{(m)}) - \tilde{f}(x^{(1)},\ \ldots,\ x^{(n)};\ c^{(0)},\ \ldots,\ c^{(m)}) \\
&\leqslant \gamma \max_{1 \leqslant k \leqslant n} |u - x^{(k)}| \leqslant \gamma d(X).
\end{aligned}$$

Разность верхних границ может быть оценена аналогичным образом. Получение этих двух оценок доказывает утверждение теоремы.

Теорема 3, как видно из ее доказательства, легко обобщается на случай функции от нескольких переменных $x^{(1)}, \ldots, x^{(n)}$. Вместо $\gamma d(X)$ имеем величину

$$\sum_{k=1}^{n} \gamma^{(k)} d(X^{(k)}) \quad \left(\leqslant \gamma \max_{1 \leqslant k \leqslant n} d(X^{(k)})\right).$$

Следующий пример иллюстрирует тот факт, что степень близости множества значений функции f и ее интервальной оценки зависит от выбора аналитического выражения $f(x;\ a^{(0)}, \ldots, a^{(m)})$.

Пример. Пусть $f(x) = x - x^2$ и $X = [0,\ 1]$. Тогда

$$W(f,\ [0,\ 1]) = \{x - x^2 \mid 0 \leqslant x \leqslant 1\} = \left[0,\ \tfrac{1}{4}\right].$$

Различные аналитические выражения для f дают следующие результаты:

$$f^{(0)}(x) = x - x^2 \Rightarrow f^{(0)}([0,\ 1]) = [0,\ 1] - [0,\ 1] = [-1,\ 1],$$

$$f^{(1)}(x) = x(1 - x) \Rightarrow f^{(1)}([0,\ 1]) = [0,\ 1]([0,\ 1]) = [0,\ 1],$$

$$f^{(2)}(x) = \tfrac{1}{4} - \left(x - \tfrac{1}{2}\right)\left(x - \tfrac{1}{2}\right) \Rightarrow$$

$$f^{(2)}([0,\ 1]) = \tfrac{1}{4} - \left([0,\ 1] - \tfrac{1}{2}\right)\left([0,\ 1] - \tfrac{1}{2}\right) = \left[0,\ \tfrac{1}{2}\right],$$

$$f^{(3)}(x) = \tfrac{1}{4} - \left(x - \tfrac{1}{2}\right)^2 \Rightarrow$$

$$f^{(3)}([0,\ 1]) = \tfrac{1}{4} - \left([0,\ 1] - \tfrac{1}{2}\right)^2 = \left[0,\ \tfrac{1}{4}\right] = W(f,\ [0,\ 1]).$$

Для некоторых классов аналитических выражений можно доказать более сильные утверждения, нежели то, которое приведено в теореме 3. К их числу относится так называемая центрированная форма записи функции. Центрированная форма представляет собой специальное выражение, предназначенное для оценки функции f на интервале X. Ограничим наше рассмотрение случаем одной вещественной

переменной. Выберем в X произвольную точку z и представим $f(x)$ в виде

$$f(x) = f(z) + (x - z)\,h(x - z), \tag{3}$$

где сомножитель $h(x — z)$ зависит от новой переменной z, равной $x — z$. Будем называть (3) формой $f(x)$, центрированной относительно z. Применительно к многочленам центрированная форма есть не что иное, как обычное тейлоровское разложение $f(x)$ в окрестности точки z, записанное с сомножителем $x — z$, имеющимся у всех членов, отличных от постоянного.

Рациональная функция $f(x) = p(x)/q(x)$ может быть, согласно Ратшеку, приведена к центрированной форме следующим образом. Пусть n — максимум из степеней многочленов $p(x)$ и $q(x)$. Для z из X определим

$$\gamma_\nu := p^{(\nu)}(z) - f(z)\,q^{(\nu)}(z), \quad 1 \leqslant \nu \leqslant n.$$

Функция

$$h(y) = \sum_{\nu=1}^{n} \gamma_\nu \frac{y^{\nu-1}}{\nu!} \Bigg/ \sum_{\nu=0}^{s} q^{(\nu)}(z) \frac{y^\nu}{\nu!}$$

является решением функционального уравнения

$$f(x) = f(z) + (x - z)\,h(x - z).$$

Теорема 4. *Пусть f — вещественная функция от вещественного аргумента x и*

$$f(x) = f(z) + (x - z)\,h(x - z)$$

— аналитическое выражение для f в центрированной форме. Кроме того, пусть имеется выражение $\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z)$, аналогичное соответствующему выражению из теоремы 3. *Допустим, что для некоторого Y из $I(R)$ существует интервальная оценка $f(Y)$ и для каждой своей переменной $\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z)$ удовлетворяет условию Липшица, подобно тому как это было в теореме* 3. *Тогда для $X \subseteq Y$ выполняется соотношение*

$$q(W(f, X)f(x)) \leqslant c(d(X))^2, \quad c \geqslant 0. \tag{4}$$

Доказательство. Поскольку

$$\tilde{h}(x - z, \ldots, x - z) = h(x - z)$$

и

$$\tilde{f}(x^{(0)}, \ldots, x^{(n)}) = f(z) + (x^{(0)} - z)\,\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z),$$

то

$$\tilde{f}(x, \ldots, x) = f(z) + (x - z)\tilde{h}(x - z, \ldots, x - z)$$
$$= f(z) + (x - z)h(x - z) = f(x).$$

Теперь можно получить интервальную оценку дли f, записанной в центрированной форме:

$$f(X) = W(\tilde{f}, X, \ldots, X).$$

Этот результат позволяет переписать (4) в виде

$$q(W(f, X), W(\tilde{f}, X, \ldots, X)) \leqslant c(d(X))^2, \quad c \geqslant 0.$$

Пусть

$$W(\tilde{f}, X, \ldots, X) = [f(z) + (x^{(0)} - z)\tilde{h}(x^{(1)} - z, \ldots, x^{(n)} - z),$$
$$f(z) + (y^{(0)} - z)\tilde{h}(y^{(1)} - z, \ldots, y^{(n)} - z)],$$
$$x^{(k)}, y^{(k)} \in X, \quad 0 \leqslant k \leqslant n.$$

Заметим, что

$$W(f, X) \subseteq W(\tilde{f}, X, \ldots, X).$$

Из (21 п. 1.2) вытекает

$$q(W(f, X), W(\tilde{f}, X, \ldots, X)) \leqslant d(W(\tilde{f}, X, \ldots, X)) - d(W(f, X)).$$

Теперь положим

$$\min_{x \in X} |h(x - z)| = |h(w - z)|.$$

Легко

$$f(z) + (X - z)h(w - z) \subseteq f(z) + \{(x - z)h(x - z) | x \in X\} = W(f, X),$$

убедиться в истинности соотношения
если проанализировать два случая, связанных со знаком $h(w—z)$. Далее, исходя из (9 п. 1.2) и (14 п. 1.2), получаем

$$d(W(f, X)) \geqslant d((X - z)h(w - z)) = d(X)|h(w - z)|, \quad w \in X.$$

Наконец,

$$
\begin{aligned}
&q(W(f,\ X),\ W(\tilde{f},\ X,\ \ldots,\ X))\\
&\leqslant (y^{(0)}-z)\,\tilde{h}(y^{(1)}-z,\ \ldots,\ y^{(n)}-z)\\
&\quad -(x^{(0)}-z)\,\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)-d(X)|\,h(w-z)\,|\\
&= (y^{(0)}-z)\,\tilde{h}(y^{(1)}-z,\ \ldots,\ y^{(n)}-z)\\
&\quad -(y^{(0)}-z)\,\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)\\
&\quad +(y^{(0)}-z)\,\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)\\
&\quad -(x^{(0)}-z)\,\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)-d(X)|\,h(w-z)\,|\\
&= (y^{(0)}-z)(\tilde{h}(y^{(1)}-z,\ \ldots,\ y^{(n)}-z)\\
&\quad -\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z))\\
&\quad +(y^{(0)}-x^{(0)})\,\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)\\
&\quad -d(X)|\,\tilde{h}(w-z,\ \ldots,\ w-z)\,|\\
&\leqslant |\,y^{(0)}-z\,|\,|\,\tilde{h}(y^{(1)}-z,\ \ldots,\ y^{(n)}-z)\\
&\quad -\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)\,|+|\,y^{(0)}-x^{(0)}|\,|\\
&\quad \times|\,\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)\,|-d(X)|\,\tilde{h}(w-z,\ \ldots,\ w-z)\,|\\
&\leqslant d(X)(|\,\tilde{h}(y^{(1)}-z,\ \ldots,\ y^{(n)}-z)-\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)\,|\\
&\quad +\|\,\tilde{h}(x^{(1)}-z,\ \ldots,\ x^{(n)}-z)\,|-|\,\tilde{h}(w-z,\ \ldots,\ w-z)\,\|)\\
&\leqslant d(X)\Big(c^{(1)}\max_{1\leqslant k\leqslant n}|\,y^{(k)}-x^{(k)}\,|+c^{(2)}\max_{1\leqslant k\leqslant n}|\,x^{(k)}-w\,|\Big)\\
&\leqslant d(X)(c^{(1)}+c^{(2)})\,d(X)=c(d(X))^2.
\end{aligned}
$$

Выполнение использованного в этом построении условия Липшица было оговорено для $\tilde{h}$ и как следствие для [$\tilde{h}$].

Утверждение теоремы 4 также можно распространить на случай функции нескольких переменных. Обобщенное таким образом соотношение (4) было дано Хансеном; его же, но с применением другой техники получили Шуба и Миллер.

Как следствие из теоремы 3 возникает

Теорема 5. *Пусть f — вещественная функция от вещественного аргумента x, f(x) — аналитическое выражение для f. Будем считать, что выполнены все предположения теоремы* 3. *Тогда для* $X\subseteq Y$ *имеет место неравенство*

$$d(f(X))\leqslant cd(X),\qquad c\geqslant 0. \tag{5}$$

Доказательство. Опираясь на теорему 3 и соотношение (21 п.1.2), получаем

$$d(f(X)) \leqslant 2q(f(X),\ W(f,\ X)) + d(W(f,\ X)) \leqslant 2c^{(1)}d(X) + d(W(f,\ X)), \quad c^{(1)} \geqslant 0.$$

Исходя из условия Липшица для функции f, можно записать неравенство

$$d(W(f,\ X)) = |f(x) - f(y)| \leqslant c^{(2)}|x - y|, \quad \text{где} \quad x,\ y \in X, \quad c^{(2)} \geqslant 0,$$

из которого следует

$$d(f(X)) \leqslant 2c^{(1)}d(X) + c^{(2)}d(X) = cd(X),$$

что и требовалось доказать.

Соответствующее обобщение на случай нескольких переменных выглядит так:

$$d(f(X^{(1)},\ X^{(2)},\ \ldots,\ X^{(n)})) \leqslant \sum_{k=1}^{n} c^{(k)}d(X^{(k)}) \leqslant c \max_{1 \leqslant k \leqslant n} d(X^{(k)}). \quad (5')$$

Теперь докажем теорему о вхождении множества *W(f, X)* в другое множество, появляющееся в результате вычисления интервального выражения на основе теоремы о среднем значении.

Теорема 6. *Пусть f — вещественная функция от вещественного аргумента x, дифференцируемая на интервале $X = [x_1, x_2]$, и пусть $f'(x)$ — аналитическое выражение для f' такое, что интервальное выражение для $f'(X)$ определено. Тогда если для f' справедливы предположения теоремы* 5, *то*

$$W(f,\ X) \subseteq f(y) + f'(X)(X - y), \quad (а)$$

$$q(W(f,\ X),\ f(y) + f'(X)(X - y)) \leqslant \tilde{c}(d(X))^2, \quad (б)$$

где $y \in X$ и константа $\tilde{c} \geqslant 0$.

Доказательство. (а) Из теоремы о среднем, примененной к x и y из X, получаем

$$f(x) = f(y) + f'(y + \theta(x - y))(x - y), \quad 0 < \theta < 1.$$

Из

$$y + \theta(x - y) \in y + [0,\ 1](X - y) = X$$

с учетом монотонности включения вытекает

$$f(x) \in f(y) + f'(X)(X - y),$$

что доказывает утверждение (а), (б) Пусть

$$W(f,\ X) = [f(u),\ f(v)], \qquad u,\ v \in X.$$

Тогда из теоремы о среднем следует, что

$$d(W(f,\ X)) = f(v) - f(u) = |f(v) - f(u)| \geqslant |f(x_1) - f(x_2)| = |f'(\xi)|\,d(X), \quad \xi \in X.$$

Формулы (12 п.1.2), (3 п.1.2) и (20 п.1.2) дают

$$d(f'(X)(X-y)) \leqslant |f'(X)|d(X) + d(f'(X))|X-y| \leqslant |f'(X)|d(X) + d(f'(X))d(X).$$

Так как $f'(\xi) \in f'(X)$, то, принимая во внимание (21 п.7.2), получаем

$$q(f'(X),\ f'(\xi)) \leqslant d(f'(X)).$$

Теперь используем неравенство

$$|f'(X)| - |f'(\xi)| \leqslant q(f'(X),\ f'(\xi)),$$

которое следует из (4 п.1.2), (5 п.1.2) и определения 6 п.1.2. Применяя к *f(X)* соотношения (а), (21 п.1.2) и теорему 5, получаем требуемый результат:

$$\begin{aligned} &q(W(f,\ X),\ f(y) + f'(X)(X-y)) \\ &\leqslant d(f(y) + f'(X)(X-y)) - d(W(f,\ X)) \\ &\leqslant d(f'(X))d(X) + (|f'(X)| - |f'(\xi)|)d(X) \\ &\leqslant d(f'(X))d(X) + q(f'(X),\ f'(\xi))d(X) \\ &\leqslant 2c(d(X))^2 = \bar{c}(d(X))^2. \end{aligned}$$

Из теоремы 6 вытекает качественный результат теоремы 4 для записанного в центрированной форме выражения

$$f(y) + f'(X)(X-y), \qquad y \in X.$$

Это важный факт, поскольку уже для многочленов получение центрированной формы требует применения полной схемы Горнера Теорема 6 также может быть обобщена на случай нескольких переменных. Детали этого мы опускаем.

Рассмотрим теперь рациональную функцию

$$f(x) = p(x)/q(x),$$

где $p(x) = \sum_{\nu=0} a_\nu x^\nu$ и $q(x) = \sum_{\nu=0} b_\nu x^\nu$.

Связав *p(x)* и *q(x)* некоторыми условиями, можно указать выражения, для которых сохраняет силу свойство

$$q(W(f,\ X),\ f(X)) \leqslant cd'(X)^2, \qquad c \geqslant 0.$$

и более простые, нежели центрированная форма или использованное в теореме 6 представление, основанное на теореме о среднем.

Пусть даны $c = m(x)$— середина интервала и тейлоровские разложения $p(x) = \sum_{\nu=0}^{r} a'_\nu (x-c)^\nu$ и $q(x) = \sum_{\nu=0}^{s} b'_\nu (x-c)^\nu$. Без потери общности допустим, что $b'_0 = 1$ и $0 \notin q(X)$, где $q(X) = 1 + \sum_{\nu=0}^{s} b'_\nu (X-c)^\nu$. Если теперь

$$\operatorname{sign}(a_1')\operatorname{sign}(b_1' \cdot a_0') \leqslant 0 \qquad (7)$$

и предполагается, что $p(x)$ и $q(x)$ удовлетворяют неравенствам

$$d(p(X)) \leqslant c_1 d(X),$$

$$d(q(X)) \leqslant c_2 d(X),$$

то для интервального выражения

$$f(X) = \sum_{\nu=0}^{r} a_\nu'(X-c)^\nu \Big/ \left(1 + \sum_{\nu=1}^{s} b_\nu'(X-c)^\nu\right)$$

выполнено свойство (6). Это утверждение справедливо для обоих вышеприведенных выражений независимо от того, вычисляются они с помощью степеней $X - c$ либо по схеме Горнера. Если мы по-прежнему находимся в условиях предположения (7) и

$$0 \notin 1 + (X-c)q'(X),$$

то в (6) можно подставить выражение

$$f(X) = \frac{a_0' + (X-c)p'(X)}{1 + (X-c)q'(X)}.$$

Сомножитель $p'(X)$ представляет собой интервальную оценку для первой производной функции $p(x)$, удовлетворяющую соотношению $d(p'(X)) \leqslant \alpha d(X)$. Аналогично, $q'(X)$ — интервальная оценка для первой производной $q(x)$, удовлетворяющая неравенству $d(q'(X)) \leqslant \beta d(X)$.

В дальнейшем мы рассмотрим методы локализации нулей, использующие включения на участках монотонности функции.

Ниже мы дадим ряд возможных включений, применимых к отношению разностей. Эти включения будут частично упорядочены Оказывается, чго оптимальное включение может быть описано просто и систематично и что вычисления с соответствующими итерациями могут быть выполнены с теми же вычислительными затратами, что и интервальное оценивание производной. Эти включения выводятся другими способами, значительно отличающимися от использовавшихся у Хансена, решавшего ту же задачу

Включения для примеров, приведенных Хансеном, в точности соответствуют оптимальным включениям для этих же примеров, полученных методами, которые изложены в данной работе. Далее в рассуждениях следуем работе Алефельда.

Пусть имеется многочлен

$$p(x) = \sum_{\nu=0}^{n} a_\nu x^\nu.$$

Справедливость двух нижеследующих равенств очевидна:

$$p(x)-p(y)=\sum_{i=0}^{n}a_i(x^i-y^i)=\left(\sum_{i=1}^{n}a_i\sum_{j=1}^{i}x^{i-j}y^{j-1}\right)(x-y) \qquad (8)$$

$$=\left(\sum_{i=1}^{n}\left(\sum_{j=i}^{n}a_jy^{j-i}\right)x^{i-1}\right)(x-y),$$

$$p(x)-p(y)=\sum_{i=0}^{n}a_i(x^i-y^i)=\left(\sum_{i=1}^{n}a_i\sum_{j=1}^{i}y^{i-j}x^{j-1}\right)(x-y) \qquad (9)$$

$$=\left(\sum_{i=1}^{n}\left(\sum_{j=i}^{n}a_jx^{j-i}\right)y^{i-1}\right)(x-y).$$

Для фиксированного y и произвольного x из X с помощью (8) и свойства монотонности включения получаем

$$\frac{p(x)-p(y)}{x-y}\in\left(\sum_{i=1}^{n}c_{i-1}X^{i-1}\right)_{\mathrm{H}}=:J_1\subseteq J_2:=\sum_{i=1}^{n}c_{i-1}X^{i-1},$$

где

$$c_{i-1}=\sum_{j=i}^{n}a_jy^{j-i}, \qquad 1\leqslant i\leqslant n.$$

Здесь и далее буква H обозначает, что выражение, которое ею помечено, вычисляется по схеме Горнера. В многочлене J_2 степени X' вычисляются по правилу $X^0=1$, $X^r=X^{r-1}X$ при $r\geq 1$.

Включение $J_1\subseteq J_2$ следует из закона субдистрибутивности. Для вещественного числа y и интервалов A_j, $0\leqslant j\leqslant n-1$, всегда

$$\sum_{i=1}^{n}A_{i-1}y^{i-1}=\left(\sum_{i=1}^{n}A_{i-1}y^{i-1}\right)_{\mathrm{H}}.$$

При фиксированном y и произвольном x из X, $x\neq y$, используя субдистрибутивность, данное равенство и формулу (9), получаем

$$\frac{p(x)-p(y)}{x-y}\in\sum_{i=1}^{n}(C_{i-1})_{\mathrm{H}}\,y^{i-1}=\left(\sum_{i=1}^{n}(C_{i-1})_{\mathrm{H}}\,y^{i-1}\right)_{\mathrm{H}}=:J_3$$

$$\subseteq J_4:=\sum_{i=1}^{n}C_{i-1}y^{i-1}=\left(\sum_{i=1}^{n}C_{i-1}y^{i-1}\right)_{\mathrm{H}},$$

где

$$(C_{i-1})_{\mathrm{H}}=\left(\sum_{j=i}^{n}a_jX^{j-i}\right)_{\mathrm{H}}, \qquad 1\leqslant i\leqslant n,$$

и

$$C_{i-1} = \sum_{j=i}^{n} a_j X^{j-i}, \qquad 1 \leqslant i \leqslant n.$$

Докажем теперь еще одну теорему.

Теорема 7. *Введенные выше выражения удовлетворяют соотношениям*

$$J_1 \subseteq J_2 \subseteq J_4, \tag{a}$$

$$J_1 \subseteq J_3 \subseteq J_4, \tag{b}$$

$$J_4 \subseteq p'(X) = \sum_{\nu=1}^{n} \nu a_\nu X^{\nu-1}. \tag{c}$$

Доказательство. Для простоты ограничимся случаем многочлена четвертой степени (n = 4). Общий случай может быть рассмотрен аналогичным образом.

(a) и (c).Нам достаточно показать, что $J_2 \subseteq J_4 \subseteq p'(X)$. Из свойства монотонности включения и соотношения (8 п. 7.1) получаем

$$\begin{aligned}
J_2 &= \sum_{i=1}^{n} c_{i-1} X^{i-1} \\
&= (a_1 + a_2 y + a_3 y^2 + a_4 y^3) X^0 + (a_2 + a_3 y + a_4 y^2) X \\
&\quad + (a_3 + a_4 y) X^2 + a_4 X^3 \\
&\subseteq a_1 + a_2 X + a_3 X^2 + a_4 X^3 + a_2 y + a_3 y X + a_4 y X^2 \\
&\quad + a_3 y^2 + a_4 y^2 X + a_4 y^3 \\
&= a_1 + a_2 X + a_3 X^2 + a_4 X^3 + (a_2 + a_3 X + a_4 X^2) y \\
&\quad + (a_3 + a_4 X) y^2 + a_4 y^3 = J_4 \\
&\subseteq a_1 + a_2 X + a_3 X^2 + a_4 X^3 + a_2 X + a_3 X^2 + a_4 X^3 \\
&\quad + a_3 X^2 + a_4 X^3 + a_4 X^3 = p'(X).
\end{aligned}$$

Достаточно показать, что

$$J_1 \subseteq J_3. \tag{b}$$

$$\begin{aligned}
J_1 &= ((c_3 X + c_2) X + c_1) X + c_0 \\
&= ((a_4 X + (a_3 + a_4 y)) X + a_2 + a_3 y + a_4 y^2) X + a_1 + a_2 y + a_3 y^2 + a_4 y^3 \\
&\subseteq ((a_4 X + a_3) X + a_4 y X + a_2 + a_3 y + a_4 y^2) X + a_1 + a_2 y + a_3 y^2 + a_4 y^3 \\
&= (((a_4 X + a_3) X + a_2) + a_4 y X + a_3 y + a_4 y^2) X + a_1 + a_2 y + a_3 y^2 + a_4 y^3 \\
&= (((a_4 X + a_3) X + a_2) + (a_4 X + a_3) y + a_4 y^2) X + a_1 + a_2 y + a_3 y^2 + a_4 y^3 \\
&\subseteq ((a_4 X + a_3) X + a_2) X + (a_4 X + a_3) y X + a_4 y^2 X + a_1 + a_2 y + a_3 y^2 + a_4 y^3 \\
&= (((a_4 X + a_3) X + a_2) X + a_1) y^0 + ((a_4 X + a_3) X + a_2) y \\
&\quad + (a_4 X + a_3) y^2 + a_4 y^3 = J_3.
\end{aligned}$$

Итак, теорема доказана.

Нельзя ответить в общем случае на вопрос о том, какое из выражений: J_2 или J_3 — дает лучшее включение. Возможно и $J_2 \subseteq J_3$, и $J_3 \subseteq J_2$. Пусть, например,

$$p(x) = x^3 - x^2, \quad X = [-1, 2], \quad y = 1.$$

Тогда имеем

$$J_2 = (a_1 + a_2 y + a_3 y^2) X^0 + (a_2 + a_3 y) X + a_3 X^2 = X^2 = [-2, 4]$$

и

$$J_3 = ((a_3 X + a_2) X + a_1) y^0 + (a_3 X + a_2) y + a_3 y^2$$
$$= (X - 1) X + (X - 1) + 1 = [-5, 4].$$

Здесь $J_2 \subset J_3$.

Если, с другой стороны, $y = 0$ и, следовательно, $c_{i-1} = a_i$, $1 \leqslant i \leqslant n$, то получаем

$$J_2 = \sum_{i=1}^{n} a_i X^{i-1} \quad \text{и} \quad J_3 = \left(\sum_{i=1}^{n} a_i X^{i-1} \right)_H.$$

Теперь $J_3 \subseteq J_2$.

Рассмотрим снова пример с $p(x) = x^3 - x^2$ при $y = 0$ и $X = [0, 2]$. В этом случае

$$J_2 = X^2 - X = [-2, 4] \quad \text{и} \quad J_3 = (X - 1) X = [-2, 2],$$

откуда имеем $J_3 \subset J_2$.

Получение интервалов J_1 и J_2 с помощью теоремы 7 требует предварительного нахождения $c_{i-1} = \sum_{j=i}^{n} a_j y^{j-i}$, $1 \leqslant i \leqslant n$. Если вычисляется значение многочлена $p(x)$ в точке y, что встречается, например, в итерационных методах , которые будут пассмотрены дальше, то нахождение c_{i-1} не требует выполнения каких-либо дополнительных арифметических операций. Значения c_{i-1} могут быть найдены в процессе вычисления $p(y)$. Пусть, как и ранее,

$$p(x) = \sum_{i=0}^{n} a_i x_i.$$

Воспользуемся схемой Горнера

$$p_n := a_n,$$

и для $i = n, n-1, \ldots, 1$

$$p_{i-1} := p_i y + a_{i-1},$$

откуда $p_0 = p(y)$. По определению

$$\begin{aligned} c_{n-1} &= a_n && (= p_n), \\ c_{n-2} &= a_n y + a_{n-1} && (= p_{n-1}), \\ &\vdots && \vdots \\ c_0 &= c_1 y + a_1 && (= p_1). \end{aligned}$$

Следовательно, $c_{i-1} = p_i$, $1 \leqslant i \leqslant n$.

Примеры.

$$p(x) = x^4 - 1, \quad x = [0.5,\ 3.5], \quad y = 2 \tag{a}$$

Имеем

$$J_1 = J_2 = J_3 = J_4 = [10.625,\ 89.375],$$
$$p'(X) = (p'(X))_{\text{H}} = [0.5,\ 171.5].$$

Оценка J_1 совпадает с той, которую получил на этих же данных Хансен.

$$p(x) = x^3 + 4x - 16, \quad X = [-1,\ 3], \quad y = 1. \tag{b}$$

Получаем

$$J_1 = J_2 = J_3 = J_4 = [1,\ 17],$$
$$p'(X) = (p'(X))_{\text{H}} = [-5,\ 31],$$

что совпадает с результатом Хансена.

$$p(x) = \sum_{i=0}^{n} a_i x^i, \quad 0 \in X, \quad y = 0. \tag{c}$$

В этом случае

$$c_0 = a_1, \quad c_1 = a_2, \ \ldots, \ c_{n-1} = a_n$$

и

$$J_1 = \left(\sum_{i=1}^{n} c_{i-1} X^{i-1}\right)_{\text{H}} = \left(\sum_{i=1}^{n} a_i X^{i-1}\right)_{\text{H}}.$$
$$p(x) = x^3 - x^2, \quad X = [1,\ 3], \quad y = 2. \tag{d}$$
$$J_1 = J_2 = J_3 = [4,\ 14] \subset [2,\ 16] = J_4 \subset (p'(X))_{\text{H}}$$
$$= [1,\ 21] \subset [-3,\ 25] = p'(X),$$

что опять совпадает со значением, вычисленным Хансеном.

Однако при $X = [-1, 2]$ и $y = 1$

$$J_1 = J_2 = [-2,4], \quad J_3 = [-5,\ 4], \quad J_4 = [-5,\ 7],$$
$$(p'(X))_{\text{H}} = [-10,\ 8], \quad p'(X) = [-10,\ 14].$$

Пусть $x_0 \in X$ и $f \in C^{n+1}(X)$. (e)

Используя тейлоровское разложение, получаем

$$f(x) = p(x) + \varphi(x),$$

где

$$\varphi(x) = \int_{x_0}^{x} \frac{(x-t)^n}{n!} f^{(n+1)}(t)\,dt$$

и

$$p(x) = \sum_{k=0}^{n} \frac{(x-x_0)^k}{k!} f^{(k)}(x_0).$$

Функция φ дифференцируема и φ

$$\varphi'(x) = \int_{x_0}^{x} \frac{(x-t)^{n-}}{(n-1)!} f^{(n+1)}(t)\,dt.$$

Интегральная теорема о среднем дает формулу

$$\varphi'(x) = f^{(n+1)}(\eta) \int_{x_0}^{x} \frac{(x-t)^{n-1}}{(n-1)!}\,dt = \frac{(x-x_0)^n}{n!} f^{(n+1)}(\eta),$$

где η лежит между x и x_0. Применяя к φ теорему о среднем, получаем

$$f(x) - f(y) = p(x) - p(y) + \varphi(x) - \varphi(y)$$
$$= \left\{ \sum_{k=1}^{n} c_{k-1}(x-x_0)^{k-1} + \varphi'(\xi) \right\} (x-y),$$

где

$$c_{k-1} = \sum_{i=k}^{n} (y-x_0)^{i-k} \frac{f^{(k)}(x_0)}{k!}, \quad 1 \leqslant k \leqslant n,$$

и

$$\varphi'(\xi) = \frac{(\xi - x_0)^n}{n!} f^{(n+1)}(\eta),$$

причем ξ лежит между x и y, а η — между x_0 и ξ. Полагая $y = x_0$, имеем

$$c_0 = f'(x_0)/1!, \ \ldots, \ c_{n-1} = f^{(n)}(x_0)/n!.$$

Если для (n+ 1)-й производной существует вычислимое интервальное выражение, то для $y = x_0$

$$\frac{f(x) - f(y)}{x - y} \in \sum_{k=1}^{n} \frac{f^{(k)}(x_0)}{k!} (X - x_0)^{k-1} + f^{(n+1)}(X) \frac{(X-x_0)^n}{n!},$$

поскольку η, ξ ∈ X. Эта оценка снова совпадает с данной Хансеном.

$$p(x) = x^7 + 3x^6 - 4x^5 - 12x^4 - x^3 - 3x^2 + 4x + 12,$$
$$X = [1.8,\ 3], \quad y = 2. \tag{f}$$

Теперь

$$J_1 = [173.2362,\ 2400], \qquad J_2 = [161.4762,\ 2411.76],$$
$$J_3 = [24.72,\ 2400], \qquad J_4 = [-870.2933,\ 3443.5296],$$
$$(p'(X))_{\text{н}} = [71.799808,\ 6520], \quad p'(X) = [-2378.791292,\ 8970.592].$$

Сделанные утверждения могут быть распространены на многомерный случай.

Замечания. В этом микромодуле было рассмотрено интервальное оценивание вещественных функций. Мы сознательно не говорили о произвольных отображениях из $I(\mathrm{R})$ в $I(\mathrm{R})$. Приложения в последующих микромодулях требуют использования многих свойств, которые возможно доказать только для интервальных оценок. Если разрешить на $I(\mathrm{R})$ отображения более общего вида, то для каждого приложения необходимо описать множество условий. Следующий пример показывает, сколь велик класс отображений из $I(\mathrm{R})$ в $I(\mathrm{R})$. Единственное ограничение здесь состоит в том, что если область определения сузить до R, то множество значений также будет принадлежать R. Итак, пусть f — вещественная функция, $f(x)$ — ее аналитическое выражение, $f(X)$ — оценивающая функция для $f(x)$. Тогда для произвольной $\varphi(x)$ такой, что $\varphi(0)=0$,

$$\Psi(X) = f(X) + \varphi(d(X))[-1,\ 1]$$

определяет отображение из $I(\mathrm{R})$ в $I(\mathrm{R})$. Очевидно, что $\Psi([x,\ x]) \in \mathrm{R}$. Если $\varphi(x) \geqslant 0$ при $x \geqslant 0$, то $W(f,\ X) \subseteq \Psi(X)$, и если $\varphi(x)$ монотонна и не убывает при $x \geq 0$, то $\Psi(X)$ обладает свойством монотонности включения в форме (1'). Этот пример показывает, что, соответствующим образом выбирая $\varphi(x)$, можно строить отображения Ψ, имеющие различные свойства. Если же потребовать от отображений из $I(\mathrm{R})$ в $I(\mathrm{R})$ выполнения всех свойств интервальных оценок, то не найдется никаких других полезных отображений, кроме в точности этих оценок.

Покажем, как можно сократить доказательство теоремы 4, используя теорему 5 примерно так же, как и при доказательстве теоремы 6. Записанная в центрированной форме интервальная оценка

$$f(X) = f(z) + (X - z)h(X - z)$$

удовлетворяет соотношению

$$W(f,\ X) \subseteq f(X).$$

В соответствии с (21 п.1.2) получаем

$$q(W(f, X), f(X)) \leqslant d(f(X)) - d(W(f, X)).$$

Пусть теперь

$$\min_{x \in X} |h(x - z)| = |h(w - z)|.$$

Тогда

$$f(z) + (X - z)h(w - z) \subseteq f(z) + \{(x - z)h(x - z) \mid x \in X\} = W(f, X).$$

С учетом (9 п.1.2), (14 п.1.2) и приведенного выше включения имеем

$$d(W(f, X)) \geqslant d((X - z)h(w - z)) = d(X)|h(w - z)|, \quad w \in X.$$

Исходя из (10 п.1.2), (12 п.1.2), (3 п.1.2) и (20 п.1.2), получаем

$$\begin{aligned} d(f(X)) &= d(f(z) + (X - z)h(X - z)) = d((X - z)h(X - z)) \\ &\leqslant |X - z| d(h(X - z)) + d(X)|h(X - z)| \\ &\leqslant d(X)d(h(X - z)) + d(X)|h(X - z)|. \end{aligned}$$

Поскольку $h(w - z) \in h(X - z)$, из (21 п.7.2) следует, что

$$q(h(X - z), h(w - z)) \leqslant d(h(X - z)).$$

Определение 6 п.1.2 и соотношения (4 п.1.2), (5 п.1.2) дают неравенство

$$|h(X - z)| - |h(w - z)| \leqslant q(h(X - z), h(w - z)).$$

На основе приведенных выше неравенств получаем

$$\begin{aligned} q(W(f, X), f(X)) &\leqslant d(f(z) + (X - z)h(X - z)) - d(W(f, X)) \\ &\leqslant d(X)d(h(X - z)) + d(X)|h(X - z)| - d(X)|h(w - z)| \\ &= d(X)d(h(X - z)) + (|h(X - z)| - |h(w - z)|)d(X) \\ &\leqslant d(X)d(h(X - z)) + q(h(X - z), h(w - z))d(X) \\ &\leqslant d(X) \cdot 2 \cdot d(h(X - z)). \end{aligned}$$

И наконец, после применения теоремы 5 к выражению $h(X - z)$ оказывается, что

$$q(W(f, X), f(X)) \leqslant d(X) \cdot 2 \cdot \bar{c} \cdot d(X) = c(d(X))^2.$$

При выполнении некоторых условий дифференцируемости для аналитических выражений можно определить общие условия, при которых соотношение (4) справедливо. При этом теоремы 4 и 6 получаются как частные случаи. У Херцбергера можно найти простую интерполяционную формулу, вычисляющую множество значений для семейства многочленов с заданными коэффициентами Им используется тот вытекающий из теоремы 2 факт, что интервальное оценивание позволяет точно вычислить множество значений функции,

когда все переменные и параметры входят в аналитическое выражение лишь по одному разу.

Корнелиус и Лонер предложили выражение *f(X),* для которого $q(W(f, X), f(X)) \leqslant cd(X)^{s+1}$, $s \geqslant 1$. При s = 1, 2, 3 *f(X)* вычисляется с помощью весьма простого алгоритма.

1.4. Машинная интервальная арифметика

Теперь мы остановимся на вопросах реализации интервальных операций на цифровой вычислительной машине Обшеизвестно, что в машине может быть представлено лишь конечное множество чисел. Чаще всего они записываются в полулогарифмической форме, а точнее — в форме с плавающей точкой:

$$x = m \cdot b^e.$$

Здесь m — мантисса, b — основание степени, e — порядок. Как правило, для внутримашинного представления выбирается основание b, равное 2, а мантисса нормализуется, т. е. ее абсолютное значение помещается в интервал [1/2, 1). Целое e принадлежит интервалу $[e_{\min}, e_{\max}]$.

Множество машинных чисел описанного типа обозначим через R_M, и всюду далее будем предполагать, что оно симметрично относительно нуля, т. е. R_M = — R_M. Для аппроксимации вещественных чисел, лежащих в интервале $[\min\limits_{y \in R_M} y;\ \max\limits_{y \in R_M} y]$, можно с успехом использовать машинные числа $\{\tilde{x} \mid \tilde{x} \in \mathbb{R}_M\}$. Аппроксимация достигается применением отображения

$$fl\colon \mathbb{R} \ni x \to \tilde{x} = fl(x) \in \mathbb{R}_M. \qquad (1)$$

Это отображение называется округлением, если выполнено свойство

$$x \leqslant y \Rightarrow fl(x) \leqslant fl(y) \quad \text{(монотонность)}. \qquad (2)$$

Округление, которое отображает R_M в R_M так, что

$$x \in R_M \Rightarrow fl(x) = x, \qquad (3)$$

называется оптимальным (приведенное определение не является стандартным. Обычно под оптимальным понимают такое округление, которое отображает округляемое число x в $\tilde{x}$, ближайшее в некотором смысле к x.)

Особый интерес представляют так называемые направленные округления. Если для округления ↓ справедлива импликация

$$x \in \mathbb{R} \Rightarrow \downarrow x \leqslant x, \tag{4}$$

то говорят об округлении вниз. Аналогично,

$$\uparrow x := -(\downarrow(-x)), \qquad x \leqslant \mathbb{R}, \tag{5}$$

определяет округление вверх. Техника выполнения этих округлений для различных способов представления чисел неоднократно освещалась в литературе. Подобно тому как вещественные числа приближаются с помощью машинных, можно вещественные интервалы приближать машинными интервалами. В этом случае интервал X из $I(\mathrm{R})$, для которого справедливо соотношение $X \subseteq [\min\limits_{y \in \mathbb{R}_M} y, \max\limits_{y \in \mathbb{R}_M} y]$, заменяется соответствующим машинным интервалом из множества

$$I(\mathbb{R}_M) = \{[x_1, x_2] \mid x_1, x_2 \in \mathbb{R}_M, x_1 \leqslant x_2\} \subset I(\mathbb{R}).$$

Для того чтобы основные свойства интервальных операций выполнялись и для их машинных аналогов, применяется округление интервалов (интервальное округление)

$$\updownarrow : I(\mathbb{R}) \ni X \to \updownarrow X \in I(\mathbb{R}_M),$$

причем

$$X \in I(\mathbb{R}) \Rightarrow X \subseteq \updownarrow X \tag{6}$$

и

$$X, Y \in I(\mathbb{R}), \quad X \subseteq Y \Rightarrow \updownarrow X \subseteq \updownarrow Y. \tag{7}$$

Если рассмотреть переход от интервала $X = [x_1, x_2]$ из $I(\mathrm{R})$ к его машинному представлению $\tilde{X} = [\tilde{x}_1, \tilde{x}_2]$, то окажется, что (7) означает, что необходимо осуществить этот переход путем округления каждой из границ X. Из (6) следует, что границы должны быть округлены направленно. Таким образом, округление интервала X состоит в нахождении $\updownarrow X$ по правилу

$$\updownarrow X = \updownarrow [x_1, x_2] = [\downarrow x_1, \uparrow x_2]. \tag{8}$$

Проведенное обсуждение показывает, что для того, чтобы округлить интервал, достаточно иметь $\downarrow$ — направленное округление вниз. С другой стороны, $\uparrow$ и $\downarrow$ не обязательно должны быть связаны соотношением (5).

Если над двумя машинными числами x и y из $\mathrm{R_M}$ производится машинная операция *, где $* \in \{+, -, \cdot, :\}$, то ее результатом оказывается новое число z из $\mathrm{R_M}$. Проигнорировав возможность выхода за пределы допустимого диапазона (переполнение и антипереполнение), можно, используя соответствующее округление *fl*, представить z в виде

$$z = fl(x * y). \tag{9}$$

Таким же образом мы можем определить результат машинной операции над интервалами.

Определение 1. Пусть $A, B \in I(\mathbb{R}_M)$, $* \in \{+, -, \cdot, :\}$, $\updownarrow$ — интервальное округление. Тогда результат операции *, выполненной над A и B с применением $\updownarrow$ есть

$$C = \updownarrow (A * B) \in I(\mathbb{R}_M). \tag{10}$$

Теперь мы покажем, что основные свойства интервальной арифметики при использовании этого определения сохраняются.

Теорема 2. *Для машинных интервальных операций, задаваемых определением* 1, *справедливо следующее утверждение:*

$$A^{(k)}, B^{(k)} \in I(\mathbb{R}_M),\ * \in \{+, -, \cdot, :\},\ A^{(k)} \subseteq B^{(k)},\ k = 1, 2, \tag{11}$$
$$\Rightarrow C^{(1)} = \updownarrow (A^{(1)} * A^{(2)}) \subseteq C^{(2)} = \updownarrow (B^{(1)} * B^{(2)}).$$

Доказательство теоремы 2 следует непосредственно из свойства интервальных округлений (7).

Утверждение (11) отражает не что иное, как свойство монотонности включения (9 п.1.1) применительно к машинным интервальным операциям.

Очередная теорема представляет интерес с точки зрения оценки погрешностей округлений.

Теорема 3. *Пусть* $\updownarrow$ *— интервальное округление, сводящееся с помощью* (8) *к направленным округлениям* $\downarrow$ *и* $\uparrow$, *и пусть* $* \in \{+, -, \cdot, :\}$. *Тогда*

$$A, B \in I(\mathbb{R}_M) \Rightarrow A * B \subseteq C = \updownarrow (A * B) \in I(\mathbb{R}_M),$$
$$a \in A,\ b \in B \Rightarrow a * b \in C = \updownarrow (A * B) \in I(\mathbb{R}_M). \tag{12}$$

Если имеется округление fl, применение которого приводит к выполнению неравенства

$$\downarrow a \leqslant fl(a) \leqslant \uparrow a, \quad a \in \mathbb{R},$$

то для x, y, z из R_M *справедливо*

$$z = fl(x * y) \in Z = \updownarrow ([x, x] * [y, y]) \in I(\mathbb{R}_M). \tag{13}$$

Доказательство свойств (12) и (13) мы опускаем, поскольку оно элементарно и следует непосредственно из соответствующих определений.

Интервальное оценивание аналитического выражения функции, проведенное с использованием операций из определения 1, дает интервалы, объемлющие значения оценивающей функции. Среди этих интервалов находятся и оценки множества значений функции. Более того, при выполнении подобных вычислений сохраняется свойство монотонности включения.

На практике машинные интервальные операции реализуются с помощью соответствующих программно-аппаратных средств Эти средства могут служить поддержкой языка программирования высокого уровня**.** Один из вариантов реализации — набор подпрограмм, написанных, скажем, на Алголе. Рассмотрим вкратце последнюю возможность. В большинстве случаев, в частности у Криста, такой набор содержит средство, с помощью которого выполняется округление ↓. Это средство может быть, например, оформлено в виде процедуры-функции LOW; через нее определяются стандартные операции интервальной арифметики — ADD, SUB, MUL и DIV, а также элементарные функции. На деталях реализации подобных подпрограмм мы остановимся в приложении В.

Посмотрим теперь на алгоритмы, описанные в терминах вещественных чисел. К их числу можно отнести, скажем, схему Горнера или метод Гаусса. Если такой алгоритм реализуется на компьютере, т. е. с использованием машинной арифметики, то даже исходные данные в общем случае не могут быть представлены точно. Возникающие при этом трудности преодолеваются применением машинной интервальной арифметики. Исходные данные просто заключаются в интервалы, имеющие своими границами машинные числа. Если теперь представить себе, что алгоритм будет выполняться без учета погрешностей округления, то, как показано в микромодуле 24, ширина реализующего интервала станет, вообще говоря, возрастать в большей степени, чем это обусловлено исходными данными**.** При наличии погрешностей округления описанное свойство проявляется еще сильнее.

Обсудим следующий вопрос: на какой рост точности результата можно рассчитывать, если перейти от алгоритма, использующего машинную интервальную арифметику с t_1 цифрами в мантиссе, к алгоритму, использующему t_2-значную арифметику, где $t_2>t_1$? Предполагается, что при таком переходе диапазон возможных значений порядка остается неизменным. Таким образом, все числа, представимые с t_1 цифрами, столь же точно записываются с t_2 цифрами.

Пусть $x \in K$, $x \neq 0$ и

$$x=\left(\sum_{\nu=-1}^{-\infty} a_\nu b^\nu\right) b^e, \quad 1 \leqslant a_{-1} \leqslant b-1, \quad 0 \leqslant a_\nu \leqslant b-1, \quad \nu \leqslant -2.$$

Чтобы гарантировать единственность представления x, мы предполагаем, что не существует ν_0 такого, что при $\nu \leq \nu_0$ все a_ν равны b — 1. Будем также считать, что число x не представимо в t_1-значной системе с плавающей точкой. (Если бы последнее допущение отсутствовало, то следующее рассуждение было бы совершенно излишним.) Пусть, кроме того, интервальное округление (8)

осуществляется через оптимальное округление границ. В соответствии с (8) при $x > 0$ получаем

$$\updownarrow x = \updownarrow [x, x] = [\downarrow x, \uparrow x],$$

где

$$\downarrow x = \left(\sum_{\nu=-1}^{-t} a_\nu b^\nu \right) b^e, \quad \uparrow x = \left(\sum_{\nu=-1}^{-t_1} a_\nu b^\nu \right) b^e + b^{-t_1+e}.$$

Очевидно, что ширина $\updownarrow x$ есть

$$d(\updownarrow x) = b^{-t_1+e}.$$

Точно такой же результат получается для ширины $\updownarrow x$ при $x < 0$. В дальнейшем зависимость результата от длины мантиссы будет отражаться с помощью записи $fl_1(x)$ (соответственно $fl_2(x)$). Следовательно, под fl мы будем понимать интервальное округление вещественного числа (а впоследствии и вещественного интервала). Предыдущее равенство теперь может быть переписано в виде

$$d(fl_1(x)) = b^{-t_1+e}.$$

Аналогично, для мантиссы длины $t_2 = t_1 + l$ получаем

$$d(fl_2(x)) \leqslant b^{-t_1+e-l}.$$

Неравенство выполняется как строгое в случае, когда x представляется точно с t_2-разрядной мантиссой. Как следствие

$$d(fl_2(x)) \leqslant b^{-l} d(fl_1(x)). \tag{14}$$

Из предположений, сформулированных для интервальных округлений, следует, что для двух машинных интервалов A и B

$$\updownarrow (A * B) = fl_1(A * B) = [(1 - \varepsilon_1)(A * B)_1, (1 + \varepsilon_2)(A * B)_2]$$

(см. определение 1). С помощью $(A*B)_1$ и $(A*B)_2$ вычисляются границы точного значения результата, причем

$$-\varepsilon_1 (A * B)_1 \leqslant 0, \quad \varepsilon_2 (A * B)_2 \geqslant 0$$

и

$$|\varepsilon_1|, |\varepsilon_2| \leqslant b^{1-t_1}.$$

Следовательно, можно записать

$$fl_1(A * B) = A * B + [-\varepsilon_1 (A * B)_1, \varepsilon_2 (A * B)_2]. \tag{15a}$$

Оценкой ширины результата служит

$$d(fl_1(A * B)) \leqslant d(A * B) + 2b^{1-t_1} |A * B|. \tag{15b}$$

Эта оценка показывает, что когда используется мантисса фиксированной длины, то рост ширины $d(fl_1(A * B))$ определяется величиной $|A * B|$. Пусть мы знаем, что x принадлежит интервалу X из $I(\mathrm{R})$. Естественно выбрать некоторое x из X в качестве приближенного

значения x. Оценим абсолютную и относительную погрешность такого приближения:

$$|x - \tilde{x}| \leqslant d(X) =: \Delta(X), \tag{16}$$

и если $0 \notin X$, $x \neq 0$, то

$$\left|\frac{x - \tilde{x}}{\tilde{x}}\right| \leqslant \frac{d(X)}{\min\{|x| \mid x \in X\}} =: \rho(X). \tag{17}$$

Теорема 4. *Пусть A, B, C и D — машинные интервалы, причем*

$$A \subseteq C, \quad B \subseteq D, \tag{18}$$

и

$$\begin{aligned} d(C) &\leqslant s_1, & d(D) &\leqslant s_2, \\ d(A) &\leqslant b^{-t} s_1, & d(B) &\leqslant b^{-t} s_2. \end{aligned} \tag{19}$$

*Предположим, что * — одна из арифметических операций над вещественными интервалами и $0 \notin fl_1(C * D)$. Тогда границы для $\Delta(fl_2(A * B))$ (соответственно $\rho(fl_2(A * B))$) оказываются в b^l раз меньше, чем границы для $\Delta(fl_1(C * D))$ (соответственно $\rho(fl_1(C * D))$).*

Доказательство. Используя (15b), (10 п.1.2), (12 п.1.2), неравенство

$$d(1/X) \leqslant |1/X|^2 d(X) \quad (0 \notin X),$$

а также первую строку из (19), сразу же получаем

$$\begin{aligned} d(fl_1(C * D)) &\leqslant d(C * D) + 2b^{1-t_1}|C * D| \\ &\leqslant \left\{\begin{array}{ll} s_1 + s_2, & * = +, - \\ |C| s_2 + s_1 |D|, & * = \cdot \\ |C||1/D|^2 s_2 + |1/D| s_1, & * = : \end{array}\right\} + 2b^{1-t_1}|C * D|. \end{aligned}$$

На основе (18) и (19) аналогичным способом доказывается, что

$$d(A * B) \leqslant b^{-t} \left\{\begin{array}{ll} s_1 + s_2, & * = +, - \\ |C| s_2 + s_1 |D|, & * = \cdot \\ |C||1/D|^2 s_2 + |1/D| s_1, & * = : \end{array}\right\}. \tag{20}$$

Из (18) и теоремы 2 следует включение

$$fl_2(A * B) \subseteq fl_2(C * D) \subseteq fl_1(C * D).$$

Оно справедливо, поскольку мы предположили, что интервальное округление задано через оптимальные округления границ. Таким образом,

$$\min\{|x| \mid x \in fl_2(A * B)\} \geqslant \min\{|x| \mid x \in fl_1(C * D)\}. \tag{21}$$

Из (15b), (20) и неравенства $|A * B| \leqslant |C * D|$ вытекает

$$d(fl_2(A * B)) \leqslant d(A * B) + 2b^{1-t_1-t}|C * D|$$
$$\leqslant \left\{ \begin{array}{ll} s_1 + s_2, & * = +, - \\ |C|s_2 + s_1|D|, & * = \cdot \\ |C||1/D|^2 s_2 + |1/D|s_1, & * = : \end{array} \right\} + 2b^{1-t_1-t}|C * D|.$$

Это доказывает утверждение теоремы для верхних границ абсолютной погрешности. Для верхних границ относительной погрешности требуемый результат получается непосредственно из (21).

Простое, но важное следствие из теоремы 4 содержит

Теорема 5. *Допустим, что справедливы все приведенные выше предположения, касающиеся машинной интервальной арифметики. Кроме того, имеется заданный в поле вещественных чисел алгоритм, который выполняется в машинной интервальной арифметике с мантиссой длины t_1. Если затем этот алгоритм выполнить в арифметике с мантиссой длины t_2, где $t_2 = t_1 + l$, $l > 0$, то границы абсолютной и относительной погрешностей уменьшатся в b^l раз.* ***(Под алгоритмом здесь понимается однозначно определенная последовательность арифметических операций вместе с конкретными входными данными.)***

Доказательство. Из (14) следует, что в нашем случае интервальное округление входных данных удовлетворяет важному предположению (19) теоремы 4. Свойства интервальной арифметики обеспечивают справедливость (18). Окончательно доказательство получаем из теоремы 4 применением полной индукции. Теорема 5 указывает способ получения результата с наперед заданной абсолютной или относительной точностью. Пусть, например, d_1 — наибольшая ширина, которую имеют результирующие интервалы, вычисленные с помощью t_1-значной мантиссы, а ε — требуемая абсолютная точность. Если $d_1 \leqslant \varepsilon$, то цель достигнута. В противном случае число цифр в мантиссе увеличивается на l, где l удовлетворяет неравенству

$$b^{-l} d_1 \leqslant \varepsilon.$$

(Такой выбор не гарантирует, что абсолютная погрешность уменьшится в b^l раз. В соответствии с теоремой 5 эта оценка верна лишь для верхней границы абсолютной погрешности.)

Факты, обсужденные и доказанные в теореме 5, были изучены Румпом; он же проиллюстрировал их числовыми примерами. Один из этих примеров — решение системы уравнений, задаваемой матрицей Гильберта размерности 7×7, причем в правой части каждого уравнения стоит 1. Результаты применения к этой системе алгоритма Гаусса, реализованного в машинной интервальной арифметике с 15, 20, 25, 30 и 35 цифрами в мантиссе, воспроизведены в табл. 1.

Таблица 1

Верхняя граница ρ *(Xi)* относительной погрешности в алгоритме Гаусса

Число цифр в мантиссе, t	15	20	25	30	35
1	$> 1^a$	0.11×10^{-5}	0.11×10^{-8}	0.11×10^{-13}	0.11×10^{-18}
2	0.34×10^{0}	0.29×10^{-5}	0.29×10^{-10}	0.29×10^{-15}	0.29×10^{-20}
3	0.18×10^{-1}	0.17×10^{-6}	0.17×10^{-11}	0.17×10^{-16}	0.17×10^{-21}
4	0.16×10^{-2}	0.16×10^{-7}	0.16×10^{-12}	0.16×10^{-17}	0.16×10^{-22}
5	0.26×10^{-3}	0.25×10^{-8}	0.25×10^{-13}	0.25×10^{-18}	0.25×10^{-23}
6	0.64×10^{-4}	0.64×10^{-9}	0.64×10^{-14}	0.64×10^{-19}	0.64×10^{-24}
7	0.58×10^{-4}	0.58×10^{-9}	0.58×10^{-14}	0.58×10^{-19}	0.58×10^{-24}

[a]) $\rho(X_1) > 1$ означает здесь, что интервал X_i содержит 0.

Следует иметь в виду, что в ней дана лишь верхняя граница $\rho(X_i)$ относительной погрешности для каждой компоненты вектора результата.

Рассмотрим следующую задачу. Пусть имеются машинные интервалы (т. е. вещественные интервалы, границами которых служат машинные числа)

$$C_0,\ A_0,\ B_0,\ D_0,\ A_1,\ B_1,\ D_1,\ \ldots,\ A_{n-1},\ B_{n-1},\ D_{n-1},$$

а также машинное число a_n. Требуется вычислить выражение

$$R_n = (1/a_n)\{C_0 - A_0(B_0 - D_0) - A_1(B_1 - D_1) - \ldots - A_{n-1}(B_{n-1} - D_{n-1})\}.$$

Теоретически можно воспользоваться таким алгоритмом:

$$\begin{aligned} &S_0 := C_0, \\ (S)\quad &S_i := S_{i-1} - A_{i-1}(B_{i-1} - D_{i-1}), \quad 1 \leqslant i \leqslant n, \\ &R_n := S_n/a_n. \end{aligned}$$

На практике, однако, этот алгоритм выполняется в виде

$$\begin{aligned} &\bar{S}_0 := S_0 := C_0, \\ (\bar{S})\quad &\bar{S}_i := fl(\bar{S}_{i-1} - fl(A_{i-1} fl(B_{i-1} - D_{i-1}))), \quad 1 \leqslant i \leqslant n, \\ &\bar{R}_n := fl(\bar{S}_n/a_n). \end{aligned}$$

Начав с (15а), установим $\mathrm{eps} := \frac{1}{2} b^{1-t}$ и для произвольных интервалов A и B получим, что

$$fl(A * B) \subseteq A * B + [-\varepsilon, \varepsilon] A * B, \tag{22}$$

где $\max\{|\varepsilon_1|, |\varepsilon_2|\} \leqslant \varepsilon, \ \varepsilon = 2\ \mathrm{eps}.$

Предположим на время, что

$$\bar{S}_0 = S_0 = C_0, \quad \bar{S}_1, \ldots, \bar{S}_{n-1}$$

уже вычислено. Тогда из (22) следует

$$\begin{aligned}
&fl(B_{n-1} - D_{n-1}) \subseteq B_{n-1} - D_{n-1} + |B_{n-1} - D_{n-1}|[-\varepsilon, \varepsilon],\\
&fl(A_{n-1} fl(B_{n-1} - D_{n-1}))\\
&\subseteq A_{n-1}(B_{n-1} - D_{n-1} + |B_{n-1} - D_{n-1}|[-\varepsilon, \varepsilon])\\
&\quad + |A_{n-1}(B_{n-1} - D_{n-1} + |B_{n-1} - D_{n-1}|[-\varepsilon, \varepsilon])|[-\varepsilon, \varepsilon]\\
&\subseteq A_{n-1}(B_{n-1} - D_{n-1}) + |A_{n-1}||B_{n-1} - D_{n-1}|[-2\varepsilon - \varepsilon^2, 2\varepsilon + \varepsilon^2],
\end{aligned}$$

а значит,

$$\begin{aligned}
\bar{S}_n &\subseteq \bar{S}_{n-1} - A_{n-1}(B_{n-1} - D_{n-1})\\
&\quad - |A_{n-1}||B_{n-1} - D_{n-1}|[-2\varepsilon - \varepsilon^2, 2\varepsilon + \varepsilon^2]\\
&\quad + |\bar{S}_{n-1} - A_{n-1}(B_{n-1} - D_{n-1})\\
&\quad - |A_{n-1}||B_{n-1} - D_{n-1}|[-2\varepsilon - \varepsilon^2, 2\varepsilon + \varepsilon^2]|[-\varepsilon, \varepsilon]\\
&\subseteq \bar{S}_{n-1} - A_{n-1}(B_{n-1} - D_{n-1}) + |\bar{S}_{n-1}|[-\varepsilon, \varepsilon]\\
&\quad + |A_{n-1}||B_{n-1} - D_{n-1}|[-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3].
\end{aligned} \tag{23}$$

При помощи математической индукции покажем, что

$$\begin{aligned}
\bar{S}_n \subseteq S_n &+ [-\varepsilon, \varepsilon] \sum_{i=0}^{n-1} |\bar{S}_i| + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]\\
&\times \sum_{i=0}^{n-1} |A_i||B_i - D_i|.
\end{aligned} \tag{24}$$

Для n=1 из (23) с учетом того, что $\bar{S}_0 = S_0 = C_0$, получаем

$$\begin{aligned}
\bar{S}_1 &\subseteq \bar{S}_0 - A_0(B_0 - D_0) + |\bar{S}_0|[-\varepsilon, \varepsilon]\\
&\quad + |A_0||B_0 - D_0|[-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]\\
&= S_1 + [-\varepsilon, \varepsilon]|\bar{S}_0| + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]|A_0||B_0 - D_0|,
\end{aligned}$$

откуда видна справедливость нашего предположения при n= 1. Если для некоторого $n \geq 1$ выполнено (24), тогда замена n на n+ 1 в (23), а также использование (5) дают

$$\begin{aligned}
\bar{S}_{n+1} &\subseteq \bar{S}_n - A_n(B_n - D_n) + [-\varepsilon, \varepsilon]|\bar{S}_n|\\
&\quad + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]|A_n||B_n - D_n|\\
&\subseteq S_{n+1} + [-\varepsilon, \varepsilon] \sum_{i=0}^{n} |\bar{S}_i| + [-3\varepsilon - 3\varepsilon^2 - \varepsilon^3, 3\varepsilon + 3\varepsilon^2 + \varepsilon^3]\\
&\quad \times \sum_{i=0}^{n} |A_i||B_i - D_i|.
\end{aligned}$$

Последнее выражение тождественно (24), у которого n заменено на n+1. Повторное применение (22) приводит к окончательному результату

$$\bar{R}_n \subseteq \bar{S}_n/a_n + (|\bar{S}_n|/|a_n|)[-\varepsilon, \varepsilon]. \qquad (25)$$

Неравенства (24) и (25) будут использованы дальше.

Замечания. Понятие округления в том виде, как оно встречается в этом микромодуле, подробно рассмотрено Миранкером и Кулишем. Неравенство (15b), явившееся исходным пунктом при обсуждении влияния погрешносгей округления, можно найти у Валлиша и Грюцманна. Оценка (24) была доказана Алефельдом и Рокном; мы еще вернемся к ней дальше. Теорема 5 была доказана Муром.

1.5. Комплексная интервальная арифметика

Теперь определим так называемую комплексную интервальную арифметику. Будет показано, что многие из свойств и результатов, полученных для вещественной интервальной арифметики, можно перенести на случай комплексной. Чтобы это проделать, определим множества комплексных чисел, которые будут использоваться в качестве комплексных интервалов. Имеются два предпочтительных подхода, к рассмотрению которых и перейдем.

А. Прямоугольники в качестве комплексных интервалов

Определение 1. Пусть A_1 и A_2 — произвольные элементы из I(R). Тогда множество комплексных чисел

$$A = \{a = a_1 + ia_2 \mid a_1 \in A_1,\ a_2 \in A_2\} \qquad (i = \sqrt{-1})$$

называется комплексным интервалом.

Определенные таким образом множества комплексных чисел могут быть изображены на комплексной плоскости в виде прямоугольников со сторонами, параллельными осям координат. Множество всех таких комплексных интервалов обозначим через R(C), а прописные буквы *A, B, C, ..., X, Y, Z* будем использовать для обозначения его элементов. Всякое A из R(C) можно записать в виде

$$A = A_1 + iA_2, \text{где } A_1,\ A_2 \in I(\mathbb{R}).$$

Комплексное число $a = a_1$-$+ia_2$ можно рассматривать как точечный комплексный интервал:

$$A = [a_1, a_1] + i[a_2, a_2] \in R(\mathbf{C}),$$

а каждый элемент A_1 из $I(\mathrm{R})$ — как сумму $A=A_1+i[0,\ 0]\in \mathrm{R(C)}$, откуда видно, что $I(\mathbb{R})\subset R(\mathbb{C})$.

Определение 2. Пусть $A=A_1+iA_2$ и $B=B_1+iB_2$ —два элемента из R(C). Тогда A и B считаются равными (запись: $A=B$), если

$$A_1=B_1 \quad \text{и} \quad A_2=B_2$$

см. также определение 1. п.1.1).

Определенное здесь отношение равенства рефлективно, симметрично и транзитивно.

Теперь мы обобщим арифметику комплексных чисел на случай R(C).

Определение 3. Пусть * из $\{+,\ -,\ \cdot,\ :\}$ — бинарная операция над элементами из $I(\mathrm{R})$ (как в определении 2 п.1.1). Тогда если

$$A=A_1+iA_2, \qquad B=B_1+iB_2\in R(\mathbb{C}),$$

то мы полагаем

$$A\pm B=A_1\pm B_1+i(A_2\pm B_2),$$
$$A\cdot B=A_1B_1-A_2B_2+i(A_1B_2+A_2B_1),$$
$$A:B=(A_1B_1+A_2B_2):(B_1^2+B_2^2)+i(A_2B_1-A_1B_2):(B_1^2+B_2^2).$$

Считается, что в случае деления $0\notin B_1^2+B_2^2$. При вычислении степеней $B_1^2=B_1B_1$ и $B_2^2=B_2B_2$ это требование может оказаться невыполненными, даже если $0\notin B_1+iB_2$ в соответствии с определением 2 п.1.1. Если при этом оказывается, что $0\in B_1^2+B_2^2$, то деление не определено.

Чтобы проиллюстрировать сказанное, рассмотрим следующий пример.

Пример. Пусть

$$B=[-1,\ 1]+i[1,\ 3].$$

Тогда

$$0\in[0,\ 10]=[-1,\ 1]+[1,\ 9]=B_1B_1+B_2B_2.$$

Поэтому мы оговариваем, что в определении 3, если производится деление двух элементов из R(C), выражение $B_1^2+B_2^2$ следует вычислять по правилу

$$B_1^2+B_2^2=\{b_1^2\,|\,b_1\in B_1\}+\{b_2^2\,|\,b_2\in B_2\}$$

(см. также определение 3 п.1.1).

Тогда в приведенном выше примере

$$B_1^2+B_2^2=[0,\ 1]+[1,\ 9]=[1,\ 10].$$

Теперь рассмотрим более внимательно свойства введенной выше комплексной интервальной арифметики.

Сразу же видно, что если $A,\ B \in R(\mathbb{C})$, то равенство

$$A \pm B = \{a \pm b \mid a \in A,\ b \in B\}$$

справедливо для сложения (соответственно вычитания) на множестве R(C). Аналогичное равенство для умножения и деления, вообще говоря, не выполняется. Это можно увидеть из следующего простого примера.

Пример. Пусть

$$A = [2,\ 4] + i[0,\ 0], \qquad B = [1,\ 1] + i[1,\ 1].$$

Из определения 3 получаем

$$AB = [2,\ 4] + i[2,\ 4].$$

С другой стороны,

$$\{ab \mid a \in A,\ b \in B\} = \{s(1+i) \mid s \in \mathbb{R},\ 2 \leqslant s \leqslant 4\} \subset AB.$$

Справедлива, однако, следующая теорема.

Теорема 4. *Операции, введенные определением* 3, *удовлетворяют соотношению*

$$\{a * b \mid a \in A,\ b \in B\} \subseteq A * B.$$

Для сложения и вычитания включение может быть заменено на равенство. Для умножения

$$AB = \inf \{X \in R(\mathbb{C}) \mid \{ab \mid a \in A,\ b \in B\} \subseteq X\},$$

где точная нижняя грань берется в смысле частичного порядка на R(C), *определяемого теоретико-множественным включением.*

Доказательство. Случай сложения и вычитания был рассмотрен ранее. Пусть теперь $a \in A$ и $b \in B$. Используя монотонность включения вещественных интервалов, для $a = a_1 + ia_2$ и $b = b_1 + ib_2$ имеем

$$\begin{aligned} ab &= a_1b_1 - a_2b_2 + i(a_1b_2 + a_2b_1) \\ &\in A_1B_1 - A_2B_2 + i(A_1B_2 + A_2B_1) = AB. \end{aligned}$$

Поскольку каждая переменная входит в выражение $a_1b_1 - a_2b_2$ лишь один раз, то получаем, что

$$\{a_1b_1 - a_2b_2 \mid a_k \in A_k,\ b_k \in B_k,\ k = 1,\ 2\} = A_1B_1 - A_2B_2.$$

По той же причине

$$\{a_1b_2 + a_2b_1 \mid a_k \in A_k,\ b_k \in B_k,\ k = 1,\ 2\} = A_1B_2 + A_2B_1.$$

Последние два соотношения показывают, что для каждого вещественного числа c_1 такого, что

$$c_1 = a_1b_1 - a_2b_2 \in A_1B_1 - A_2B_2,$$

где $a_k \in A_k$, $b_k \in B_k$, $k = 1, 2$, можно найти другое вещественное число c_2 такое, что

$c_2 = a_2b_1 + a_1b_2 \in A_2B_1 + A_1B_2$, где $a_k \in A_k$, $b_k \in B_k$, $k = 1, 2$ и $c_1 + ic_2 \in AB$. Это и требовалось доказать.

Соотношение

$$\{a : b \mid a \in A,\ b \in B\} \subseteq A : B$$

также следует из монотонности включения.

Утверждение теоремы 4, касающееся умножения, не допускает, вообще говоря, распространения на случай деления. Тем не менее можно получить «уточнение» (в смысле сужения объемлющего интервала), если определить, что

$$A : B = A \cdot \frac{1}{B},$$

а затем вычислять $1/B$ по формуле

$$1/B = \inf \{X \in R(\mathbb{C}) \mid \{1/b \mid b \in B\} \subseteq X\}.$$

Данная возможность была предложена Рокном и Ланкастером в виде набора формул, требующих значительной вычислительной работы.

В. Круги в качестве комплексных интервалов

Определение 5. Пусть a из С — комплексное число, и пусть $r \geq 0$. Мы называем множество

$$Z = \{z \in \mathbb{C} \mid |z - a| \leqslant r\}$$

кругом или круговым интервалом (или просто комплексным интервалом, когда не опасаемся спутать его с прямоугольным интервалом).

Множество всех кругов обозначим через К(С), а прописные буквы *A, B, C, …, X, Y,* Z используем для обозначения его элементов. Круг Z с центром a и радиусом r будем записывать в виде

$$Z = \langle a, r \rangle.$$

Комплексные числа можно рассматривать как специальные элементы из К(С), имеющие вид $\langle a, 0 \rangle$. Ясно, что $\mathrm{C} \subset \mathrm{K}(\mathrm{C})$.

Определение 6. Два круга $A = \langle a, r_1 \rangle$ и $B = \langle b, r_2 \rangle$ называются равными (обозначение: *A = B),* если они равны в теоретико-множесгвенном смысле. В этом случае $a = b$ и $r_1 = r_2$.

Это отношение равенства также рефлексивно, симметрично и транзитивно.

Операции на К(С) вводятся как обобщения операций над вещественными числами следующим образом.

Определение 7. Пусть * из $\{+, -, \cdot, :\}$ — бинарная операция над комплексными числами. Тогда если $A = \langle a, r_1\rangle$ и $B = \langle b, r_2\rangle$, то

$$A \pm B = \langle a \pm b,\ r_1 + r_2\rangle,$$

$$A \cdot B = \langle ab,\ |a| r_2 + |b| r_1 + r_1 r_2\rangle,$$

$$\frac{1}{B} = \left\langle \frac{\bar{b}}{b\bar{b} - r_2^2},\ \frac{r_2}{b\bar{b} - r_2^2}\right\rangle \quad \text{при условии, что} \quad 0 \notin B,$$

$$A : B = A \cdot \frac{1}{B} \quad \text{при условии, что} \quad 0 \notin B.$$

Здесь $|a| = \sqrt{a_1^2 + a_2^2}$ обозначает евклидову норму комплексного числа $a = a + ia_2$, а $\bar{b} = b_1 + ib_2$ — сопряженное с $b = b_1 + ib_2$.

Очевидно, что для сложения и вычитания кругов выполнено равенство

$$A \pm B = \{a \pm b \mid a \in A,\ b \in B\}.$$

То же справедливо для операции обращения круга: если мы применим теорию конформных отображений к отображению вида $w = 1/z$ для не содержащего нуля круга, то получим другой круг. Иными словами,

$$1/B = \{1/b \mid b \in B\}.$$

Элементарными преобразованиями легко проверить формулы определения 7 для центра и радиуса области $\{1/b \mid b \in B\}$.

Для умножения (а следовательно, и для деления) двух элементов из К (С) по правилам определения 7 верно, вообще говоря, только то, что

$$\{z_1 z_2 \mid z_1 \in A,\ z_2 \in B\} \subseteq AB.$$

Это вытекает из следующих неравенств:

$$\begin{aligned} |z_1 z_2 - ab| &= |a(z_2 - b) + b(z_1 - a) + (z_1 - a)(z_2 - b)| \\ &\leqslant |a||z_2 - b| + |b||z_1 - a| + |z_1 - a||z_2 - b| \\ &\leqslant |a| r_2 + |b| r_1 + r_1 r_2. \end{aligned}$$

Аналогично теореме 4 п.1.1, соберем теперь вместе наиболее важные свойства операций на *R*(С) и *K* (С). Если не оговорено противное, то *I*(С) можно понимать и как обозначение множества *R*(С) с операциями из определения 3, и как обозначение множества *K* (С) с операциями из определения 7.

Теорема 8. *Пусть* $A,\ B,\ C \in I(\mathbb{C})$. *Тогда*

$$A + B = B + A, \quad AB = BA \quad (\textit{коммутативность}); \tag{1}$$

$$(A + B) + C = A + (B + C), \tag{2}$$

$(AB)C = A(BC)$ *для* $A,\ B,\ C$ *из* $K(\mathbb{C})$ (*ассоциативность*);

$$[0,\ 0] + i[0,\ 0] \in R(\mathbb{C}) \ (\textit{соответственно}\ \langle 0,\ 0\rangle \in K(\mathbb{C})) \ \textit{и} \tag{3}$$

$$[1,\ 1] + i[0,\ 0] \in R(\mathbb{C}) \ (\textit{соответственно}\ \langle 1,\ 0\rangle \in K(\mathbb{C}))$$

— *определенные единственным образом нейтральные элементы сложения (нуль) и умножения (единица);*

$$I(\mathbb{C})\ \textit{не имеет делителей нуля}; \tag{4}$$

элемент Z множества $I(\mathbb{C})$ *имеет противоположный и обратный элементы, если и только если* $Z \in C$ u *в случае мультипликативных операций* $Z \neq 0$. (5)

Однако $0 \in A - A$ *и* $1 \in A : A$;

$$A(B + C) \subseteq AB + AC \quad (\textit{субдистрибутивность}), \tag{6}$$

$$a(B + C) = aB + aC \quad \textit{для} \quad a \in \mathbb{C}.$$

Доказательство. Доказательство этих утверждений следует из определений операций 3 и 7. В качестве примера докажем (6) для K(C). Если $A = \langle a,\ r_1\rangle$, $B = \langle b,\ r_2\rangle$, $C = \langle c,\ r_3\rangle \in K(\mathbb{C})$, то

$$\begin{aligned} A(B + C) &= \langle a,\ r_1\rangle\langle b + c,\ r_2 + r_3\rangle \\ &= \langle a(b + c),\ |a|(r_2 + r_3) + |b + c|r_1 + r_1(r_2 + r_3)\rangle \\ &\subseteq \langle ab + ac,\ |a|r_2 + |a|r_3 + |b|r_1 + |c|r_1 + r_1r_2 + r_1r_3\rangle \\ &= \langle ab,\ |a|r_2 + |b|r_1 + r_1r_2\rangle + \langle ac,\ |a|r_3 + |c|r_1 + r_1r_3\rangle \\ &= AB + AC. \end{aligned}$$

В случае $A = \langle a,\ 0\rangle$, т. е. равенства нулю r_1, доказательство приводит к соотношению

$$a(B + C) = aB + aC.$$

Необходимо особо отметить, что ассоциативный закон (2) в общем случае не выполняется при перемножении элементов R(C). Это можно увидеть из следующего примера. Пример.

$$A = [2,\ 4] + i[0,\ 0], \quad B = [1,\ 1] + i[1,\ 1], \quad C = [1,\ 1] + i[1,\ 1],$$

$$(AB)C = ([2,\ 4] + i[2,\ 4])([1,\ 1] + i[1,\ 1]) = [-2,\ 2] + i[4,\ 8],$$

$$A(BC) = ([2,\ 4] + i[0,\ 0])([0,\ 0] + i[2,\ 2]) = [0,\ 0] + i[4,\ 8].$$

$I(\mathbb{C})$ обладает также монотонностью включения.

Теорема 9. *Пусть* $A^{(k)},\ B^{(k)} \in I(\mathbb{C}),\ k = 1,\ 2,$ *таковы, что*

$$A^{(k)} \subseteq B^{(k)}, \quad k = 1, 2.$$

Тогда соотношение

$$A^{(1)} * A^{(2)} \subseteq B^{(1)} * B^{(2)}$$

*выполняется для операций * из* $\{+, -, \cdot, :\}$.

Доказательство. Это верно для $R(C)$, поскольку монотонность включения выполняется для элементов $I(R)$ (см. теорему 5 п.1.1).
В случае сложения и вычитания элементов $K(C)$ имеем

$$A^{(1)} \pm A^{(2)} = \{z = x \pm y \mid x \in A^{(1)}, \ y \in A^{(2)}\}$$
$$\subseteq \{w = u \pm v \mid u \in B^{(1)}, \ v \in B^{(2)}\} = B^{(1)} \pm B^{(2)}.$$

Рассмотрим теперь умножение на множестве $K(C)$. Пусть

$$A^{(k)} = \langle a^{(k)}, r^{(k)} \rangle, \quad B^{(k)} = \langle b^{(k)}, s^{(k)} \rangle, \quad k = 1, 2.$$

Предположение $A^{(k)} \subseteq B^{(k)}$, $k = 1, 2$, эквивалентно тому, что

$$|a^{(k)} - b^{(k)}| \leqslant s^{(k)} - r^{(k)}, \quad k = 1, 2.$$

Далее,

$$A^{(1)}A^{(2)} = \langle a^{(1)}a^{(2)}, \ |a^{(1)}|r^{(2)} + |a^{(2)}|r^{(1)} + r^{(1)}r^{(2)} \rangle,$$
$$B^{(1)}B^{(2)} = \langle b^{(1)}b^{(2)}, \ |b^{(1)}|s^{(2)} + |b^{(2)}|s^{(1)} + s^{(1)}s^{(2)} \rangle.$$

Требуется доказать, что

$$|a^{(1)}a^{(2)} - b^{(1)}b^{(2)}| \leqslant |b^{(1)}|s^{(2)} + |b^{(2)}|s^{(1)} + s^{(1)}s^{(2)}$$
$$- (|a^{(1)}|r^{(2)} + |a^{(2)}|r_1 + r^{(1)}r^{(2)}).$$

Из неравенства треугольника получаем

$$-|b^{(2)}| \leqslant -|a^{(2)}| + |a^{(2)} - b^{(2)}|,$$
$$-|b^{(1)}| \leqslant -|a^{(1)}| + |a^{(1)} - b^{(1)}|,$$

а поскольку

$$|a^{(k)} - b^{(k)}| \leqslant s^{(k)} - r^{(k)}, \quad k = 1, 2,$$

имеем

$$-|b^{(2)}|r^{(1)} \leqslant -|a^{(2)}|r^{(1)} + r^{(1)}(s^{(2)} - r^{(2)})$$
$$= -|a^{(2)}|r^{(1)} + r^{(1)}s^{(2)} - r^{(1)}r^{(2)},$$
$$-|b^{(1)}|r^{(2)} \leqslant -|a^{(1)}|r^{(2)} + r^{(2)}(s^{(1)} - r^{(1)})$$
$$= -|a^{(1)}|r^{(2)} + r^{(2)}s^{(1)} - r^{(1)}r^{(2)}.$$

Отсюда

$$\begin{aligned}|a^{(1)}a^{(2)} - b^{(1)}b^{(2)}| &\leqslant |b^{(2)}||a^{(1)} - b^{(1)}| + |b^{(1)}||a^{(2)} - b^{(2)}| \\ &\quad + |a^{(1)} - b^{(1)}||a^{(2)} - b^{(2)}| \\ &\leqslant |b^{(2)}|(s^{(1)} - r^{(1)}) + |b^{(1)}|(s^{(2)} - r^{(2)}) \\ &\quad + (s^{(1)} - r^{(1)})(s^{(2)} - r^{(2)}) \\ &\leqslant |b^{(2)}|s^{(1)} + |b^{(1)}|s^{(2)} + s^{(1)}s^{(2)} \\ &\quad - (|a^{(2)}|r^{(1)} + |a^{(1)}|r^{(2)} + r^{(1)}r^{(2)}),\end{aligned}$$

что доказывает теорему для случая умножения. Из того, что

$$1/A^{(2)} = \{z = 1/x \mid x \in A^{(2)}\} \subseteq \{w = 1/u \mid u \in B^{(2)}\} = 1/B^{(2)},$$

следует, что

$$A^{(1)} : A^{(2)} = A^{(1)} \cdot \frac{1}{A^{(2)}} \subseteq B^{(1)} \cdot \frac{1}{B^{(2)}} = B^{(1)} : B^{(2)}.$$

Теорема доказана.

Частным случаем теоремы 9 является

Следствие 10. *Пусть* $A,\ B \in I(\mathbb{C})$ *и* $a \in A,\ b \in B$. *Тогда*

$$a * b \in A * B,$$

где $* \in \{+,\ -,\ \cdot,\ :\}$.

Замечания. Круги в качестве комплексных интервалов впервые ввели в систематическое употребление Гаргантиии и Энричи. Рассмотренная в этом разделе арифметика кругов была предложена ими и была использована для одновременной локализации корней многочлена. Другие приложения арифметики кругов будут изложены дальше. Важное свойство монотонности включения в том виде, как оно сформулировано в теореме 9, впервые доказано Г. Алефельдом и Ю. Херцбергом. Как уже было отмечено, решена проблема такого определения умножения кругов, которое приводит к получению меньших множеств. Арифметика на *R*(C), сводящаяся при выполнении некоторых условий к вещественной, предложена Алефельдом. Свойства этой арифметики исследовали также Бош, Рокн и Ланкастер.

Как уже неоднократно указывалось, в основе почти всех приложений интервальной арифметики лежит монотонность включения (теорема 9). Умножение, предложенное Криером, не обладает этим свойством, что было показано там же на примере.

Приближенная реализация арифметики прямоугольников на цифровой вычислительной машине не вызывает проблем, поскольку операции на *R*(C) сводятся к операциям на *I*(R). Ранее было показано, как сделать возможной реализацию приближенных операций над элементами *I*(R) на ЦВМ без потери наиболее важных свойств арифметики, что позволяет проделать то же самое для *R*(C).

1.6. Метрика, абсолютная величина и ширина в $I(\mathbf{C})$

В этом разделе q будет обозначать метрику на $I(\mathrm{R})$, задаваемую определением 1 п.1.2. Введем теперь метрику на $R(\mathrm{C})$.

Определение 1. Пусть $A = A_1 + iA_2$ и $B = B_1 + iB_2$ принадлежат $R(C)$. Тогда расстояние p между A и B определяется формулой

$$p(A, B) = q(A_1, B_1) + q(A_2, B_2).$$

Если p используется в пространстве $I(\mathrm{R})$, то оно принимает те же самые значения, что и q из определения 1 п.1.2. Поэтому в дальнейшем расстояние на $R(\mathrm{C})$ будет обозначаться через q, так что

$$q(A, B) = q(A_1, B_1) + q(A_2, B_2).$$

Поскольку q является метрикой на $I(\mathrm{R})$, легко доказать, что q останется ею и при переходе к $R(\mathrm{C})$. Введение на $R(\mathrm{C})$ метрики q делает его топологическим пространством. Если теперь обычным для метрических пространств способом определить сходимость, то окажется, что последовательность $\{A^{(k)}\}_{k=0}^{\infty}$, где $A^{(k)} = A_1^{(k)} + iA_2^{(k)} \in R(\mathbb{C})$, сходится к $A = A_1 + iA_2$ — элементу $R(\mathrm{C})$ тогда и только тогда, когда

$$\lim_{k\to\infty} A_1^{(k)} = A_1 \quad \text{и} \quad \lim_{k\to\infty} A_2^{(k)} = A_2. \tag{1}$$

Из факта замкнутости метрического пространства $(I(\mathbb{R}), q)$ вытекает, что $R(\mathrm{C})$ с заданной на нем метрикой q также образует замкнутое метрическое пространство.

Определение 2. Пусть $A = A_1 + iA_2 \in R(\mathbb{C})$. Тогда

$$|A| = q(A, 0) = |A_1| + |A_2| = q(A_1, 0) + q(A_2, 0)$$

называется абсолютной величиной A.

Если, в частности, $A = [a_1, a_1] + i[a_2, a_2] = a_1 + ia_2 = a$, то

$$|A| = |a| = |a_1| + |a_2|. \tag{2}$$

Итак, абсолютная величина из определения 2 не совпадает с евклидовой нормой комплексного числа. В дальнейшем из контексга всегда будет ясно, какая из них используется в конкретном случае. Кроме того, заметим, что из (2) следует справедливость соотношения

$$|A| = \max_{a \in A} |a|.$$

Пусть d в соответствии с определением 8 п.1.2 служит обозначением ширины вещественного интервала. Тогда имеем

Определение 3. Если $A = A_1 + iA_2 \in R(\mathbb{C})$, то шириной A будем называть

$$d(A) = d(A_1) + d(A_2).$$

Теперь перейдем к множеству $K(\mathbb{C})$.

Определение 4. Пусть

$$A = \langle a, r_1 \rangle, \; B = \langle b, r_2 \rangle \in K(\mathbb{C}).$$

Тогда назовем

(a) $q(A, B) = |a - b| + |r_1 - r_2|$ расстоянием между A и B,
(b) $|A| = |a| + r_1$ абсолютной величиной A и
(c) $d(A) = 2r_1$ шириной A.

Для того чтобы определить расстояние между двумя круговыми интервалами на комплексной плоскости, в определении 4 используется евклидова метрика. Когда абсолютная величина кругового интервала применяется к обычным комплексным числам, она совпадает с евклидовой нормой. Обратим внимание, что и на этот раз выполняется соотношение

$$|A| = \max_{a \in A} |a|.$$

Если сходимость последовательности, состоящей из элементов $K(\mathbb{C})$, определена, как обычно, с помощью соответствующей метрики, то можно легко убедиться в замкнутости метрического пространства $(K(\mathbb{C}), q)$. Опираясь на такого рода определение, получаем, что

$$\lim_{k \to \infty} A^{(k)} = A \Leftrightarrow \lim_{k \to \infty} a^{(k)} = a, \quad \lim_{k \to \infty} r^{(k)} = r, \tag{3}$$

где $\{A^{(k)}\}_{k=0}^{\infty} = \{\langle a^{(k)}, r^{(k)} \rangle\}_{k=0}^{\infty}$, $A = \langle a, r \rangle$.

В следующей теореме собраны наиболее важные свойства метрики, абсолютной величины и ширины на множествах $R(\mathbb{C})$ и $K(\mathbb{C})$.

Теорема 5. *Пусть имеются A, B, C, D из* $I(\mathbb{R})$. *Тогда*

$$q(A + B, A + C) = q(B, C), \tag{4}$$

$$q(A + B, C + D) \leqslant q(A, C) + q(B, D), \tag{5}$$

$$q(aB, aC) \leqslant |a| q(B, C), \quad a \in \mathbb{C}. \tag{6}$$

Если B и C принадлежат $K(\mathbb{C})$, *то в* (6) *всегда имеет место равенство.*

$$q(AB, AC) \leqslant |A| q(B, C), \tag{7}$$

$$|A| \geqslant 0, \quad |A| = 0 \Leftrightarrow A = 0, \tag{8}$$

$$|A + B| \leqslant |A| + |B|, \tag{9}$$

$$|aB| \leqslant |a| |B|, \quad a \in \mathbb{C}. \tag{10}$$

Если $B \in K(\mathbb{C})$, *то* (10) *превращается в равенство.*

$$|AB| \leqslant |A||B|, \tag{11}$$

$$d(aB) = |a| d(B), \quad a \in \mathbb{C}, \tag{12}$$

$$d(AB) \leqslant |A| d(B) + |B| d(A), \tag{13}$$

$$d(A) = |A - A|, \tag{14}$$

$$d(AB) \geqslant |A| d(B), \tag{15}$$

$$d(A \pm B) = d(A) + d(B), \tag{16}$$

$$A \subseteq B \Rightarrow \frac{1}{2}(d(B) - d(A)) \leqslant q(A, B) \leqslant d(B) - d(A). \tag{17}$$

Доказательство. Сначала докажем перечисленные свойства из $R(\mathbb{C})$. Справедливость (4) — (7) следует непосредственно из соответствующих свойств (4 п.1.2) — (7 п.1.2), сформулированных в теореме 7 п.1.2 применительно к вещественным интервалам.

Пусть

$$A = A_1 + iA_2, \quad B = B_1 + iB_2,$$
$$C = C_1 + iC_2, \quad D = D_1 + iD_2 \in R(\mathbb{C}).$$

(4): $q(A + B, A + C) = q(A_1 + B_1 + i(A_2 + B_2), A_1 + C_1 + i(A_2 + C_2))$
$= q(A_1 + B_1, A_1 + C_1) + q(A_2 + B_2, A_2 + C_2)$
$= q(B_1, C_1) + q(B_2, C_2) = q(B, C).$

(5): $q(A + B, C + D) = q(A_1 + B_1, C_1 + D_1) + q(A_2 + B_2, C_2 + D_2)$
$\leqslant q(A_1, C_1) + q(B_1, D_1) + q(A_2, C_2) + q(B_2, D_2)$
$= q(A, C) + q(B, D).$

(6), (7):

$q(AB, AC) = q(A_1B_1 - A_2B_2, A_1C_1 - A_2C_2)$
$+ q(A_1B_2 + A_2B_1, A_1C_2 + A_2C_1)$
$\leqslant |A_1| q(B_1, C_1) + |A_2| q(B_2, C_2) + |A_1| q(B_2, C_2) + |A_2| q(B_1, C_1)$
$= (|A_1| + |A_2|) q(B, C) = |A| q(B, C).$

Свойства (8)—(11) доказываются на основании определения $|A|$.

(8): $|A| = q(A, 0) = q(A_1, 0) + q(A_2, 0) = |A_1| + |A_2| \geqslant 0,$
$|A| = 0 \Leftrightarrow |A_1| = |A_2| = 0 \Leftrightarrow A = 0.$

(9). $|A + B| = q(A + B, 0) \leqslant q(A, 0) + q(B, 0) = |A| + |B|$
(в соответствии с 5)).

(10), (11):

$$|AB| = q(AB, 0) = q(AB, A \cdot 0) \leqslant |A| q(B, 0) = |A||B|$$

(из 6) и (7)).

(12): Пусть $a = a_1 + ia_2 \in \mathbb{C}$. Из определения 5.3 п. 1.4 следует

$$aB = a_1B_1 - a_2B_2 + i(a_1B_2 + a_2B_1),$$

и с помощью (2) получаем

$$\begin{aligned} d(aB) &= d(a_1B_1 - a_2B_2) + d(a_1B_2 + a_2B_1) \\ &= d(a_1B_1) + d(a_2B_2) + d(a_1B_2) + d(a_2B_1) \\ &= |a_1|d(B_1) + |a_2|d(B_2) + |a_1|d(B_2) + |a_2|d(B_1) \\ &= (|a_1| + |a_2|)(d(B_1) + d(B_2)) = |a|d(B). \end{aligned}$$

(13): $$\begin{aligned} d(AB) &= d(A_1B_1 - A_2B_2) + d(A_1B_2 + A_2B_1) \\ &= d(A_1B_1) + d(A_2B_2) + d(A_1B_2) + d(A_2B_1) \\ &\leqslant |A_1|d(B_1) + |B_1|d(A_1) + |A_2|d(B_2) + |B_2|d(A_2) \\ &\quad + |A_1|d(B_2) + |B_2|d(A_1) + |A_2|d(B_1) + |B_1|d(A_2) \\ &= (|A_1| + |A_2|)(d(B_1) + d(B_2)) \\ &\quad + (|B_1| + |B_2|)(d(A_1) + d(A_2)) \\ &= |A|d(B) + |B|d(A). \end{aligned}$$

(14): $$d(A) = d(A_1) + d(A_2) = |A_1 - A_1| + |A_2 - A_2| = |A - A|.$$

(15): $$\begin{aligned} d(AB) &= d(A_1B_1 - A_2B_2) + d(A_1B_2 + A_2B_1) \\ &\geqslant |A_1|d(B_1) + |A_2|d(B_2) + |A_1|d(B_2) + |A_2|d(B_1) \\ &= (|A_1| + |A_2|)(d(B_1) + d(B_2)) = |A|d(B). \end{aligned}$$

(16): $$\begin{aligned} d(A \pm B) &= d(A_1 \pm B_1) + d(A_2 \pm B_2) \\ &= d(A_1) + d(A_2) + d(B_1) + d(B_2) \\ &= d(A) + d(B). \end{aligned}$$

(17): Это свойство является прямым следствием (21 п.1.2.). Теперь пусть

$$A=\langle a,\ r_1\rangle,\quad B=\langle b,\ r_2\rangle,$$
$$C=\langle c,\ r_3\rangle,\quad D=\langle d,\ r_4\rangle\in K(\mathbb{C}).$$

(4): $q(A+B,\ A+C)=|a+b-(a+c)|+|r_1+r_2-(r_1+r_3)|$
$=|b-c|+|r_2-r_3|=q(B,\ C).$

(5): $q(A+B,\ C+D)=|a+b-(c+d)|+|r_1+r_2-(r_3+r_4)|$
$\leqslant|a-c|+|r_1-r_3|+|b-d|+|r_2-r_4|$
$=q(A,\ C)+q(B,\ D).$

(6): $q(aB,\ aC)=|ab-ac|+||a|r_2-|a|r_3|$
$=|a|\{|b-c|+|r_2-r_3|\}=|a|q(B,\ C).$

(7): $q(AB,\ AC)=|ab-ac|+||a|r_2+|b|r_1+r_1r_2$
$-(|a|r_3+|c|r_1+r_1r_3)|$
$\leqslant|a||b-c|+|a||r_2-r_3|+r_1||b|-|c||+r_1|r_2-r_3|$
$\leqslant(|a|+r_1)(|b-c|+|r_2-r_3|)=|A|q(B,\ C).$

(8): $|A|=|a|+r_1\geqslant 0,\quad |A|=0\Leftrightarrow(a=0,\ r_1=0).$

(9): $|A+B|=|a+b|+|r_1+r_2|\leqslant|a|+r_1+|b|+r_2=|A|+|B|.$

(10): $|aB|=|ab|+|a|r_2=|a||B|.$

(11): $|AB|=q(AB,\ 0)=q(AB,\ A\cdot 0)\leqslant|A|q(B,\ 0)=|A||B|$
(из (7)).

(12): $d(aB)=2|a|r_2=|a|d(B).$

(13): $d(AB)=2\{|a|r_2+|b|r_1+r_1r_2\}=2\{(|a|+r_1)r_2+|b|r_1\}$
$\leqslant 2\{(|a|+r_1)r_2+(|b|+r_2)r_1\}=|A|d(B)+|B|d(A).$

(14): $d(A)=2r_1=|\langle 0,\ 2r_1\rangle|=|A-A|.$

(15): $d(AB)=2\{|a|r_2+|b|r_1+r_1r_2\}$
$=2\{(|a|+r_1)r_2+|b|r_1\}$
$\geqslant 2(|a|+r_1)r_2=|A|d(B).$

(16): $d(A\pm B)=d(\langle a\pm b,\ r_1+r_2\rangle)=2(r_1+r_2)=d(A)+d(B).$

(17): $A\subseteq B\Leftrightarrow|a-b|\leqslant r_2-r_1$. Следовательно,

$$\frac{1}{2}(d(B)-d(A))=|r_2|-|r_1|\leqslant|r_2-r_1|\leqslant|a-b|+|r_1-r_2|$$
$$=q(A,\ B)\leqslant r_2-r_1+|r_2-r_1|=d(B)-d(A).$$

С теоремой 4 п.1.1 связана

Теорема 6. *Операции* $\{+,\ -,\ \cdot,\ :\}$, *заданные на* $R(\mathbb{C})$ *определением* 3 п.1.4, *а на* $K(\mathbb{C})$ *— определением* 7 п.1.4, *непрерывны.*

Доказательство. Пусть $\{A^{(k)}\}_{k=0}^{\infty}$, $\{B^{(k)}\}_{k=0}^{\infty}$ — последовательности, у которых

$$A^{(k)} = A_1^{(k)} + iA_2^{(k)}, \qquad B^{(k)} = B_1^{(k)} + iB_2^{(k)} \in R(\mathbb{C})$$

и

$$\lim_{k\to\infty} A^{(k)} = A = A_1 + iA_2, \qquad \lim_{k\to\infty} B^{(k)} = B = B_1 + iB_2.$$

Докажем непрерывность умножения:

$$\lim_{k\to\infty} A^{(k)}B^{(k)} = \lim_{k\to\infty} \left\{A_1^{(k)}B_1^{(k)} - A_2^{(k)}B_2^{(k)} + i\left(A_1^{(k)}B_2^{(k)} + A_2^{(k)}B_1^{(k)}\right)\right\}$$

$$= \lim_{k\to\infty} \left(A_1^{(k)}B_1^{(k)} - A_2^{(k)}B_2^{(k)}\right) + i \lim_{k\to\infty} \left(A_1^{(k)}B_2^{(k)} + A_2^{(k)}B_1^{(k)}\right)$$

$$= A_1B_1 - A_2B_2 + i(A_1B_2 + A_2B_1) = AB,$$

поскольку операции отделения вещественной и мнимой частей комплексного числа непрерывны на *I*(C).

Подобное доказательство можно провести и для остальных операций на *R*(C) и *K*(C).

Аналогично (22 п.1.1) вводится еще одна бинарная операция на *R*(C). Теоретико-множественное пересечение *F* и *B* из *R*(C), задаваемое формулой

$$A \cap B = \{c \mid c \in A,\ c \in B\}, \tag{18}$$

называется пересечением *A* и *B*. Это пересечение принадлежит *R*(C), если оно не пусто. При $A = A_1 + iA_2$ и $B = B_1 + iB_2$ имеем

$$A \cap B = A_1 \cap B_1 + i(A_2 \cap B_2), \tag{19}$$

где $A_i \cap B_i$ находится по формуле (23 п.1.1).

Со следствием 12 п.1.1 связано

Следствие 7. *Пусть* $A,\ B,\ C,\ D \in R(\mathbb{C})$. *Тогда имеем*

$A \subseteq C,\ B \subseteq D \Rightarrow A \cap B \subseteq C \cap D$ *(монотонность включения).* (20)

При этом операция пересечения непрерывна, если ее результат остается в R(C).

Данное следствие можно доказать с помощью следствия 12 п.1.1, если последовательно применить последнее к вещественной и мнимой частям.

2. Методы локализации

2.1. Локализация нулей функций одной вещественной переменной

В этом разделе мы рассмотрим методы локализации нулей вещественной функции *f* одной вещественной переменной *x*. Эти методы позволят найти множество интервалов наименьшей возможной

ширины, таких что каждый интервал содержит один или несколько нулей функции f из заданного интервала $X^{(0)} \in I(\mathrm{R})$. Особый интерес представляет случай одного изолированного нуля в $X^{(0)}$. При разработке таких методов будет обращено особое внимание на два обстоятельства. С одной стороны, методы должны быть применимы к широким классам функций при легко проверяемых условиях. С другой стороны, должна быть гарантирована локализация нулей и в том случае, когда рассматриваемые методы реализуются на вычислительной машине, где вместо обычной интервальной арифметики возникает машинная интервальная арифметика, описанная ранее. Поэтому такие методы радикально отличаются от методов для конкретных классов функций и от других процедур общего назначения.

Простые реализации таких методов задаются с помощью так называемых методов деления. Это — интервальные варианты метода двоичного поиска или других методов поиска. Мы кратко опишем такую процедуру. Для нее требуется лишь существование интервального вычисления функции f в интервале $X^{(0)}$. Чтобы улучшить локализацию нулей в $X^{(0)}$\ мы делим $X^{(0)}$ пополам точкой

$$m(X^{(0)}) = \frac{1}{2}\left(x_1^{(0)} + x_2^{(0)}\right)$$

на интервалы $U^{(0)}$ и $V^{(0)}$, такие что

$$X^{(0)} = U^{(0)} \cup V^{(0)} = \left[x_1^{(0)},\ m(X^{(0)})\right] \cup \left[m(X^{(0)}),\ x_2^{(0)}\right].$$

Если $0 \in f(U^{(0)})$, то $U^{(0)}$ может содержать нуль функции f, и потому мы повторяем ту же процедуру для $U^{(0)}$. Если $0 \in f(V^{(0)})$, то мы аналогичным образом повторяем процедуру для $V^{(0)}$. Если же мы имеем $0 \notin f(U^{(0)})$ или $0 \notin f(V^{(0)})$, то игнорируем соответствующий подынтервал, так как ввиду (1 микромодуля 24) он не может содержать нуля функции f. Поэтому такой подынтервал исключается из дальнейших вычислений. Описанный итерационный процесс порождает последовательность подынтервалов, содержащихся в $X^{(0)}$ и «подозрительных на наличие нуля функции f». Ширина этих интервалов стремится к нулю, так как она уменьшается вдвое на каждом шаге. Ввиду (5 микромодуля 24) эти постепенно вычисляемые интервалы сходятся к нулям функции f в интервале $X^{(0)}$.

Чтобы предотвратить слишком сильный рост количества «подозрительных» интервалов, мы можем ввести следующую модификацию. На каждом шаге мы исследуем либо только правую половину интервала, либо только левую. Если на некотором шаге мы имеем $0 \notin f(Y)$ для этого полуинтервала Y, то процедура повторяется снова, начиная с интервала $\left[x_1^{(0)}, y_1\right] \subset X^{(0)}$

(соответственно $[y_2, x_2^{(0)}] \subset X^{(0)}$). Таким образом, мы последовательно вычисляем отдельные нули функции f в порядке справа налево (соответственно слева направо) и не сталкиваемся с проблемой хранения большого количества «подозрительных» интервалов.

2.1.1. А. Методы ньютоновского типа

В этом и следующем разделах мы исследуем интервальные модификации метода Ньютона. Для этого рассмотрим непрерывную функцию f, имеющуюнуль в данном интервале $X^{(0)} = [x_1^{(0)}, x_2^{(0)}]$:

$$f(\xi) = 0$$

для некоторого $\xi \in X^{(0)}$. Пусть

$$f(x_1^{(0)}) < 0 \quad \text{и} \quad f(x_2^{(0)}) > 0 \tag{1}$$

в граничных точках интервала $X^{(0)}$. Пусть, далее, m_1, m_2 — границы разностных отношений

$$0 < m_1 \leqslant \frac{f(x) - f(\xi)}{x - \xi} = \frac{f(x)}{x - \xi} \leqslant m_2 < \infty, \quad \xi \neq x \in X^{(0)}. \tag{2}$$

Эти границы определяют интервал $M = [m_1, m_2] \in I(\mathbb{R})$. (Аналогичные соображения справедливы, если предположить, что $f(x_1^{(0)}) > 0$, $f(x_2^{(0)}) < 0$ и $m_2 < 0$.) Очевидно, что при сделанных предположениях функция f не имеет других корней в $X^{(0)}$. Начав с исходного локализующего интервала $X^{(0)} \ni \xi$, мы вычисляем итерационно новые интервалы $X^{(k)}$, $k \geqslant 1$, согласно следующей процедуре:

$$X^{(k+1)} = \{m(X^{(k)}) - f(m(X^{(k)}))/M\} \cap X^{(k)}, \quad k \geqslant 0, \tag{3}$$

где

$$m(X^{(k)}) \in X^{(k)}.$$

Этот шаг изображен на рис. 1.

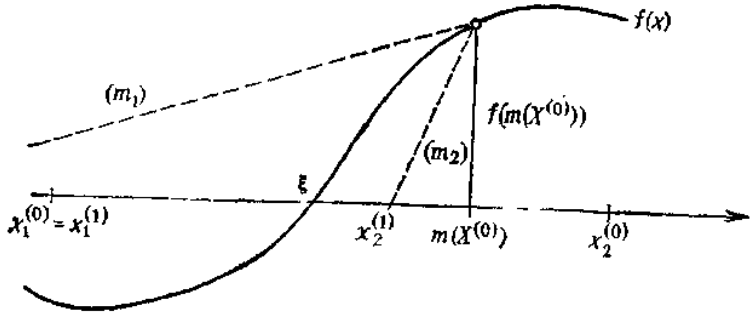


Рис. 1

Не пользуясь интервальными операциями, можно записать итерации (3) в виде

$$x_1^{(k+1)} = \begin{cases} \max\{x_1^{(k)}, m(X^{(k)}) - f(m(X^{(k)}))/m_1\}, & \text{если } f(m(X^{(k)})) \geqslant 0, \\ m(X^{(k)}) - f(m(X^{(k)}))/m_2, & \text{если } f(m(X^{(k)})) \leqslant 0, \end{cases}$$

$$x_2^{(k+1)} = \begin{cases} m(X^{(k)}) - f(m(X^{(k)}))/m_2, & \text{если } f(m(X^{(k)})) \geqslant 0, \\ \min\{x_2^{(k)}, m(X^{(k)}) - f(m(X^{(k)}))/m_1\}, & \text{если } f(m(X^{(k)})) \leqslant 0. \end{cases} \quad (3')$$

В обеих формулировках (3) и (3')

$$m: I(\mathbb{R}) \to \mathbb{R}$$

обозначает некоторую процедуру выбора вещественного числа m из данного интервала. Часто используют его середину

$$m(X) = \frac{1}{2}(x_1 + x_2). \quad (4)$$

Теперь мы перечислим важнейшие свойства последовательности итераций $\{X^{(k)}\}_{k=0}^{\infty}$.

Теорема 1. *Пусть f — непрерывная функция, ξ — нуль функции f в интервале $X^{(0)}$, причем имеют место* (1) *и* (2) *для интервала* $M = [m_1, m_2]$, $m_1 > 0$. *Тогда последовательность* $\{X^{(k)}\}_{k=0}^{\infty}$, *вычисленная по формулам* (3), *обладает следующими свойствами:*

$$\xi \in X^{(k)}, \quad k \geqslant 0, \quad (5)$$

$$X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots, \quad \text{где} \quad \lim_{k \to \infty} X^{(k)} = \xi, \quad (6)$$

либо эта последовательность стабилизируется через конечное число шагов на точке $[\xi, \xi]$;

$$d(X^{(k+1)}) \leqslant (1 - m_1/m_2)\, d(X^{(k)}). \quad (7)$$

Доказательство. (5): Из (2) и следствия 1.п.1.5 получаем

$$\xi = m(X^{(0)}) - \frac{f(m(X^{(0)}))}{f(m(X^{(0)}))/(m(x^{(0)}) - \xi)}$$
$$\in \{m(X^{(0)}) - f(m(X^{(0)}))/M\} \cap X^{(0)} = X^{(1)}.$$

Для $k > 1$ применяем метод математической индукции.

(6), (7): Предположим, что $f(m(X^{(k)})) > 0$. Теперь если имеет место $f(m(X^{(k)})) \geqslant (m(X^{(k)}) - x_1^{(k)})\, m_1$, то с помощью (3') мы получаем

$$d(X^{(k+1)}) = x_2^{(k+1)} - x_1^{(k+1)} = m(X^{(k)}) - f(m(X^{(k)}))/m_2 - x_1^{(k)}$$
$$\leqslant (m(X^{(k)}) - x_1^{(k)}) - (m(X^{(k)}) - x_1^{(k)})\, m_1/m_2$$
$$= (m(X^{(k)}) - x_1^{(k)})(1 - m_1/m_2) \leqslant d(X^{(k)})(1 - m_1/m_2).$$

Если же имеет место $f(m(X^{(k)})) \leqslant (m(X^{(k)}) - x_1^{(k)})\, m_1$, то из (3′) получаем

$$d(X^{(k+1)}) = x_2^{(k+1)} - x_1^{(k+1)}$$
$$= m(X^{(k)} - f(m(X^{(k)}))/m_2 - m(X^{(k)}) + f(m(X^{(k)}))/m_1$$
$$= \frac{f(m(X^{(k)}))}{m_1}(1 - m_1/m_2) \leqslant (m(X^{(k)}) - x_1^{(k)})(1 - m_1/m_2)$$
$$\leqslant d(X^{(k)})(1 - m_1/m_2).$$

Случай $f(m(X^{(k)})) < 0$ доказывается аналогично. Если, однако, $f(m(X^{(k)})) = 0$, то $m(X^{(k)}) = \xi$, и потому $d(X^{(k+1)}) = 0$ и $X^{(k+1)} = \xi$, $i \geqslant 1$. Это доказывает (7). Ввиду $m_1 \leq m_2$

$$d(X^{(k+1)}) \leqslant \gamma^{k+1} d(X^{(0)}), \qquad 0 \leqslant \gamma = (1 - m_1/m_2) < 1,$$

откуда следует

$$\lim_{k \to \infty} d(X^{(k+1)}) = 0.$$

Отсюда и из (5) следует $\lim\limits_{k \to \infty} X^{(k)} = \xi$, если только элементы последовательности не вырождаются в точку: $X^{(k_0+1)} = \xi$, $i \geqslant 1$, для некоторого k_0. Первая часть соотношения (6) следует из формул (3).

Таким образом, теорема 1 гарантирует, что в ее предположениях последовательные приближения $X^{(k)}$, $k \geqslant 0$, сходятся к нулю ξ функции f, причем каждый из этих интервалов содержит искомый нуль. Если же мы применяем (3), начав с интервала $X^{(0)}$, такого что $\xi \notin X^{(0)}$, то найдется индекс k_0, для которого пересечение в (3) пусто. Это легко доказать от противного, используя (7) и предположение, что пересечение непусто. Этот итерационный метод в общей постановке был глубоко исследован. Он связан также с

итерацией монотонных функций для нахождения неподвижной точки. Аналогичные процедуры для многочленов были использованы Бауэром еще в 1917 г.

Рассмотрим теперь два уточнения формул (3), возникающие при конкретном выборе точки *m.* Взяв в качестве *m* середину интервала, мы получаем следующее утверждение.

Следствие 2. *Если в предположениях теоремы* 1 *сделан выбор*

$$m(X^{(k)}) = \frac{1}{2}(x_1^{(k)} + x_2^{(k)}), \quad k \geqslant 0,$$

то для последовательности приближений $\{X^{(k)}\}_{k=0}^{\infty}$ *верно неравенство*

$$d(X^{(k+1)}) \leqslant \frac{1}{2}(1 - m_1/m_2)\, d(X^{(k)}), \tag{8}$$

уточняющее (7).

Доказательство. В доказательстве соотношения (7) из теоремы 1 мы имеем при нашем конкретном выборе точки $m(X^w)$, что

$$m(X^{(k)}) - x_1^{(k)} = \frac{1}{2} d(X^{(k)}).$$

Отсюда получаем (8).

Таким образом, при выборе середины интервала в качестве $m(X^{(k)})$ нам гарантировано уменьшение ширины локализующего интервала по крайней мере вдвое.

Рассматривались и другие выборы $m(X^{(k)})$. Например, используется

$$m(X^{(k)}) = m(X^{(k-1)}) - f(m(X^{(k-1)}))/m_0 \quad \text{для} \quad m_0 \in M$$

и

$$m(X^{(k)}) \in \{x_1^{(k)},\ x_2^{(k)}\},$$

если значение $m(X^{ik})$ из предыдущей формулы не принадлежит интервалу $X^{(k)}$, $k \geqslant 1$.

Интервал *M,* содержащий разностное отношение (2), требуется и в теореме 1, и в следствии 2. Если функция f непрерывно дифференцируема и $f'(x) \neq 0$ для

$$x \in X^{(0)},$$

то в силу теоремы о среднем можно положить

$$M = [\inf_{y \in X^{(0)}} f'(y),\ \sup_{y \in X^{(0)}} f'(y)].$$

В общем случае можно лишь локализовать этот интервал, например, путем интервального оценивания f', т. е. полагая

$$M = f'(X^{(0)}).$$

Условие $m_1 > 0$ может быть гарантировано в случае необходимости с помощью априорной оценки величины

$$\inf_{y \in X^{(0)}} f'(y).$$

2.1.2. В. Определение оптимального метода

В методе итераций (3), рассмотренном в разд. А, имеется определенная степень свободы при выборе $m(X^{(k)}) \in X^{(k)}$. В зависимости от выбора точек $m(X^{(k)})$ в интервалах $X^{(k)}$ мы получаем различные последовательности локализующих интервалов $\{X^{(k)}\}_{k=0}^{\infty}$. Эти последовательности в общем случае несравнимы почленно в смысле включения интервалов. Поэтому, очевидно, необходимо попытаться найти методы выбора $m(X^{(k)}) \in X^{(k)}$, порождающие последовательности $\{X^{(k)}\}_{k=0}^{\infty}$, в которых ширина отдельных элементов наименьшая возможная. Уточним это требование. Обозначим через $\varphi[X]$ класс функций f, обладающих следующими свойствами для данного интервала

$$X = [x_1, x_2]:$$

(α) $f(x_1) < 0$ и $f(x_2) > 0$;

(β) для интервала $M = [m_1, m_2]$, такого что $m_1 > 0$, имеет место

$$m_1 \leqslant (f(x) - f(y))/(x - y) \leqslant m_2 \text{ для } x \neq y,\ x,\ y \in X.$$

Очевидно, что любая функция $f \in \varphi[X]$ имеет один и только один корень ξ в интервале X. Поэтому выполнены все условия применимости метода итераций (3) и справедливы все утверждения теоремы 1.

Процесс выбора подходящего $m(X^{(k)}) \in X^{(k)}$ мы разобьем на шаги. Обозначим последовательные приближения (3) через $\{X^{(k)}\}_{k=0}^{\infty}$. Для вычисления нового приближения $X^{(k+1)}$ нам нужны величины $m(X^{(k)})$ и $f(m(X^{(k)}))$. Если мы зафиксируем величину $m(X^{(k)}) = x \in X^{(k)}$ из $X^{(k)}$, то $X^{(k+1)}$ будет зависеть только от $f(m(X^{(k)}))$. Но это значение функции f может меняться лишь между $y_1^{(k)}$ и $y_2^{(k)}$, ибо $f \in \varphi[X]$ и значения $f(m(X^{(i)}))$, $0 \leqslant i < k$, зафиксированы. Это позволяет нам определить наибольшую возможную ширину

$$\max \{d(X^{(k+1)}) \mid m(X^{(k)}) = x,\ y_1^{(k)} \leqslant f(m(X^{(k)})) \leqslant y_2^{(k)}\}.$$

Это «наихудший» случай для функции $f \in \varphi[X]$.

Теперь мы определим $\tilde{x} = m(X^{(k)}) \in X^{(k)}$ таким образом, чтобы минимизировать эту наибольшую ширину. Иными словами, вычисляется величина

$$\min_{x \in X^{(k)}} \{\max d(X^{(k+1)}) \mid m(X^{(k)}) = x,\ y_1^{(k)} \leqslant f(m(X^{(k)})) \leqslant y_2^{(k)}\}$$

и соответствующее значение x выбирается в качестве $m(X^{(k)})$. Таким образом, определение $m(X^{(k)})$ производится путем минимизации «наихудшего» случая.

Теперь опишем эту процедуру подробно. Не умаляя общности, рассмотрим случай $f(m(X^{(k-1)})) > 0$. На рис. 2 отмечена область, куда могут попасть значения $f(m(X^{(k)}))$ в предположении, что $f \in \varphi[X]$, причем $f(m(X^{(k-1)})) > 0$.

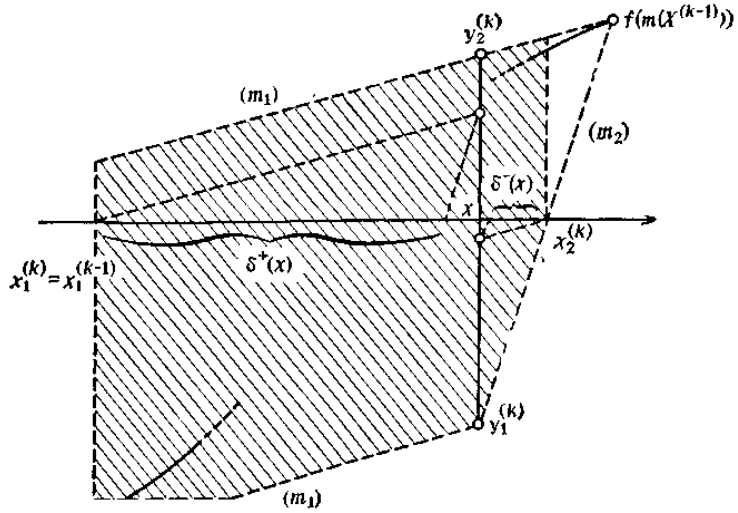


Рис. 2

Нижняя ограничивающая прямая (m_1) может отсутствовать на рис. 2, например, в случае, когда $f(m(X^{(i)})) > 0,\ 0 \leqslant i < k-1$.

Остальные значения $f(m(X^{(i)})),\ 0 \leqslant i < k-1$, не налагают никаких дополнительных ограничений на эту область.

Теперь мы оценим возможные значения $d(X^{(k+1)})$ при заданном $m(X^{(k)}) = x \in X^{(k)}$. Пусть сначала $f(m(X^{(k)})) \geqslant 0$. Для всех значений

$$0 \leqslant f(x) \leqslant (x - x_1^{(k)}) m_1$$

мы получаем из (3'), что

$$d(X^{(k+1)}) = x - \frac{f(x)}{m_2} - x + \frac{f(x)}{m_1} = f(x)\left(\frac{1}{m_1} - \frac{1}{m_2}\right).$$

Аналогично, для всех значений

$$\left(x - x_1^{(k)}\right) m_1 \leqslant f(x) \leqslant y_2^{(k)}$$

(3') дает

$$d(X^{(k+1)}) = x - f(x)/m_2 - x_1^{(k)}.$$

(Отметим, что ввиду $x_1^{(k)} = \max\{x^{(k-1)}, \; m(X^{(k-1)}) - f(m \times (X^{(k-1)}))/m_1\}$ мы всегда имеем $y_2^{(k)} \geqslant \left(x - x_1^{(k)}\right) m_1$.) В первом случае $d(X^{(k+1)})$ — монотонно возрастающая функция от $f(x)$, во втором — монотонно убывающая. Для $f(x) = \left(x - x_1^{(k)}\right) m_1$ мы имеем максимум

$$\delta^+(x) = \left(x - x_1^{(k)}\right)(1 - m_1/m_2).$$

Оставшиеся случаи $f(m(X^{(k)})) \leqslant 0$ рассматриваются аналогично и дают максимум $d(X^{(k+1)})$, равный

$$\delta^-(x) = \left(x_2^{(k)} - x\right)(1 - m_1/m_2).$$

На рис. 2 показаны два варианта вычисления $X^{(k+1)}$, приводящие к максимальной ширине $\delta^+(x)$ (и соответственно $\delta^-(x)$).

Теперь мы определим минимум

$$\min_{x \in X^{(k)}} \max\{\delta^+(x), \; \delta^-(x)\}.$$

Выражения $\delta^+(x)$ и $\delta^-(x)$ удовлетворяют условию

$$\delta^+\left(\frac{1}{2}\left(x_1^{(k)} + x_2^{(k)}\right) - t\right) = \delta^-\left(\frac{1}{2}\left(x_1^{(k)} + x_2^{(k)}\right) + t\right)$$

для $0 \leqslant |t| \leqslant \frac{1}{2}\left(x_2^{(k)} - x_1^{(k)}\right)$. Поэтому минимум равен

$$d(X^{(k+1)}) = \frac{1}{2}\, d(X^{(k)})(1 - m_1/m_2)$$

и достигается в точке

$$\bar{x} = \frac{1}{2}\left(x_1^{(k)} + x_2^{(k)}\right).$$

(Ср. со следствием 2.)

Распространим теперь принцип оптимизации, который мы применили к вычислению $X^{(k+1)}$, на определение величин $m(X^{(i)})$, $0 \leq i < k$. Мы хотим определять значения $m(X^{(0)}) = x^{(0)}$, $\ldots$, $m(X^{(k)}) = x^k$ таким образом, чтобы при этом достигались величины

$$\min_{x^{(0)} \in X^{(0)}} \max_{y_1^{(0)} \leqslant f(x^{(0)}) \leqslant y_2^{(0)}} \cdots \min_{x^{(k)} \in X^{(k)}} \max_{y_1^{(k)} \leqslant f(x^{(k)}) \leqslant y_2^{(k)}} d(X^{(k+1)}).$$

Оказывается, что это просто, так как оптимальная величина $d(X^{(k+1)})$ пропорциональна величине $d(X^{(k)})$ при фиксированном $m(X^{(k-1)})$. Область допустимых значений функции $f(m(X^{(k-1)}))$ определяется лишь значением $f(m(X^{(k-2)}))$. Поэтому можно провести для $m(X^{(k-1)})$ те же рассуждения, что и для $m(X^{(k)})$, и получить оптимальное значение

$$m(X^{(k-1)}) = \frac{1}{2}(x_1^{(k-1)} + x_2^{(k-1)}).$$

Соответствующим образом мы получаем результаты для

$$m(X^{(i)}), \quad i = k-2,\ k-3,\ \ldots,\ 0$$

в этом порядке.

Теорема 3. *Пусть метод итерации* (3) *применяется к функции* $f \in \varphi[X]$. *Если спользуется правило*

$$m(X^{(k)}) = \frac{1}{2}(x_1^{(k)} + x_2^{(k)}), \quad 0 \leqslant k \leqslant i, \quad i \geqslant 0,$$

то максимальная ширина $d(X^{(i+1)})$ *для функций* $f \in \varphi[X]$ *меньше, чем для любых других выборов точки* $m(X^{(k)})$. *Если* $f \in \varphi[X]$, *то мы имеем*

$$d(X^{(i+1)}) \leqslant \frac{1}{2^{i+1}}(1 - m_1/m_2)^{i+1}\, d(X^{(0)}).$$

Существует $g \in \varphi[X]$, *для которой в последнем соотношении выполняется равенство.*

Доказательство этой теоремы содержится в только что проведенном рассуждении. Следует отметить, что $g \in \varphi[X]$ можно выбрать в виде кусочно-линейной функции, проходящей через точки $(m(X^{(k)}),\ f(m(X^{(k)})))$, $0 \leqslant k \leqslant i$.

2.1.3. C. Квадратично сходящиеся методы

В методе (3) мы используем фиксированную пару m_1, m_2 границ для разностных отношений функции f. Эта процедура соответствует интервальному варианту упрощенного метода Ньютона. Если мы предположим, что f непрерывно дифференцируема и для производной f имеется интервальная оценка $f'(X)$, то мы можем определить интервальный вариант и для обычного метода Ньютона. Эта новая процедура получается, если мы модифицируем метод (3), заменяя интервал M на интервал

$$M^{(k)} = f'(X^{(k)}) \tag{9}$$

на *k*-м шаге итерации. Если известны априорные оценки

$$0 < l_1 \leqslant f'(x) \leqslant l_2. \qquad x \in X^{(0)},$$

то можно гарантировать оценку $m_1 > 0$ и использовать выражение

$$M^{(k)} = [m_1^{(k)},\ m_2^{(k)}] = f'(X^{(k)}) \cap L, \qquad L = [l_1,\ l_2]. \tag{10}$$

Таким образом, мы получаем

$$X^{(k+1)} = \{m(X^{(k)}) - f(m(X^{(k)}))/M^{(k)}\} \cap X^{(k)}, \quad k \geqslant 0, \quad \text{где} \quad m(X^{(k)}) \in X^{(k)}. \tag{11}$$

С помощью (11) порождается последовательность интервалов $\{X^{(k)}\}_{k=0}^{\infty}$, для которой можно доказать утверждение, аналогичное теореме 1.

Теорема 4. *Пусть f — непрерывно дифференцируемая функция и f удовлетворяет в интервале $X^{(0)}$ условиям теоремы* 3 п.1.4. *Пусть, далее, для $X^{(0)}$ выполнено* (1), *нуль функции f в $X^{(0)}$ обозначен через ξ и интервалы $M^{(k)}$ определены формулами* (9) *или* (10). *Тогда последовательность $\{X^{(k)}\}_{k=0}^{\infty}$ либо удовлетворяет соотношениям*

(5) $\xi \in X^{(k)}, \ k \geqslant 0,$

(6) $X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots \quad и \quad \lim\limits_{k \to \infty} X^{(k)} = \xi,$

либо стабилизируется на значении $[\xi,\ \xi]$ через конечное число шагов

$$d(X^{(k+1)}) \leqslant (1 - m_1^{(k)}/m_2^{(k)})\, d(X^{(k)}) \leqslant \beta (d(X^{(k)}))^2, \quad \beta \geqslant 0, \tag{12}$$

т. е. R — порядок метода итераций (11) *(см. приложение* А) *удовлетворяет условию*

$$O_R((11), \xi) \geqslant 2.$$

Доказательство. Для $x \in X^{(k)}$ верно

$$\frac{f(x)}{x - \xi} = \frac{f(x) - f(\xi)}{x - \xi} = f'(\eta) \in M^{(k)}, \quad \eta = x + \theta(\xi - x), \quad 0 < \theta < 1.$$

Поэтому аналогичное утверждение для $M^{(k)}$ можно теперь доказать так же, как в теореме 1.

Остается установить (12). Как и в доказательстве теоремы 1, получаем

$$d(X^{(k+1)}) \leqslant \left(1 - \frac{m_1^{(k)}}{m_2^{(k)}}\right) d(X^{(k)}) = \frac{m_2^{(k)} - m_1^{(k)}}{m_2^{(k)}}\, d(X^{(k)}),$$

откуда с помощью 9 п. 1.2 и теоремы 5 микромодуля 24 имеем

$$d\left(X^{(k+1)}\right) \leqslant \frac{d\left(M^{(k)}\right)}{m_1^{(0)}}\, d\left(X^{(k)}\right) \leqslant \frac{d\left(f'\left(X^{(k)}\right)\right)}{m_1^{(0)}}\, d\left(X^{(k)}\right)$$

$$\leqslant \left(c/m_1^{(0)}\right)\left(d\left(X^{(k)}\right)\right)^2, \quad c/m_1^{(0)} \geqslant 0.$$

Метод итераций (11) и теорема 4 соответствуют известным формулировкам. Продолжим дальше модифицикаю этого метода Заметим, что если $f(m(X^{(k)}))>0$ (соответственно $f(m(X^{(k)}))<0$. то искомый нуль ξ должен лежать в интервале $\left[x_1^{(k)},\ m\left(X^{(k)}\right)\right]$ (соответственно $\left[m\left(X^{(k)}\right),\ x_2^{(k)}\right]$. Если $f(m(X^{(k)}))=0$, то $m\left(X^{(k)}\right)=\xi$, и процесс итерации заканчивается. Поэтому в (11) достаточно положить

$$M^{(k)} = f\left(Y^{(k)}\right) \cap L, \quad L \text{ взято из (10)},$$

где

$$Y^{(k)} = \begin{cases} \left[x_1^{(k)},\ m\left(X^{(k)}\right)\right], & \text{если } f\left(m\left(X^{(k)}\right)\right) > 0, \\ \left[m\left(X^{(k)}\right),\ x_2^{(k)}\right], & \text{если } f\left(m\left(X^{(k)}\right)\right) < 0, \\ X^{(k)} & \text{в противном случае.} \end{cases} \qquad (13)$$

Тогда имеем $f'\left(Y^{(k)}\right) \subseteq f'\left(X^{(k)}\right)$ и $d\left(Y^{(k)}\right) \leqslant d\left(X^{(k)}\right)$, и, таким образом, легче выполнить условие $m_1^{(k)} > 0$. Теорема (4) справедлива и для метода (13)

По поводу выбора точек $m\left(X^{(k)}\right) \in X^{(k)}$ для метода (11) имеем утверждение, аналогичное следствию 2, и можем провести те же рассуждения, что и разд. В. Мы не будем углубляться в это.

Теперь поясним интервальный метод Ньютона на числовых примерах.

Примеры, (а) Функция

$$f(x) = x^2\left(\frac{1}{3}x^2 + \sqrt{2}\sin x\right) - \sqrt{3}/19$$

имеет нуль ξ в интервале $X^{(0)} = [0.1,\ 1]$. Производную

$$f'(x) = x\left(\frac{4}{3}x^2 + \sqrt{2}(2\sin x + x\cos x)\right)$$

можно оценить в $X^{(0)}$:

$$l_1 = 0{,}0013 \leqslant f'(x) \leqslant l_2, \qquad x \in X^{(0)}.$$

Локализующие интервалы

$$X^{(k)}, \quad k \geqslant 0, \quad \text{по формулам (13)},$$

$$Y^{(k)}, \quad k \geqslant 0, \quad \text{по формулам (10)}$$

вычислялись на компьютере в соответствии с (11) до тех пор, пока не переставали происходить изменения. Результаты приведены в табл. 1.

(β) Многочлен

$$p(x) = x(x^9 - 1) - 1$$

имеет единственный нуль β в интервале $X^{(0)}$=[1, 1.5]. Интервальное вычисление *p'(x)* его производной $p'(x) = 10x^9 - 1$

Таблица 1

$X^{(k)}$	
[1.000000000000, 1.153909281002]	
[1.074525733152, 1.075772270022]	
[1.075764355129, 1.075767749943]	
[1.075766066086, 1.075766066088]	
$Y^{(k)}$	$d(X^{(k)})/d(Y^{(k)})$
[1.000000000000, 1.231579011696]	0.665
[1.018539065305, 1.102153489956]	0.015
[1.071809768336, 1.084762444669]	3×10^{-4}
[1.075647094319, 1.075931180877]	6×10^{-9}
[1.075766039501, 1.075766097327]	...
[1.075766066085, 1.075766066090]	...
[1.075766066085, 1.075766066088]	...

Дает $0 \notin p'(x)$ для $X \subseteq X^{(0)}$. Итерированные локализующие интервалы

$$X^{(k)}, \quad k \geqslant 0, \quad \text{по формулам (13) с } L = p'(X^{(0)}),$$
$$Y^{(k)}, \quad k \geqslant 0, \quad \text{по формулам (9)}$$

вычислялись в соответствии с (11) с использованием (4). Полученные значения, приведены в табл. 2.

Таблица 2

$X^{(k)}$		
[0.09999999999990, 0.4384388546433]		
[0.3382030708107, 0.4384388546433]		
[0.3915056049954, 0.3924484948316]		
[0.3923789206719, 0.3923799504692]		
[0.3923795071350, 0.3923795071378]		
$Y^{(k)}$		$d(X^{(k)})/d(Y^{(k)})$
[0.09999999999990, 0.5181776715881]		0.809
[0.3455588336928, 0.5181776715881]		0.581
[0.3739864679691, 0.4075613703040]		0.028
[0.3922481030413, 0.3925441306206]		0.004
[0.3923794945039, 0.3923795211850]		0.001
[0.3923795071350, 0.3923795071378]		--

На практике определяющим условием для итераций (11) является $m_1 > 0$. Показано, что можно добиться его выполнения с помощью (10), используя известную нижнюю оценку l_1 производной $f'(x)$ в интервале $X^{(0)}$. Если такая оценка l_1 неизвестна или если $0 \in f'(X^{(0)})$, то процедуру (11) нельзя даже начать. Поэтому перед началом итераций следует выполнить несколько шагов метода разбиения интервалов, описанного во введении к этому модулю. Таким образом, мы найдем интервал $Y^{(0)} \subset X^{(0)}$, для которого верно $0 \notin f'(Y^{(0)})$.

Имеется еще одна модификация метода Ньютона, применимая даже в случае $0 \in f'(X^{(0)})$. Рассмотрим этот метод, работающий даже при наличии нескольких нулей функции f в $X^{(0)}$. Если $0 \notin f'(X^{(0)})$, то этот метод совпадает с методом (11). Предположим поэтому, что $0 \in f(X^{(0)})$. Разобьем $X^{(0)}$ на подынтервалы

$$U^{(1)} = \left[x_1^{(0)},\ m(X^{(0)}) - \left|f(m(X^{(0)}))\right|/m_2^{(0)}\right],$$

$$V^{(1)} = \left[m(X^{(0)}) + \left|f(m(X^{(0)}))\right|/m_2^{(0)},\ x_2^{(0)}\right],$$

предполагая, что $f(m(X^{(0)})) \neq 0$. Все нули функции f в $X^{(0)}$ должны лежать также и в $U^{(1)} \cup V^{(1)}$. Действительно, нуль $\xi \in X^{(0)}$ должен удовлетворять неравенству

$$\left|\frac{f(m(X^{(0)}))}{\xi - m(X^{(0)})}\right| \leqslant m_2^{(0)},$$

откуда следует

$$|f(m(X^{(0)}))|/m_2^{(0)} \leqslant |\xi - m(X^{(0)})|$$

и

$$\xi \geqslant m(X^{(0)}) + \frac{|f(m(X^{(0)}))|}{m_2^{(0)}}$$

или

$$\xi \leqslant m(X^{(0)}) - \frac{|f(m(X^{(0)}))|}{m_2^{(0)}}$$

Из последнего неравенства, однако, следует, что $\xi \in U^{(1)} \cup V^{(1)}$. Кроме того, при условии $f(m(X^{(0)})) \neq 0$ имеет место

$$d(U^{(1)} \cup V^{(1)}) = d(X^{(0)}) - 2|f(m(X^{(0)}))|/m_2^{(0)} < d(X^{(0)}).$$

Теперь эту процедуру можно повторить для подынтервалов $U^{(1)}$ и $V^{(1)}$ и т. д. Суммарная ширина этих интервалов стремится к нулю. Если f имеет в $X^{(0)}$ только простые нули, то после некоторого шага итерации все они окажутся в непересекающихся подынтервалах. Далее, после некоторого шага k процедура превращается в итерацию вида (11). После этого либо подынтервалы стремятся к интервалу, содержащему нуль, либо в какой-то момент получается пустое пересечение.

Вместо того чтобы использовать $M^{(k)} := f'(X^{(k)})$ в (11), в соответствии с (3) можно для многочленов использовать интервалы J_1, J_2, J_3 и J_4 из теоремы 7 микромодуля 24 с $y := m(X^{(k)})$ и $X := X^{(k)}$ для локализации производной. Все утверждения теоремы 4 остаются справедливыми. Так как в теореме 7 микромодуля 24 было показано, что J_1 — оптимальная локализация, то для получения наилучшей локализации на каждом шаге разумно использовать именно этот интервал для локализации производной.

В этой связи рассмотрим следующий многочлен.

Пример. Пусть

$$p(x) = x^7 + 3x^6 - 4x^5 - 12x^4 - x^3 - 3x^2 + 4x + 12.$$

Этот многочлен имеет корень ξ в $X^{(0)} = [1.8, 2.4]$. Используя формулы (11), мы находим локализующие интервалы для корня, вычисляя $M^{(k)} := p'(X^{(k)})$ по схеме Горнера. Таблица 3 содержит вычисленные интервалы.

Таблица 3

k	$X^{(k)}$
0	[1.8, 2.4]
1	[1.8, 2 0727618077482]
2	[1.9742900052812, 2 0727618077842]
3	[1.9948757147483, 2 0059215482353]
4	[1.9999888234200, 2 0000115390070]
5	[1.9999999999894, 2 0000000000107]
6	[2.0, 2 0]

Таблица 4

k	$X^{(k)}$
0	[1.8, 2 4]
1	[1.9419538108826, 2 0566964050488]
2	[1.9999999975872, 2 0001112993369]
3	[1 9999999975872, 2 0000000029595]
4	[2 0, 2 0]

Если $p'(X^{(k)})$ заменено на J_1, мы аналогичным образом получаем табл. 4. В табл. 5 мы приводим для каждого шага итерации частное $d_1^{(k)}/d_2^{(k)}$ от деления ширины первого итерированного интервала на ширину второго итерированного интервала.

Таблица 5

k	0	1	2	3
$d_1^{(k)}/d_2^{(k)}$	1	2.37	492 35	3.7×10^6

2.1.4. D. Методы более высоких порядков

Рассмотрим методы более высоких порядков для нахождения в интервале $X^{(0)} = [x_1^{(0)}, x_2^{(0)}]$ нуля ξ строго монотонно возрастающей или монотонно убывающей вещественной функции, обладающей непрерывными производными достаточно высокого порядка. Эти методы всегда сходятся. Идея описываемого построения принадлежит Эрманну. С помощью этой идеи и методов интервального анализа можно разработать методы, для которых обязательно имеет место

сходимость. Как и в разд. A, B, C, предположим, не умаляя общности, что

$$f\left(x_1^{(0)}\right) < 0 \quad \text{и} \quad f\left(x_2^{(0)}\right) > 0. \tag{1}$$

Пусть, далее, m_1 и m_2 снова обозначают границы разностных отношений

$$0 < m_1 \leqslant \frac{f(x) - f(\xi)}{x - \xi} = \frac{f(x)}{x - \xi} \leqslant m_2 < \infty, \quad \xi \neq x \in X^{(0)}. \tag{2}$$

Границы m_1 и m_2 определяют интервал $M = [m_1, m_2]$. Предполагаем, что функция f $(p+1)$-кратно непрерывно дифференцируема и для $F_i \in I(\mathbb{R})$, $2 \leqslant i \leqslant p+1$, имеет место

$$f^{(i)}(x) \in F_i, \quad x \in X^{(0)}. \tag{14}$$

Интервалы F_i могут быть найдены, например, с помощью интервального оценивания производных функций f на интервале $X^{(0)}$. Если интервальные выражения для производных не определены (например, из-за деления на X при $0 \in X$), то можно подразбить интервал $X^{(0)}$, а затем построить интервалы F_i, взяв объединение оценок, полученных для частичных интервалов.

Рассмотрим теперь следующий метод итераций:

$$\begin{cases} x^{(k)} = m(X^{(k)}) \in X^{(k)}, \\ X^{(k+1,\,0)} = \{x^{(k)} - f(x^{(k)})/M\} \cap X^{(k)}, \\ X^{(k+1,\,i)} = \left\{ x^{(k)} - \dfrac{1}{f'(x^{(k)})} \left[f(x^{(k)}) \right.\right. \\ \qquad + \displaystyle\sum_{\nu=2}^{i} \frac{f^{(\nu)}(x^{(k)})}{\nu!} (X^{(k+1,\,i-1)} - x^{(k)})^{\nu} \\ \qquad \left.\left. + \dfrac{1}{(i+1)!} F_{i+1} (X^{(k+1,\,i-1)} - x^{(k)})^{i+1} \right] \right\} \cap X^{(k+1,\,i-1)}, \\ \qquad\qquad 1 \leqslant i \leqslant p, \\ X^{(k+1)} = X^{(k+1,\,p)}, \qquad k \geqslant 0. \end{cases} \tag{15}$$

Как и в разд. A, $m(X)$ означает произвольный выбор вещественного числа в интервале X. Формулы, приведенные выше, требуют на каждом шаге вычисления значений $f(x^{(k)})$, $f'(x^{(k)})$, ..., $f^{(p)}(x^{(k)})$ и обладают следующими свойствами.

Теорема 5. *Пусть функция f имеет* $(p+1)$ *непрерывную производную,* $p \geq 1$, *и пусть для* $X^{(0)}$ *верно соотношение* (1). *Пусть* ξ *— нуль функции f в* $X^{(0)}$, *интервал* $M = [m_1, m_2]$ *определен соотношением* (2) *и имеет место* (14). *Тогда для итераций* (15) *верно, что*

$$\xi \in X^{(k)}, \quad k \geqslant 0, \tag{16}$$

и либо

$$X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots \quad и \quad \lim_{k \to \infty} X^{(k)} = \xi, \tag{17}$$

либо эти итерации стабилизируются через конечное число шагов на точке $[\xi, \xi]$;

$$d(X^{(k+1)}) \leqslant \gamma (d(X^{(k)}))^{p+1} \tag{18}$$

для некоторого $\gamma \geqslant 0$. *Таким образом, согласно теореме* 2 *из приложения* А, *R-порядок последовательности* $\{X^{(k)}\}_{k=0}^{\infty}$ *не меньше, чем* p +1.

Доказательство. (16): Допустим, что $\xi \in X^{(k)}$ для некоторого $k \geq 0$. Это верно для $k = 0$ в силу условия теоремы. Как и в теореме 1, показываем, что

$$\xi \in X^{(k+1, 0)}.$$

Пусть $\xi \in X^{(k+1, i)}$ для некоторого $i \geq 0$. Это верно для i =0,
Теперь имеем

$$\xi - x^{(k)} \in X^{(k+1, i)} - x^{(k)}.$$

Из формулы Тейлора получаем

$$0 = f(\xi) = f(x^{(k)}) + f'(x^{(k)})(\xi - x^{(k)}) + \ldots + \frac{1}{(i+1)!} f^{(i+1)}(x^{(k)})(\xi - x^{(k)})^{i+1} + \frac{1}{(i+2)!} f^{(i+2)}(\eta_{i+2})(\xi - x^{(k)})^{i+2}$$

для некоторого η_{i+2}, лежащего между $x^{(k)}$ и ξ. Вместе с монотонностью это дает нам соотношение

$$\xi = x^{(k)} - \frac{1}{f'(x^{(k)})} \left[f(x^{(k)}) + \sum_{\nu=2}^{i+1} \frac{f^{(\nu)}(x^{(k)})}{\nu!} (\xi - x^{(k)})^{\nu} + \frac{f^{(i+2)}(\eta_{i+2})}{(i+2)!} (\xi - x^{(k)})^{i+2} \right]$$

$$\in \left\{ x^{(k)} - \frac{1}{f'(x^{(k)})} \left[f(x^{(k)}) + \sum_{\nu=2}^{i+1} \frac{f^{(\nu)}(x^{(k)})}{\nu!} (X^{(k+1, i)} - x^{(k)})^{\nu} + \frac{F_{i+2}}{(i+2)!} (X^{(k+1, i)} - x^{(k)})^{i+2} \right] \right\} \cap X^{(k+1, i)} = X^{(k+1, i+1)}.$$

Поэтому имеем $\xi \in X^{(k+1,\,i)}$, $0 \leqslant i \leqslant p$, и $\xi \in X^{(k+1)} = X^{(k+1,\,p)}$.

(17): Тем же методом, что и в доказательстве теоремы 1, можно показать, что $X^{(k)} \supset X^{(k+1,\,0)}$, откуда следует $X^{(k)} \supset X^{(k+1)}$, $k \geqslant 0$, так как в формулах (15) берутся пересечения. Затем, как и в теореме 1, можно показать, что

$$d(X^{(k+1,\,0)}) \leqslant (1 - m_1/m_2)\, d(X^{(k)}).$$

Так как в формулах (15) берутся пересечения, получаем

$$d(X^{(k+1)}) \leqslant (1 - m_1/m_2)\, d(X^{(k)}), \qquad k \geqslant 0.$$

Тем же методом, что и в теореме 1, получаем сходимость $\lim\limits_{k \to \infty} X^{(k)} = \xi$. Остающаяся часть утверждения (17) доказывается так же, как в теореме 1.

(18): Мы имеем $d(X^{(k+1,\,0)}) \leqslant d(X^{(k)})$, откуда

$$\begin{aligned} d(X^{(k+1,\,1)}) &\leqslant d\left(x^{(k)} - \frac{1}{f'(x^{(k)})}\left(f(x^{(k)}) + \frac{1}{2} F_2 (X^{(k+1,\,0)} - x^{(k)})^2\right)\right) \\ &\leqslant \frac{1}{2}\, d\left(\frac{F_2}{f'(x^{(k)})} (X^{(k)} - X^{(k)})^2\right) \\ &\leqslant \frac{1}{2}\, d\left(\frac{F_2}{M} [-(d(X^{(k)}))^2,\ (d(X^{(k)}))^2]\right) \\ &= |F_2/M|\, (d(X^{(k)}))^2 = \gamma_1 (d(X^{(k)}))^2 \end{aligned}$$

с константой $\gamma_1 = |F_2/M|$, не зависящей от k.

Предположим, что для некоторого $i \geq 1$ имеем

$$d(X^{(k+1,\,i)}) \leqslant \gamma_i (d(X^{(k)}))^{i+1},$$

где γ_i не зависит от k. Это только что доказано для $i = 1$ Для $i > 1$ имеем из формул (15) с помощью правил из п.1.2 для вычисления ширины, что

$$
\begin{aligned}
d(X^{(k+1,\,i+1)}) &\leqslant d\left(\sum_{\nu=2}^{i+1}\frac{f^{(\nu)}(x^{(k)})}{\nu!\,f'(x^{(k)})}(X^{(k+1,\,i)}-x^{(k)})^{\nu}\right.\\
&\quad\left.+\frac{1}{(i+2)!}\,\frac{F_{i+2}}{f'(x^{(k)})}(X^{(k+1,\,i)}-x^{(k)})^{i+2}\right)\\
&\leqslant\sum_{\nu=2}^{i+1}\frac{1}{\nu!}\left|\frac{f^{(\nu)}(x^{(k)})}{f'(x^{(k)})}\right|d\left((X^{(k+1,\,i)}-x^{(k)})^{\nu}\right)\\
&\quad+\frac{1}{(i+2)!}\,d\left(\frac{F_{i+2}}{f'(x^{(k)})}(X^{(k+1,\,i)}-x^{(k)})^{i+2}\right)\\
&\leqslant\sum_{\nu=2}^{i+1}\frac{1}{\nu!}\left|\frac{F_{\nu}}{M}\right|\nu\,|X^{(k+1,\,i)}-x^{(k)}|^{\nu-1}\,d(X^{(k+1,\,i)}-x^{(k)})\\
&\quad+\frac{1}{(i+2)!}\,d\left(\frac{F_{i+2}}{f'(x^{(k)})}(X^{(k+1,\,i)}-x^{(k)})^{i+2}\right)\\
&\leqslant\sum_{\nu=2}^{i+1}\frac{1}{(\nu-1)!}\left|\frac{F_{\nu}}{M}\right||X^{(k)}-x^{(k)}|^{\nu-1}\,d(X^{(k+1,\,i)})\\
&\quad+\frac{1}{(i+2)!}\,d\left(\frac{F_{i+2}}{M}(X^{(k)}-x^{(k)})^{i+2}\right)\\
&\leqslant\sum_{\nu=2}^{i+1}\frac{1}{(\nu-1)!}\left|\frac{F_{\nu}}{M}\right|(d(X^{(k)}))^{\nu-1}\,\gamma_i(d(X^{(k)}))^{i+1}\\
&\quad+\frac{1}{(i+2)!}\,d\left(\frac{F_{i+2}}{M}[-(d(X^{(k)}))^{i+2},\,(d(X^{(k)}))^{i+2}]\right)\\
&=(d(X^{(k)}))^{i+2}\sum_{\nu=2}^{i+1}\frac{1}{(\nu-1)!}\left|\frac{F_{\nu}}{M}\right|\gamma_i(d(X^{(k)}))^{\nu-2}\\
&\quad+\frac{2}{(i+2)!}\left|\frac{F_{i+2}}{M}\right|(d(X^{(k)}))^{i+2}\\
&\leqslant\left(\sum_{\nu=2}^{i+1}\frac{1}{(\nu-1)!}\left|\frac{F_{\nu}}{M}\right|\gamma_i(d(X^{(0)}))^{\nu-2}\right.\\
&\quad\left.+\frac{2}{(i+2)!}\left|\frac{F_{i+2}}{M}\right|\right)(d(X^{(k)}))^{i+2}\\
&=\gamma_{i+1}(d(X^{(k)}))^{i+2}
\end{aligned}
$$

с константой

$$\gamma_{i+1}=\gamma_i\sum_{\nu=2}^{i+1}\frac{1}{(\nu-1)!}\left|\frac{F_{\nu}}{M}\right|(d(X^{(0)}))^{\nu-2}+\frac{2}{(i+2)!}\left|\frac{F_{i+2}}{M}\right|,$$

не зависящей от k. Поэтому соотношение

$$d(X^{(k+1,\,i)}) \leqslant \gamma_i (d(X^{(k)}))^{i+1}$$

справедливо для $1 \leqslant i \leqslant p$.

Далее

$$d(X^{(k+1)}) = d(X^{(k+1,\,p)}) \leqslant \gamma_p (d(X^{(k)}))^{p+1}$$

с константой γ_p, не зависящей от k. Это доказывает формулу (18) для $\gamma = \gamma_p$, что и завершает доказательство теоремы.

Теперь исследуем случай $p = 1$ несколько подробнее. Для $p = 1$ формулы (15) можно переписать в виде

$$\begin{cases} x^{(k)} = m(X^{(k)}) \in X^{(k)}, \\ X^{(k+1,\,0)} = \{x^{(k)} - f(x^{(k)})/M\} \cap X^{(k)}, \\ X^{(k+1,\,1)} = \left\{x^{(k)} - (1/f'(x^{(k)}))(f(x^{(k)}) + \frac{1}{2} F_2(X^{(k+1,\,0)} - x^{(k)})^2)\right\} \cap X^{(k+1,\,0)}, \\ X^{(k+1)} = X^{(k+1,\,1)}, \quad k \geqslant 0. \end{cases}$$

Этот метод имеет те же свойства, что и метод Мура, приведенный в разд. С. Если не считать некоторых дополнительных арифметических операций, он менее трудоемок, чем метод Мура, так как значение функции и производной приходится вычислять в одной и той же точке $x^{(k)}$. Для метода Мура производную приходится вычислять с использованием интервала $X^{(k)}$. Это в общем случае требует больше затрат, чем вычисление значений в одной точке $x^{(k)}$. Если интервал F_2 вычисляется легко, то метод (15) с $p = 1$ предпочтительнее метода Мура. Эти результаты полностью справедливы лишь в теории, когда вычисления считаются точными. Если мы хотим гарантировать локализацию нулей при реализации метода на компьютере, то должны учесть влияние погрешностей округления. Это делается путем реализации всех действий в виде машинных интервальных операций. В частности, нам требуется вычислять $f'(x^{(k)})$, используя машинную интервальную арифметику. В этом случае метод (15) при $p = 1$ требует, если не считать нескольких арифметических операций, того же объема вычислений, что и метод Мура. Так как приходится вычислять еще и интервал F_2, следует, видимо, предпочесть метод Мура, если нужно учитывать погрешности округления.

Следует упомянуть, что существует следующий метод:

$$\begin{cases} x^{(k)} = m(X^{(k)}) \in X^{(k)}, \\ X^{(k+1)} = \left\{x^{(k)} - (1/f'(x^{(k)}))(f(x^{(k)}) \right. \\ \left. + \frac{1}{2} f''(X^{(k)})(X^{(k)} - x^{(k)})^2)\right\} \cap X^{(k)}, \quad k \geqslant 0, \end{cases}$$

в предположении, что f дважды дифференцируема. Мы имеем $\xi \in X^{(k)}$, $k \geqslant 0$. Условия сходимости $\lim\limits_{k \to \infty} X^{(k)} = \xi$ не приводятся. Если метод сходится, то последовательность $d(X^{(k)})$ сходится к нулю квадратично, если $f'(\xi) \neq 0$. По сравнению с методом (15) для p=1 мы должны на каждом шаге производить интервальное оценивание второй производной $f''(X^{(k)})$. Это уменьшает постоянную сходимость, но не улучшает порядок сходимости. (То же верно для метода (15) при p= 1, если на каждом шаге заменить постоянный интервал F_2 на $f''(X^{(k)})$.) Использование этого метода на практике, т. е. с учетом погрешностей округления при вычислении $f(x^{(k)})$ и $f'(x^{(k)})$, увеличивает объем вычислений еще примерно на треть. Так как сходимость не гарантирована, этот метод несколько менее привлекателен. Применяя метод (15), мы должны выбрать конкретный порядок p. Отметим еще, что при определенных предположениях метод (15) оптимален при p = 2, т. е. он является методом третьего порядка.

2.1.5. E. Интерполяционные методы

Как и в предыдущем разделе мы рассматриваем сходящиеся методы высшего порядка. На этот раз в основу положен хорошо известный интерполяционный метод нахождения нулей функции. Изменим его с помощью приемов из интервального анализа таким образом, чтобы всегда получать монотонную локализацию корней. Так же, как и в разд. D, нам нужны интервальные оценки старших производных функции f.

Различные методы определяются с помощью (n+1)-элементного множества неотрицательных параметров

$$m_0, m_1, \ldots, m_n.$$

Положим

$$r = \sum_{i=0}^{n} m_i$$

и допустим, что

$$m_0 m_n > 0,$$

откуда следует, что $r > 0$ Мы хотим найти нуль $\xi \in X^{(0)} = |x_1^{(0)}, x_2^{(0)}|$ функции f, которая предполагается дифференцируемой нужное число раз в $X^{(0)}$.

Теперь находятся интервалы H и K, такие что имеет место

$$f'(x) \in H,\ x \in X^{(0)} \quad \text{и} \quad 0 \notin H = [h_1, h_2],$$
$$f^{(r)}(x) \in K,\ x \in X^{(0)}.$$

Чтобы описать очередной $(k+1)$-й шаг итерации, допустим, что мы уже имеем $n+1$ попарно различных приближений к нулю ξ:

$$x^{(k)},\ x^{(k-1)},\ \ldots,\ x^{(k-n)} \text{ в } X^{(0)}$$

и что последний из найденных локализующих интервалов $X^{(k)}$ имеет вид

$$X^{(k)} = [x^{(k)} - \varepsilon^{(k)},\ x^{(k)} + \varepsilon^{(k)}]$$

для некоторого $\varepsilon^{(k)} > 0$. Допустим еще, что

$$\xi \neq x_1^{(0)} \text{ и } \xi \neq x_2^{(0)}$$

После исполнения описываемых ниже шагов (S1)—(S5) определяется улучшенный локализующий интервал — новая аппроксимация $X^{(k+1)}$

(S1) *Нахождение единственного интерполяционного многочлена Эрмита*

$$p_k(x) = p_{(m_0, m_1, \ldots, m_n)}(x;\ x^{(k)},\ \ldots,\ x^{(k-n)}),$$

удовлетворяющего интерполяционным условиям

$$p_k^{(j)}(x^{(k-i)}) = f^{(j)}(x^{(k-i)}),\quad 0 \leqslant i \leqslant n,\quad 0 \leqslant j \leqslant m_i - 1.$$

(Мы полагаем $f^{(0)} = f$, *и если* $m_i = 0$, *то условия в точке* $x^{(k-i)}$ *отсутствуют.) Определяется интервал* $Z^{(k)} \subset X^{(k)}$ *по формулам*

$$Z^{(k)} = \begin{cases} [x^{(k)} - \varepsilon^{(k)},\ x^{(k)}], & \text{если} \quad f(x^{(k)})h_1 > 0, \\ [x^{(k)},\ x^{(k)} + \varepsilon^{(k)}], & \text{если} \quad f(x^{(k)})h_1 < 0. \end{cases}$$

(S2) *Нахождение вещественного корня* y^k *многочлена* $p^k(x)$ *в интервале*

$$[x^{(k)} - 2\varepsilon^{(k)},\ x^{(k)} + 2\varepsilon^{(k)}] \cap X^{(0)}.$$

Если такого корня нет, то переходим непосредственно к шагу (S5), *положив*

$$\tilde{X}^{(k+1)} := [\tilde{x}_1^{(k+1)},\ \tilde{x}_1^{(k+1)}] = Z^{(k)}.$$

(S3) *Вычисление интервала* $F^{(k)}$, *локализующего значение* $f(y^{(k)})$, *с помощью выражения*

$$F^{(k)} = \frac{K}{r!} \prod_{i=0}^{n} (y^{(k)} - x^{(k-i)})^{m_i}.$$

(S4) *Вычисление улучшенного локализующего интервала по формуле*

$$\tilde{X}^{(k+1)} = \{y^{(k)} - F^{(k)}/H\} \cap Z^{(k)}.$$

(S5) *Нахождение нового приближения*

$$x^{(k+1)} = \left(\tilde{x}_1^{(k+1)} + \tilde{x}_2^{(k+1)}\right)/2,$$

нового значения

$$\varepsilon^{(k+1)} = \left(\tilde{x}_2^{(k+1)} - \tilde{x}_1^{(k+1)}\right)/2$$

и нового интервала

$$X^{(k+1)} = \left[x^{(k+1)} - \varepsilon^{(k+1)},\ x^{(k+1)} + \varepsilon^{(k+1)}\right] = \tilde{X}^{(k+1)}.$$

(Если оказывается, что некоторые из точек $x^{(k+1)}$, $x^{(k)}$, $\ldots$, $x^{(k-n+1)}$ *совпали между собой, то можно взять* $x^{(k+1)} \in \tilde{X}^{(k+1)}$ *по формулам*

$$x^{(k+1)} = \begin{cases} \frac{1}{2}\left(\tilde{x}_1^{(k+1)} + \tilde{x}_2^{(k+1)}\right) + \tilde{\varepsilon}^{(k)}, & \text{если} \quad f\left(x^{(k)}\right) h_1 > 0, \\ \frac{1}{2}\left(\tilde{x}_1^{(k+1)} + \tilde{x}_2^{(k+1)}\right) - \tilde{\varepsilon}^{(k)}, & \text{если} \quad f\left(x^{(k)}\right) h_1 < 0, \end{cases}$$

с подходящим $\tilde{\varepsilon}^{(k)}$, *гарантирующим несовпадение этих точек и соотношение* $x^{(k+1)} \in \tilde{X}^{(k+1)}$. *Такой выбор возможен всегда, когда* $x^{(k-i)} \neq \xi$, $i = 0, 1, \ldots, n$. *Наконец, новое значение* $\varepsilon^{(k+1)}$ *выбирается по формуле*

$$\varepsilon^{(k+1)} = \max\left\{\tilde{x}_2^{(k+1)} - x^{(k+1)},\ x^{(k+1)} - \tilde{x}_1^{(k+1)}\right\},$$

что дает

$$X^{(k+1)} = \left[x^{(k+1)} - \varepsilon^{(k+1)},\ x^{(k+1)} + \varepsilon^{(k+1)}\right] \supset \tilde{X}^{(k+1)}.)$$

Следует отметить, что определение локализующих интервалов $\{X^{(k)}\}$ не использует перемен знака рассматриваемой функции. Это означает, что новый локализующий интервал будет вычислен всегда, даже если несколько последовательно вычисленных значений функции имеют один и тот же знак. Будет также показано, что шаги (S3) и (S4) могут быть пропущены лишь конечное число раз. Не требуется, чтобы локализующий интервал $X^{(i)}$ содержал какое либо из предыдущих приближений, кроме $x^{(i)}$. Свойства определенного выше алгоритма собраны в следующей теореме.

Теорема 6. *Пусть* f *— вещественная функция, имеющая нуль* ξ, *для которого задан локализующий интервал*

$$X^{(0)} = \left\{x \mid \left| x^{(0)} - x \right| \leqslant \varepsilon^{(0)}\right\} \ni \xi,$$

такой что

$$\xi \neq x_1^{(0)}, \quad \xi \neq x_2^{(0)} \quad \left(X^{(0)} = \left[x_1^{(0)},\ x_2^{(0)}\right]\right).$$

Пусть далее для ξ *заданы попарно различные приближения*

$$x^{(-n)},\ x^{(-n+1)},\ \ldots,\ x^{(0)} \in X^{(0)},$$

и производные функции f *удовлетворяют условиям*

$$f'(x) \in H, \text{ где } 0 \notin H, \ x \in X^{(0)}$$

и

$$f^{(r)}(x) \in K, \quad x \in X^{(0)}$$

для интервалов H, K и всех x в интервале $X^{(0)}$ *при заданных неотрицательных параметрах*

$$m_0, \ m_1 \ \ldots, \ m_n, \ \text{где } r = \sum_{i=0}^{n} m_i, \quad m_0 m_n > 0.$$

Тогда для итераций, заданных шагами (S1—S5), *верны следующие утверждения:*

$$\xi \in X^{(k)}, \quad k \geqslant 0, \tag{19}$$

$$X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots \ u \lim_{k \to \infty} X^{(k)} = \xi \tag{20}$$

или после конечного числа шагов последовательность стабилизируется на точке $[\xi, \xi]$.

R-порядок итераций (S1) — (S5) *(см. приложение* А) (21)

равен $O_R((S1) - (S5), \xi) \geqslant s$, *где s — единственный положительный корень многочлена*

$$p(s) = s^{n+1} - \sum_{i=0}^{n} m_i s^{n-i}.$$

Доказательство (19): В силу (S1) мы имеем $\xi \in Z^{(k)}$. Остаточный член интерполяционной формулы Эрмита дает

$$f(x) = p_k(x) + \frac{f^{(r)}(\eta)}{r!} \prod_{i=0}^{n} (x - x^{(k-i)})^{m_i},$$

где число η лежит в интервале, образованном точками

$$x, \ x^{(k)}, \ \ldots, \ x^{(k-n)},$$

лежащими в $X^{(k-n)}$. Так как $p_k(y^{(k)}) = 0$, получаем, используя свойство включения интервальной арифметики, что

$$f(y^{(k)}) \in \frac{K}{r!} \prod_{i=0}^{n} (y^{(k)} - x^{(k-i)})^{m_i} = F^{(k)}.$$

Так как (S4)—шаг упрощенного метода Ньютона, мы имеем $\xi \in X^{(k+1)}$, и потому $\xi \in X^{(k+1)}$.

(20): Непосредственно следует из определения шага (S1) вместе с (19) и определением шага (S5), где $X^{(k+1)}$ всегда можно выбрать таким образом, что

$$d(X^{(k+1)}) \leqslant c \cdot d(X^{(k)}),$$

где $(1/2) \leq c < 1$.

(21): Можно показать, что всегда имеется такой индекс k^1, что для всех шагов с номерами $i \geq k_1$ имеется вещественный корень в интервале $[x^{(k)} - 2\varepsilon^{(k)},\ x^{(k)} + 2\varepsilon^{(k)}]$. Доказательство этого утверждения основано на оценке, показывающей, что если $\varepsilon^{(k)}$ достаточно мало, то $p_k(x)$ меняет знак либо между $x^{(k)}$ и $x^{(k)} + 2\varepsilon^{(k)}$, либо между $x^{(k)}$ и $x^{(k)} - 2\varepsilon^{(k)}$. Таким образом, при оценке порядка сходимости мы всегда можем считать, что шаги (S3) и (S4) выполняются. Из (S4) получаем оценку ширины

$$d(\tilde{X}^{(k+1)}) \leqslant d(y^{(k)} - F^{(k)}/H) = d(F^{(k)}/H).$$

Используя (S3) и эту оценку, мы получаем

$$d(\tilde{X}^{(k+1)}) \leqslant \frac{1}{r!} \prod_{j=0}^{n} |y^{(k)} - x^{(k-j)}|^{m_j} d\left(\frac{K}{H}\right).$$

Так как имеет место соотношение

$$y^{(k)} \in [x^{(k-j)} - 2\varepsilon^{(k-j)},\ x^{(k-j)} + 2\varepsilon^{(k-j)}],\quad 0 \leqslant j \leqslant n,$$

получаем, наконец, что

$$d(\tilde{X}^{(k+1)}) \leqslant \tilde{c} \prod_{j=0}^{n} d(X^{(k-j)})^{m_j}$$

и

$$d(X^{(k+1)}) \leqslant c \prod_{j=0}^{n} d(X^{(k-j)})^{m_j}.$$

Окончательный результат получаем, применив теорему 3 из приложения А.

Обсудим теперь еще одно модифицированное семейство методов локализации. Оно получается путем модификации шага (S2) в случае, когда не существует нужного корня $y^{(k)}$. Чтобы сделать возможным выполнение шага (S4), необходимо тогда определить $y^{(k)}$ иначе. Это делается так.

(S2') *Нахождение вещественного корня $y^{(k)}$ многочлена $p_k(x)$ в интервале $[x^{(k)} - 2\varepsilon^{(k)},\ x^{(k)} + 2\varepsilon^{(k)}] \cap \bar{X}^{(0)}$. Если такого корня не существует, то полагаем*

$$y^{(k)} = \begin{cases} x^{(k)} - \varepsilon^{(k)}, & \text{если} \quad f(x^{(k)})\, h_1 > 0, \\ x^{(k)} + \varepsilon^{(k)}, & \text{если} \quad f(x^{(k)})\, h_1 < 0. \end{cases}$$

(S3') *Вычисление локализующего интервала $F^{(k)}$ для величины $f(y^{(k)})$ с помощью выражения*

$$F^{(k)} = \begin{cases} \dfrac{K}{r!} \prod_{l=0}^{n} (y^{(k)} - x^{(k-l)})^{m_l}, & \text{если } p_k(y^{(k)}) = 0, \\ p_k(y^{(k)}) + \dfrac{K}{r!} \prod_{l=0}^{n} (y^{(k)} - x^{(k-l)})^{m_l} & \text{в противном случае.} \end{cases}$$

Непосредственно очевидно, что теорема 6 истинна без каких-либо изменений и для итераций (SI), (S2'), (S3'), (S4), (S5). Дальнейшие предложенные модификациитносятся к шагу (S1).

(S1') *Построение единственного интерполяционного многочлена Эрмита*

$$p_k(x) = p_{(m_0, m_1, \ldots, m_n)}(x; x^{(k)}, \ldots, x^{(k-n)}),$$

удовлетворяющего условиям

$$p_k^{(j)}(x^{(k-i)}) = f^{(j)}(x^{(k-i)}), \quad 0 \leqslant i \leqslant n, \quad 0 \leqslant j \leqslant m_i - 1.$$

(Полагаем $f^{(0)}=f$ и считаем, что при $m_i=0$ условия в соответствующей точке отсутствуют.) Теперь вычисляем интервал $Z^{(k)} \subset X^{(k)}$ по формуле

$$Z^{(k)} = \{x^{(k)} - f(x^{(k)})/H\} \cap X^{(k)}.$$

Очевидно, что теорема 6 верна без изменений также и для итераций (S1'), (S2'), (S3'), (S4), (S5).

Среди рассмотренных выше методов вычисления нулей содержался при $n=1$ и $m_0 = m_1 = 1$ интервальный вариант метода секущих. В случае $n = 2$ и $m_0 = m_1 = m_2 = 1$ получаем интервальный вариант метода Мюллера.

Рассмотрим пример применения интервального метода секущих.

Пример. Рассмотрим функцию

$$f(x) = 2xe^{-5} + 1 - 2e^{-5x}$$

на интервале $X^{(0)} = [0, 1]$. Интервальный метод секущих в применении к этой формуле порождает приближения, приведенные в табл. 6.

Таблица 6

$X^{(k)}$
[0.0000000000000, 1.0000000000000]
[0.0000000000000, 0.2703167011351]
[0.1351583505675, 0.1417860581860]
[0.1380542457667, 0.1382588014849]
[0.1382505159257, 0.1382572086567]
[0.1382571542348, 0.1382572086567]
[0.1382571550288, 0.1382571550581]
[0.1382571550553, 0.1382571550581]

Кроме описанного выше интервального метода секущих, существует так называемый интервальный метод regula falsi (ложного основания) Этот метод предполагает примерно те же свойства функции *f,* что и интервальный метод секущих. Он также оказывается интерполяционным методом. В противоположность описанным ранее методам он использует разделенные разности Ньютона. Опишем его кратко.

Предположим, что функция f дважды непрерывно дифференцируема в интервале X и имеет в X единственный и притом простой нуль ξ Далее, мы имеем интервалы H, K, удовлетворяющие условиям

$$f'(x) \in H, \quad x \in H, \quad \text{где} \quad 0 \notin H,$$
$$f''(x) \in K, \quad x \in X.$$

Интервальный *метод regula falsi,* сокращенно RF, описывается теперь так:

$$X^{(0)} = X, \; x^{(0)} = m(X^{(0)}) \;\; (\text{середина интервала } X^{(0)}),$$
$$X^{(1)} = \{x^{(0)} - f(x^{(0)})/H\} \cap X^{(0)},$$
$$\text{RF} \begin{cases} x^{(k)} = m(X^{(k)}) \;\; (\text{середина интервала } X^{(k)}), \\ Z^{(k+1)} = \{x^{(k)} - f(x^{(k)})/H\} \cap X^{(k)}, \\ X^{(k+1)} = \begin{cases} \left\{x^{(k)} - \dfrac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)}) + \dfrac{1}{2}K(Z^{(k+1)} - x^{(k)})\right.\right. \\ \left.\left. \times (Z^{(k+1)} - x^{(k-1)})\right)\right\} \cap Z^{(k+1)}, \quad \text{если} \quad f(x^{(k)}) \neq 0, \\ Z^{(k+1)} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{в противном случае} \end{cases} \end{cases}$$

при $k \geq 1$.

Свойства алгоритма RF резюмированы в следующем утверждении.

Теорема 7. *Пусть дважды дифференцируемая функция f имеет простой нуль ξ в интервале X. Допустим далее, что выполнены условия*

$$f'(x) \in H, \quad x \in X \quad и \quad 0 \notin H,$$
$$f''(x) \in K, \quad x \in X.$$

Тогда последовательность $\{X^{(k)}\}$, вычисленная согласно процедуре RF, *либо удовлетворяет условиям*

$$\xi \in X^{(k)}, \quad k \geqslant 0, \tag{22}$$

$$X^{(0)} \supset X^{(1)} \supset X^{(2)} \supset \ldots \; и \lim_{k \to \infty} X^{(k)} = \xi, \tag{23}$$

либо стабилизируется на точке $[\xi, \; \xi]$ через конечное число шагов.

Для некоторого $\gamma \geqslant 0$ имеет место соотношение

$$d(X^{(k+1)}) \leqslant \gamma d(X^{(k)}) d(X^{(k-1)}), \tag{24}$$

т. е. (см. приложение А)

$$O_R((\mathrm{RF}), \xi) \geqslant \frac{1}{2}(1+\sqrt{5}).$$

Доказательство. (22): Доказывается с помощью математической индукции по *k*. Для $k = 1$ утверждение $\xi \in X^{(1)}$ очевидно. Допустим теперь, что для фиксированного *k* мы имеем $x^{(k)} = m(X^{(k)}) \neq \xi$ и $\xi \in X^{(k)}$. Тогда имеет место $\xi \in Z^{(k+1)}$ и $x \neq x_{k-1}$. Рассмотрение интерполяционной формулы Ньютона показывает, что

$$f(\xi) = f(x^{(k)}) + \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}(\xi - x^{(k)}) + \frac{1}{2} f''(\eta)(\xi - x^{(k)})(\xi - x^{(k-1)}),$$

где η находится в интервале, образованном точками $x^{(k)}$, $x^{(k-1)}$ и ξ. Из $f(\xi) = 0$ следует, что

$$\xi = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)}) + \frac{f''(\eta)}{2}(\xi - x^{(k)})(\xi - x^{(k-1)})\right).$$

Из предположения $f''(\eta) \in K$ и того, что $\xi \in Z^{(k+1)}$, а также свойства включения для интервальной арифметики следует, что

$$\xi \in x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)}) + \frac{K}{2}(Z^{(k+1)} - x^{(k)})(Z^{(k+1)} - x^{(k-1)})\right),$$

откуда в силу $\xi \in Z^{(k+1)}$ мы получаем $\xi \in X^{(k+1)}$.

(23): Следует из построения $Z^{(k)}$ и того, что $X^{(k)} \subset Z^{(k)}$.

(24): Пусть $x^{(k)} \neq \xi$. Тогда мы получаем

$$d(X^{(k+1)}) \leqslant d\left(x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)}) + \frac{1}{2}K(Z^{(k+1)} - x^{(k)})(Z^{(k+1)} - x^{(k-1)})\right)\right)$$

$$= d\left(-\frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} \cdot \frac{1}{2}K(Z^{(k+1)} - x^{(k)})(Z^{(k+1)} - x^{(k-1)})\right).$$

Из

$$\frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} = (f'(\tau))^{-1} \in \frac{1}{H}, \quad \tau \in X$$

следует, что

$$d(X^{(k+1)}) \leqslant \frac{1}{2} d((K/H)(Z^{(k+1)} - x^{(k)})(Z^{(k+1)} - x^{(k-1)})).$$

Ввиду

$$X^{(k-1)} \supset X^{(k)} \supset Z^{(k+1)} \supseteq X^{(k+1)},$$

получаем, наконец,

$$d(X^{(k+1)}) \leqslant \frac{1}{2} d((K/H)(X^{(k)} - x^{(k)})(X^{(k-1)} - x^{(k-1)})).$$

Интервал $A \in I(\mathbb{R})$ с центром $m(A)$ удовлетворяет условию

$$A - m(A) = [-d(A)/2,\ d(A)/2],$$

откуда следует соотношение

$$(X^{(k)} - x^{(k)})(X^{(k-1)} - x^{(k-1)}) = [-d(X^{(k)})\,d(X^{(k-1)})/4,\ d(X^{(k)})\,d(X^{(k-1)})/4].$$

В применении к предыдущему неравенству оно дает

$$d(X^{(k+1)}) \leqslant \frac{1}{2} d((K/H)[-d(X^{(k)})\,d(X^{(k-1)})/4,\ \ d(X^{(k)})\,d(X^{(k-1)})/4]),$$

откуда получается

$$d(X^{(k+1)}) \leqslant \frac{1}{2} |K/H| \cdot 2 \cdot \frac{1}{4} d(X^{(k)})\,d(X^{(k-1)})$$

$$= \gamma\, d(X^{(k)})\,d(X^{(k-1)}),$$

где $\gamma = \frac{1}{4} |K/H|$. Из теоремы 3 (приложение А) мы получаем затем требуемое неравенство для *R*-порядка.

Используя для данной функции *f* разделенные разности более высокого порядка, мы можем построить новые методы, порядок сходимости которых тоже лежит между 1 и 2 и которые требуют лишь по одному новому значению функции на каждый шаг итерации.

Можно также построить интервальные варианты некоторых методов, используя интервальный метод regula falsi (RF). Эти методы имеют порядок сходимости выше первого, хотя они используют значения только самой функции *f*. Мы опишем здесь такой метод, представляющий собой прямое обобщение интервального метода regula falsi.

Снова предполагается, что функция *f* имеет простой нуль ξ в интервале *X*, а интервалы *H*, *K* удовлетворяют условиям теоремы 7. Задан параметр *p* — целое число ≥ 1. Теперь *параметрический метод* regula falsi, короче *p*-RF, формулируется следующим образом:

$$
p\text{-RF}\left\{\begin{array}{l}
X^{(0)}=X,\quad x^{(0)}=m(X^{(0)}),\\
X^{(1)}=\{x^{(0)}-f(x^{(0)})/H\}\cap X^{(0)}.\\
\text{Для } k\geqslant 1 \text{ вычисляем приближения по следующим формулам:}\\
x^{(k)}=m(X^{(k)}) \text{ (середина интервала } X^{(k)}),\\
X^{(k+1,\,0)}=\{x^{(k)}-f(x^{(k)})/H\}\cap X^{(k)},\\
X^{(k+1,\,1)}=\begin{cases}\left\{x^{(k)}-\dfrac{x^{(k)}-x^{(k-1)}}{f(x^{(k)})-f(x^{(k-1)})}\left(f(x^{(k)})\right.\right.\\
\left.\left.+\dfrac{1}{2}K(X^{(k+1,\,0)}-x^{(k)})(X^{(k+1,\,0)}-x^{(k-1)})\right)\right\}\\
\cap X^{(k+1,\,0)}, \quad \text{если} \quad f(x^{(k)})\neq 0,\\
X^{(k+1,\,0)} \qquad \text{в противном случае.}\end{cases}\\
\text{После этого производим вычисления для } i=2,\ 3,\ \ldots,\ p \text{ (только для } p>1).\\
z^{(i)}=m(X^{(k+1,\,i-1)}),\\
X^{(k+1,\,i)}=\begin{cases}\left\{z^{(i)}-\dfrac{z^{(i)}-x^{(k)}}{f(z^{(i)})-f(x^{(k)})}\left(f(z^{(i)})\right.\right.\\
\left.\left.+\dfrac{1}{2}K(X^{(k+1,\,i-1)}-z^{(i)})(X^{(k+1,\,i-1)}-x^{(k)})\right)\right\}\\
\cap X^{(k+1,\,i-1)}, \text{ если } f(x^{(k)})\neq 0,\\
X^{(k+1,\,i-1)} \qquad \text{в противном случае,}\end{cases}\\
X^{(k+1)}=X^{(k+1,\,p)}.
\end{array}\right.
$$

Метод *p-RF* обладает следующими свойствами.

Теорема 8. *Пусть функция f дважды непрерывно дифференцируема в интервале X и имеет там нуль* ξ. *Пусть выполнены условия*

$$f'(x)\in H,\quad x\in H,\quad где\quad 0\notin H,$$
$$f''(x)\in K,\quad x\in X.$$

Тогда последовательность $\{X^{(k)}\}$, *вычисленная по формулам p-RF, удовлетворяет для* $p\geq 1$ *условиям*

$$\xi\in X^{(k)},\quad k\geqslant 0, \tag{25}$$

$$X^{(0)}\supset X^{(1)}\supset X^{(2)}\supset\ldots,\ \text{где}\ \lim_{k\to\infty}X^{(k)}=\xi, \tag{26}$$

или стабилизируется через конечное число шагов на точке $[\xi,\ \xi]$.

Для некоторого $\gamma\geq 0$ *справедлива оценка* (27)

$$d(X^{(k+1)})\leqslant\gamma d(X^{(k)})^p\,d(X^{(k-1)})$$

и

$$O_R((p\text{-RF}),\ \xi)\geqslant\frac{1}{2}(p+\sqrt{p^2+4}).$$

Доказательство. При $p = 1$ теорема 8 сводится к теореме 7, поэтому мы предполагаем далее, что $p \geq 2$.

Установим соотношение (25). Мы наметим доказательство методом математической индукции. Доказываемое утверждение очевидно для $k = 0, 1$. Мы допустим, что оно верно для фиксированного k и докажем его для $k+1$. Из нашего индукционного предположения следует, что

$$\xi \in X^{(k+1,\,0)} \text{ и } X^{(k+1,\,0)} = [\xi, \xi] \text{ для } x^{(k)} = \xi.$$

В случае когда $x^{(k)} = \xi$, мы имеем $f(x^{(k)}) = 0$, откуда следует, что

$$X^{(k+1,\,i)} = [\xi, \xi]$$

для $i = 2, 3, \ldots, p$. Поэтому имеем

$$\xi \in X^{(k+1)} = X^{(k+1,\,p)} = [\xi, \xi].$$

Если теперь $x^{(k)} \neq \xi$, то $f(x^{(k)}) \neq 0$ и получаем соотношение

$$0 = f(\xi) = f(x^{(k)}) + \frac{f(x^{(k)}) - f(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}(\xi - x^{(k)}) + \frac{1}{2} f''(\eta)(\xi - x^{(k)})(\xi - x^{(k-1)}),$$

где $x^{(k)} \neq x^{(k-1)}$, $x^{(k-1)} \neq \xi$.

Тем же методом, что и в доказательстве теоремы 7, отсюда получаем, что

$$\xi \in x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})}\left(f(x^{(k)}) + \frac{1}{2} K (X^{(k+1,\,0)} - x^{(k)}) \times (X^{(k+1,\,0)} - x^{(k-1)})\right)$$

и окончательно $\xi \in X^{(k+1,1)}$. Остается показать что $\xi \in X^{(k+1,i)}$, $i=2,3,\ldots,p$.

Для $x^{(k)} \neq \xi$ и $z^{(i)} \neq \xi$ снова показываем, используя остаточный член интерполяционной формулы Ньютона, что из $\xi \in X^{(k+1,i-1)}$ всегда следует $\xi \in X^{(k+1,i)}$.

Это снова очевидно для $x^{(k)} \neq \xi$ и $z^{(i)} = \xi$. Так как $\xi \in X^{(k+1,\,1)}$, мы получаем

$$\xi \in X^{(k+1,\,i)}, \quad i = 1, 2, \ldots, p,$$

а потому и

$$\xi \in X^{(k+1)} = X^{(k+1,\,p)}.$$

Соотношение (26) немедленно следует из формул, по которым вычисляется $X^{(k+1,0)}$. Мы не приводим здесь элементарного обоснования этого факта.

(27): Не умаляя общности, предположим, что $x^{(k)} \neq \xi$. Тогда из определения процедуры *p*-RF немедленно следует, что

$$d(X^{(k+1,\,0)}) < \frac{1}{2} d(X^{(k)}).$$

Аналогично тому, как доказано утверждение (24) в теореме (7), получаем, что

$$d(X^{(k+1,\,1)}) \leqslant \frac{1}{4} |K/H| d(X^{(k)}) d(X^{(k-1)})$$

и аналогичным образом

$$d(X^{(k+1,\,i)} \leqslant \frac{1}{4} |K/H| d(X^{(k+1,\,i-1)}) d(X^{(k)}), \quad i = 2,\ 3,\ \ldots,\ p$$

Эта простая рекурсия дает соотношение

$$d(X^{(k+1,\,i)}) \leqslant \beta_i d(X^{(k)})^i d(X^{(k-1)}),$$

$$\beta_i = \left(\frac{1}{4} |K/H|\right)^i, \quad i = 1,\ 2,\ \ldots,\ p.$$

Используя равенство $X^{(k+1)} = X^{(k+1,\,p)}$, получаем соотношение

$$d(X^{(k+1)}) \leqslant \beta_p d(X^{(k)})^p d(X^{(k-1)}),$$

из которого снова получаем требуемое неравенство для *R*-порядка с помощью теоремы 3 из приложения А.

Замечания. Были предложены и исследовались различные модификации методов (3) и (11). Процедура, аналогичная (3'), используется для уточнения оценок всех нулей в предписанном интервале $\left[x_1^{(0)},\ x_2^{(0)}\right]$. Уточнения нижней границы $x_1^{(k+1)}$ вычисляются отдельно от уточнений верхней границы $x_2^{(k+1)}$. Для этой процедуры требуются как значение функции $f(x_1^{(k)})$ (соответственно $f\left(x_2^{(k)}\right)$), так и оценка $m=|\mathrm{M}|$ величины $|f(x)|$ для $x \in \left[x_1^{(k)},\ x_2^{(k)}\right]$. Тогда последовательность нижних границ $x_1^{(0)},\ x_1^{(1)},\ldots$ сходится к наименьшему нулю, принадлежащему интервалу $\left[x_1^{(0)},\ x_2^{(0)}\right]$. Соответственно последовательность верхних границ сходится к наибольшему нулю функции $f(x)$ в $\left[x_1^{(0)},\ x_2^{(0)}\right]$. Существует аналогичный метод для многочленов.

Используется неявное представление метода (3) (соответственно (11)). Функция $f(x)$ локализуется интервальным выражением

$$I(x) = f\left(x^{(k)}\right) + \left(x - x^{(k)}\right) M^{(k)}, \quad x^{(k)} \in \left[x_1^{(k)},\ x_2^{(k)}\right] = X^{(k)},$$

где $M^{(k)}$ определяется аналогично (9). Тогда интервалы $\left[x_1^{(k+1)},\ x_2^{(k+1)}\right]$ вычисляются с помощью требования

$$\left[x_1^{(k+1)},\ x_2^{(k+1)}\right] := \left\{x \mid x \in X^{(k)},\ 0 \in I(x)\right\} \cap X^{(k)}.$$

При этом возникают различные случаи в зависимости от того, верно ли, что $0 \in M^{(k)}$. Этот метод сходится при определенных условиях, и

если имеется несколько нулей, то возникает несколько подпоследовательностей. Найдена рекурсивная процедура для этого случая. Этот метод был эскизно описан в конце разд. С. Он был применен к нахождению нулей производной дважды непрерывно дифференцируемой функции на интервале, что используется при нахождении глобального минимума функции.

Приведенные в теореме 1 (соответственно в теореме 4) результаты о методе (3) (соответственно о методе (11)) можно обобщить следующим образом. Предполагается, что для функции f на интервале $X^{(0)}$ имеется интервал M, такой что $0 \notin M$ и

$$\frac{f(x) - f(y)}{x - y} \in M \quad \text{для} \quad x, y \in X^{(0)}, \quad x \neq y. \tag{28}$$

Если

$$X^{(1)} = m(X^{(0)}) - f(m(X^{(0)}))/M \in X^{(0)},$$

то найдется $\xi \in X^{(1)}$, такое что $f(\xi) = 0$. Это будет в том случае, если мы предположим, не умаляя общности, что $f(m(X^{(0)})) > 0$, $m_1 > 0$, а затем $f(x_1^{(1)}) > 0$. Из этого предположения получается противоречие, если рассмотреть

$$x_1^{(1)} > m(X^{(0)}) - \frac{f(m(X^{(0)}))}{(f(m(X^{(0)})) - f(x_1^{(1)}))/(m(X^{(0)}) - x_1^{(1)})} \geqslant x_1^{(1)}.$$

Поэтому мы должны иметь $f(x_1^{(1)}) \leqslant 0$, откуда следует, что $X^{(0)}$, а тогда в силу теоремы 1 также и $X^{(1)}$ содержит нуль ξ.

Мы хотим теперь, используя средства интервальной арифметики, дать короткое доказательство того, что уравнение $f(x) = 0$ имеет корень в интервале $X^{(0)} = [x^{(0)} - r, x^{(0)} + r]$. Предположим, что f дважды непрерывно дифференцируема и что

$$|f''(x)| \leqslant \gamma, \quad x \in X^{(0)}.$$

Допустим еще, что $f'(x^{(0)}) \neq 0$, и положим

$$\left| \frac{f(x^{(0)})}{f'(x^{(0)})} \right| = \eta, \quad \left| \frac{1}{f'(x^{(0)})} \right| = \beta.$$

Тогда для любого $y \in X^{(0)}$, $x^{(0)} \neq y$ и некоторого θ, лежащего между $x^{(0)}$ и y, имеем

$$\frac{f(x^{(0)}) - f(y)}{x^{(0)} - y} = f'(x^{(0)}) + \frac{1}{2} f''(\theta)(y - x^{(0)})$$

$$\in f'(x^{(0)}) + \frac{1}{2} \gamma [-r, r] =: M^{(0)}.$$

Если теперь $0 \notin M^{(0)}$ и

$$X^{(1)} = x^{(0)} - f(x^{(0)})/M^{(0)} \subset X^{(0)},$$

то мы показываем тем же методом, что и раньше, что имеется нуль $\xi \in X^{(1)} \subseteq X^{(0)}$. Требование $X^{(1)} \subseteq X^{(0)}$ выполнено тогда и только тогда, когда

$$\frac{1}{2}\beta\gamma r^2 - r + \eta \leqslant 0.$$

Это эквивалентно неравенству

$$\beta\gamma\eta \leqslant \frac{1}{2} \tag{29}$$

вместе с

$$(1 - \sqrt{1 - 2\beta\gamma\eta})/\beta\gamma \leqslant r \leqslant (1 + \sqrt{1 - 2\beta\gamma\eta})/\beta\gamma. \tag{30}$$

Неравенства (29) и (30) гарантируют выполнение условий теоремы Канторовича о существовании нуля в интервале $X^{(0)}$.

Если $X^{(1)} \cap X^{(0)} = \varnothing$, то f не имеет нулей в $X^{(0)}$. Условие $X^{(1)} \cap X^{(0)} = \varnothing$ выполнено тогда и только тогда, когда

$$\frac{1}{2}\beta\gamma r^2 + r - \eta < 0,$$

т. е. при $\eta \neq 0$ тогда и только тогда, когда

$$0 \leqslant r < (-1 + \sqrt{1 + 2\beta\gamma\eta})/\beta\gamma.$$

Это уверждение может быть аналогичным образом использовано для доказательства теорем об исключении в банаховых пространствах.

2.2. Методы одновременной локализации вещественных корней многочленов

В этом разделе мы рассмотрим интервальные методы ньютоновского типа для вычисления интервалов, локализующих все вещественные корни вещественного многочлена. Сначала рассматривается случай, когда все корни вещественны. Комплексные корни рассматриваются дальше. Для случая, когда все корни вещественные и простые, строим короткошаговый метод, сходящийся быстрее, чем квадратично. В качестве приложения используем этот метод для вычисления всех собственных значений симметрической трехдиагональной матрицы.

Дан вещественный многочлен

$$p(x) = a^{(n)}x^n + a^{(n-1)}x^{n-1} + \ldots + a^{(0)} \tag{1}$$

и мы предполагаем в дальнейшем, что

$$a^{(n)} = 1.$$

Далее предполагается, что этот многочлен имеет n вещественных корней $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(n)}$, которые собраны в вектор $(\xi^{(i)})$, причем кратные корни выписаны столько раз, какова их кратность Предполагается, что для всех корней известны локализующие интервалы

$$\xi^{(j)} \in X^{(0,\,j)} = \left[x_1^{(0,\,j)},\ x_2^{(0,\,j)}\right], \quad 1 \leqslant j \leqslant n.$$

Предположим сначала, что все эти локализующие интервалы попарно не пересекаются, т. е.

$$X^{(0,\,j)} \cap X^{(0,\,k)} = \varnothing, \qquad 1 \leqslant j < k \leqslant n. \qquad (2)$$

Многочлен $p(x)$ можно записать в виде

$$p(x) = \prod_{j=1}^{n} (x - \xi^{(j)})$$

или

$$p(x) = (x - \xi^{(i)}) \prod_{j=1,\ j \neq i}^{n} (x - \xi^{(j)}),$$

откуда следует

$$\xi^{(i)} = x - p(x) \Big/ \prod_{j=1,\ j \neq i}^{n} (x - \xi^{(j)}).$$

Если мы выберем $x = x^{(0,\,i)} \in X^{(0,\,i)}$, то получим, что

$$0 \notin \prod_{j=1,\ j \neq i}^{n} (x^{(0,\,i)} - X^{(0,\,j)}),$$

и из (9 п.1.1) следует, что

$$\xi^{(i)} \in X^{(1,\,i)} = \left\{ x^{(0,\,i)} - p(x^{(0,\,i)}) \Big/ \prod_{j=1,\ j \neq i}^{n} (x^{(0,\,i)} - X^{(0,\,j)}) \right\} \cap X^{(0,\,i)}.$$

Таким образом, интервальное выражение, стоящее в правой части этого равенства, задает новый локализующий интервал $X^{(1,i)}$, для которого верно

$$\xi^{(i)} \in X^{(1,\,i)} \subseteq X^{(0,\,i)}.$$

Это соотношение порождает следующую итерационную схему:

$$X^{(k+1,\,i)} = \left\{ x^{(k,\,i)} - p(x^{(k,\,i)}) \Big/ \prod_{j=1,\ j \neq i}^{n} (x^{(k,\,i)} - X^{(k,\,j)}) \right\} \cap X^{(k,\,i)}, \quad (3)$$

где

$$x^{(k,\,i)} \in X^{(k,\,i)}, \qquad 1 \leqslant i \leqslant n, \qquad k \geqslant 0.$$

Для интервального выражения в знаменателе вводится сокращение

$$Q^{(k,\,i)} = \prod_{j=1,\ j \neq i}^{n} (x^{(k,\,i)} - X^{(k,\,j)}).$$

Итерационная схема (3) задает полношаговый метод локализации корней многочлена $\xi^{(i)}$, $1 \leqslant i \leqslant n$. Если в $Q^{(k,\,i)}$ мы всегда используем последнее вычисленное значение локализующих интервалов, то получаем

$$R^{(k,\,i)} = \prod_{j=1}^{i-1} (x^{(k,\,i)} - X^{(k+1,\,j)}) \prod_{j=i+1}^{n} (x^{(k,\,i)} - X^{(k,\,j)}),$$

что приводит к соответствующей короткошаговой итерации. Теперь мы хотим провести для этой короткошаговой итерации такие же рассуждения, как для метода (13 микромодуль 26). Локализующий интервал $X^{(k+1,\,i)}$ урезается до $Y^{(k+1,\,i)}$ в зависимости от знаков выражений $p(x^{(k+1,\,i)})$ и $R^{(k,\,i)}$. Функция sign для интервалов определяется соотношением

$$\operatorname{sign}(X) = \begin{cases} 1, & \text{если} \quad x_1 > 0, \\ -1, & \text{если} \quad x_2 < 0, \\ 0 & \text{в противном случае.} \end{cases} \tag{4}$$

Интервалы $Y^{(k+1,\,i)}$, также содержащие корни $\xi^{(i)}$, определяются тогда следующим образом:

$$Y^{(k+1,\,i)} = \begin{cases} [x_1^{(k+1,\,i)},\ x^{(k+1,\,i)}], & \text{если } \operatorname{sign}(R^{(k,\,i)})\operatorname{sign}(p(x^{(k+1,\,i)})) > 0, \\ [x^{(k+1,\,i)},\ x_2^{(k+1,\,i)}], & \text{если } \operatorname{sign}(R^{(k,\,i)})\operatorname{sign}(p(x^{(k+1,\,i)})) < 0, \\ X^{(k+1,\,i)} & \text{в противном случае.} \end{cases}$$

Заметим, что всегда имеет место соотношение

$$\operatorname{sign}(R^{(0,\,i)}) = \operatorname{sign}(R^{(1,\,i)}) = \ldots, \qquad 1 \leqslant i \leqslant n.$$

Используя только что введенные новые локализующие интервалы, вычисляем теперь новое значение знаменателя с помощью выражения

$$S^{(k+1,\,i)} = \prod_{j=1}^{i-1} (x^{(k+1,\,i)} - Y^{(k+2,\,j)}) \cdot \prod_{j=i+1}^{n} (x^{(k+1,\,i)} - Y^{(k+1,\,j)}).$$

Применяя его, приходим к следующему модифицированному короткошаговому методу:

$$
\begin{cases}
Y^{(0,\,i)} = X^{(0,\,i)}, \quad x^{(0,\,i)} \in X^{(0,\,i)}, \\
X^{(k+1,\,i)} = \{x^{(k,\,i)} - p(x^{(k,\,i)})/S^{(k,\,i)}\} \cap X^{(k,\,i)}, \quad \text{где} \\
S^{(k,\,i)} = \prod_{j=1}^{i-1} (x^{(k,\,i)} - Y^{(k+1,\,j)}) \prod_{j=i+1}^{n} (x^{(k,\,i)} - Y^{(k,\,j)}), \\
x^{(k+1,\,i)} \in X^{(k+1,\,i)}, \\
Y^{(k+1,\,i)} = \begin{cases} [x_1^{(k+1,\,i)},\ x^{(k+1,\,i)}], \\ \qquad \text{если} \quad \operatorname{sign}(S^{(k,\,i)}) \operatorname{sign}(p(x^{(k+1,i)})) > 0, \\ [x^{(k+1,\,i)},\ x_2^{(k+1,\,i)}], \\ \qquad \text{если} \quad \operatorname{sign}(S^{(k\ \ i)}) \operatorname{sign}(p(x^{(k+1,\ i)})) < 0, \\ X^{(k+1,\ i)} \quad \text{в противном случае.} \end{cases}
\end{cases} \tag{5}
$$

Оба метода (3), (5) можно считать интервальными вариантами известных методов одновременного нахождения корней вещественных многочленов. Преимущество интервальных вариантов этих методов состоит в том, что они не только дают локализующие интервалы для корней, но и всегда сходятся при сделанных выше допущениях. Это устанавливается в следующей теореме.

Теорема 1. *Пусть дан многочлен* (1), *имеющий n простых корней* $\xi^{(i)}$, $1 \leqslant i \leqslant n$, *причем известны локализующие интервалы* $X^{(0,\,i)} \ni \xi^{(i)}$, *для которых верно* (2). *Тогда последовательность приближений* $\{X^{(k,\,i)}\}_{k=0}^{\infty}$, $1 \leqslant i \leqslant n$, *вычисленная по формулам* (3) *(соответственно* (5)), *либо удовлетворяет условиям*

$$\xi^{(i)} \in X^{(k,\,i)}, \qquad k \geqslant 0,$$

и

$$X^{(0,\,i)} \supset X^{(1,\,i)} \supset X^{(2,\,i)} \supset \ldots, \quad \text{где} \quad \lim_{k \to \infty} X^{(k,\,i)} = \xi^{(i)},$$

либо стабилизируется за конечное число шагов на точке $[\xi^{(i)}, \xi^{(i)}]$.

Теорема 1 получается тем же методом, что и соответствующее утверждение в теореме 1 микромодуль 26, так как каждое из интервальных выражений $Q^{(k,\,i)}$, $S^{(k,\,i)}$ обладает либо свойством (2 микромодуль 26), либо соответствующим свойством в интервале $X^{(k,\,i)}$ для $m_2 < 0$.

Выбирая в обоих методах

$$x^{(k,\,i)} = \frac{1}{2}\left(x_1^{(k,\,i)} + x_2^{(k,\,i)}\right)$$

и рассматривая структуру выражений (3) и (5), немедленно получаем, что ширина локализующего интервала для каждого из корней уменьшается на каждом шаге итерации по крайней мере вдвое.

Теорема 1 частично сохраняется и в случае, когда многочлен имеет кратные корни. Если выпишем эти кратные корни вместе

$$\xi^{(m)},\ \xi^{(m+1)},\ \ldots,\ \xi^{(n)},$$

то оба метода (3), (5) нужно изменить таким образом, чтобы вычисление локализующих интервалов производилось только для индексов $1 \leqslant i < m$. Если эти интервалы для простых корней перевычисляются на каждом шаге, а остальные интервалы остаются неизменными, то теорема 1 сохраняется для простых корней.

Можно обобщить метод (3) таким образом, что в теореме 1 удастся заменить предположение (2), касающееся локализующих интервалов $X^{(0,\,i)}$, $1 \leqslant i \leqslant n$, на более слабое условие. Для этого следует полностью использовать возможность варьирования значения $x^{(k,\,i)} \in X^{(k,\,i)}$, а не применять систематическую процедуру выбора, где это значение полагают равным, например, среднему арифметическому границ интервала.

Теперь рассмотрим подробнее поведение последовательности

$$\{d(X^{(k,\,i)})\}_{k=0}^{\infty}, \qquad 1 \leqslant i \leqslant n.$$

Для метода (3) получаем оценку

$$\begin{aligned} d(X^{(k+1,\,i)}) &\leqslant d(\{x^{(k,\,i)} - p(x^{(k,\,i)})/Q^{(k,\,i)}\}) \\ &= d(p(x^{(k,\,i)})/Q^{(k,\,i)}) = |p(x^{(k,\,i)})|\, d(1/Q^{(k,\,i)}) \end{aligned}$$

с помощью (9 п.1.2), (10 п.1.2) и (14 п.1.2). Так как

$$\begin{aligned} |p(x^{(k,\,i)})| &= |p(x^{(k,\,i)}) - p(\xi^{(i)})| = |(x^{(k,\,i)} - \xi^{(i)})\, p'(\tilde{\eta}^{(k,\,i)})| \\ &\leqslant d(X^{(k,\,i)})\, |p'(\tilde{\eta}^{(k,\,i)})| \leqslant d(X^{(k,\,i)})\, |p'(X^{(0,\,i)})|, \end{aligned}$$

отсюда следует, что

$$d(X^{(k+1,\,i)}) \leqslant d(X^{(k,\,i)})\, |p'(X^{(0,\,i)})|\, d(1/Q^{(k,\,i)}).$$

Применяя теорему 5 микромодуль 24, мы получаем оценку

$$d(1/Q^{(k,\,i)}) \leqslant \gamma^{(k,\,i)} d(Q^{(k,\,i)})$$

и, так как

$$Q^{(k,\,i)} \subseteq \prod_{j=1,\ j \neq i}^{n} (X^{(0,\,i)} - X^{(0,\,j)}),$$

получаем далее

$$d\left(\frac{1}{Q^{(k,\,i)}}\right) \leqslant \gamma^{(i)} d(Q^{(k,\,i)}) = \gamma^{(i)} d\left(\prod_{j=1,\ j \neq i}^{n} (x^{(k,\,i)} - X^{(k,\,j)})\right)$$

с константами $\gamma^{(i)}$, зависящими только от $X^{(0,\,j)}$, $1 \leqslant j \leqslant n$. Многократно применяя (12 п.1.2), мы получаем далее

$$d\left(\frac{1}{Q^{(k,\ i)}}\right)\leqslant\gamma^{(i)}\sum_{j=1,\ j\neq i}^{n}\eta^{(i,\ j)}d\left(X^{(k,\ j)}\right)$$

с подходящими константами $\eta^{(i,j)}$, зависящими только от $X^{(0,j)}$, $1\leq j\leq n$, так как $X^{(k,\ j)}\subseteq X^{(0,\ j)}$ и применимы (3 п.1.2) и (9 п.1.2). Резюмируя все это, получаем следующее неравенство:

$$d\left(X^{(k+1,\ i)}\right)\leqslant|\ p'\left(X^{(0,\ i)}\right)|\,\gamma^{(i)}d\left(X^{(k,\ i)}\right)\sum_{j=1,\ j\neq i}^{n}\eta^{(i,\ j)}d\left(X^{(k,\ j)}\right),\quad(6)$$

$$1\leqslant i\leqslant n.$$

То же самое рассуждение можно провести для метода (5); единственное дополнительное обстоятельство, которое нужно учесть — это соотношение $Y^{(k,\ i)}\subseteq X^{(k,\ i)}$. Это приводит к неравенству

$$d\left(X^{(k+1,\ i)}\right)\leqslant|\ p'\left(X^{(0,\ i)}\right)|\,\gamma^{(i)}d\left(X^{(k\ \ i)}\right)\left(\sum_{j=1}^{i-1}\eta^{(i\ \ j)}d\left(X^{(k+1,\ j)}\right)\right.\quad(7)$$

$$\left.+\sum_{j=i+1}^{n}\eta^{(i,\ j)}d\left(X^{(k,\ j)}\right)\right),\qquad 1\leqslant i\leqslant n.$$

В следующем утверждении оценивается R-порядок методов (3) и (5).

Теорема 2. *В предположениях и обозначениях теоремы* 1 *для R-порядка методов* (3) *и* (5) *имеет место*

$$Q_R((3),\ (\xi^{(i)}))\geqslant 2\quad(8)$$

и

$$Q_R((5),\ (\xi^{(i)}))\geqslant 1+\sigma^{(n)},\quad(9)$$

где $\sigma^{(n)}>1$ — единственный положительный корень многочлена

$$\tilde{q}^{(n)}(y)=y^n-y-1.$$

Доказательство. (8): Из (6) немедленно получаем, что

$$d\left(X^{(k+1,\ i)}\right)\leqslant|\ p'\left(X^{(0,\ i)}\right)|\,\gamma^{(i)}\left(\sum_{j=1,\ j\neq i}^{n}\eta^{(i,\ j)}\right)\left(d^{(k)}\right)^2$$

$$\leqslant\max_{1\leqslant i\leqslant n}\left\{|\ p'\left(X^{(0,\ i)}\right)|\,\gamma^{(i)}\left(\sum_{j=1\ \ j\neq i}^{n}\eta^{(i,\ j)}\right)\right\}\left(d^{(k)}\right)^2$$

$$\leqslant\gamma\left(d^{(k)}\right)^2,\qquad 1\leqslant i\leqslant n,$$

где

$$d^{(k)}=\max_{1\leqslant i\leqslant n}\left\{d\left(X^{(k,\ i)}\right)\right\}.$$

Отсюда следует, что

$$d^{(k+1)}=\max_{1\leqslant i\leqslant n}\left\{d\left(X^{(k+1,\ i)}\right)\right\}\leqslant\gamma\left(d^{(k)}\right)^2.$$

Тогда из теоремы 2 приложения А следует, что

$$O_R((3), (\xi^{(l)})) \geqslant 2$$

(9): Доказательство соотношения (9) требует больших усилий.

Пусть

$$\gamma = \max_{1 \leqslant i, j \leqslant n} \{\eta^{(i, j)} | p'(X^{(0, i)}) | \gamma^{(j)}\}.$$

Тогда мы можем переписать (7) в виде

$$d(X^{(k+1, i)}) \leqslant \gamma d(X^{(k, i)}) \left(\sum_{j=1}^{i-1} d(X^{(k+1, j)}) + \sum_{j=i+1}^{n} d(X^{(k, j)}) \right).$$

Применяя подстановку

$$d(X^{(k, i)}) = \frac{1}{(n-1)\gamma} h^{(k, i)}, \qquad 1 \leqslant i \leqslant n, \qquad \hat{e} = \frac{1}{n-1}$$

это соотношение можно далее переписать в виде

$$h^{(k+1, i)} \leqslant \hat{e} h^{(k, i)} \left(\sum_{j=1}^{i-1} h^{(k+1, j)} + \sum_{j=i+1}^{n} h^{(k, j)} \right).$$

Предположим без потери общности, что

$$h^{(0, i)} \leqslant h < 1, \qquad 1 \leqslant i \leqslant n.$$

Тогда мы имеем

$$h^{(k+1, i)} \leqslant h^{u^{(k+1, i)}}, \qquad 1 \leqslant i \leqslant n, \qquad k \geqslant 0.$$

Вектор $u_p^{(k+1)} = (u^{(k+1, 1)}, \ldots, u^{(k+1, n)})^T$, компоненты которого — целые числа, может быть вычислен по правилу

$$u_p^{(k+1)} = \mathscr{A}_p u_p^{(k)}, \qquad k \geqslant 0,$$

с помощью начального вектора $u_p^{(0)} = (1, 1, \ldots, 1)^T$. (Здесь векторы с вещественными компонентами, называемые точечными векторами, обозначаются буквами a_p, b_p, ... с индексами p (point), чтобы отличить их от интервальных векторов. То же делается для «точечных матриц».

Тогда матрица $\mathscr{A}_p$ имеет вид

$$\mathscr{A}_p = \begin{pmatrix} 1 & 1 & & & & & \\ & 1 & 1 & & 0 & & \\ & & 1 & 1 & & & \\ & & & & & & \\ & 0 & & & & & \\ & & & & & 1 & 1 \\ 1 & 1 & 0 & & \cdot\cdot & 0 & 1 \end{pmatrix}$$

Доказательство этого соотношения может быть проведено методом математической индукции, и мы не приводим его здесь. Матрица $\mathscr{A}_p$ неотрицательна, и ее направленный граф очевидным образом является сильно связным. Отсюда следует, что $\mathscr{A}_p$ неприводима. Тогда из теоремы Перрона — Фробениуса следует, что $\mathscr{A}_p$ имеет собственное число $\lambda^{(1)}$, равное ее спектральному радиусу $\rho(\mathscr{A}_p)$. Матрица $\mathscr{A}_p$ примитивна. Поэтому остальные ее собственные числа удовлетворяют соотношению

$$\lambda^{(1)} = \rho(\mathscr{A}_p) > |\lambda^{(2)}| \geqslant \ldots \geqslant |\lambda^{(n)}|.$$

Так как матрица $\mathscr{A}_p$ примитивна, мы имеем для некотопого натурального числа $k^{(0)}$, что

$$\mathscr{A}_p^k = (a_{ij}^{(k)}) > O_p, \qquad k \geqslant k^{(0)}.$$

Для матриц, обладающих двумя последними свойствами, имеем

$$\lim_{k\to\infty} (a_{ij}^{(k+1)}/a_{ij}^{(k)}) = \lambda^{(1)}.$$

Поэтому для данного $\varepsilon > 0$ верно, что

$$a_{ij}^{(k+1)}/a_{ij}^{(k)} \geqslant \rho(\mathscr{A}_p) - \varepsilon, \qquad k \geqslant k(\varepsilon) \geqslant k^{(0)},$$

или

$$a_{ij}^{(k+1)} \geqslant \alpha(\rho(\mathscr{A}_p) - \varepsilon), \qquad 1 \leqslant i, \qquad j \leqslant n$$

где

$$\alpha = \min_{1\leqslant i\ j\leqslant n} a_{ij}^{(k)} > 0.$$

Отсюда следует, что

$$a_{ij}^{(k+2)} \geqslant a_{ij}^{(k+1)}(\rho(\mathscr{A}_p) - \varepsilon) \geqslant \alpha(\rho(\mathscr{A}_p) - \varepsilon),$$

что дает

$$a_{ij}^{(k+r)} \geqslant \alpha(\rho(\mathscr{A}_r) - \varepsilon)^r, \qquad 1 \leqslant i, \qquad j \leqslant n, \qquad r \geqslant 0.$$

Используя правило вычисления векторов

$$u_p^{(k)},$$

мы получаем тогда

$$u_p^{(k+r)} = \mathscr{A}_p^{k+r} u_p^{(0)} = \left(\sum_{i=1}^{n} a_{ij}^{(k+r)}\right) \geqslant (n\alpha(\rho(\mathscr{A}_p) - \varepsilon)^r)\, e_p,$$

где $e_p = (1, 1, \ldots, 1)'$. Поэтому получаем

$$h^{(k+r,\,i)} \leqslant h^{a^{i}(k+r,\,i)} \leqslant h^{na\,(\rho\,(\mathscr{A}_p)-\varepsilon)^r},$$
$$1 \leqslant i \leqslant n, \quad r \geqslant 0, \quad k \geqslant k(\varepsilon) \geqslant k^{(0)}.$$

Это легко переформулировать в виде

$$d\left(X^{(k+r,\,i)}\right) \leqslant (\hat{\varepsilon}/\gamma)\, h^{na\,(\rho\,(\mathscr{A}_p)-\varepsilon)^r}.$$

Пусть теперь

$$d^{(k)} = \max_{1 \leqslant i \leqslant n} \{d(X^{(k,\,i)})\}.$$

Тогда получаем

$$d^{(k+r)} \leqslant (\hat{\varepsilon}/\gamma)\, h^{an\,(\rho\,(\mathscr{A}_p)-\varepsilon)^r}.$$

Поэтому мы можем заключить, что R-фактор удовлетворяет соотношению

$$R_{\rho(\mathscr{A}_p)-\varepsilon}\{d^{(k)}\} = \limsup_{r\to\infty} (d^{(k+r)})^{[1/(\rho\,(\mathscr{A}_p)-\varepsilon)^r]}$$
$$\leqslant \limsup_{r\to\infty} \left(\frac{\hat{\varepsilon}}{\gamma} h^{an\,(\rho\,(\mathscr{A}_p)-\varepsilon)^r}\right)^{[1/(\rho\,(\mathscr{A}_p)-\varepsilon)^r]} = h^{an} < 1.$$

Отсюда следует, что

$$O_R((5), (\xi^{(i)})) \geqslant \rho(\mathscr{A}_p) - \varepsilon$$

для всех $\varepsilon > 0$, а потому

$$O_R((5), (\xi^{(i)})) \geqslant \rho(\mathscr{A}_p).$$

Рассмотрим теперь характеристический многочлен

$$q^{(n)}(\lambda)$$

матрицы $\mathscr{A}_p$:

$$q^{(n)}(\lambda) = (\lambda - 1)^n - (\lambda - 1) - 1.$$

Полагая $\tau = \lambda - 1$, мы можем написать его в виде

$$\tilde{q}^{(n)}(\tau) = \tau^n - \tau - 1.$$

Так как

$$\tilde{q}^{(n)}(1) = -1 < 0 \quad \text{и} \quad \tilde{q}^{(n)}(2) = 2^n - 3 \geqslant 1 > 0$$

для $n \geq 2$, то по правилу Декарта многочлен $\tilde{q}^{(n)}(\tau)$ имеет ровно один положительный корень $\sigma^{(n)}$, для которого

$$1 < \sigma^{(n)} < 2.$$

Поэтому спектральный радиус матрицы $\mathscr{A}_p$ удовлетворяет соотношению

$$\rho(\mathscr{A}_p) = 1 + \sigma^{(n)} > 2,$$

откуда получается

$$O_R((5), (\xi^{(n)})) \geqslant 1 + \sigma^{(n)}.$$

Мы опишем теперь одно приложение метода (5). Пусть дана вещественная симметрическая матрица $\mathscr{A}'_p = (a_{ij})$ размерности $n \times n$. Нужно определить собственные числа этой матрицы, т. е. числа λ, для которых выполнено равенство

$$\mathscr{A}'_p x_p = \lambda x_p \quad \text{при} \quad x_p \neq o_p.$$

Чтобы сделать это, применяем конечное число раз ортогональные преобразования подобия

$$\mathscr{A}_p = \mathscr{U}_p^T \mathscr{A}_p \mathscr{U}_p,$$

преобразующие, вообще говоря, неразреженную матрицу $\mathscr{A}'_p$ в матрицу $\mathscr{A}_p$, имеющую вид

$$\mathscr{A}_p = \begin{pmatrix} a^{(1)} & b^{(1)} & & \\ b^{(1)} & a^{(2)} & b^{(2)} & 0 \\ & \ddots & \ddots & \ddots \\ & 0 & & b^{(n-1)} a^{(n)} \end{pmatrix}$$

Затем собственные числа матрицы $\mathscr{A}_p$ (а тем самым и матрицы $\mathscr{A}'_p$) вычисляются как корни характеристического многочлена

$$p(\lambda) = \det(\lambda \mathscr{J}_p - \mathscr{A}_p)$$

матрицы $\mathscr{A}_p$. Значение $p(\lambda)$ может быть найдено с помощью следующей рекуррентной формулы:

$$\begin{cases} f^{(0)}(\lambda) = 1, \quad f^{(1)}(\lambda) = \lambda - a^{(1)}, \\ f^{(k)}(\lambda) = (\lambda - a^{(k)}) f^{(k-1)}(\lambda) - (b^{(k-1)})^2 f^{(k-2)}(\lambda), \quad 2 \leqslant k \leqslant n, \\ p(\lambda) = f^{(n)}(\lambda). \end{cases} \quad (10)$$

Если теперь $\mathscr{A}_p$ имеет только простые собственные числа и для них найдены непересекающиеся локализующие интервалы (например, с помощью теоремы Гершгорина), то мы можем применить (5). Это иллюстрирует следующий пример.

Пример, (а) Рассмотрим матрицу

$$\mathscr{A}_p = \begin{pmatrix} 15 & 1 & & & & & & & \\ 1 & 10 & 1 & & & & & 0 & \\ & 1 & 7 & 1 & & & & & \\ & & 1 & -4 & 1 & & & & \\ & & & 1 & 0 & 1 & & & \\ & & & & 1 & -4 & 1 & & \\ & & & & & 1 & -7 & 1 & \\ & 0 & & & & & 1 & -10 & 1 \\ & & & & & & & 1 & -15 \end{pmatrix}$$

Применяя теорему Гершгорина, мы получаем для собственных чисел матрицы $\mathscr{A}_p$ следующие локализующие интервалы:

$$X^{(0,1)} = [+ 13.99999999995, + 16.00000000005],$$
$$X^{(0,2)} = [+ 7.999999999974, + 12.00000000005],$$
$$X^{(0,3)} = [+ 4.999999999981, + 9.000000000015],$$
$$X^{(0,4)} = [+ 1.999999999992, + 6.000000000022],$$
$$X^{(0,5)} = [- 2.000000000008, + 2.000000000008],$$
$$X^{(0,6)} = [- 6.000000000022, - 1.999999999992],$$
$$X^{(0,7)} = [- 9.000000000015, - 4.999999999981],$$
$$X^{(0,8)} = [- 12.00000000005, - 7.999999999974],$$
$$X^{(0,9)} = [- 17.00000000005, - 12.99999999995].$$

Исходя из этих интервалов, строим итерации по формулам (5) и получаем следующие результаты:

$$X^{(5,1)} = [+ 15.19709300868, \quad + 15.19709300872],$$
$$X^{(4,2)} = [+ 10.13174515464, \quad + 10.13174515471],$$
$$X^{(4,3)} = [+ 7.001927580904, \quad + 7.001927580960],$$
$$X^{(4,4)} = [+ 3.920346203678, \quad + 3.920346203715],$$
$$X^{(5,5)} = [- 0.1096791595101 \times 10^{-10}, \quad + 0.1096791595101 \times 10^{-10}],$$
$$X^{(4,6)} = [- 3.920346203719, \quad - 3.920346203674],$$
$$X^{(4,7)} = [- 7.001927580969, \quad - 7.001927580895],$$
$$X^{(3,8)} = [- 10.13174515473, \quad - 10.13174515463],$$
$$X^{(3,9)} = [- 15.19709300876, \quad - 15.19709300866].$$

Эти интервалы невозможно улучшить, используя имеющуюся программу (см. приложение С). Подчеркнуты знаки, совпадающие в верхней и нижней границах.

(β) Теперь рассмотрим матрицу

$$\mathscr{A}_p = \begin{pmatrix} 12 & 1 & & & 0 \\ 1 & 9 & 1 & & \\ & 1 & 6 & 1 & \\ & & 1 & 3 & 1 \\ 0 & & & 1 & 0 \end{pmatrix}$$

Снова применяя теорему Гершгорина, мы находим для собственных чисел матрицы $\mathscr{A}_p$ следующие локализуюшяе интервалы:

$$X^{(0,1)} = [+ 10.99999999998, + 13.00000000003],$$
$$X^{(0,2)} = [+ 6.999999999970, + 11.00000000003],$$
$$X^{(0,3)} = [+ 3.999999999989, + 8.000000000021],$$
$$X^{(0,4)} = [+ 0.9999999999945, + 5.000000000019],$$
$$X^{(0,5)} = [- 1.000000000004, + 1.000000000004].$$

Следующие уточненные интервалы были вычислены с помощью итерационного метода (5) (ср. с замечаниями, сделанными после теоремы 1):

$$X^{(1,1)} = [+ 12.11013986010, + 12.55506993010],$$
$$X^{(1,2)} = [+ 9.006328989416, + 9.048379503166],$$
$$X^{(1,3)} = [+ 5.999999999958, + 6.000000000041],$$
$$X^{(1,4)} = [+ 2.979804773200, + 2.987022580008],$$
$$X^{(1,5)} = [- 0.3230758693540, - 0.3162523763767].$$

$$X^{(2,1)} = [+ 12.31617201370, + 12.31774922532],$$
$$X^{(2,2)} = [+ 9.016110401580, + 9.016149094187],$$
$$X^{(2,3)} = [+ 5.999999999958, + 6.000000000013],$$
$$X^{(2,4)} = [+ 2.983860239266, + 2.983864788268],$$
$$X^{(2,5)} = [- 0.3168759526293, - 0.3168750526051],$$

$$X^{(3,1)} = [+ 12.31687595112, + 12.31687595546],$$
$$X^{(3,2)} = [+ 9.016136303134, + 9.016136303198],$$
$$X^{(3,3)} = X^{(2,3)}$$
$$X^{(3,4)} = [+ 2.983863696823, + 2.983863696853],$$
$$X^{(3,5)} = [- 0.3168759526293, - 0.3168759526051],$$
$$X^{(4,1)} = [+ 12.31687595258, + 12.31687595266],$$
$$X^{(4,2)} = [+ 9.016136303134, + 9.016136303181],$$
$$X^{(4,3)} = X^{(3,3)}$$
$$X^{(4,4)} = X^{(3,4)}$$
$$X^{(4,5)} = [- 0.3168759526284, - 0.3168759526051].$$

2.3. Методы одновременной локализации комплексных корней многочленов

В этом разделе обсудим метод одновременной локализации комплексных (в общем случае) корней многочленов, предложенный Гаргантини и Энричи. Пусть задан многочлен

$$p(z) = a^{(n)}z^n + a^{(n-1)}z^{n-1} + \ldots + a^{(1)}z + a^{(0)}, \qquad (1)$$

где $a^{(i)} \in \mathbb{C}$, $0 \leqslant i \leqslant n$, $n \geqslant 2$. Предположим далее, что заданы n интервалов

$$W^{(0,\,i)} = \langle z^{(0,\,i)},\ r^{(0,\,i)} \rangle \in K(\mathbb{C}),$$

для которых

$$\zeta^{(i)} \in W^{(0,\,i)}, \quad p(\zeta^{(i)}) = 0, \quad 1 \leqslant i \leqslant n, \qquad (2)$$

$$W^{(0,\,i)} \cap W^{(0,\,j)} = \varnothing, \quad 1 \leqslant i < j \leqslant n. \qquad (3)$$

Элемент $Z \in K(\mathbb{C})$ представляется в дальнейшем **в** виде

$$Z = \langle m(Z),\ r(Z) \rangle.$$

Рассмотрим следующий метод итерации:

$$\begin{cases} z^{(k,\,i)} = m(W^{(k,\,i)}), \\ C^{(k,\,i)} = \sum\limits_{j=1,\ j \neq i}^{n} \dfrac{1}{z^{(k,\,i)} - W^{(k,\,j)}}, \\ q(z^{(k,\,i)}) = \dfrac{p'(z^{(k,\,i)})}{p(z^{(k,\,i)})} \quad \text{для} \quad p(z^{(k,\,i)}) \neq 0, \\ W^{(k+1,\,i)} = \langle z^{(k+1,\,i)},\ r^{(k+1,\,i)} \rangle = -\dfrac{1}{q(z^{(k,\,i)}) - C^{(k,\,i)}}, \\ 1 \leqslant i \leqslant n, \quad k \geqslant 0, \end{cases} \qquad (4)$$

и пусть

$$r^{(k)} = \max_{1 \leqslant i \leqslant n} \{r^{(k,\,i)}\}, \qquad (5)$$

$$\rho^{(k)} = \min_{1 \leqslant i < j \leqslant n} \{\min\{|z| \mid z \in z^{(k,\,i)} - W^{(k,\,j)}\}\}. \qquad (6)$$

Для $i \neq j$ из (3) следует, что

$$\min\{|z| \mid z \in z^{(0,\,i)} - W^{(0,\,j)}\} = |z^{(0,\,i)} - z^{(0,\,j)}| - r^{(0,\,j)} \geqslant \rho^{(0)}. \qquad (7)$$

Определим еще величины $\eta^{(k)}$ соотношением

$$\rho^{(k)} = (n-1)\eta^{(k)}. \qquad (8)$$

Тогда для итерационной схемы (4) верно следующее.

Теорема 1. *Пусть p(z) есть многочлен* (1) *и его корни* $\xi^{(i)}$, $1 \le i \le n$, *удовлетворяют условиям* (2) *и* (3). *В обозначениях* (5), (6), (8) *пусть*

$$6r^{(0)} \leqslant \eta^{(0)}. \tag{9}$$

Тогда

(a) *итерация* (4) *всегда осуществима, причем*

$$\zeta^{(i)} \in W^{(k,\ i)}, \qquad 1 \leqslant i \leqslant n, \qquad k \geqslant 0;$$

(b) *имеет место неравенство*

$$r^{(k+1)} \leqslant \frac{1}{\rho^{(0)}\left(\eta^{(0)} - 4r^{(0)}\right)} \left(r^{(k)}\right)^3 \leqslant \frac{1}{12\,(n-1)}\, r^{(k)}, \qquad k \geqslant 0.$$

Замечание. Из (6) следует, что $\lim\limits_{k \to \infty} r^{(k)} = 0$. Поэтому в силу (a) получаем, что

$$\lim_{k \to \infty} W^{(k,\ i)} = \zeta^{(i)}, \qquad 1 \leqslant i \leqslant n.$$

Из (6) следует с помощью теоремы 2 из приложения А, что *R*-порядок итераций (4) удовлетворяет условию $O_R((4), (\xi^{(i)})) \geqslant 3$.

Доказательство, (a): Из

$$\begin{aligned} &|z^{(0,\ i)} - \zeta^{(i)}| \leqslant r^{(0,\ i)} \leqslant r^{(0)}, \\ &|z^{(0,\ i)} - \zeta^{(j)}| \geqslant |z^{(0,\ i)} - z^{(0,\ j)}| - |z^{(0,\ j)} - \zeta^{(j)}| \\ &\qquad \geqslant |z^{(0,\ i)} - z^{(0,\ j)}| - r^{(0,\ j)} \geqslant \rho^{(0)} \end{aligned}$$

следует, что

$$\begin{aligned} |q(z^{(0,\ i)})| &= \left| \sum_{j=1}^{n} \frac{1}{z^{(0,\ i)} - \zeta^{(j)}} \right| \\ &\geqslant \left| \frac{1}{z^{(0,\ i)} - \zeta^{(i)}} \right| - \sum_{j=1,\ j \neq i}^{n} \left| \frac{1}{z^{(0,\ i)} - \zeta^{(j)}} \right| \\ &\geqslant \frac{1}{r^{(0)}} - \frac{1}{\eta^{(0)}} \quad \text{для} \quad z^{(0,\ i)} \neq \zeta^{(i)}. \end{aligned}$$

Из

$$|z^{(0,\ i)} - z^{(0,\ j)}| - r^{(0,\ j)} \geqslant \rho^{(0)} > 0 \tag{10}$$

имеем

$$0 \notin z^{(0,\ i)} - W^{(0,\ j)},$$

а также

$$\frac{1}{z^{(0,\,i)}-W^{(0,\,i)}}\subset\left\langle 0,\ \frac{1}{\rho^{(0)}}\right\rangle.$$

$$C^{(0,\,i)}=\sum_{j=1,\ j\neq i}^{n}\frac{1}{z^{(0,\,i)}-W^{(0,\,i)}}\subset\sum_{j=1,\ j\neq i}^{n}\left\langle 0,\ \frac{1}{\rho^{(0)}}\right\rangle=\left\langle 0,\ \frac{1}{\eta^{(0)}}\right\rangle,$$

$$q(z^{(0,\,i)})-C^{(0,\,i)}\subset\langle q(z^{(0,\,i)}),\ 1/\eta^{(0)}\rangle. \tag{11}$$

Так как

$$|q(z^{(0,\,i)})|-1/\eta^{(0)}\geqslant 1/r^{(0)}-2/\eta^{(0)}>0,$$

то ясно, что

$$0\notin q(z^{(0,\,i)})-C^{(0,\,i)},$$

и потому определены

$$W^{(1,\,i)},\qquad 1\leqslant i\leqslant n.$$

Ввиду

$$\frac{p'(z^{(0,\,i)})}{p(z^{(0,\,i)})}=\sum_{j=1}^{n}\frac{1}{z^{(0,\,i)}-\zeta^{(j)}}$$

из (2) и монотонности включения следует, что

$$\zeta^{(i)}=z^{(0,\,i)}-p(z^{(0,\,i)})\Big/\left[p'(z^{(0,\,i)})-p(z^{(0,\,i)})\sum_{j=1,\ j\neq i}^{n}\frac{1}{z^{(0,\,i)}-\zeta^{(j)}}\right]$$

$$\in z^{(0,\,i)}-\frac{1}{q(z^{(0,\,i)})-C^{(0,\,i)}}=W^{(1,\,i)},\qquad 1\leqslant i\leqslant n.$$

Это доказывает (a) для $k=1$.

(b): Из

$$|z^{(0,\,i)}-z^{(0,\,j)}|^2-(r^{(0,\,j)})^2\geqslant(\rho^{(0)}+r^{(0,\,j)})^2-(r^{(0,\,j)})^2\geqslant(\rho^{(0)})^2$$

получаем, что

$$r\left(\frac{1}{z^{(0,\,i)}-W^{(0,\,j)}}\right)=\frac{r^{(0,\,j)}}{|z^{(0,\,i)}-z^{(0,\,j)}|^2-(r^{(0,\,j)})^2}\leqslant\frac{r^{(0)}}{(\rho^{(0)})^2},$$

а потому

$$r(C^{(0,\,i)})\leqslant\frac{n-1}{\rho^{(0)}}\cdot\frac{r^{(0)}}{\rho^{(0)}}=\frac{r^{(0)}}{\eta^{(0)}\rho^{(0)}}.$$

Используя это неравенство и (11), получаем теперь

$$r(q(z^{(0,\,i)})-C^{(0,\,i)})=r(C^{(0,\,i)}),$$

$$|m(q(z^{(0,\,i)})-C^{(0,\,i)})|\geqslant 1/r^{(0)}-2/\eta^{(0)}+r(q(z^{(0,\,i)})-C^{(0,\,i)})$$

$$=1/r^{(0)}-2/\eta^{(0)}+r(C^{(0,\,i)}),$$

а отсюда и неравенство

$$r(W^{(1,\,i)}) = r\left(\frac{1}{q(z^{(0,\,i)}) - C^{(0,\,i)}}\right)$$
$$= \frac{r(q(z^{(0,\,i)}) - C^{(0,\,i)})}{|m(q(z^{(0,\,i)}) - C^{(0,\,i)})|^2 - (r(q(z^{(0,\,i)} - C^{(0,\,i)}))^2}$$
$$\leqslant \frac{(r^{(0)})^3}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})},$$

т. е.

$$r^{(1)} \leqslant \frac{(r^{(0)})^3}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}. \tag{12}$$

Применяя (9), мы получаем из предыдущей оценки неравенство

$$r^{(1)} \leqslant \frac{1}{12(n-1)} r^{(0)}.$$

Пусть

$$\delta^{(0)} = \max_{1 \leqslant i \leqslant n} \{|z^{(0,\,i)} - z^{(1,\,i)}|\}.$$

Тогда, применяя (6), мы получаем

$$\rho^{(1)} \geqslant \rho^{(0)} - \delta^{(0)} - 2r^{(1)}. \tag{13}$$

Чтобы оценить $\delta^{(0)}$, используем (10), (11) и соотношение

$$z^{(1,\,i)} - z^{(0,\,i)} \in \frac{1}{q(z^{(0,\,i)}) - C^{(0,\,i)}},$$

чтобы получить

$$|z^{(1,\,i)} - z^{(0,\,i)}| \leqslant \left|\frac{1}{\langle q(z^{(0,\,i)}),\ 1/\eta^{(0)}\rangle}\right| = \frac{1}{|q(z^{(0,\,i)})| - 1/\eta^{(0)}} \leqslant \frac{r^{(0)}\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}},$$

что дает, наконец,

$$\delta^{(0)} \leqslant \frac{r^{(0)}\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}}. \tag{14}$$

Используя (12), (13) и (14), получаем из (9), что

$$\eta^{(1)} - 6r^{(1)} = \rho^{(1)}/(n-1) - 6r^{(1)} \tag{15}$$
$$\geqslant \eta^{(0)} - r^{(0)}\left(\frac{\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{8(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}\right)$$
$$\geqslant \eta^{(0)} - 3r^{(0)} \geqslant 0,$$

т. е.

$$\eta^{(1)} \geqslant 6r^{(1)}.$$

Отсюда можно так же, как и раньше, получить, что

$$r^{(2)} \leqslant \frac{1}{\rho^{(1)}(\eta^{(1)} - 4r^{(1)})}(r^{(1)})^3 \leqslant \frac{1}{12(n-1)} r^{(1)}.$$

Из (13) тем же способом, каким было получено (15), выводим, что

$$\eta^{(1)} - 4r^{(1)} \geqslant \eta^{(0)} - r^{(0)}\left(\frac{\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{6(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}\right) \geqslant 0 \quad (16)$$

и

$$\eta^{(1)} \geqslant \eta^{(0)} - r^{(0)}\left(\frac{\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{2(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}\right) \geqslant 0. \quad (17)$$

Используя оба последних неравенства, мы получаем из (9)

$$\eta^{(1)}(\eta^{(1)} - 4r^{(1)}) \geqslant (\eta^{(0)})^2 - \eta^{(0)} r^{(0)}\left(\frac{2\eta^{(0)}}{\eta^{(0)} - 2r^{(0)}} + \frac{8(r^{(0)})^2}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}\right)$$
$$\geqslant \eta^{(0)}(\eta^{(0)} - 4r^{(0)}),$$

а потому

$$r^{(2)} \leqslant \frac{1}{\rho^{(0)}(\eta^{(0)} - 4r^{(0)})}(r^{(1)})^3.$$

Оставшаяся часть доказательства получается методом математической индукции.

Теперь проиллюстрируем итерационный метод (4). Для этого рассмотрим задачу вычисления собственных чисел гессен-берговой матрицы с помощью последовательности локализаций. Нужные при этом значения характеристического многочлена и его производных можно вычислить с помощью метода Хаймана. В качестве конкретного примера рассмотрим матрицу

$$\mathcal{H}_p = \begin{pmatrix} 12 + 16i & 1 & 0 & 0 \\ 0 & 9 + 12i & 1 & 0 \\ 0 & 0 & 6 + 8i & 1 \\ 1 & 0 & 0 & 3 + 4i \end{pmatrix},$$

Где $i = \sqrt{-1}$. Применяя теорему Гершгорина, мы получаем, что каждый из кругов

$$W^{(0,1)} = \langle 12 + 16i,\ 1\rangle, \qquad W^{(0,2)} = \langle 9 + 12i,\ 1\rangle,$$
$$W^{(0,3)} = \langle 6 + 8i,\ 1\rangle, \qquad W^{(0,4)} = \langle 3 + 4i,\ 1\rangle$$

содержит в точности одно собственное число матрицы H_P. С помощью (4) мы получаем уточненные локализующие множества $W^{(k,i)}$ для собственных чисел матрицы H_P. Ниже в табл. 1 используется представление

$$W^{(k,\ i)} = \langle m(W^{(k,\ i)}),\ r(W^{(k,\ i)})\rangle,$$

где

$$m(W^{(k,\ i)}) = \mathrm{Re}(m(W^{(k\ \ i)})) + i\,\mathrm{Im}(m(W^{(k,\ i)})).$$

Таблица 1

k	i	Re	Im	r
1	1	+ 11.99875131516	+ 15.99953080496	0.1001255×10^{-6}
	2	+ 9.003742419628	+ 12.00140833328	0.1494005×10^{-5}
	3	+ 5.996257580383	+ 7.998591666711	0.1493969×10^{-5}
	4	+ 3.001248654837	+ 4.000469195035	0.1000782×10^{-6}
2	1	+ 11.99875136181	+ 15.99953080159	0.1019500×10^{-9}
	2	+ 9.003742437190	+ 12.00140832752	$0.8760740 \times 10^{-10}$
	3	+ 5.996257562811	+ 7.998591672458	$0.3665239 \times 10^{-10}$
	4	+ 3.001248638204	+ 4.000469198423	$0.2555951 \times 10^{-10}$
3	1	+ 11.99875136181	+ 15 99953080159	0.1019496×10^{-9}
	2	+ 9.003742437190	+ 12 00140832752	$0.8760740 \times 10^{-10}$
	3	+ 5.996257562811	+ 7.998591672458	$0.3665353 \times 10^{-10}$
	4	+ 3.001248638204	+ 4 000469198423	$0.2556093 \times 10^{-10}$

Замечания. Итерационный метод, исследованный в теореме 1, можно назвать полношаговым методом. Если на каждом шаге использовать только что уточненные значения приближений, то получим метод, который можно назвать короткошаговым. Можно показать, что этот метод имеет более чем кубический порядок сходимости к нулю радиуса аппроксимирующего круга.

2.4. Операции над интервальными матрицами

Множество вещественных матриц размерности $m \times n$ обозначается через $M_{mn}(\mathbb{R})$, а множество комплексных матриц размерности $m \times n$ — через $M_{mn}(C)$. Элементы множества $M_{mn}(\mathbb{R})$, $M_{mn}(\mathbb{C})$ обозначаются через $\mathcal{A}_p, \mathcal{B}_p, \mathcal{C}_p, \ldots, \mathcal{X}_p, \mathcal{Y}_p, \mathcal{Z}_p$. Матрицы-столбцы, т. е. вещественные или комплексные векторы, обозначаются через $a_p, b_p, c_p, \ldots, x_p, y_p, z_p$. Множество вещественных n-мерных векторов обозначается через $V_n(\mathbb{R})$, множество комплексных векторов — через $V_n(C)$. Аналогичным образом обозначаем через $M_{mn}(I(\mathbb{R}))$ множество матриц, элементами которых являются вещественные интервалы, а через $M_{mn}(I(\mathbb{C}))$ — множество матриц, элементами которых являются комплексные интервалы; здесь $I(C)$ может обозначать как $R(C)$, так и $K(C)$. Элементы множества $M_{mn}(I(R))$ (соответственно $M_{mn}(I(\mathbb{C}))$ обозначаются через $\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots, \mathcal{X}, \mathcal{Y}, \mathcal{Z}$, и мы называем их вещественными (соответственно комплексными) интервальными матрицами.

Интервальные матрицы-столбцы, т. е. вещественные или комплексные интервальные векторы, обозначаются через $a,\ b,\ c,\ \ldots,\ x,\ y,\ z$. Множество вещественно-интервальных векторов-столбцов обозначается через $V_n(I(\mathbb{R}))$, а множество комплексно-интервальных векторов-столбцов — через $V_n(I(\mathbb{C}))$. Интервальные векторы и матрицы записываются, как обычно, в виде $\mathcal{A}=(A_{ij})$ в случае матриц и $a=(A_i)$ в случае векторов. Интервальная матрица, все компоненты которой являются точечными интервалами, называется *точечной матрицей. Точечные векторы* определяются аналогично. Упомянем очевидные соотношения

$$M_{mn}(I(\mathbb{R}))\subset M_{mn}(R(\mathbb{C}))$$

и

$$V_{mn}(I(\mathbb{R}))\subset V_n(R(\mathbb{C})).$$

Определение 1. Две интервальные матрицы $\mathcal{A}=(A_{ij})$ и $\mathcal{B}=(B_{ij})$ размерности $m\times n$ равны (это записывается, как обычно, в виде $\mathcal{A}=\mathcal{B}$), если равны их соответствующие компоненты. Иными словами, $\mathcal{A}=\mathcal{B}\leftrightarrow A_{ij}=B_{ij},\ 1\leqslant i\leqslant m,\ 1\leqslant j\leqslant n$.

Введем частичный порядок на множестве интервальных матриц.

Определение 2. Пусть $\mathcal{A}=(A_{ij})$ и $\mathcal{B}=(B_{ij})$ — интервальные матрицы размерности $m\times n$. Тогда полагаем

$$\mathcal{A}\subseteq\mathcal{B}\Leftrightarrow A_{ij}\subseteq B_{ij},\quad 1\leqslant i\leqslant m,\quad 1\leqslant j\leqslant m.\ \blacksquare$$

Отношение $\mathcal{A}\subset\mathcal{B}$ вводится аналогичным поэлементным определением. Если при этом $\mathcal{A}_n=(a_{ij})$ — точечная матрица, то пишем также $\mathcal{A}_p\in\mathcal{B}$. Каждую интервальную матрицу можно рассматривать как множество точечных матриц. Отношения $\subseteq$ и $\subset$ между множествами точечных матриц понимаются в обычном теоретико-множественном смысле.

Следующая цель — определить операции над интервальными матрицами, формально соответствующие операциям над точечными матрицами.

Определение 3. (а) Пусть $\mathcal{A}=(A_{ij}),\ \mathcal{B}=(B_{ij})$ — две интервальные матрицы размерности $m\times n$. Тогда соотношения

$$\mathcal{A}\pm\mathcal{B}:=(A_{ij}\pm B_{ij})$$

определяют соответственно сложение и вычитание интервальных матриц

(b) Пусть $\mathcal{A} = (A_{ij})$ — интервальная матрица размерности $m \times r$ и $\mathcal{B} = (B_{ij})$ — интервальная матрица размерности $r \times n$. Тогда соотношение

$$\mathcal{A}\mathcal{B} := \left(\sum_{\nu=1}^{r} A_{i\nu} B_{\nu j} \right)$$

определяет умножение интервальных матриц В частности, для интервальной матрицы $\mathcal{A} = (A_{ij})$ размерности $n \times r$ и интервального вектора $u = (U_i)$ размерности r мы имеем

$$\mathcal{A}u = \left(\sum_{\nu=1}^{r} A_{i\nu} U_{\nu} \right).$$

(c) Пусть $\mathcal{A} = (A_{ij})$ — интервальная матрица и X — интервал. Тогда полагаем

$$X\mathcal{A} = \mathcal{A}X := (XA_{ij}).$$

В дальнейшем предполагается, что интервальные матрицы, участвующие в интервальной операции, имеют нужное для этой операции число строк и столбцов, и это обстоятельство не будет специально оговариваться. Далее предполагается, что интервальные операнды (т. е. аргументы операций) имеют подходящие элементы Если, например, мы имеем $\mathcal{A} \in M_{mn}(K(\mathbb{C}))$, то произведение $\mathcal{A}\mathcal{B}$ определено, только если $\mathcal{B} \in M_{nr}(K(\mathbb{C}))$.

Операции над интервальными матрицами и векторами были формально введены в определения 3. Для вещественных интервальных операций мы имеем простое определение 2 п.1.1. Для интервальных матриц аналогичное определение невозможно, однако в общем случае имеет место

$$\{\mathcal{A}_p \mathcal{B}_p \mid \mathcal{A}_p \in \mathcal{A},\ \mathcal{B}_p \in \mathcal{B}\} \subseteq \{\mathcal{C}_p \mid \mathcal{C}_p \in \mathcal{A}\mathcal{B}\}.$$

Доказательство получается с помощью монотонности отношения включения для интервальных операций. Следующий пример показывает, что равенство не имеет места в общем случае. Пусть

$$\mathcal{A}_p = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \qquad u_p = \begin{pmatrix} [0, & 1] \\ [0, & 1] \end{pmatrix}$$

Тогда мы имеем

$$\mathcal{A}_p u = \begin{pmatrix} [& 0, & 2] \\ [-1, & & 1] \end{pmatrix}$$

и если возьмем

$$x_p = \begin{pmatrix} 2 \\ -1 \end{pmatrix} \in \mathcal{A}_p u,$$

то видим, чш не найдется $y_p \in u$, такого что $\mathcal{A}_p y_p = x_p$.

Пусть $\mathcal{A},\ \mathcal{B} \in M_{mn}(I(\mathbb{R}))$ и $c_p \in V_n(\mathbb{R})$. Тогда справедливы соотношения

$$\left\{\begin{array}{c}\{\mathcal{A}_p \pm \mathcal{B}_p \mid \mathcal{A}_p \in \mathcal{A},\ \mathcal{B}_p \in \mathcal{B}\} = \mathcal{A} \pm \mathcal{B}, \\ \{\mathcal{A}_p c_p \mid \mathcal{A}_p \in \mathcal{A}\} = \mathcal{A} c_p,\end{array}\right\} \tag{1}$$

которыми будем пользоваться в дальнейшем.

Множество интервальных матриц замкнуто относительно операций из определения 3. Множество вещественных или комплексных матриц изоморфно соответствующему множеству точечных матриц. Именно по этой причине в определении 3 использованы те же символы операций, что и для соответствующих вещественных и комплексных операций.

Теперь сформулируем некоторые свойства введенных операций.

Теорема 4. *Пусть $\mathcal{A}, \mathcal{B}, \mathcal{C}$ — интервальные матрицы. Тогда*

$$\mathcal{A} + \mathcal{B} = \mathcal{B} + \mathcal{A}, \tag{2}$$

$$\mathcal{A} + (\mathcal{B} + \mathcal{C}) = (\mathcal{A} + \mathcal{B}) + \mathcal{C}, \tag{3}$$

$$\mathcal{A} + O_p = O_p + \mathcal{A} = \mathcal{A}, \text{ где } O_p - \text{нулевая матрица}, \tag{4}$$

$$\mathcal{A}\mathcal{I}_p = \mathcal{I}_p\mathcal{A} = \mathcal{A}, \text{ где } \mathcal{I}_p - \text{единичная матрица}, \tag{5}$$

$$\left\{\begin{array}{l}(\mathcal{A} + \mathcal{B})\mathcal{C} \subseteq \mathcal{A}\mathcal{C} + \mathcal{B}\mathcal{C} \\ \mathcal{C}(\mathcal{A} + \mathcal{B}) \subseteq \mathcal{C}\mathcal{A} + \mathcal{C}\mathcal{B}\end{array}\right\} \quad (\text{субдистрибутивность}), \tag{6}$$

$$(\mathcal{A} + \mathcal{B})\mathcal{C}_p = \mathcal{A}\mathcal{C}_p + \mathcal{B}\mathcal{C}_p, \tag{7}$$

$$\mathcal{C}_p(\mathcal{A} + \mathcal{B}) = \mathcal{C}_p\mathcal{A} + \mathcal{C}_p\mathcal{B}, \tag{8}$$

$$\left\{\begin{array}{c}\mathcal{A}(\mathcal{B}_p\mathcal{C}_p) \subseteq (\mathcal{A}\mathcal{B}_p)\mathcal{C}_p, \\ (\mathcal{A}_p\mathcal{B})\mathcal{C} \subseteq \mathcal{A}_p(\mathcal{B}\mathcal{C}) \text{ для } \mathcal{C} = -\mathcal{C}, \\ \mathcal{A}_p(\mathcal{B}\mathcal{C}_p) = (\mathcal{A}_p\mathcal{B})\mathcal{C}_p, \\ \mathcal{A}(\mathcal{B}\mathcal{C}) = (\mathcal{A}\mathcal{B})\mathcal{C} \text{ для } \mathcal{A},\ \mathcal{B},\ \mathcal{C} \in M_{mn}(I(\mathbb{R})) \\ \text{и } \mathcal{B} = -\mathcal{B},\ \mathcal{C} = -\mathcal{C}.\end{array}\right. \tag{9}$$

Доказательство. Соотношения (2)—(8) доказываются поэлементно с использованием формул из теорем 4 п.1.1 и 8 п.1.4. Докажем (9) для квадратных матриц. Из дистрибутивности (8 п.1.1) и формулы (6 п.1.4) получаем

$$\mathcal{A}(\mathcal{B}_p\mathcal{C}_p) = \sum_{j=1}^{n} A_{ij}\left(\sum_{l=1}^{n} b_{jl}c_{lk}\right)\Bigg) \subseteq \left(\sum_{j=1}^{n}\sum_{l=1}^{n} A_{ij}b_{jl}c_{lk}\right)$$

$$= \left(\sum_{j=1}^{n}\left(\sum_{j=1}^{n} A_{ij}b_{jl}\right)c_{lk}\right) = (\mathcal{A}\mathcal{B}_p)\mathcal{C}_p,$$

что доказывает первое из соотношений (9). Равенства

$$(\mathscr{A}_{\mathrm{p}}\mathscr{B})\mathscr{C}=\left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}a_{ik}B_{kl}\right)C_{lj}\right)=\left(\sum_{l=1}^{n}\left|\sum_{k=1}^{n}a_{ik}B_{kl}\right|C_{lj}\right)$$
$$\subseteq\left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}|a_{ik}||B_{kl}|\right)C_{lj}\right)\subseteq\left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}|a_{ik}||B_{kl}|C_{lj}\right)\right)$$
$$=\left(\sum_{k=1}^{n}|a_{ik}|\left(\sum_{l=1}^{n}|B_{kl}|C_{lj}\right)\right)$$
$$=\left(\sum_{k=1}^{n}a_{ik}\left(\sum_{l=1}^{n}B_{kl}C_{lj}\right)\right)=\mathscr{A}_{\mathrm{p}}(\mathscr{B}\mathscr{C}).$$

дают второе соотношение. Из равенств

$$(\mathscr{A}_{\mathrm{p}}\mathscr{B})\mathscr{C}_{\mathrm{p}}=\left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}a_{ik}B_{kl}\right)c_{lj}\right)=\left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}a_{ik}B_{kl}c_{lj}\right)\right)$$
$$=\left(\sum_{k=1}^{n}a_{ik}\left(\sum_{l=1}^{n}B_{kl}c_{lj}\right)\right)$$

мы получаем третье соотношение.

Последнее соотношение получается следующим образом с помощью третьей из формул (8 п.1.1.):

$$\mathscr{A}(\mathscr{B}\mathscr{C})=\left(\sum_{k=1}^{n}A_{ik}\left(\sum_{l=1}^{n}B_{kl}C_{lj}\right)\right)$$
$$=\left(\sum_{k=1}^{n}A_{ik}\left(\sum_{l=1}^{n}|B_{kl}|C_{lj}\right)\right)$$
$$=\left(\sum_{k=1}^{n}|A_{ik}|\left(\sum_{l=1}^{n}|B_{kl}|C_{lj}\right)\right)$$
$$=\left(\sum_{k=1}^{n}\left(\sum_{l=1}^{n}|A_{ik}||B_{kl}|C_{lj}\right)\right)$$
$$=\left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}|A_{ik}||B_{kl}|C_{lj}\right)\right)$$
$$=\left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}|A_{ik}||B_{kl}|\right)C_{lj}\right)$$
$$=\left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}|A_{ik}|B_{kl}\right)C_{lj}\right)$$
$$=\left(\sum_{l=1}^{n}\left(\sum_{k=1}^{n}A_{ik}B_{kl}\right)\right)C_{lj}=(\mathscr{A}\mathscr{B})\mathscr{C}.$$

В общем случае ассоциативный закон не имеет места для интервальных матриц. Это показывает следующий

Пример.

$$\begin{bmatrix} [-1,\ 1] & 1 \\ -1 & [0,\ 1] \end{bmatrix}\left\{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}\right\} = \begin{bmatrix} 1 & [-2,\ 0] \\ [0,\ 1] & [0,\ 1] \end{bmatrix},$$

$$\left\{\begin{bmatrix} [-1,\ 1] & 1 \\ -1 & [0,\ 1] \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\right\}\begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} [-1,\ 3] & [-2,\ 0] \\ [0,\ 1] & [0,\ 1] \end{bmatrix}.$$

Основное свойство монотонности включения справедливо и для интервальных матричных операций.

Теорема 5. *Пусть* $\mathcal{A}^{(k)}$, $\mathcal{B}^{(k)}$, $k = 1, 2$, *— интервальные матрицы. Далее, пусть X, Y—интервалы и*

$$\mathcal{A}^{(k)} \subseteq \mathcal{B}^{(k)}, \quad k = 1, 2 \quad u \quad X \subseteq Y$$

Тогда соотношения

$$\begin{cases} \mathcal{A}^{(1)} * \mathcal{A}^{(2)} \subseteq \mathcal{B}^{(1)} * \mathcal{B}^{(2)}, \\ X\mathcal{A}^{(1)} \subseteq Y\mathcal{B}^{(1)} \end{cases} \tag{10}$$

имеют место для $* \in \{+, -, \cdot\}$.

Доказательство соотношений (10) проводится покомпонентно с использованием (9 п.1.1) и теоремы 9 п.1.4. Имеет место частный случай соотношений (10):

$$\mathcal{A}_p \in \mathcal{A}, \quad \mathcal{B}_p \in \mathcal{B} \Rightarrow \mathcal{A}_p * \mathcal{B}_p \in \mathcal{A} * \mathcal{B}, \ * \in \{+, -, \cdot\},$$
$$x \in X, \quad \mathcal{A}_p \in \mathcal{A} \Rightarrow x\mathcal{A}_p \in X\mathcal{A}.$$

Введем теперь понятия ширины и абсолютной величины интервальных матриц.

Определение 6. Пусть $\mathcal{A} = (A_{ij})$ — интервальная матрица. Тогда

(a) вещественная неотрицательная матрица

$$d(\mathcal{A}) := (d(A_{ij}))$$

называется шириной матрицы $\mathcal{A}$;

(b) вещественная неотрицательная матрица

$$|\mathcal{A}| := (|A_{ij}|)$$

называется матрицей абсолютных величин или абсолютной величиной матрицы $\mathcal{A}$.

Соберем теперь в одном месте некоторые свойства ширины и абсолютной величины интервальных матриц. Частичный порядок

$$\mathcal{X}_p \leqslant \mathcal{Y}_p \Leftrightarrow x_{ij} \leqslant y_{ij}, \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant j \leqslant n$$

используется здесь для вещественных интервальных матриц $\mathcal{X}_p$ и $\mathcal{Y}_p$ размерности $m \times n$. Перечислим эти свойства.

$$\mathcal{A} \subseteq \mathcal{B} \Rightarrow d(\mathcal{A}) \leqslant d(\mathcal{B}), \tag{11}$$

$$d(\mathcal{A} \pm \mathcal{B}) = d(\mathcal{A}) + d(\mathcal{B}), \tag{12}$$

$$d(\mathcal{A}) = \sup_{\mathcal{A}'_p, \mathcal{A}_p \in \mathcal{A}} |\mathcal{A}_p - \mathcal{A}'_p|, \tag{13}$$

$$|\mathcal{A}| = \sup_{\mathcal{A}_p \in \mathcal{A}} |\mathcal{A}_p|, \tag{14}$$

$$\mathcal{A} \subseteq \mathcal{B} \Rightarrow |\mathcal{A}| \leqslant |\mathcal{B}|, \tag{15}$$

$$|\mathcal{A}| \geqslant O_p \quad \text{и} \quad |\mathcal{A}| = O_p \Leftrightarrow \mathcal{A} = O_p,$$

$$|\mathcal{A} + \mathcal{B}| \leqslant |\mathcal{A}| + |\mathcal{B}|,$$

$$|x\mathcal{A}| = |\mathcal{A}x| = |x||\mathcal{A}|, \quad x \in \mathbb{C}, \tag{16}$$

$$\mathcal{A} \in M_{mn}(I(\mathbb{R})) \quad \text{или} \quad \mathcal{A} \in M_{mn}(K(\mathbb{C})),$$

$$|\mathcal{A}\mathcal{B}| \leqslant |\mathcal{A}||\mathcal{B}|,$$

$$d(\mathcal{A}\mathcal{B}) \leqslant d(\mathcal{A})|\mathcal{B}| + |\mathcal{A}|d(\mathcal{B}), \tag{17}$$

$$d(\mathcal{A}\mathcal{B}) \geqslant |\mathcal{A}|d(\mathcal{B}), \; d(\mathcal{A}\mathcal{B}) \geqslant d(\mathcal{A})|\mathcal{B}| \tag{18}$$

$$\begin{cases} d(a\mathcal{B}) = |a|d(\mathcal{B}), \; a \in \mathbb{C}, \\ d(\mathcal{A}_p\mathcal{B}) = |\mathcal{A}_p|d(\mathcal{B}), \quad d(\mathcal{B}\mathcal{A}_p) = d(\mathcal{B})|\mathcal{A}_p|. \end{cases} \tag{19}$$

Для вещественных интервальных матриц $\mathcal{A}$, $\mathcal{B}$ имеем

$$O_p \in \mathcal{A} \Rightarrow |\mathcal{A}| \leqslant d(\mathcal{A}) \leqslant 2|\mathcal{A}|, \tag{20}$$

$$\mathcal{A} = (-1)\mathcal{A} \Rightarrow \mathcal{A}\mathcal{B} = \mathcal{A}|\mathcal{B}|, \tag{21}$$

$$O_p \in \mathcal{A}, \quad 0 \notin \mathcal{B}_{ij} \quad \text{для} \quad \mathcal{B} = (\mathcal{B}_{ij}) \Rightarrow d(\mathcal{A}\mathcal{B}) = d(\mathcal{A})|\mathcal{B}|. \tag{22}$$

Доказательство этих свойств проводится покомпонентно с использованием свойств $I(\mathbb{R})$ из п.1.2 и свойств $I(\mathbb{C})$ из п.1.5.

Мы заметим также, что соотношения (20)—(22) неверны для комплексных интервальных матриц. Рассмотрим, например, (21) для матрицы размерности 1×1 с элементами из $R(C)$, т. е. интервал $A = A_1 + iA_2 \in R(\mathbb{C})$. Утверждение $A=(-1)A$ эквивалентно равенствам

$$A_1 = (-1)A_1, \quad A_2 = (-1)A_2$$

для $A = A_1 + iA_2$. Используя (18 п.7.2), мы получаем для $B = B_1 + iB_2$, что

$$AB = (A_1B_1 - A_2B_2) + i(A_1B_2 + A_2B_1)$$

$$= A_1|B_1| + A_2|B_2| + i(A_1|B_2| + A_2|B_1|).$$

С другой стороны, мы имеем

$$A|B| = A_1(|B_1| + |B_2|) + iA_2(|B_1| + |B_2|).$$

Эти два интервала различны в общем случае, например при $B_1=0$. Так как соотношения (20)—(22) не потребуются нам для комплексных интервалов, не будем рассматривать случаев, в которых они справедливы.

Теперь введем понятие матрицы расстояний для пары интервальных матриц.

Определение 7. Пусть $\mathcal{A} = (A_{ij})$ и $\mathcal{B} = (B_{ij})$ — интервальные матрицы. Тогда вещественная неотрицательная матрица

$$q(\mathcal{A}, \mathcal{B}) := (q(A_{ij}, B_{ij}))$$

называется матрицей расстояний или расстоянием между матрицами A и B.

Соотношения

$$q(\mathcal{A}, \mathcal{B}) = \mathcal{O}_{\mathrm{p}} \Leftrightarrow \mathcal{A} = \mathcal{B},$$

$$q(\mathcal{A}, \mathcal{B}) \leqslant q(\mathcal{A}, \mathcal{C}) + q(\mathcal{B}, \mathcal{C})$$

очевидным образом справедливы для расстояний между интервальными матрицами вместе с соотношениями

$$q(\mathcal{A} + \mathcal{C}, \mathcal{B} + \mathcal{C}) = q(\mathcal{A}, \mathcal{B}), \tag{23}$$

$$q(\mathcal{A} + \mathcal{B}, \mathcal{C} + \mathcal{D}) = q(\mathcal{A}, \mathcal{C}) + q(\mathcal{B}, \mathcal{D}), \tag{24}$$

$$q(\mathcal{A}\mathcal{B}, \mathcal{A}\mathcal{C}) \leqslant |\mathcal{A}|\, q(\mathcal{B}, \mathcal{C}). \tag{25}$$

Доказательства последних свойств проводятся покомпонентно с использованием соответствующих свойств $I(\mathrm{R})$ (см. п.1.2) или $I(\mathrm{C})$ (см. п.1.5). С помощью понятия расстояния между интервальными матрицами, введенного в определении 7 можно, используя монотонную норму матриц $\|\cdot\|$, определить метрику на множестве интервальных матриц как $\|q(\mathcal{A},\mathcal{B})\|$. Множество всех интервальных матриц размерности $m \times n$ можно также рассматривать как $m \cdot n$-крат-ное произведение полного метрического пространства $I(\mathrm{C})$ на себя Известные теоремы топологии показывают, что это произведение снова является полным метрическим пространством. Сходимость в произведении пространств эквивалентна сходимости отдельных компонент. Поэтому справедливы следующие утверждения.

Сходимость последовательности $\{\mathcal{A}^{(k)}\}_{k=0}^{\infty}$ интервальных матриц размерности $m \times n$ к матрице A, т. е. $\lim\limits_{k \to \infty} \mathcal{A}^{(k)} \quad \mathcal{A}$, эквивалентна (26)

$$\lim_{k \to \infty} A_{ij}^{(k)} = A_{ij}, \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant j \leqslant n.$$

Следствие 8. *Любая последовательность интервальных матриц* $\{\mathcal{A}^{(k)}\}_{k=0}^{\infty}$ *размерности* $m \times n$, *для которой имеет место*

$$\mathcal{A}^{(0)} \supseteq \mathcal{A}^{(1)} \supseteq \mathcal{A}^{(2)} \supseteq \ldots,$$

сходится к интервальной матрице $A = (A_{ij})$, *где*

$$A_{ij} = \bigcap_{k=0}^{\infty} A_{ij}^{(k)}, \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant j \leqslant n.$$

Это утверждение следует из (26) определения 2 и утверждения об интервалах, аналогичного следствию (8).

Следствие 9. *Операции, введенные в определении* 3, *непрерывны.*

Доказательство получается из того, что непрерывность операций на элементах влечет за собой непрерывность операций в целом. В силу определения 3 и теорем 6 п.1.2, 6 п.1.5 элементы результата операции непрерывно зависят от операндов.

Следующее соотношение справедливо ввиду (21 п.1.2) и (17. п.1.4).

$$\mathscr{X} \subseteq \mathscr{Y} \Rightarrow \frac{1}{2}(d(\mathscr{Y}) - d(\mathscr{X})) \leqslant q(\mathscr{X}, \mathscr{Y}) \leqslant d(\mathscr{Y}) - d(\mathscr{X}). \quad (27)$$

Теперь введем операцию пересечения на $M_{mn}(I(\mathbb{R}))$ и $M_{mn}(I(\mathbb{C}))$ таким же образом, как это было сделано в п.1.2 для элементов множества I(R) и в п.1.5 для элементов множества R(C). Так как $M_{mn}(I(\mathbb{R})) \subset M_{mn}(R(\mathbb{C}))$, достаточно определить эту операцию на $M_{mn}(R(\mathbb{C}))$. Пусть

$$\mathscr{A} = (A_{ij}), \mathscr{B} = (B_{ij}) \in M_{mn}(R(\mathbb{C})).$$

Тогда определяем пересечение $\mathscr{A}$ и $\mathscr{B}$ как теоретико-множественное пересечение

$$\mathscr{A} \cap \mathscr{B} = \{\mathscr{C}_p | \mathscr{C}_p \in \mathscr{A}, \mathscr{C}_p \in \mathscr{B}\}.$$

Пересечение двух интервальных матриц $\mathscr{A}$ и $\mathscr{B}$ принадлежит множеству $M_{mn}(R(\mathbb{C}))$ тогда и только тогда, когда это теоретико-множественное пересечение непусто. В этом случае мы имеем

$$\mathscr{A} \cap \mathscr{B} = (A_{ij} \cap B_{ij}),$$

где $A_{ij} \cap B_{ij}$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$, строится согласно (23 п.1.2) (соответственно (19 п.1.5)).

Аналогично следствиям 12 п.1.2 и 7 п.1.5 получаем

Следствие 10. *Пусть* $\mathscr{A}, \mathscr{B}, \mathscr{C}, \mathscr{D} \in M_{mn}(R(\mathbb{C}))$. *Тогда имеем*

$$\mathscr{A} \subseteq \mathscr{C}, \mathscr{B} \subseteq \mathscr{D} \Rightarrow \mathscr{A} \cap \mathscr{B} \subseteq \mathscr{C} \cap \mathscr{D} \quad (\textit{монотонность включения})$$

и пересечение, если оно не выводит за пределы $M_{mn}(R(\mathbb{C}))$, *является непрерывной операцией.*

Как и в случае следствий 12 п.1.2 и 7 п.1.5 , утверждение следует из того, что непрерывность операций на элементах влечет за собой непрерывность операций в целом.

Чтобы обобщить понятие билинейности на операторы из $V_n(\mathbb{C}) \times V_n(\mathbb{C})$ в V_n(C), мы рассмотрим теперь трехмерные

массивы с интервальными элементами. Множество всех таких массивов обозначается через $M_{n^3}(I(\mathbb{C}))$. Имеем

$$\mathscr{B} = (B_{ijk}) \in M_{n^3}(I(\mathbb{C})),$$

где

$$B_{ijk} \in I(\mathbb{C}), \quad 1 \leqslant i,\ j, \quad k \leqslant n$$

Равенство, включение и сложение определяются поэлементно, т. е. так же, как для интервальных матриц. Аналогично определяются ширина, расстояние и абсолютная величина. Например, определение

$$|\mathscr{B}| := (|\mathscr{B}_{ijk}|)$$

вводит билинейный оператор из $V_n(\mathbb{R}) \times V_n(\mathbb{R})$ в $V_n(\mathbb{R})$. Множество всех билинейных операторов из $V_n(\mathbb{R}) \times V_n(\mathbb{R})$ в $V_n(\mathbb{R})$ обозначается через $M_{n^3}(\square)$.

Определение 11. Пусть $\mathscr{B} = (B_{ijk}) \in M_{n^3}(I(\mathbb{C}))$. $x = (X_i)$ $y = (Y_i) \in V_n(I(\mathbb{C}))$ и $\mathscr{A} \in M_{nn}(I(\mathbb{C}))$.

Тогда полагаем (a) $\mathscr{C} := \mathscr{B}x \in M_{nn}(I(\mathbb{C}))$, где

$$C_{ij} = \sum_{k=1}^{n} B_{ijk} X_k, \quad 1 \leqslant i,\ j \leqslant n,$$

(b) $z := (\mathscr{B}x)\,y \in V_n(I(\mathbb{C}))$, где

$$Z_i = \sum_{j=1}^{n} C_{ij} Y_j = \sum_{j=1}^{n} \left(\sum_{k=1}^{n} B_{ijk} X_k \right) Y_j, \quad 1 \leqslant i < n.$$

Вместо $\mathscr{B}(x)\,y$ будем иногда писать $\mathscr{B}xy$.

(c) Далее полагаем

$$\mathscr{C} := \mathscr{A}\mathscr{B} \in M_{n^3}(I(\mathbb{C})),$$

где

$$C_{ijk} = \sum_{\nu=1}^{n} A_{i\nu} B_{\nu jk}, \quad 1 \leqslant i,\ j, \quad k \leqslant n.$$

Следующая теорема содержит некоторые соотношения, нужные в дальнейшем.

Теорема 12. *Пусть*

$$\mathscr{A}_{\mathrm{p}} = (a_{ij}) \in M_{nn}(\mathbb{C}), \quad \mathscr{B} = (B_{ijk}) \in M_{n^3}(I(\mathbb{C})),$$
$$x = (X_i), \quad y = (Y_i) \in V_n(I(\mathbb{C})).$$

Тогда

(a) $$(\mathscr{A}_{\mathrm{p}}\mathscr{B})\,xy \subseteq \mathscr{A}_{\mathrm{p}}(\mathscr{B}xy).$$

Если $x = -x$, то

(b) $$\mathscr{B}xx = \frac{1}{2}|\mathscr{B}|\,d(x)\,x$$

и

(c) $$d(\mathscr{B}xx) = \frac{1}{2}|\mathscr{B}|\,d(x)\,d(x).$$

Доказательство.

(a) Пусть $\mathscr{C} = \mathscr{A}_p\mathscr{B} = (C_{i,jk})$ и $z = (Z_i) = \mathscr{B}xy$. Тогда с помощью определения 11 и субдистрибутивности получаем

$$\begin{aligned}
(\mathscr{A}_p\mathscr{B})\,xy &= \left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n} C_{ijk}X_k\right)Y_j\right) = \left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n}\left(\sum_{\nu=1}^{n} a_{i\nu}B_{\nu jk}\right)X_k\right)Y_j\right)\\
&\subseteq \left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n}\left(\sum_{\nu=1}^{n} a_{i\nu}B_{\nu jk}X_k\right)\right)Y_j\right)\\
&= \left(\sum_{j=1}^{n}\left(\sum_{\nu=1}^{n}\left(\sum_{k=1}^{n} a_{i\nu}B_{\nu jk}X_k\right)\right)Y_j\right)\\
&= \left(\sum_{j=1}^{n}\left(\sum_{\nu=1}^{n} a_{i\nu}\left(\sum_{k=1}^{n} B_{\nu jk}X_k\right)\right)Y_j\right)\\
&\subseteq \left(\sum_{j=1}^{n}\left(\sum_{\nu=1}^{n} a_{i\nu}\left(\sum_{k=1}^{n} B_{\nu jk}X_k\right)Y_j\right)\right)\\
&= \left(\sum_{\nu=1}^{n}\left(\sum_{j=1}^{n} a_{i\nu}\left(\sum_{k=1}^{n} B_{\nu jk}X_k\right)Y_j\right)\right)\\
&= \left(\sum_{\nu=1}^{n} a_{i\nu}\left(\sum_{j=1}^{n}\left(\sum_{k=1}^{n} B_{\nu jk}X_k\right)Y_j\right)\right)\\
&= \left(\sum_{\nu=1}^{n} a_{i\nu}Z_\nu\right) = \mathscr{A}_p(\mathscr{B}xy).
\end{aligned}$$

(b) Полагая $z = \mathscr{B}xx$ и используя симметричность по x, получаем в обозначениях определения 11 (b), что

$$\begin{aligned}
Z_i &= \sum_{j=1}^{n}\left(\sum_{k=1}^{n} B_{ijk}X_k\right)X_j = \sum_{j=1}^{n}\left|\sum_{k=1}^{n} B_{ijk}X_k\right|X_j\\
&= \sum_{j=1}^{n}\left|\left(\sum_{k=1}^{n}|B_{ijk}|\,X_k\right) = \sum_{j=1}^{n}\left(\sum_{k=1}^{n}\frac{1}{2}|B_{ijk}|\,d(X_k)\right)X_j.\right.
\end{aligned}$$

Поэтому $\mathscr{B}xx = \frac{1}{2}|\mathscr{B}|\,d(x)\,x$.

(c) Используя соотношения из п. (b), получаем

$$d(Z_i) = \sum_{j=1}^{n}\left(\sum_{k=1}^{n}\frac{1}{2}|B_{ijk}|\,d(X_k)\right)d(X_j),$$

т. е.

$$d(\mathscr{B}xx) = \frac{1}{2}|\mathscr{B}|\,d(x)\,d(x).$$

3. Интервальная арифметика для решения систем уравнений

3.1.Итерационная локализация неподвижной точки для систем нелинейных уравнений

Рассмотоим теперь функцию $f(x_p)$ от векторной переменной $x_p = (x_1, \ldots, x_n)^T \in V_n(\mathbb{C})$, принимающую значения в С. Будем предполагать, что функция f построена при помощи основных арифметических операций, а также стандартных функций синус, косинус и т. д. Это означает, что ее можно вычислять и как интервальную функцию Предположим еще, что функция f зависит от m параметров $a_1, a_2, \ldots, a_m \in \mathbb{C}$. Таким образом, мы можем записать f в виде

$$f(x_p) = f(x_1, x_2, \ldots, x_n; a_1, a_2, \ldots, a_m).$$

Пусть теперь заданы n функции такого вида

$$f_i(x_p), \quad 1 \leqslant i \leqslant n.$$

Тогда соотношение

$$y_p = f_p(x_p) = (f_i(x_p))$$

определяет отображение из $V_n(\mathbb{C})$ в $V_n(\mathbb{C})$, а соотношение

$$y = f_p(x) = (f_i(x))$$

определяет отображение на множестве n-компонентных интервальных векторов

$$f_p: V_n(I(\mathbb{C})) \to V_n(I(\mathbb{C})), \quad \text{где} \quad f_p(x) = (f_i(x)).$$

Ниже будет использоваться интервальная арифметика для локализации решений системы уравнений

$$f_p(x_p) = o_p,$$

где

$$f_p(x_p) = (f_i(x_p))$$

и

$$f_i(x_p) = f_i(x_1, \ldots, x_n; a_{i1}, \ldots, a_{im_i}) \qquad 1 \leqslant i \leqslant n,$$

в предположении, что параметры a_{ij} независимо изменяются в некоторых комплексных интервалах.

Много возможностей дня решения этой задачи дает метод итераций. Заметим, что данное уравнение всегда может быть преобразовано к виду

$$x_p = f_p(x_p).$$

Вычисление правой части этого уравнения для произвольного интервального вектора $x^{(0)}$ дает интервальный вектор $x^{(1)}$. Продолжая в том же духе, получаем метод итераций

$$x^{(k+1)} = f_p(x^{(k)}), \qquad k \geq 0.$$

Возникают следующие вопросы, (a) Когда существует последовательность $\{x^{(k)}\}_{k=0}^{\infty}$? (b) Когда эта последовательность сходится? (c) Когда предел x^* единствен? (d) Какое отношение имеет предел x^* к решению сформулированной выше задачи?

Сначала докажем теорему о неподвижной точке, опирающуюся на моноточность интервального оценивания функций относительно включения.

Теорема 1. *Пусть дано отображение*

$$f_p: V_n(\mathbb{C}) \to V_n(\mathbb{C}), \quad \text{где} \quad f_p(x_p) = (f_i(x_p)),$$

причем функции $f_i(x_p)$ *имеют указанный выше вид. Рассмотрим метод итераций в* $V_n(I(\mathbb{C}))$, *заданный соотношением*

$$x^{(k+1)} = f_p(x^{(k)}), \qquad k \geq 0,$$

и удовлетворяющий условию

$$x^{(1)} \subseteq x^{(0)}.$$

Тогда имеет место следующее.

(1) *Последовательность результатов итерации* $\{x^{(k)}\}_{k=0}^{\infty}$ *сходится к пределу x, такому что* $x = f_p(x)$.

(2) *Любой вектор удовлетворяющий* $x_p \in x^{(0)}$, *уравнению* $x_p = f_p(x_p)$, *содержится в x, т. е имеет место*

$$\{x_p \mid x_p \in x^{(0)}, x_p = f_p(x)\} \subseteq x.$$

Доказательство. (1) По предположению имеем $x^{(1)} \subseteq x^{(0)}$. Так как интервальные вычисления монотонны относительно включения, мы получаем

$$x^{(2)} = f_p(x^{(1)}) \subseteq f_p(x^{(0)}) = x^{(1)} \subseteq x^{(0)}.$$

С помощью математической индукции можно показать, что

$$\ldots \subseteq x^{(3)} \subseteq x^{(2)} \subseteq x^{(1)} \subseteq x^{(0)}.$$

Из следствия 8 п. 2.3 вытекает, что эта последовательность сходится к некоторому элементу $x \in V_n(I(\mathbb{C}))$. Из непрерывности интервальных оценок следует, что

$$x = \lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} \mathcal{f}_{\mathrm{p}}(x^{(k)}) = \mathcal{f}_{\mathrm{p}}(x).$$

(2). Пусть $x_{\mathrm{p}} \in x^{(0)}$ и $x_{\mathrm{p}} = \mathcal{f}_{\mathrm{p}}(x_{\mathrm{p}})$.

Из монотонности интервальных операций относительно включения снова следует, что

$$x_{\mathrm{p}} = \mathcal{f}_{\mathrm{p}}(x_{\mathrm{p}}) \in \mathcal{f}_{\mathrm{p}}(x^{(0)}) = x^{(1)},$$

и по индукции получаем

$$x \in x^{(k)}, \qquad k \geqslant 0,$$

откуда следует, что $x_{\mathrm{p}} \in x$.

Условия теоремы 1 гарантируют существование неподвижной точки, но не ее единственность. Это показывает следующий пример.

Пример. Рассмотрим уравнение

$$X = X \cdot X \cdot X$$

в $R(\mathbb{C})$.Очевидно, что этому уравнению удовлетворяют следующие элементы $R(\mathbb{C})$:

$$X = [-1, 1], [1, 1], [-1, -1], [0, 1], [-1, 0], [0, 0], i[-1, 1].$$

Докажем теперь другую теорему о неподвижной точке, имеющую несколько иные условия и использующую несколько иной итерационный процесс.

Теорема 2. *Пусть задано отображение*

$$\mathcal{f}_{\mathrm{p}}: V_n(\mathbb{C}) \to V_n(\mathbb{C}), \quad \text{где} \quad \mathcal{f}_{\mathrm{p}}(x_{\mathrm{p}}) = (f_i(x_{\mathrm{p}}))$$

с функциями $f_i(x_{\mathrm{p}})$ указанного выше вида.

Рассмотрим итерационный процесс

$$x^{(k+1)} = \mathcal{f}_{\mathrm{p}}(x^{(k)}) \cap x^{(k)}, \qquad k \geqslant 0,$$

в $V_n(R(\mathbb{C}))$. Предположим, что существует элемент $\tilde{x}_{\mathrm{p}} \in x^{(0)}$, удовлетворяющий уравнению $\tilde{x}_{\mathrm{p}} = \mathcal{f}_{\mathrm{p}}(\tilde{x}_{\mathrm{p}})$. Тогда верно следующее.

(3) *Последовательные приближения $\{x^{(k)}\}_{k=0}^{\infty}$ удовлетворяют условию $\lim_{k\to\infty} x^{(k)} = x$, причем $x = \mathcal{f}_{\mathrm{p}}(x) \cap x$.*

(4) *Любой вектор $x_{\mathrm{p}} \in x^{(0)}$, удовлетворяющий уравнению $x_{\mathrm{p}} = \mathcal{f}_{\mathrm{p}}(x_{\mathrm{p}})$, содержитсяв x, т, е.*

$$\{x_{\mathrm{p}} \mid x_{\mathrm{p}} \in x^{(0)}, \ x_{\mathrm{p}} = \mathcal{f}_{\mathrm{p}}(x_{\mathrm{p}})\} \subseteq x.$$

Доказательство. Рассуждением, аналогичным доказательству теоремы 1, из $\tilde{x}_{\mathrm{p}} \in x^{(0)}$ мы получаем соотношени

$x_p \in f_p(x^{(0)}) \cap x^{(0)} = x^{(1)}$. Применяя индукцию, получаем $\tilde{x}_p \in x^{(k)}$ $k \geq 0$. Так как пересечение здесь всегда непусто, мы получаем последовательность интервалов

$$x^{(0)} \supseteq x^{(1)} \supseteq \ldots,$$

которая в силу следствия 8 п. 2.3 сходится к некоторому пределу x. Ввиду непрерывности интервальных вычислений и пересечений мы имеем $x = f_p(x) \cap x$ для этого предела, а также $\hat{x}_p \in x$ для всех $\hat{x}_p = f_p(\hat{x}_p) \in x^{(0)}$.

По поводу единственности неподвижной точки в теореме 2 можно сказать то же самое, что и в случае теоремы 1. Приведем теперь две теоремы о неподвижной точке, которые будут применены в дальнейшем к двум конкретным итерационным процедурам. В отличие от теорем 1 и 2, имеющих довольно общий характер, эти новые теоремы обеспечивают единственность неподвижной точки. Сначала введем одно понятие.

Определение 3. Пусть

$$f_p: V_n(\mathbb{C}) \to V_n(\mathbb{C}), \qquad f_p(x_p) = (f_i(x_p)),$$

— отображение указанного выше вида. f_p называется $\mathcal{P}_p$-сжатием, если существует неотрицательная матрица $\mathcal{P}_p$, такая что

$$q(f_p(x), f_p(y)) \leqslant \mathcal{P}_p q(x, y) \quad \text{для всех} \quad x, y \in V_n(I(\mathbb{C})),$$

где

$$\rho(\mathcal{P}_p) < 1.$$

Здесь ρ обозначает спектральный радиус матрицы $\mathcal{P}_p$, а q обозначает расстояние между двумя интервальными векторами, определенное в п.2.3.. Докажем следующее утверждение.

Теорема 4. *Если* $f_p: V_n(\mathbb{C}) \to V_n(\mathbb{C})$ *есть* $\mathcal{P}_p$*-сжатие, то уравнение* $x = f_p(x)$ *имеет единственную неподвижную точку* $x^* \in V_n(I(\mathbb{C}))$. *При этом для любого* $x^{(0)} \in V_n(I(\mathbb{C}))$ *итерации сходятся к* x^*.

Доказательство Из того, что $\rho(\mathcal{P}_p) < 1$ и $\mathcal{P}_p \geqslant O_p$, следует существование матрицы $(\mathcal{I}_p - \mathcal{P}_p)^{-1}$ и соотношение

$$(\mathcal{I}_p - \mathcal{P}_p)^{-1} = \sum_{i=1}^{\infty} \mathcal{P}_p^i \geqslant \sum_{i=0}^{m-1} \mathcal{P}_p^i \geqslant O_p.$$

Тогда мы получаем для любого k и $m \geq 1$

$$q(x^{(k+m)}, x^{(k)}) \leqslant \sum_{i=0}^{m-1} \mathcal{P}_p^i q(x^{(k+1)}, x^{(k)}) \leqslant (\mathcal{I}_p - \mathcal{P}_p)^{-1} \mathcal{P}_p^k q(x^{(1)}, x^{(0)}).$$

Так как $\lim_{k\to\infty} \mathscr{P}_p^k = \mathcal{O}_p$, каждая компонента последовательности $\{x^{(k)}\}_{k=0}^{\infty}$, а значит, и сама эта последовательность удовлетворяют условию сходимости Коши. Так как пространство $V_n(I(\mathbb{C}))$ полно, а отображение $\mathscr{P}_p$ является сжатием и потому непрерывно, мы получаем

$$\lim_{k\to\infty} x^{(k)} = x^* \quad \text{и} \quad x^* = f_p(x^*).$$

Единственность неподвижной точки следует из соотношений

$$q(x^*, y^*) = q(f_p(x^*), f_p(y^*)) \leqslant \mathscr{P}_p q(x^*, y^*)$$

и $(\mathcal{I}_p - \mathscr{P}_p)^{-1} \geqslant \mathcal{O}_p$.

Эта теорема — частный случай более общего результата, доказанного Шредером.

Вот еще одна теорема о неподвижной точке, которая будет использована в дальнейшем.

Теорема 5. *Пусть*

$$f_p: V_n(\mathbb{C}) \to V_n(\mathbb{C})$$

и

$$g_p: V_n(\mathbb{C}) \times V_n(\mathbb{C}) \to V_n(\mathbb{C}),$$

где f_p *и* g_p *имеют описанный выше вид, причем*

$$g_p(x, x) = f_p(x) \quad \textit{для всех} \quad x \in V_n(I(\mathbb{C})),$$

$$q(g_p(x, z), g(y, z)) \leqslant Q_p q(x, y),$$

$$q(g(z, x), g(z, y)) \leqslant \mathscr{R}_p q(x, y) \quad \textit{для всех} \quad x, y, z \in V_n(I(\mathbb{C}))$$

и

$$Q_p \geqslant \mathcal{O}_p, \quad \mathscr{R}_p \geqslant \mathcal{O}_p, \quad \rho(Q_p) < 1, \quad \rho((\mathcal{I}_p - Q_p)^{-1}\mathscr{R}_p) < 1$$

Тогда уравнение $f_p(x) = x$ *имеет единственную неподвижную точку* $x^* \in V_n(I(\mathbb{C}))$, *и для любого* $x^{(0)}$ *существует единственная последовательность* $\{x^{(k)}\}_{k=0}^{\infty}$, *удовлетворяющая уравнению*

$$x^{(k+1)} = g_p(x^{k+1}, x^k), \qquad k \geqslant 0$$

При этом

$$\lim_{k\to\infty} x^{(k)} = x^*.$$

Доказательство. Матрицу $(\mathcal{I} - Q)^{-1}\mathscr{R}_p$ можно рассматривать как итерационную матрицу, соответствующую матрице $\mathcal{I}_p - Q_p - \mathscr{R}_p$. Из $\mathscr{P}_p = Q_p + \mathscr{R}_p \geqslant \mathcal{O}_p$ и того, что $\rho(\mathscr{P}_p) < 1$,

следует $\rho((\mathcal{I}_p - Q_p)^{-1}\mathcal{R}_p) < 1$ Отсюда с помощью (23 п.2.3) и (10 п.2.3) получаем для произвольных $x, y \in V_n(I_n(\mathbb{C}))$, что

$$q(f_p(x), f_p(y)) \leqslant q(g_p(x, x), g_p(x, y)) + q(g_p(x, y), g_p(y, y)) \leqslant (\mathcal{R}_p + Q_p)q(x, y)$$

т. е. что f_P есть ρ_P-сжатие. В силу предыдущей теоремы уравнение $f_p(x) = x$ имеет единственную неподвижную точку $x^* \in V_n(I_n(\mathbb{C}))$. Из наших предположений следует, что отображение $g(\cdot, z)$ для фиксированного $z \in V_n(I_n(\mathbb{C}))$ является Q_p-сжа-тием. Применяя предыдущую теорему, получаем, что $g(\cdot, x^{(k)})$ имеет единственную неподвижную точку $x^{(k+1)}$. Таким образом, существование последовательности $\{x^{(k)}\}_{k=0}^{\infty}$ установлено для произвольного $x^{(0)} \in V_n(I(\mathbb{C}))$. Из того, что

$$q(x^{(k+1)}, x^*) \leqslant q(g_p(x^{(k+1)}, x^{(k)}), g(x^*, x^{(k)})) + q(g(x^*, x^{(k)}), g(x^*, x^*)) \leqslant Q_p q(x^{(k+1)}, x^*) + \mathcal{R}_p q(x^{(k)}, x^*)$$

или

$$q(x^{(k+1)}, x^*) \leqslant (\mathcal{I}_p - Q_p)^{-1}\mathcal{R}_p q(x^{(k)}, x^*) \leqslant \leqslant ((\mathcal{I}_p - Q_p)^{-1}\mathcal{R})^{k+1} q(x^{(0)}, x^*),$$

следует, что

$$\lim_{k \to \infty} x^{(k)} = x^*,$$

так как $\rho((\mathcal{I}_p - Q_p)^{-1}\mathcal{R}_p) < 1$.

Это завершает доказательство теоремы.

Результат из теоремы 5 был сначала получен для отображений из $V_n(\mathbb{R})$ в $V_n(\mathbb{R})$.

В связи с двумя последними теоремами продемонстрируем соотношение между единственной неподвижной точкой и потенциальными решениями уравнения

$$x_p = f_p(x_p), \quad x_p \in V_n(\mathbb{C}).$$

Следствие 6. *Пусть задано отображение*

$$f: V_n(\mathbb{C}) \to V_n(\mathbb{C}), \quad f_p(x_p) = (f_i(x_p)),$$

причем

$$f_i(x) = f_i(X_1, X_2, \ldots, X_n; A_{i1}, \ldots, A_{im_i}), \quad 1 \leqslant i \leqslant n.$$

Пусть выполнены условия одной из теорем 4, 5, *и пусть* x^* — *единственная неподвижная точка уравнения* $x = f_p(x)$, *существование которой доказано там. Тогда*

$$\{x_p \mid x_p = f_p(x_p), \quad a_{ij} \in A_{ij}, \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant m_i\} \subseteq x^*$$

Доказательство. Рассмотрим уравнение

$$x_{\mathrm{p}} = f_{\mathrm{p}}(x_{\mathrm{p}}), \qquad x_{\mathrm{p}} \in V_n(\mathbb{C}),$$

при фиксированном выборе элементов $a_{ij} \in A_{ij}$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m_i$, и допустим, что ему удовлетворяет элемент $x_{\mathrm{p}}^* \in V_n(\mathbb{C})$. Мы можем тогда начать итерации в теореме 4 или 5 со значения $x^{(0)} = x_{\mathrm{p}}^*$ и в пределе получим неподвижную точку x^*. Так же, как это было сделано в доказательстве п. (2) теоремы 1, мы можем использовать монотонность включения, чтобы показать, что всегда имеет место

$$x_{\mathrm{p}}^* \in x^{(k)}, \qquad k \geqslant 0.$$

Отсюда следует $x_{\mathrm{p}}^* \in x^*$.

Теперь рассмотрим практическое нахождение констант Липшица для интервальных вычислений. Мы увидим, что константы Липшица для интервальных вычислений будут мажорировать константы Липшица для соответствующих точечных функций. Это означает, что каждая из систем $x_{\mathrm{p}} = f(x_{\mathrm{p}})$, рассмотренных в следствии 6, удовлетворяет условиям теорем 4 и 5 при ограничении на множество $V_n(\mathbb{C})$, а потому имеет единственное решение x_{p}^*.

С помощью найденных констант Липшица мы сможем проверить выполнены ли в конкретных условиях предположения теоремы 4 или 5. Для простоты мы ограничимся пространством $V_n(I(\mathbb{R}))$ Аналогичные формулы для $V_n(I(\mathbb{C}))$ могут быть получены без труда. Для подготовки докажем одно свойство метрики q, которое следует из того, что она является метрикой Хаусдорфа.

Лемма 7. *Пусть* $Y, Z \in I(\mathbb{R})$ *и* $\alpha \geqslant 0$. *Тогда*

$$q(Y, Z) \leqslant \alpha \Leftrightarrow \{\text{для любого } y \in Y \text{ существует } z \in Z \text{ со свойством } |y - z| \leqslant \alpha, \text{ и для любого } z \in Z \text{ существует } y \in Y \text{ со свойством } |z - y| \leqslant \alpha\}$$

Доказательство. Если $Y = [y_1, y_2]$ и $Z = [z_1, z_2]$, то $q(Y, Z) = \max\{|y_1 - z_1|, |y_2 - z_2|\}$. Докажем сначала импликацию $\Rightarrow$.

Пусть $q(Y, Z) \leqslant \alpha$ и зафиксирован $y \in Y$. Если $z \notin Z$, то можно взять $z=y$, что дает первую половину правой части $\Rightarrow$. Если же $y \notin Z$, то при $z_1 > y$ мы получаем для $z = z_1$, что

$$\alpha \geqslant |z_1 - y_1| = z_1 - y_1 \geqslant z_1 - y = |z_1 - y|,$$

а в случае $z_2 < y$ получаем для $z = z_2$, что

$$\alpha \geqslant |z_2 - y_2| = y_2 - z_2 \geqslant y - z_2 = |y - z_2|.$$

Так как в этом рассуждении y и z равноправны, оно дает и вторую половину правой части $\Rightarrow$ для фиксированного $z \in Z$,

Докажем обратную импликацию $\Leftarrow$. Пусть сначала $y_1 \leqslant z_1$. Зафиксируем $y = y_1$. Тогда найдется $z \geq z_1$, такое что

$$\alpha \geqslant |z - y_1| = z - y_1 \geqslant z_1 - y_1 = |z_1 - y_1|.$$

Если же $z_1 < y_1$, то зафиксируем $z = z_1$. Тогда найдется $y \geqslant y_1$, для которого

$$\alpha \geqslant |y - z_1| = y - z_1 \geqslant y - z_1 = |y_1 - z_1|.$$

Таким образом, всегда получается $|y_1 - z_1| \leqslant \alpha$. Соотношение $|y_2 - z_2| \leqslant \alpha$ доказывается тем же методом. Это и дает

$$q(Y, Z) \leqslant \alpha.$$

Теперь используем эту лемму, чтобы получить утверждение о константах Липшица для интервальных вычислений

Теорема 8. *Пусть f — вещественная функция вещественной переменной x. Из выражения $f(x; a_1, \ldots, a_m)$, принадлежащего функции f, построим выражение $\tilde{f}(x_1, x_2, \ldots, x_n; a_1, \ldots, a_m)$, заменяя каждое вхождение переменной x на новую переменную x_i, $1 \leqslant i \leqslant n$.*

Допустим, что это новое выражение удовлетворяет условию Липшица

$$|\tilde{f}(x_1, \ldots, y_i, \ldots, x_n; a_1, \ldots, a_m) - \tilde{f}(x_1, \ldots, z_i, \ldots, x_n; a_1, \ldots, a_m)| \leqslant l_i |y_i - z_i|$$

при каждом i, $1 \leqslant i \leqslant n$ и фиксированных

$$x_k \in X, \quad 1 \leqslant k \leqslant n, \quad i \neq k \quad u \quad a_k \in A, \quad 1 \leqslant k \leqslant m.$$

Если существует интервальная оценка $f(X; A_1, \ldots, A_m)$, то для любых интервалов $Y \subseteq X$ и $Z \subseteq X$ выполнено следующее условие Липшица:

$$q(f(Y; A_1, \ldots, A_m), f(Z; A_1, \ldots, A_m)) \leqslant \left(\sum_{i=1}^{n} l_i\right) q(Y, Z). \quad (5)$$

Доказательство. Из способа построения выражения $\tilde{f}(x_1, \ldots, x_n; a_1, \ldots, a_m)$ следует, что

$$f(X; A_1, \ldots, A_m) = \tilde{f}(X, \ldots, X; A_1, \ldots, A_m) = W(\tilde{f}, X, \ldots, X; A_1, \ldots, A_m).$$

Если произвольным образом выбрано

$$u \in f(Y; A_1, \ldots, A_m),$$

то мы имеем

$$u = \tilde{f}(y_1, \ldots, y_n; a_1, \ldots, a_m) \quad \text{для} \quad y_i \in Y, \quad 1 \leqslant i \leqslant n,$$
$$a_k \in A_k, \quad 1 \leqslant k \leqslant m.$$

Из леммы 7 следует, что для произвольного y_i существует $z_i \in Z$, для которого

$$|y_i - z_i| \leqslant q(Y, Z).$$

Рассматривая значение функции

$$\tilde{f}(z_1, \ldots, z_n; a_1, \ldots, a_m) \in f(Z; A_1, \ldots, A_m),$$

получаем с помощью сделанных предположений и многократного использования неравенства треугольника, что

$$\begin{aligned}
&|\tilde{f}(y_1, \ldots, y_n; a_1, \ldots, a_m) - \tilde{f}(z_1, \ldots, z_n; a_1, \ldots, a_m)| \\
&\quad \leqslant |\tilde{f}(y_1, \ldots, y_n; a_1, \ldots, a_m) \\
&\qquad - \tilde{f}(z_1, y_2, \ldots, y_n; a_1, \ldots, a_m)| + \ldots \\
&\qquad + |\tilde{f}(z_1, \ldots, z_{n-1}, y_n; a_1, \ldots, a_m) \\
&\qquad - \tilde{f}(z_1, \ldots, z_n; a_1, \ldots, a_m)| \\
&\quad \leqslant \sum_{i=1}^{n} l_i |y_i - z_i| \leqslant \left(\sum_{i=1}^{n} l_i \right) q(Y, Z) = \alpha.
\end{aligned}$$

Аналогичное неравенство можно установить для произвольного $v \in f(Z, A_1, \ldots, A_m)$. Применяя лемму (7), получаем неравенство (5).

Условия теоремы 8 на практике почти всегда выполнены. Рассматриваемые здесь функциональные выражения составлены из основных арифметических операций и стандартных функций. Поэтому они почти всегда дифференцируемы на своей области определения.

В качестве приложения доказанной теоремы приведем несколько конкретных примеров.

Примеры. (a) $f(x; a) = ax$. Интервальная оценка этой функции для $A \in I(\mathbb{R})$ и произвольных $Y, Z \in I(\mathbb{R})$ удовлетворяет в силу (5) неравенству

$$q(AY, AZ) \leqslant |A| q(Y, Z).$$

(b) $$f(x; a_0, \ldots, a_n) = \sum_{k=0}^{n} a_k x^k.$$

В силу (5) интервальная оценка этой функции удовлетворяет для $A_k \in I(\mathbb{R})$, $0 \leqslant k \leqslant n$, и произвольных $Y, Z \in I(\mathbb{R})$ неравенству

$$q\left(\sum_{k=0}^{n} A_k Y^k,\ \sum_{k=0}^{n} A_k Z^k\right) \leqslant \left(\sum_{k=1}^{n} k\,|A_k|\,|X|^{k-1}\right) q(Y,\ Z).$$

Здесь X— наименьший интервал, содержащий $Y \cup Z$. Полученное неравенство верно независимо от того, как вычисляется X^k — как произведение $X \cdot \cdots \cdot X$ или как одноместная операция согласно определению 3 п.1.1.

Читателю предоставляется в качестве простого упражнения получить аналогичную формулу для рациональных выражений.

(c) $$f(x;\ a) = x/2 - 2e^{ax}.$$

В силу (5) интервальная оценка этой функции удовлетворяет для $A \in I(\mathbb{R})$ и произвольных $Y,\ Z \in I(\mathbb{R})$ неравенству

$$q(Y/2 - 2e^{AY},\ Z/2 - 2e^{AZ}) \leqslant \left(\frac{1}{2} + 2\,|A|\,e^{|A|\,|X|}\right) q(Y,\ Z).$$

Здесь X снова обозначает наименьший интервал, содержащий $Y \cup Z$.

Теорема 8 без труда переносится на функции нескольких переменных. Таким образом, мы получаем формулы для интервальных вычислений отображений f_P, описанных в начале этого микромодуля. Нужно только применить эти неравенства покомпонентно. Это без труда следует из уже доказанных результатов, и мы опускаем подробности. Такие формулы в свою очередь позволяют вычислять практически матрицу P_p, участвующую в определении 3. Приведем простой пример.

Пример. Дано n функций

$$f_i(x_1,\ \ldots,\ x_n;\ a_{i1},\ \ldots,\ a_{in}) = \sum_{\nu=1}^{n} \sin(a_{i\nu} x_\nu), \quad 1 \leqslant i \leqslant n.$$

Интервальная оценка существует для произвольных интервалов $X_\nu \in I(\mathbb{R}),\ 1 \leqslant \nu \leqslant n$ и $A_{ij} \in I(\mathbb{R}),\ 1 \leqslant i,\ j \leqslant n$. Кроме того, каждая из данных функций удовлетворяет условиям теоремы 8 для данных интервалов $A_{ij},\ 1 \leqslant i,\ j \leqslant n$, а именно

$$\begin{aligned} &|f_i(x_1,\ \ldots,\ x_\nu,\ \ldots,\ x_n;\ a_{i1},\ \ldots,\ a_{in}) \\ &\qquad - f_i(x_1,\ \ldots,\ y_\nu,\ \ldots,\ x_n;\ a_{i1},\ \ldots,\ a_{in})| \\ &\leqslant |A_{i\nu}|\,|x_\nu - y_\nu|, \quad x_j \in X_j, \quad 1 \leqslant j \leqslant n, \quad j \neq \nu, \quad x_\nu, \quad y_\nu \in X_\nu. \end{aligned}$$

Если мы теперь применим теорему 8 к каждой компоненте, то получим

$$f_p(x) = (f_i(x)) = \left(\sum_{\nu=1}^{n} \sin(A_{i\nu} X_\nu)\right),$$

$$q(f_p(x),\ f(y)) \leqslant \mathcal{P}_p q(x,\ y), \quad \text{где} \quad \mathcal{P}_p = (|A_{ij}|).$$

Если теперь $\rho(\mathscr{P}_p) < 1$, то, согласно теореме 4, уравнение

$$x = f_p(x) = (f_i(x))$$

имеет единственную неподвижную точку, которая может быть найдена методом итераций. Применяя теорему 5, можно показать, что короткошаговый метод также сходится.

В теореме 4 мы предположим для простоты, что отображение f_P определено на всем пространстве $V_n(\mathbb{C})$, что f_P можно интервально оценить для всех элементов множества $V_n(I(\mathbb{C}))$ и что f_P есть P_p -сжатие. Простые примеры [в частности, пример (b)] показывают, однако, что матрица P_p в определении 3 может зависеть от x и y и что условие $\rho(\mathscr{P}_p) < 1$ может нарушаться для некоторых x, y. В этом случае может помочь следующее утверждение.

Теорема 9. *Пусть* f_p: $\theta \subseteq V_n(\mathbb{C}) \to V_n(\mathbb{C})$ *есть* P_p *-сжатие для всех* x, y *из замкнутого множества* $I(\theta) = \{z \in V_n(I(\mathbb{C})) \mid z \subseteq \theta\}$. *Если* $f_p(x) \in I(\theta)$ *для всех* $x \in I(\theta)$, *то уравнение* $x = f_p(x)$ *имеет единственную неподвижную точку* $x^* \in f(\theta)$, *причем, итерации* $x^{(k+1)} = f_p(x^{(k)})$, $k \geqslant 0$, *сходятся к этой неподвижной точке* x^* при любом начальном векторе $x^{(0)} \in I(\theta)$.

Доказательство можно провести аналогично доказательству теоремы 4.

Заметим, что из неравенства

$$q(x^{(k+m)}, x^{(k)}) \leqslant \sum_{i=1}^{m} \mathscr{P}_p^i q(x^{(k)}, x^{(k-1)}) \leqslant (\mathscr{I}_p - \mathscr{P}_p)^{-1} \mathscr{P}_p q(x^{(k)}, x^{(k-1)}),$$

которое следует из теоремы 4, мы получаем оценку погрешности

$$q(x^{(k)}, x^*) \leqslant (\mathscr{I}_p - \mathscr{P}_p)^{-1} \mathscr{P}_p q(x^{(k)}, x^{(k-1)})$$

для теорем 4 и 9, переходя к пределу при $m \to \infty$ в предыдущем неравенстве.

Сформулируем еще одну лемму, которая будет использована позднее.

Лемма 10. *Пусть* f_p: $x \subseteq \theta \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *—непрерывное отображение.*

Пусть отображение $\mathscr{Y}_p$: $x \subseteq \theta \subseteq V_n(\mathbb{R}) \to V_n(\mathbb{R})$ *также непрерывно и для всех* $x_p \in x$ *существует обращение* $\mathscr{Y}_p(x_p)^{-1}$. *Если* $f_p(x_p) \in x$ *при всех* $x_p \in x$ *верно для отображения* f_p: $x \to V_n(\mathbb{R})$, *определенного формулой*

$$f_p(x_p) := x_p - \mathscr{Y}_p(x_p) f_p(x_p),$$

то f_p *имеет нуль в интервале* x.

Простое доказательство этого утверждения использует теорему Брауэра о неподвижной точке. Действительно, f_p отображает выпуклое

компактное множество $\{x_p \mid x_p \in x\} \subset V_n(\mathbb{R})$ в себя и поэтому имеет неподвижную точку. Так как $\mathcal{Y}_p(x_p)$ несингулярно, отображение f_p имеет нуль в x.

Замечания. Лемма 10 сформулирована и доказана Алефельдом. Она обобщает утверждение для постоянного $\mathcal{Y}_p$, доказанное Муром.

Эта лемма важна, так как мы можем проверить выполнение ее условий с помощью конечного числа арифметических операций над интервальным вектором, найдя мажоранту для отображения $\mathcal{Y}_p(x_p)$ в виде интервального выражения, зависящего от интервального вектора x.

Подробности такой проверки описаны в последующих микромодулях,.

Прямая проверка условий теоремы Брауэра о неподвижной точке представляет собой, напротив, очень трудную задачу.

3.2. Системы линейных уравнений, поддающиеся методу итерации

Мы предполагаем, что рассматриваемая здесь система линейных уравнений уже имеет вид

$$x_p = \mathcal{A}_p x_p + b_p, \tag{1}$$

где $\mathcal{A}_p = (a_{ij})$ и $b_p = (b_i)$.

Пусть известно, что элементы a_{ij} матрицы $\mathcal{A}_p$ лежат в интервалах A_{ij}, а компоненты b_i вектора b_p лежат в интервалах B_i. Нас интересует множество решений, получающихся, когда входные данные изменяются в данных интервалах. Поэтому мы вводим интервальную матрицу $\mathcal{A} = (A_{ij})$ размерности $n \times n$, содержащую интервальные коэффициенты системы, и интервальный вектор $b = (B_i)$, содержащий интервальные правые части. Рассмотрим теперь отображение

$$f_p\colon V_n(\mathbb{C}) \to V_n(\mathbb{C}),$$

определяемое формулой

$$f_p(x_p) = \mathcal{A}_p x_p + b_p.$$

Прежде всего мы имеем следующее утверждение.

Теорема 1. *Метод последовательных приближений*

$$(x)^{(k+1)} = \mathcal{A}x^{(k)} + b, \quad k \geqslant 0, \tag{2}$$

сходится к единственной неподвижной точке x^* уравнения

$$x = \mathcal{A}x + b$$

Для любого $x^{(0)} \in V_n(I(\mathbb{C}))$ *тогда и только тогда, когда*

$$\rho(|\mathscr{A}|) < 1.$$

Доказательство. Покажем, что f_P есть $|\mathscr{A}|$-сжатие. Пусть $x,\ y \in V_n(I(\mathbb{C}))$. Применяя (23, п.2.3) и (25, п.2.3), получаем

$$q(f_p(x),\ f_p(y)) = q(\mathscr{A}x + b,\ \mathscr{A}y + b) \leqslant |\mathscr{A}|\, q(x,\ y).$$

Из теоремы 4 п.3.1 следует, что условие $\rho(|\mathscr{A}|) < 1$ достаточно для сходимости метода и единственности неподвижной точки. Для доказательства обратного утверждения допустим, что последовательные приближения $x^{(k+1)} = \mathscr{A}x^{(k)} + b,\ k \geqslant 0$, сходятся для каждого $x^{(0)} \in V_n(I(\mathbb{C}))$ к неподвижной точке x^*. Мы должны показать, что $\rho(|\mathscr{A}|) < 1$. Из теоремы Перрона и Фробениуса следует, что вещественная неотрицательная матрица $|\mathscr{A}|$ имеет неотрицательный собственный вектор, соответствующий собственному числу $\lambda = \rho(|\mathscr{A}|)$. Из сходимости приближений $x^{(k+1)} = \mathscr{A}x^{(k)} + b,\ k \geqslant 0$, к x^* при любом начальном векторе $x^{(0)}$ следует, что последовательность $\{d(x^{(k)})\}_{k=0}^{\infty}$ сходится к $d(x^*)$. Выберем теперь $x^{(0)}$ таким образом, чтобы $d(x^{(0)})$ был собственным вектором, соответствующим собственному числу $\lambda = \rho(|\mathscr{A}|)$ матрицы $|\mathscr{A}|$, причем хотя бы одна компонента вектора $d(x^{(0)})$ была строго больше, чем соответствующая компонента вектора $d(x^*)$. Тогда из (2) следует в силу (12,п.2.3) и (18, п.2.3), что в предположении $\lambda = \rho(|\mathscr{A}|) \geqslant 1$ верно

$$d(x^{(1)}) = d(\mathscr{A}x^{(0)} + b) = d(\mathscr{A}x^{(0)}) + d(b)$$
$$\geqslant d(\mathscr{A}x^{(0)}) \geqslant |\mathscr{A}|\, d(x^{(0)}) = \lambda d(x^{(0)}) \geqslant d(x^{(0)}),$$
$$d(x^{(2)}) \geqslant |\mathscr{A}|\, d(x^{(1)}) \geqslant \lambda\, |\mathscr{A}|\, d(x^{(0)}) = \lambda^2 d(x^{(0)}) \geqslant d(x^{(0)}).$$

Для произвольного k получаем

$$d(x^{(k+1)}) \geqslant |\mathscr{A}|\, d(x^{(k)}) \geqslant \ldots \geqslant \lambda^{(k+1)} d(x^{(0)}) \geqslant d(x^{(0)}).$$

Переходя к пределу при $k \to \infty$, получаем

$$d(x^*) \geqslant d(x^{(0)}),$$

что противоречит выбору $x^{(0)}$. Поэтому верно $\rho(|\mathscr{A}|) < 1$.

Установим связь этой теоремы с задачей, сформулированной во введении к этому микромодулю. (См. также следствие 6, п.3.1)

Теорема 2. *Пусть* $\mathscr{A}$ *— интервальная матрица, такая что* $\rho(|\mathscr{A}|) < 1$. *Тогда для неподвижной точки* x^* *уравнения*

$v^* = \mathcal{A}x^* + \mathcal{b}$ *(которая существует и единственна в силу теоремы* 1) *верно соотношение*

$$\{y_p = (\mathcal{J}_p - \mathcal{A}_p)^{-1}\, \mathcal{b}_p \mid \mathcal{A}_p \in \mathcal{A},\ \mathcal{b}_p \in \mathcal{b}\} \subseteq \{x_p \mid x_p \in x^*\}.$$

Если $\mathcal{A} = (A_{ij}) \in M_{nn}(I(\mathbb{R}))$, $\mathcal{b} \in V_n(I(\mathbb{R}))$ *и неравенство* $i(A_{ij}) \geqslant 0$ *справедливо для* $A_{ij}=[i(A_{ij}), s(A_{ij})]$, *то* x^* *оптимальна в следующем смысле. Не существует интервального вектора* $x \in V_n(I(\mathbb{R}))$ *такого, что* $v \subseteq x^*$, $x \neq x^*$, *но*

$$\{y_p = (\mathcal{J}_p - \mathcal{A}_p)^{-1}\, \mathcal{b}_p \mid \mathcal{A}_p \in \mathcal{A},\ \mathcal{b}_p \in \mathcal{b}\} \subseteq \{x_p \mid x_p \in x\}.$$

Доказательство. Покажем сначала, что система линейных уравнений $y_p = \mathcal{A}_p y_p + \mathcal{b}_p$ имеет решение

$$y_p = (\mathcal{J}_p - \mathcal{A}_p)^{-1}\, \mathcal{b}_p$$

для $\mathcal{A}_p \in \mathcal{A}$ и $\mathcal{b}_p \in \mathcal{b}$. Так как верно $\mathcal{A}_p \in \mathcal{A}$, имеем $|\mathcal{A}_p| \leqslant |\mathcal{A}|$ и из теоремы Перрона и Фробениуса получаем

$$\rho(\mathcal{A}_p) \leqslant \rho(|\mathcal{A}_p|) \leqslant \rho(|\mathcal{A}|) < 1.$$

Отсюда следует, что матрица $\mathcal{J}_p - \mathcal{A}_p$ несингулярна, что и требовалось.

Рассмотрим теперь последовательные приближения

$$x^{(k+1)} = \mathcal{A}x^{(k)} + \mathcal{b}, \qquad k \geqslant 0,$$

где $x^{(0)} = y_p = \mathcal{A}_p y_p + \mathcal{b}_p$. Из монотонности включения следует, что

$$y_p = \mathcal{A}_p y_p + \mathcal{b}_p \in \mathcal{A}x^{(0)} + \mathcal{b} = x^{(1)},$$

и для произвольного k

$$y_p = \mathcal{A}_p y_p + \mathcal{b}_p \in \mathcal{A}x^{(k)} + \mathcal{b} = x^{(k+1)}.$$

Из $\rho(|\mathcal{A}|) < 1$ следует, что $\lim_{k \to \infty} x^{(k)} = x^*$, а потому и $y_p \in x^*$.

Так как x^* не зависит от начального вектора, мы получаем первую часть теоремы.

Для доказательства второй части теоремы построим вектор $u_p \in V_n(\mathbb{R})$ из n нижних границ компонент вектора x^*. Из верхних границ аналогичным образом строится вектор $v_p \in V_n(\mathbb{R})$. Тогда из $x^* = \mathcal{A}x^* + \mathcal{b}$ следует по правилам интервальной арифметики, что

$$u_p = \mathcal{A}_p^{*} u_p + s_p \quad \text{и} \quad v_p = \mathcal{A}_p^{**} v_p + s_p,$$

где $u_{pi} = (u_i)$, $v_{pi} = (v_i)$ и

$$\mathcal{A}_p^* = (a_{ij}^*), \quad a_{ij}^* = \begin{cases} \iota(A_{ij}), & u_j > 0, \\ s(A_{ij}), & u_j \leqslant 0, \end{cases}$$

$$\mathcal{A}_p^{**} = (a_{ij}^{**}), \quad a_{ij}^{**} = \begin{cases} s(A_{ij}), & v_j < 0 \\ i(A_{ij}), & v_j \leqslant 0, \end{cases}$$

$$\iota_p = (i(B_i)), \quad s_p = (s(B_i)).$$

Из этих равенств следует, что u_p и v_p являются членами множества

$$\{y_p = (\mathcal{I}_p - \mathcal{A}_p)^{-1} b_p \mid \mathcal{A}_p \in \mathcal{A}, \ b_p \in \mathcal{b}\},$$

что и завершает доказательство.

Метод итерации, рассмотренный в теореме 1, можно назвать полношаговым (T) по аналогии с соответствующим методом для «точечной системы уравнений». Аналогичный короткошаговый метод (S) получается разложением интервальной матрицы $\mathcal{A}$ в сумму

$$\mathcal{A} = \mathcal{L} + \mathcal{D} + \mathcal{U},$$

где $\mathcal{L}$—строго нижняя треугольная матрица, $\mathcal{D}$ — диагональная матрица и $\mathcal{U}$— строго верхняя треугольная матрица. Тогда короткошаговый итерационный метод определяется формулами

$$x^{(k+1)} = \mathcal{L} x^{(k+1)} + (\mathcal{D} + \mathcal{U}) x^{(k)} + b, \quad k \geqslant 0. \qquad (3)$$

Следующее утверждение касается сходимости этого короткошагового метода.

Теорема 3. *Итерационный метод*

$$x^{(k+1)} = \mathcal{L} x^{(k+1)} + (\mathcal{D} + \mathcal{U}) x^{(k)} + b, \quad k \geqslant 0,$$

с произвольным начальным вектором $x^{(0)} \in V_n(I(\mathbb{C}))$ *сходится к единственной неподвижной точке* x^* *тогда и только тогда, когда*

$$\rho((\mathcal{I}_p - |\mathcal{L}|)^{-1}(|\mathcal{D}| + |\mathcal{U}|)) < 1.$$

Доказательство. Мы собираемся применить теорему 5 из п.3.1 и с этой целью полагаем

$$f: V_n(\mathbb{C}) \to V_n(\mathbb{C}),$$

где

$$f_p(x_p) = \mathcal{L}_p x_p + (\mathcal{D}_p + \mathcal{U}_p) x_p + b_p$$

и

$$g_p: V_n(\mathbb{C}) \times V_n(\mathbb{C}) \to V_n(\mathbb{C}),$$

где

$$g_p(x_p, y_p) = \mathcal{L}_p x_p + (\mathcal{D}_p + \mathcal{U}_p) y_p + b_p.$$

Мы имеем тогда

$$g_p(x, x) = f_p(x_p) \quad \text{для всех} \quad x \in V_n(I(\mathbb{C})),$$

и из (23, п.2.3) и (25, п.2.3) следует, что

$$q(g_p(x, z), g_p(y, z)) = q(\mathcal{L}x + (\mathcal{D} + \mathcal{U})z + b, \mathcal{L}y + (\mathcal{D} + \mathcal{U})z + b) \leqslant |\mathcal{L}|\, q(x, y),$$

$$q(g_p(z, x), g_p(z, y)) = q(\mathcal{L}z + (\mathcal{D} + \mathcal{U})x + b, \mathcal{L}z + (\mathcal{D} + \mathcal{U})y + b) \leqslant (|\mathcal{D}| + |\mathcal{U}|)\, q(x, y)$$

для всех x, y_p, $z \in V_n(I(\mathbb{C}))$. Мы имеем $\rho(|\mathcal{L}|) = 0$, так как $|\mathcal{L}|$ — строго нижняя треугольная матрица. Поэтому, полагая $Q_p := |\mathcal{L}|$, $\mathcal{R}_p := |\mathcal{D}| + |\mathcal{U}|$, мы оказываемся в условиях теоремы 5, из микромодуля 30, так что условие

$$\rho((\mathcal{I}_p - |\mathcal{L}|)^{-1}(|\mathcal{D}| + |\mathcal{U}|)) < 1$$

достаточное. Доказательство необходимости этого условия для сходимости при любом начальном векторе проводится так же, как соответствующее доказательство для полношагового метода в теореме 1.

Теорема Штейна и Розенберга, а также ее обобщение утверждают, что для $\mathcal{A} = \mathcal{L} + \mathcal{D} + \mathcal{U}$ верна эквивалентность $\rho(|\mathcal{A}|) < 1$ тогда и только тогда, когда

$$\rho((\mathcal{I}_p - |\mathcal{L}|)^{-1}(|\mathcal{D}| + |\mathcal{U}|)) < 1.$$

Так как условия сходимости полношагового и короткошагового методов необходимы и достаточны, получаем следующее утверждение.

Теорема 4. *Полношаговый метод* (2) *сходится для любого начального значени я* $x^{(0)} \in V_n(I(\mathbb{C}))$ *к единственной неподвижной точке тогда и только тогда, когда короткошаговый метод* (3) *сходится к единственной неподвижной точке для любого начального значения* $x^{(0)} \in V_n(I(\mathbb{C}))$.

Этот результат существенно отличается от соответствующего результата для точечных систем уравнений, где сходимость или расходимость полношагового метода не обязательно означает сходимость или расходимость короткошагового метода.

Поскольку умножение интервальных матриц на интервальные векторы в общем случае не дистрибутивно, то даже для случая $\rho(|\mathcal{A}|) < 1$ не очевидно, что неподвижная точка полношаговой итерации, удовлетворяющая равенству

$$x^* = \mathcal{A}x^* + b,$$

совпадает с неподвижной точкой короткошаговой итерации, удовлетворяющей равенству

$$\tilde{x}^* = \mathcal{L}\tilde{x}^* + (\mathcal{D} + \mathcal{U})\tilde{x}^* + \mathcal{b}.$$

Однако, используя специальный вид матриц $\mathcal{L}$, $\mathcal{D}$ и $\mathcal{U}$, мы получаем с помощью определения действий над интервальными матрицами и векторами, что

$$\tilde{x}^* = \mathcal{L}x^* + (\mathcal{D} + \mathcal{U})\tilde{x}^* + \mathcal{b} = (\mathcal{L} + \mathcal{D} + \mathcal{U})x^* + \mathcal{b} = \mathcal{A}\tilde{x}^* + \mathcal{b}.$$

Отсюда получается, что $x^* = x^*$, и приходим к следующему утверждению.

Следствие 5. *Если* $\rho(|\mathcal{A}|) < 1$, *то полношаговый и короткошаговый методы сходятся к неподвижной точке* x^* *уравнения*

$$x = \mathcal{A}x + \mathcal{b}.$$

Рассмотрим теперь симметрический короткошаговый метод (SS), в котором матрица $\mathcal{A}$ раскладывается в сумму

$$\mathcal{A} = \mathcal{L} + \mathcal{U},$$

где $\mathcal{L}$ — строго нижняя, а $\mathcal{U}$ — строго верхняя треугольные матрицы. Метод (SS) определяется соотношениями

$$\begin{cases} x^{(k+1/2)} = \mathcal{L}x^{(k+1/2)} + \mathcal{U}x^{(k)} + \mathcal{b}, \\ x^{(k+1)} = \mathcal{L}x^{(k+1/2)} + \mathcal{U}x^{(k+1)} + \mathcal{b}, \quad k \geqslant 0. \end{cases} \tag{SS}$$

Если не все диагональные элементы матрицы $\mathcal{A}$ обращаются в нуль, то вместо этого мы должны рассмотреть итерационный метод

$$\begin{cases} x^{(k+1/2)} = \mathcal{L}x^{(k+1/2)} + \mathcal{D}x^{(k)} + \mathcal{U}x^{(k)} + \mathcal{b}, \\ x^{(k+1)} = \mathcal{L}x^{(k+1/2)} + \mathcal{D}x^{(k+1/2)} + \mathcal{U}x^{(k+1)} + \mathcal{b}, \quad k \geqslant 0. \end{cases} \tag{SS'}$$

Для этого итерационного метода можно доказать утверждения, аналогичные тем, которые будут доказаны ниже для метода (SS).

Сходимость метода (SS) будет получена из следующего общего результата.

Теорема 6. *Пусть интервальная матрица* $\mathcal{A} \in M_{nn}(I(\mathbb{C}))$ *разложена в сумму* $\mathcal{A} = \mathcal{M} + \mathcal{N}$ *двух интервальных матриц* $\mathcal{M}$ *и* $\mathcal{N}$, *для которых верно* $\rho(|\mathcal{M}|) < 1$ *и* $\rho(|\mathcal{N}|) < 1$. *Тогда для произвольного вектора* $\mathcal{b} \in V_n(I(\mathbb{C}))$ *верно следующее:*

(a) *Для любого интервального вектора* $x^{(0)} \in V_n(I(\mathbb{C}))$ *существует последовательность* $\{x^{(k)}\}_{k=0}^{\infty}$, *которая идовлетворяет итерационным формулам*

$$\text{(V)} \begin{cases} x^{(k+1/2)} = \mathcal{M}x^{(k+1/2)} + \mathcal{N}x^{(k)} + \mathcal{b}, \\ x^{(k+1)} = \mathcal{M}x^{(k+1/2)} + \mathcal{N}x^{(k+1)} + \mathcal{b}, \quad k = 0, 1, 2, \ldots. \end{cases}$$

(b) *Если* $\rho\left((\mathcal{I}_p - |\mathcal{N}|)^{-1}|\mathcal{M}|(\mathcal{I}_p - |\mathcal{M}|)^{-1}|\mathcal{N}|\right) < 1$, *то уравнение* $x = \mathcal{A}x + b$ *имеет единственную неподвижную точку* x^*. *Если сверх того*

$$\mathcal{A}x^* = (\mathcal{M} + \mathcal{N})x^* = \mathcal{M}x^* + \mathcal{N}x^*,$$

то последовательность, вычисленная по формулам (V), *сходится к* x^* *для любого начального вектора* $x^{(0)}$. *(Как мы уже видели, для интервальных матриц не выполнен дистрибутивный закон. См. п. 2.3 формулы* (6).)

(c) *Обратно, если уравнение* $x = \mathcal{A}x + b$ *имеет единственную неподвижную точку* x^* *и последовательность* (V) *сходится к* x^* *для любого начального приближения* $x^{(0)}$, *то*

$$\mathcal{A}x^* = (\mathcal{M} + \mathcal{N})x^* = \mathcal{M}x^* + \mathcal{N}x^*$$

и

$$\rho\left((\mathcal{I}_p - |\mathcal{N}|)^{-1}|\mathcal{M}|(\mathcal{I}_p - |\mathcal{M}|)^{-1}|\mathcal{N}|\right) < 1$$

Доказательство. (a) Для произвольного интервального вектора z из (23, п.2.3) и (25, п.2.3) следует, что для любых векторов x, y имеет место

$$q(\mathcal{M}x + \mathcal{N}z + b,\ \mathcal{M}y + \mathcal{N}z + b) = q(\mathcal{M}x, \mathcal{M}y) \leqslant |\mathcal{M}|q(x, y).$$

Ввиду $\rho(|\mathcal{M}|) < 1$ мы получаем по теореме 1, что для любого k уравнение

$$x^{(k+1/2)} = \mathcal{M}x^{(k+1/2)} + \mathcal{N}x^{(k)} + b$$

имеет единственную неподвижную точку $x^{(k+1/2)}$. Аналогично можно показать, что для любого k уравнение

$$x^{(k+1)} = \mathcal{M}x^{(k+1/2)} + \mathcal{N}x^{(k+1)} + b$$

имеет единственную неподвижную точку $x^{(k+1)}$. Тем самым доказаны существование и единственность последовательности $\{x^{(k)}\}_{k=0}^{\infty}$ при данном начальном векторе $x^{(0)}$.

(b) Сначала покажем, что $\rho(|\mathcal{A}|) < 1$. Так как $\rho(|\mathcal{M}|) < 1$, $\rho(|\mathcal{N}|) < 1$, то обратные матрицы

$$(\mathcal{I}_p - |\mathcal{M}|)^{-1} \text{ и } (\mathcal{I}_p - |\mathcal{N}|)^{-1},$$

как известно, существуют и неотрицательны. Поэтому вещественная матрица

$$(\mathcal{I}_p - |\mathcal{N}|)^{-1}|\mathcal{M}|(\mathcal{I}_p - |\mathcal{M}|)^{-1}|\mathcal{N}| = (\mathcal{I} - |\mathcal{N}|)^{-1} \times (\mathcal{I}_p - |\mathcal{M}|)^{-1}|\mathcal{M}||\mathcal{N}|$$

также неотрицательна. Применяя известную теорему, получаем

$$\mathcal{O}_p \leqslant (\mathcal{I}_p - (\mathcal{I}_p - |\mathcal{N}|)^{-1}(\mathcal{I}_p - |\mathcal{M}|)^{-1}(|\mathcal{M}| |\mathcal{N}|)^{-1}$$
$$= (\mathcal{I}_p - (|\mathcal{M}| + |\mathcal{N}|))^{-1}(\mathcal{I}_p - |\mathcal{M}|)(\mathcal{I}_p - |\mathcal{N}|),$$

а используя

$$(\mathcal{I}_p - |\mathcal{M}|)^{-1} \geqslant \mathcal{O}_p, \quad (\mathcal{I}_p - |\mathcal{N}|)^{-1} \geqslant \mathcal{O}_p,$$

имеем, наконец,

$$(\mathcal{I}_p - (|\mathcal{M}| + |\mathcal{N}|))^{-1} \geqslant \mathcal{O}_p.$$

Отсюда по известной теореме следует неравенство.

$$\rho(|\mathcal{M}| + |\mathcal{N}|) < 1.$$

Из соотношения

$$|\mathcal{A}| = |\mathcal{M} + \mathcal{N}| \leqslant |\mathcal{M}| + |\mathcal{N}|$$

ввиду теоремы Перрона и Фробениуса следует, что

$$\rho(|\mathcal{A}|) \leqslant \rho(|\mathcal{M}| + |\mathcal{N}|) < 1.$$

Поэтому ввиду теоремы 1 уравнение $x = \mathcal{A}x + b$ имеет единственную неподвижную точку x^*. Из равенства

$$x^* = \mathcal{A}x^* + b = \mathcal{M}x^* + \mathcal{N}x^* + b$$

с помощью (23, п.2.3)—(25, п.2.3) и (V) следует, что

$$q(x^{(k+1)}, x^*) = q(\mathcal{M}x^{(k+1/2)} + \mathcal{N}x^{(k+1)} + b, \mathcal{M}x^* + \mathcal{N}x^* + b)$$
$$\leqslant q(\mathcal{M}x^{(k+1/2)} + \mathcal{N}x^{(k+1)}, \mathcal{M}x^{(k+1/2)} + \mathcal{N}x^*)$$
$$+ q(\mathcal{M}x^{(k+1/2)} + \mathcal{N}x^*, \mathcal{M}x^* + \mathcal{N}x^*)$$
$$\leqslant |\mathcal{N}| q(x^{(k+1)}, x^*) + |\mathcal{M}| q(x^{(k+1/2)}, x^*),$$

а так как $(\mathcal{I}_p - |\mathcal{N}|)^{-1} \geqslant \mathcal{O}_p$, это дает

$$q(x^{(k+1)}, x^*) \leqslant (\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| q(x^{(k+1/2)}, x^*).$$

Аналогичным образом получаем

$$q(x^{(k+1/2)}, x^*) \leqslant (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}| q(x^{(k)}, x^*);$$

откуда, наконец,

$$q(x^{(k+1)}, x^*) \leqslant (\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}| q(x^k, x^*)$$
$$\leqslant \{(\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|\}^{k+1} q(x^{(0)}, x^*).$$

Из того, что спектральный радиус выражения в фигурных скобках меньше 1, получаем, что $\lim_{k \to \infty} x^{(k)} = x^*$.

(c): Пусть уравнение $x = \mathcal{A}x + b$ имеет единственную неподвижную точку x^*. Из неравенства

$$q(x^{(k+1/2)}, x^*) \leqslant (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}| q(x^{(k)}, x^*),$$

которое выводится так же, как в доказательстве п. (b), следует, что последовательность $\{x^{(k+1/2)}\}_{k=0}^{\infty}$ сходится к x^* для любого $x^{(0)}$. Отсюда и из верхнего равенства (V) следует при $k \to \infty$, что

$$x^* = \mathcal{M}x^* + \mathcal{N}x^* + b,$$

т. е.

$$o_p = q(x^*, x^*) = q(\mathcal{M}x^* + \mathcal{N}x^* + b, \mathcal{A}x^* + b) = q(\mathcal{M}x^* + \mathcal{N}x^*, \mathcal{A}x^*)$$

или

$$\mathcal{A}x^* = (\mathcal{M} + \mathcal{N})x^* = \mathcal{M}x^* + \mathcal{N}x^*.$$

Мы должны еще доказать, что

$$\rho((\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|) < 1.$$

Чтобы сделать это, поступаем так же, как в доказательстве теоремы 1. Из теоремы Перрона и Фробениуса известно, что матрица

$$(\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|$$

имеет неотрицательный собственный вектор, соответствующий неотрицательному собственному числу

$$\tilde{\lambda} = \rho((\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|).$$

Теперь мы выбираем $x^{(0)}$ так, чтобы $d(x^{(0)})$ был собственным вектором, соответствующим собственному числу $\tilde{\lambda}$, и при этом хотя бы одна компонента вектора $d(x^{(0)})$ была больше, чем соответствующая компонента вектора $d(x^*)$. Тогда из (V) с помощью (12, п.2.3) и (18, п.2.3) следует, что

$$d(x^{(k+1/2)}) \geqslant |\mathcal{M}| d(x^{(k+1/2)}) + |\mathcal{N}| d(x^{(k)})$$

или

$$d(x^{(k+1/2)}) \geqslant (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}| d(x^{(k)}),$$

а также

$$d(x^{(k+1/2)}) \geqslant (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}| d(x^{(k)})$$

или

$$d(x^{(k+1)}) \geqslant (\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| d(x^{(k+1/2)}).$$

Наконец, мы получаем

$$\begin{aligned} d(x^{(k+1)}) &\geqslant (\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}| d(x^{(k)}) \\ &\geqslant \{(\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|\}^{(k+1)} d(x^{(0)}) \\ &= \tilde{\lambda}^{(k+1)} d(x^{(0)}). \end{aligned}$$

Из сходимости итераций (V) к x^* следует сходимость последовательности $\{d(x^{(k)})\}_{k=0}^{\infty}$ к $d(x^*)$. Предположение $\tilde{\lambda} \geqslant 1$ приводит к неравенству

$$d(x^{(k+1)}) \geqslant \tilde{\lambda}^{(k+1)} d(x^{(0)}) \geqslant d(x^{(0)}), \quad k \geqslant 0,$$

что в пределе при $k \to \infty$ дает $d(x^*) \geqslant d(x^{(0)})$, а это противоречит выбору вектора $x^{(0)}$. Поэтому $\tilde{\lambda} < 1$, и теорема доказана.

Теперь возьмем в теореме 6

$$\mathcal{M} := \mathcal{L}, \quad \mathcal{N} := \mathcal{U}.$$

Тогда имеем $\rho(|\mathcal{M}|) = \rho(|\mathcal{N}|) = 0$. Ввиду специального выбора матриц $\mathcal{L}$ и $\mathcal{U}$ равенство

$$\mathcal{A}x = \mathcal{L}x + \mathcal{U}x$$

справедливо для всех интервальных векторов. Это дает приводимое ниже следствие теоремы 6.

Следствие 7. *Симметрический короткошаговый метод* (SS) *сходится к единственной неподвижной точке* x^* *уравнения* $x = \mathcal{A}x + b$ *для любого начального вектора* $x^{(0)}$ *тогда и только тогда, когда спектральный радиус матрицы*

$$(\mathcal{I}_p - |\mathcal{U}|)^{-1} |\mathcal{L}| (\mathcal{I}_p - |\mathcal{L}|)^{-1} |\mathcal{U}|$$

меньше единицы.

Мы уже установили в доказательстве теоремы 6 (b), что неравенство

$$\rho((\mathcal{I}_p - |\mathcal{N}|)^{-1} |\mathcal{M}| (\mathcal{I}_p - |\mathcal{M}|)^{-1} |\mathcal{N}|) < 1$$

влечет за собой $\rho(|\mathcal{A}|) < 1$. Теперь мы хотим показать, что в предположении

$$|\mathcal{A}| = |\mathcal{M}| + |\mathcal{N}|$$

верна и обратная импликация, если выполнены условия $\rho(|\mathcal{M}|) < 1$ и $\rho(|\mathcal{N}|) < 1$ из теоремы 6. Итак, предположим, что

$$\rho(|\mathcal{A}|) = \rho(|\mathcal{M}| + |\mathcal{N}|) < 1.$$

В силу известной теоремы матрица, обратная к $\mathcal{I}_p - |\mathcal{A}|$, существует, и мы имеем $(\mathcal{I}_p - |\mathcal{A}|)^{-1} \geqslant \mathcal{O}_p$. Рассмотрим следующее разложение:

$$\mathcal{I}_p - |\mathcal{A}| = (\mathcal{I}_p - |\mathcal{M}|)(\mathcal{I}_p - |\mathcal{N}|) - |\mathcal{M}||\mathcal{N}|.$$

Из $\rho(|\mathcal{M}|) < 1$, $\rho(|\mathcal{N}|) < 1$ следует, что обращения матриц $\mathcal{I}_p - |\mathcal{M}|$ и $\mathcal{I}_p - |\mathcal{N}|$ существуют и неотрицательны. Поэтому неотрицательно и произведение $(\mathcal{I}_p - |\mathcal{N}|)^{-1} \times (\mathcal{I}_p - |\mathcal{M}|)^{-1}$. Тем самым рассматриваемое разложение матрицы $\mathcal{I}_p - |\mathcal{A}|$ регулярно, так как матрица $|\mathcal{M}||\mathcal{N}|$ неотрицательна. Отсюда с помощью известной теоремы мы получаем соотношение

$$\rho((\mathcal{I}_p - |\mathcal{N}|)^{-1}(\mathcal{I}_p - |\mathcal{M}|)^{-1}|\mathcal{M}||\mathcal{N}|) = \rho((\mathcal{I}_p - |\mathcal{N}|)^{-1} \times |\mathcal{M}|(\mathcal{I}_p - |\mathcal{M}|)^{-1}|\Lambda| < 1.$$

Собирая все вместе и применяя теорему 6 (b), получаем следующее утверждение.

Теорема 8. *Пусть интервальная матрица $\mathcal{A}$ разложена в сумму $\mathcal{A} = \mathcal{M} + \mathcal{N}$ двух интервальных матриц, для которых выполнено*

$$|\mathcal{A}| = |\mathcal{M}| + |\mathcal{N}| \quad и \quad \rho(|\mathcal{M}|) < 1, \quad \rho(|\mathcal{N}|) < 1.$$

Тогда неравенство

$$\rho((\mathcal{I}_p - |\mathcal{N}|)^{-1}|\mathcal{M}|(\mathcal{I}_p - |\mathcal{M}|)^{-1}|\mathcal{N}|) < 1$$

эквивалентно неравенству

$$\rho(|\mathcal{A}|) < 1.$$

Равенство $|\mathcal{A}| = |\mathcal{M}| + |\mathcal{N}|$ будет выполнено, например, в случае, когда $\mathcal{A} = (A_{ij})$ разложена в сумму $\mathcal{A} = \mathcal{M} + \mathcal{N}$ с $\mathcal{M} = (M_{ij})$, $\mathcal{N} = (N_{ij})$ таким образом, что для любых $1 \le i, j \le n$ по крайней мере одна из компонент M_{ij} и N_{ij} равна нулю.

Это последнее условие выполнено, например, для разложения, с которого начинается описание симметрического короткошагового метода (SS). Теорема 8 немедленно дает приводимое ниже утверждение.

Следствие 9. *Симметрический короткошаговый метод сходится к единственной неподвижной точке x^* уравнения $x = \mathcal{A}x + \mathcal{b}$ для любого начального вектора $x^{(0)} \in V_n(I(\mathbb{C}))$ тогда и только тогда, когда $\rho(|\mathcal{A}|) < 1$, т.е. когда полношаговый метод (а потому и короткошаговый метод) сходится к x^* для любого $x^{(0)} \in V_n(I(\mathbb{C}))$*

Теперь мы рассмотрим скорость, с которой сходится к x^* последовательность $\{x^{(k)}\}_{k=0}^{\infty}$ интервальных векторов, порожденных итерационной процедурой

$$x^{(k+1)} = f_p(x^{(k)}), \quad k \ge 0. \tag{4}$$

Определение 10. *Пусть $x^* = f_p(x^*)$ и пусть $\mathfrak{S}$ —множество всех последовательностей $\{x^{(k)}\}_{k=0}^{\infty}$, вычисленных по формуле* (4) *и удовлетворяющих условию $\lim_{k\to\infty} x^{(k)} = x^*$. Тогда величина*

$$\alpha = \sup\left\{\lim_{k\to\infty} \sup \|q(x^{(k)}, x^*\|^{1/k} \{x^{(k)}\}_{k=0}^{\infty} \in \mathfrak{S}\right\}$$

называется асимптотическим фактором сходимости итерации (4) *к точке x^*.*

Пусть $\{x^{(k)}\}_{k=0}^{\infty}$ — последовательность, сходящаяся к x^*. Положим

$$\beta = \lim_{k\to\infty} \sup \| q(x^{(k)}, x^*)\|^{1/k}.$$

Так как $\lim_{k\to\infty} \| q(x^{(k)}, x^*)\| = 0$, мы получаем, что $0 \leqslant \beta < 1$, а значит, и $0 \leqslant \alpha \leqslant 1$. Из определения β следует, что для любого $\varepsilon > 0$ найдется k_0, такое что

$$\| q(x^{(k)}, x^*)\| \leqslant (\beta + \varepsilon)^k, \quad k \geqslant k_0. \qquad (5)$$

Если $\beta < 1$, то можно выбрать $\varepsilon > 0$ таким образом, что $\beta + \varepsilon < 1$. Тогда неравенство (5) показывает, что последовательность $\| q(x^{(k)}, x^*)\|$ асимптотически сходится к нулю не хуже, чем геометрическая прогрессия со знаменателем $\beta + \varepsilon$. Точная верхняя грань по всем последовательностям из $\mathfrak{C}$ характеризует асимптотически наихудший выбор вектора $x^{(0)}$. Определение 10 — это непосредственное обобщение определения фактора асимптотической сходимости для последовательностей точечных векторов.

Для дальнейшего важно, что α не зависит от нормы. Чтобы убедиться в этом, достаточно показать, что β не зависит от нормы Пусть $\|\cdot\|$ и $\|\cdot\|'$ — две нормы на векторах. Из теоремы об эквивалентности норм следует, что существуют вещественные числа $d \geqslant c > 0$, такие, что $c\| x_p\| \leqslant \| x_p\|' \leqslant d\| x_p\|$ для любых точечных векторов. Отсюда мы получаем

$$\begin{aligned}\lim_{k\to\infty} \sup \| q(x^{(k)}, x^*)\|^{1/k} &\leqslant \lim_{k\to\infty} \left(\frac{1}{c}\right)^{1/k} \lim_{k\to\infty} \sup \| q(x^{(k)}, x^*)\|'^{1/k} \\ &\leqslant \lim_{k\to\infty} \left(\frac{d}{c}\right)^{1/k} \lim_{k\to\infty} \sup \| q(x^{(k)}, x^*)\|^{1/k} \\ &= \lim_{k\to\infty} \sup \| q(x^{(k)}, x^*)\|^{1/k}.\end{aligned}$$

Перед тем как применить определение 10 к методам, рассмотренным в этом микромодуле, докажем одну теорему о точечных матрицах.

Рассмотрим положительный точечный вектор $h_p = (h_i)$, $h_i > 0$ и диагональную точечную матрицу $\mathscr{H}_p = \operatorname{diag}(1/h_i)$. Тогда неравенство

$$\| x_p\| = \max_{1 \leqslant i \leqslant n} (|x_i|/h_i)$$

определяет монотонную норму на векторах (короче, векторную норму), т .е. из

$$|x_p| \leqslant |y_p|$$

следует, что

$$\| x_p\| \leqslant \| y_p\|$$

(в частности, из $o_p \leqslant x_p \leqslant y_p$ следует, что справедливо неравенство $\|x_p\| \leqslant \|y_p\|$).

Полагая

$$\|\mathscr{A}_p\| = \sup_{\|x_p\|=1} \|\mathscr{A}_p x_p\|,$$

мы получаем норму на матрицах, подчиненную векторной норме $\|x_p\|$, причем

$$\|\mathscr{A}_p\| = \max_{1 \leqslant i \leqslant n} \left(\frac{1}{h_i} \sum_{j=1}^{n} |a_{ij}| h_j \right).$$

Теперь докажем следующее утверждение.

Теорема 11. *Для любой точечной матрицы* $\mathscr{A}_p \geqslant O_p$ *и любого* $\varepsilon > 0$ *существует монотонная векторная норма* $\|\cdot\|$, *такая что имеет место*

$$\|\mathscr{A}_p\| = \sup_{\|x_p\|=1} \|\mathscr{A}_p x_p\| \leqslant \rho(\mathscr{A}_p) + \varepsilon.$$

Доказательство. Пусть матрица $\mathscr{A}_p \geqslant O_p$ неприводима. Тогда $\mathscr{A}_p$ имеет положительный собственный вектор $c_p = (c_i)$, $c > 0$, соответствующий собственному числу $\lambda = \rho(\mathscr{A}_p)$. Из равенства $\mathscr{A}_p c_p = \lambda c_p$ следует, что

$$\rho(\mathscr{A}_p) = \lambda = \frac{1}{c_i} \sum_{j=1}^{n} a_{ij} c_j = \|\mathscr{A}_p\| = \sup_{\|x_p\|=1} \|\mathscr{A}_p x_p\|,$$

где

$$\|x_p\| = \max_{1 \leqslant i \leqslant n} \frac{|x_i|}{c_i},$$

т. е.

$$\|\mathscr{A}_p\| \leqslant \rho(\mathscr{A}_p) + \varepsilon.$$

Если $\mathscr{A}_p \geqslant O_p$ приводима, то определим неприводимую матрицу $\tilde{\mathscr{A}}_p \geqslant O_p$ равенствами

$$\tilde{\mathscr{A}}_p = (\tilde{a}_{ij}), \quad \text{где } \tilde{a}_{ij} = \begin{cases} a_{ij}, & \text{если } a_{ij} > 0 \\ a > 0, & \text{если } a_{ij} = 0. \end{cases}$$

Очевидно, что $\tilde{\mathscr{A}}_p \geqslant \mathscr{A}_p \geqslant O_p$. Если теперь $\tilde{c}_p = (\tilde{c}_i)$ — положительный собственный вектор матрицы $\tilde{\mathscr{A}}_p$, соответствующий собственному числу $\lambda = \rho(\mathscr{A}_p)$, то из неравенства $|\lambda| \leqslant \|\tilde{\mathscr{A}}_p\|$ (которое верно для любой матричной нормы) следует, что

$$\rho(\tilde{\mathscr{A}}_p)=\frac{1}{\tilde{c}_i}\sum_{j=1}^{n}\tilde{a}_{ij}\tilde{c}_j=\|\tilde{\mathscr{A}}_p\|\geqslant\max_{1\leqslant i\leqslant n}\frac{1}{c_i}\sum_{j=1}^{n}a_{ij}\tilde{c}_j=\|\mathscr{A}_p\|\geqslant\rho(\mathscr{A}_p),$$

где $\|x_p\|=\max_{1\leqslant i\leqslant n}|x_i|/\tilde{c}_i$ и $\|\mathscr{A}_p\|=\sup_{\|x_p\|=1}\|\mathscr{A}_p x_p\|$. Так как спектральный радиус $\rho(\mathscr{A}_p)$ — непрерывная функция от элементов матрицы $\tilde{\mathscr{A}}_p$, получаем, что для каждого $\varepsilon>0$ существует $\delta=\delta(\varepsilon)>0$, такое что

$$\rho(\tilde{\mathscr{A}}_p)-\rho(\mathscr{A}_p)\leqslant\varepsilon$$

имеет место для всех $a\leqslant\delta(\varepsilon)$. Так как

$$\rho(\tilde{\mathscr{A}}_p)\geqslant\|\mathscr{A}_p\|,$$

получаем отсюда, что

$$\|\mathscr{A}_p\|-\rho(\mathscr{A}_p)\leqslant(\tilde{\mathscr{A}}_p)-\rho(\mathscr{A}_p)\leqslant\varepsilon$$

или

$$\|\mathscr{A}_p\|\leqslant\rho(\mathscr{A}_p)+\varepsilon.$$

После этой подготовки мы легко докажем следующее утверждение.

Теорема 12. *Пусть дано уравнение*

$$x=\mathscr{A}x+b,$$

где b *— интервальный вектор и* $\mathscr{A}$ *— интервальная матрица, для которой* $\rho(|\mathscr{A}|)<1$. *Тогда асимптотический фактор сходимости* α_T *для полношагового метода удовлетворяет неравенству*

$$\alpha_T\leqslant\rho(|\mathscr{A}|),$$

фактор α_s *для короткошагового метода удовлетворяет неравенству*

$$\alpha_S\leqslant\rho((\mathscr{I}_p-|\mathscr{L}|)^{-1}(|\mathscr{D}|+|\mathscr{U}|)),$$

а фактор α_{SS} *для симметрического короткошагового метода удовлетворяет неравенству*

$$\alpha_{SS}\leqslant\rho((\mathscr{I}_p-|\mathscr{U}|)^{-1}|\mathscr{L}|(\mathscr{I}_p-|\mathscr{L}|)^{-1}|\mathscr{U}|).$$

Доказательство. Проведем доказательство для полношагового метода. Если $\rho(|\mathscr{A}|)<1$, то, согласно теореме 1, полношаговый метод сходится для любого начального вектора $x^{(0)}$ к единственной неподвижной точке x^* уравнения $x=\mathscr{A}x+b$. Выбрав $x^{(0)}$ произвольным образом, мы получаем из свойств расстояния q, что

$$q(x^{(1)},x^*)\leqslant|\mathscr{A}|q(x^{(0)},x^*),$$

$$q(x^{(2)},x^*)\leqslant|\mathscr{A}|^2q(x^{(0)},x^*),$$

и для произвольного $k\geq1$

$$q(x^{(k)},x^*)\leqslant|\mathscr{A}|^kq(x^{(0)},x^*).$$

Используя монотонную векторную норму, которая существует по теореме 11, и подчиненную ей матричную норму, имеем

$$\| q(x^{(k)}, x^*) \| \leqslant (\rho(|\mathscr{A}|) + \varepsilon^k \| q(x^{(0)}, x^*) \|,$$

т. е.

$$\lim_{k \to \infty} \sup \| q(x^{(k)}, x^*) \|^{1/k} \leqslant \rho(|\mathscr{A}|) + \varepsilon.$$

Так как $\varepsilon > 0$ было произвольным и $\alpha_т$ не зависит от нормы, получаем неравенство $\alpha_т \leqslant \rho(|\mathscr{A}|)$, т. е. утверждение теоремы для этого случая. Остальные случаи рассматриваются аналогично.

Неизвестно, можно ли поставить знак равенства в наших оценках для $\alpha_т$, α_s, α_{ss}. Чтобы доказать, что это верно, скажем, для полношагового метода, нужно было бы указать начальный вектор $x^{(0)}$, для которого имеет место

$$\lim_{k \to \infty} \sup \| q(x^{(k)}, x^*) \| = \rho(|\mathscr{A}|).$$

Это можно довольно легко сделать в конкретных случаях, но доказательства для общего случая нет.

Замечания. Непосредственное применение интервального анализа к итерационному решению систем уравнений было впервые рассмотрено и показано, что полношаговый метод (2) для вещественной интервальной матрицы $\mathscr{A}$ и вещественного интервального вектора $\mathscr{b}$ сходится к единственной неподвижной точке, если $\| |\mathscr{A}| \| < 1$. Необходимые и достаточные условия из теорем 1 и 3 были найдены для вещественных интервальных матриц и для интервальных матриц с элементами из $R(\mathbb{C})$. Заметим, не вдаваясь в подробности, что более общую задачу о неподвижной точке

$$\mathscr{X} = \mathscr{A}\mathscr{X} + \mathscr{B}, \quad \mathscr{A}, \mathscr{B}, \mathscr{X} \in M_{nn}(I(\mathbb{C}))$$

можно исследовать тем же методом, который был использован в этом микромодуле. Например, если $\rho(\mathscr{A}) < 1$, то для итерации

$$\mathscr{X}^{(k+1)} = \mathscr{A}\mathscr{X}^{(k)} + \mathscr{B}, \quad k \geqslant 0,$$

имеется единственная неподвижная точка $\mathscr{X}^*$ при любой начальной матрице $\mathscr{X}^{(0)} \in M_{nn}(I(\mathbb{C}))$. Мы имеем в этом случае

$$\{\mathscr{Y}_p = (\mathscr{I}_p - \mathscr{A}_p)^{-1} \mathscr{B}_p \mid \mathscr{A}_p \in \mathscr{A}, \mathscr{B}_p \in \mathscr{B}\} \subseteq \mathscr{X}^*.$$

Следует также сделать некоторые замечания по поводу условия $\rho(|\mathscr{A}|) < 1$, которое в силу теоремы 1 необходимо и достаточно для сходимости полношагового метода. В частном случае, когда $\mathscr{A}$ — точечная матрица и $\mathscr{b}$ — точечный вектор, мы получаем, что последовательные приближения

$$x^{(k+1)} = \mathscr{A}_p x^{(k)} + \mathscr{b}_p, \quad k \geqslant 0, \tag{6}$$

сходятся к решению

$$x_{\rho} = (\mathcal{J}_{\rho} - \mathcal{A}_{\rho})^{-1} \ell_{\rho}$$

для любого интервального вектора $x^{(0)}$ тогда и только тогда, когда $\rho(|\mathcal{A}_{\rho}|) < 1$. С другой стороны, последовательные приближения

$$x_{\rho}^{(k+1)} = \mathcal{A}_{\rho} x_{\rho}^{(k)} + \ell_{\rho}, \quad k \geqslant 0, \tag{7}$$

сходятся для всех $x_{\rho}^{(0)} \in V_n(\mathbb{C})$ к

$$x_{\rho} = (\mathcal{J}_{\rho} - \mathcal{A}_{\rho})^{-1} \ell,$$

если $\rho(\mathcal{A}_{\rho}) < 1$.

Если мы проводим итерацию (6) в $V_n(K(\mathbb{C}))$, то в обозначениях $\mathcal{A}_{\rho} = (a_{rs})$, $a_{rs} = a_{rs}^{(1)} + i a_{rs}^{(2)}$, $1 \leqslant r, \ s \leqslant n$, мы имеем:

$$|\mathcal{A}_{\rho}| = |\mathcal{A}_{\rho}|_2 := \left(\sqrt{(a_{rs}^{(1)})^2 + (a_{rs}^{(2)})^2}\right).$$

Если, с другой стороны, мы проводим итерацию (6) в $V_n(R(\mathbb{C}))$, то имеем

$$|\mathcal{A}_{\rho}| = |\mathcal{A}_{\rho}|_1 := (|a_{rs}^{(1)}| + |a_{rs}^{(2)}|).$$

В силу неравенства

$$\sqrt{a_1^2 + a_2^2} \leqslant |a_1| + |a_2|$$

имеем

$$|\mathcal{A}_{\rho}|_2 \leqslant |\mathcal{A}_{\rho}|_1.$$

Из

$$O_{\rho} \leqslant |\mathcal{A}_{\rho}|_2 \leqslant |\mathcal{A}_{\rho}|_1$$

следует, что

$$\rho(\mathcal{A}_{\rho}) \leqslant \rho(|\mathcal{A}_{\rho}|_2) \leqslant \rho(|\mathcal{A}_{\rho}|_1). \tag{8}$$

Последнее неравенство показывает, что (6) требует в общем случае более сильных предположений, чем (7). Это происходит потому, что необходимое и достаточное условие для (6) гарантирует сходимость для любых интервальных векторов, а множество всех точечных векторов, которые допускаются в (7) в качестве начальных, является собственным подмножеством множества всех интервальных векторов.

Вторая часть неравенства (8) показывает, что при реализации итерации (6) в $V_n(K(\mathbb{C}))$ от $\mathcal{A}_{\rho}$ требуется выполнение не менее сильных условий, чем в случае $V_n(R(\mathbb{C}))$. Это верно и для систем уравнений, коэффициенты которых — невырожденные интервалы. С

этой точки зрения арифметические операции в $K(\mathbb{C})$ имеют некоторые преимущества перед операциями в $R(\mathbb{C})$.

3.3. Методы релаксации

Мы уже рассмотрели полношаговый, короткошаговый и симметрический короткошаговый методы. Существует много других методов решения линейных систем точечных уравнений вида

$$x_p = \mathcal{A}_p x_p + b_p,$$

для которых можно уменьшить асимптотический фактор сходимости путем введения одного или нескольких параметров. Большую часть этих приемов можно перенести на итерационные методы, использующие интервальные векторы. В качестве примера мы рассмотрим метод релаксации для короткошаговою случая.

Как и в короткошаговом методе, разложим матрицу $\mathcal{A}$ в сумму $\mathcal{A} = \mathcal{L} + \mathcal{D} + \mathcal{U}$, где $\mathcal{L}$ — нижняя строго треугольная матрица, $\mathcal{U}$ — верхняя строго треугольная матрица и $\mathcal{D}$ —диагональная матрица. Затем строим последовательные приближения

$$\tilde{X}_i^{(k+1)} = \sum_{j=1}^{i-1} A_{ij} X_j^{(k+1)} + \sum_{j=i}^{n} A_{ij} X_j^{(k)} + B_i,$$

$$X_i^{(k+1)} = (1-\omega) X_i^{(k)} + \omega \tilde{X}_i^{(k+1)}, \quad 1 \leqslant i \leqslant n, \quad k \geqslant 0,$$

начиная с произвольного интервального вектора $x^{(0)}$. С помощью векторных обозначений эти формулы можно записать в виде

$$x^{(k+1)} = (1-\omega) x^{(k)} + \omega \{\mathcal{L} x^{(k+1)} + (\mathcal{D} + \mathcal{U}) x^{(k)} + b\}, \quad k \geqslant 0,$$

где ω>0 — параметр. Так же, как это делалось для полношагового или короткошагового метода, можно показать, что

$$\rho((\mathcal{I}_p - \omega |\mathcal{L}|)^{-1} \{|1-\omega| \mathcal{I}_p + \omega (|\mathcal{D}| + |\mathcal{U}|)\}) < 1$$

является необходимым и достаточным условием сходимости метода к единственной неподвижной точке для произвольного начального вектора.

Можно далее показать, что при $\rho(|\mathcal{A}|) < 1$ это условие выполнено для всех значений ω, которые удовлетворяют неравенству

$$0 < \omega < 2/(1 + \rho(|\mathcal{A}|)).$$

Если полношаговый метод сходится и ω удовлетворяет приведенному неравенству, то метод релаксации тоже сходится к неподвижной точке $\tilde{x}^*$, удовлетворяющей уравнению

$$\tilde{x}^* = (1-\omega)\tilde{x}^* + \omega(\mathcal{A}\tilde{x}^* + \mathcal{b}).$$

Эта неподвижная точка, вообще говоря, отлична от неподвижной точки x^* уравнения $x = \mathcal{A}x + \mathcal{b}$. Чтобы показать это, заметим сначала, что для всех вещественных чисел a, b и невырожденных интервалов Z (т. е. таких Z, что $d(Z) > 0$) при $ab>0$ выполнено соотношение

$$(a+b)Z = aZ + bZ.$$

Пусть $0 < \omega < 1$. Тогда имеем

$$(1-\omega)x^* + \omega(\mathcal{A}x^* + \mathcal{b}) = (1-\omega)x^* + \omega x^* = (1-\omega+\omega)x^* = x^*;$$

т. е. $x^* = \tilde{x}^*$, так как неподвижная точка $\tilde{x}^*$, вычисленная методом релаксации, единственна.

Если же, напротив, мы имеем $\omega > 1$, то

$$\begin{aligned}\mathcal{A}x^* + \mathcal{b} = x^* &= (1-\omega+\omega)x^* \subseteq (1-\omega)x^* + \omega x^* \\ &= (1-\omega)x^* + \omega(\mathcal{A}x^* + \mathcal{b}) \\ &= (1-\omega)x^* + \omega\{\mathcal{L}x^* + (\mathcal{D}+\mathcal{U})x^* + \mathcal{b}\}.\end{aligned}$$

Если мы возьмем начальное приближение $\tilde{x}^{(0)} = x^*$ для метода релаксации, то из этого включения и монотонности включения следует, что

$$x^* \subseteq (1-\omega)\tilde{x}^{(0)} + \omega\{\mathcal{L}\tilde{x}^{(1)} + (\mathcal{D}+\mathcal{U})\tilde{x}^{(0)} + \mathcal{b}\} =: \tilde{x}^{(1)}.$$

С помощью математической индукции можно показать, что

$$x^* \subseteq \tilde{x}^{(k)},\ k \geqslant 0,\ \text{т. е. } x^* \subseteq \tilde{x}^*.$$

Из простых примеров видно, что включение здесь собственное. Поэтому при $\omega > 1$ мы должны учитывать, что применение метода релаксации «увеличивает» неподвижную точку уравнения $x^* = \mathcal{A}x^* + \mathcal{b}$. Это нежелательный эффект, так как задача состоит в нахождении интервального вектора, который содержит множество

$$\{y_p = (\mathcal{I}_p - \mathcal{A}_p)^{-1}\mathcal{b}_p \mid \mathcal{A}_p \in \mathcal{A},\ \mathcal{b}_p \in \mathcal{b}\}$$

и дает достаточно хорошую локализацию.

В частном случае, когда $\mathcal{A}$ — точечная матрица и $\mathcal{b}$ — точечный вектор, мы всегда имеем $x^* = x^*$. Множество точечных векторов — это подмножество множества интервальных векторов. Если мы выберем начальный вектор точечным, то все последовательные приближения и неподвижная точка также будут точечными векторами. Поэтому мы имеем для полношагового метода равенство

$$x_p^* = \mathcal{A}_p x_p^* + \mathcal{b}_p,$$

а для метода релаксации — равенство

$$\hat{x}_p^* = (1-\omega)\hat{x}_p^* + \omega(\mathcal{A}_p\hat{x}_p^* + \mathcal{b}_p),$$

откуда следует, что $\hat{x}_p^* = x_p^*$. Таким образом, для случая точечной матрицы и точечного вектора оба метода сходятся к решению

$$x_p^* = (\mathcal{I}_p - \mathcal{A}_p)^{-1}\mathcal{b}_p$$

уравнения $x_p = \mathcal{A}_p x_p + \mathcal{b}_p$.

Мы хотим теперь исследовать этот случай несколько подробнее. Последовательность $\{d(x^{(k)})\}_{k=0}^{\infty}$, полученная вычислением ширины из последовательности $\{x^{(k)}\}_{k=0}^{\infty}$, сходится к нулевому вектору, так как $\lim_{k\to\infty} x^{(k)} = x_p^* = (\mathcal{I}_p - \mathcal{A}_p)^{-1}\mathcal{b}_p$. Поэтому кажется, что естественно характеризовать скорость сходимости величиной $\tilde{\alpha}$, вводимой следующим определением.

Определение 1. Пусть $x_p^* = \mathcal{f}_p(x_p^*)$, и пусть 6 обозначает множество всех последовательностей $\{x^{(k)}\}_{k=0}^{\infty}$, которые получаются применением итерационного метода

$$x^{(k+1)} = \mathcal{f}_p(x^{(k)}), \quad k \geqslant 0,$$

и для которых $\lim_{k\to\infty} x^{(k)} = x_p^*$. Тогда величина

$$\tilde{\alpha} = \sup\left\{\limsup_{k\to\infty} \left\| d\left(x_p^{(k)}\right)\right\|^{1/k} \mid \{x^{(k)}\}_{k=0}^{\infty} \in \mathfrak{G}\right\}$$

называется асимптотическим фактором сходимости этого итерационного метода в точке x_p^*.

По аналогии с величиной α (определение 10, п. 3.2**)** мы можем сказать, что $\tilde{\alpha}$ характеризует асимптотически наихудший случай при произвольном выборе $x^{(0)}$. Точно так же, как для α, можно показать, что $\tilde{\alpha}$ не зависит от используемой нормы.

Докажем теперь следующее утверждение.

Теорема 2. *Пусть задано уравнение*

$$x_p = \mathcal{A}_p x_p + \mathcal{b}_p$$

для точечной матрицы $\mathcal{A}_p$, *такой что*

$$\rho(|\mathcal{A}_p|) < 1,$$

и точечного вектора $\mathcal{b}_p$. *Тогда асимптотический фактор сходимости* $\tilde{\alpha}_T$ *для полношагового метода* (2, п.3.2) *удовлетворяет равенству*

$$\tilde{\alpha}_T = \rho(|\mathcal{A}_p|),$$

а асимптотический фактор сходимости $\tilde{\alpha}_R$ *для метода релаксации — равенству*

$$\tilde{\alpha}_R = \rho((\mathcal{I}_p - \omega|\mathcal{L}_p|)^{-1}\{|1-\omega|\mathcal{I}_p + \omega(|\mathcal{D}_p| + |\mathcal{U}_p|)\})$$

для

$$0 < \omega < 2/(1+\rho(|\mathcal{A}_p|)).$$

Доказательство. Проведем рассуждение для метода релаксации. Начав с произвольного интервального вектора $x^{(0)}$, мы с помощью (12, п.2.3) и (19, 2.3) получаем из нашей итерационной формулы

$$x^{(k+1)} = (1-\omega)x^{(k)} + \omega\{\mathcal{L}_p x^{(k+1)} + (\mathcal{D}_p + \mathcal{U}_p)x^{(k)} + \mathcal{E}_p\}, \quad k \geqslant 0$$

что

$$d(x^{(k+1)}) = |1-\omega|d(x^{(k)}) + \omega|\mathcal{L}_p|d(x^{(k+1)}) + \omega(|\mathcal{D}_p| + |\mathcal{U}_p|)d(x^{(k)})$$

или

$$d(x^{(k+1)}) = (\mathcal{I}_p - \omega|\mathcal{L}_p|)^{-1}\{|1-\omega|\mathcal{I}_p + \omega(|\mathcal{D}_p| + |\mathcal{U}_p|)\}d(x^{(k)})$$
$$= \{(\mathcal{I}_p - \omega|\mathcal{L}_p|)^{-1}(|1-\omega|\mathcal{I}_p + \omega(|\mathcal{D}_p| + |\mathcal{U}_p|))\}^{k+1} d(x^{(0)}).$$

Отсюда непосредственно получаем, что

$$\tilde{\alpha}_R \leqslant \rho((\mathcal{I}_p - \omega|\mathcal{L}_p|)^{-1}\{|1-\omega|\mathcal{I}_p + \omega(|\mathcal{D}_p| + |\mathcal{U}_p|)\}).$$

Если теперь выбрать конкретный вектор $x^{(0)}$ так, что $d(x^{(0)})$ — собственный вектор неотрицательной матрицы $(\mathcal{I}_p - \omega|\mathcal{L}_p|)^{-1} \times \times\{|1-\omega|\mathcal{I}_p + \omega(|\mathcal{D}_p| + |\mathcal{U}_p|)\}$, соответствующий собственному числу λ, равному спектральному радиусу этой матрицы, то из уравнения для $d(x^{(k+1)})$ следует, что

$$d(x^{(k+1)}) = \lambda^{k+1} d(x^{(0)}),$$

откуда получается нужное утверждение. Доказательство для полношагового метода можно провести аналогично.

Только что доказанная теорема позволяет сформулировать утверждение об асимптотически наискорейшем (в смысле определения 1) методе.

Теорема 3. *В условиях теоремы 2*

$$\min\{\tilde{\alpha}_R | 0 < \omega < 2/(1+\rho(|\mathcal{A}_p|))\} = \tilde{\alpha}_R|_{\omega=1}$$
$$= \rho((\mathcal{I}_p - |\mathcal{L}_p|)^{-1}(|\mathcal{D}_p| + |\mathcal{U}_p|)) = \tilde{\alpha}_S,$$

а также

$$\tilde{\alpha}_S \leqslant \tilde{\alpha}_T.$$

Доказательство. Рассмотрим вещественную точечную матрицу

$$\mathcal{P}_p = ((1 - |1-\omega|)/\omega)\mathcal{I}_p - |\mathcal{A}_p|$$

и ее разложение $\mathscr{P}_p = \mathscr{M}_{p\omega} - \mathscr{N}_{p\omega}$, где

$$\mathscr{M}_{p\omega} = (1/\omega)(\mathscr{I}_p - \omega|\mathscr{L}_p|), \quad \mathscr{N}_{p\omega} = (1/\omega)\{|1 - \omega|\mathscr{I}_p + \omega(|\mathscr{D}_p| + |\mathscr{U}_p|)\}.$$

Это разложение регулярно, так как $\mathscr{M}_{p\omega}^{-1}$ существует и $\mathscr{M}_{p\omega}^{-1} \geqslant \mathscr{O}_p$, $\mathscr{N}_{p\omega} \geqslant \mathscr{O}_p$. Если ω удовлетворяет неравенству

$0 < \omega < 2/(1 + \rho|\mathscr{A}_p|)$, то $\mathscr{P}_p^{-1} \geqslant \mathscr{O}_p$,

и в силу неравенства $\mathscr{N}_{p\omega} \geqslant \mathscr{N}_{p1}$ мы получаем , что

$$\rho(\mathscr{M}_{p1}^{-1}\mathscr{N}_{p1}) = \rho((\mathscr{I}_p - |\mathscr{L}_p|)^{-1}(|\mathscr{D}_p| + |\mathscr{U}_p|)) \leqslant \rho(\mathscr{M}_{p\omega}^{-1}\mathscr{N}_{p\omega}) < 1.$$

Этим доказана первая часть теоремы. Ее вторая часть следует из теоремы Штейна и Розенберга и ее обобщения, утверждающего, что из $\rho(|\mathscr{A}_p|) < 1$ следует

$$\rho((\mathscr{I}_p - |\mathscr{L}_p|)^{-1}(|\mathscr{D}_p| + |\mathscr{U}_p|)) \leqslant \rho(|\mathscr{A}_p|).$$

В предыдущей теореме утверждается, что в случае системы точечных уравнений невозможно асимптотически (в смысле определения 1) ускорить сходимость короткошагового метода, используя метод релаксации. Кроме того, короткошаговый метод сходится не медленнее, чем полношаговый.

Теперь рассмотрим связь между описанным выше итерационным методом в пространстве интервальных векторов и принципом локализации решений, который получен иным способом.

Рассмотрим систему линейных уравнений

$$x_p = \mathscr{A}_p x_p + b_p,$$

где $\mathscr{A}_p$ — вещественная точечная матрица и b_p — вещественный точечный вектор. Мы вводим естественный (т. е. покомпонентный) порядок на точечных матрицах и точечных векторах. Точечная матрица $\mathscr{P}_p$ называется изотонной (соответственно антитонной), если из $x_p \geqslant o_p$ следует $\mathscr{P}_p x_p \geqslant o_p$ (соответственно o_p). Теперь матрица $\mathscr{A}_p$ раскладывается в сумму изотонной и антитонной точечных матриц:

$$\mathscr{A}_p = \mathscr{T}_{p1} + \mathscr{T}_{p2}.$$

Начиная с пары точечных векторов $v_p^{(0)}$ и $w_p^{(0)}$, для которых верно $v_p^{(0)} \leqslant w_p^{(0)}$, наш итерационный метод вычисляет две последовательности $\{v_p^{(k)}\}_{k=0}^{\infty}$ и $\{w_p^{(k)}\}_{k=0}^{\infty}$ по формулам

$$\begin{cases} v_p^{(k+1)} = \mathscr{T}_{p1} v_p^{(k)} + \mathscr{T}_{p2} w_p^{(k)} + b_p, \\ w_p^{(k+1)} = \mathscr{T}_{p1} w_p^{(k)} + \mathscr{T}_{p2} v_p^{(k)} + b_p, \quad k \geqslant 0. \end{cases} \tag{1}$$

Если теперь $v_p^{(0)} \leqslant v_p^{(1)} \leqslant w_p^{(1)} \leqslant w_p^{(0)}$, то с помощью математической индукции можно показать, что

$$v_p^{(0)} \leqslant v_p^{(1)} \leqslant \ldots \leqslant v_p^{(k)} \leqslant v_p^{(k+1)} \leqslant w_p^{(k+1)} \leqslant w_p^{(k)} \leqslant \ldots \leqslant w_p^{(1)} \leqslant w_p^{(0)}.$$

Поэтому последовательности $\{v_p^{(k)}\}_{k=0}^{\infty}$ и $\{w_p^{(k)}\}_{k=0}^{\infty}$ сходятся, и простые рассуждения показывают, что решение x_p^* уравнения $x_p = \mathcal{A}_p x_p + b_p$ существует и находится между граничными точками. Если $\rho(|\mathcal{A}_p|) = \rho(\mathcal{T}_{p1} - \mathcal{T}_{p2}) < 1$, то $v_p^* = w_p^* = x_p^*$.

Рассматривая вместе с методом итераций для

$$\{v_p^{(k)}\}_{k=0}^{\infty}$$

и $\{w_p^{(k)}\}_{k=0}^{\infty}$ также итерационный метод

$$x^{(k+1)} = \mathcal{A}_p x^{(k)} + b_p, \quad k \geqslant 0, \tag{2}$$

с интервалом $x^{(0)} = ([v_i^{(0)}, w_i^{(0)}])$ в качестве начального приближения и учитывая правила умножения интервальных матриц на интервальные векторы, мы сразу видим, что границы компонент последовательности интервальных векторов $\{x^{(k)}\}_{k=0}^{\infty}$ совпадают с компонентами последовательностей $\{v_p^{(k)}\}_{k=0}^{\infty}$ и $\{w_p^{(k)}\}_{k=0}^{\infty}$.

Сказанное только что об итерационном методе (2) показывает, что в методе (1) можно избавиться от предположения $v_p^{(0)} \leqslant v_p^{(1)} \leqslant w_p^{(1)} \leqslant w_p^{(0)}$, если выполнено условие $v_p^{(0)} \leqslant x_p^* \leqslant w_p^{(0)}$, которое гарантирует локализацию решения x_p^*. Последовательности $\{v_p^{(k)}\}_{k=0}^{\infty}$ и $\{w_p^{(k)}\}_{k=0}^{\infty}$ в этом случае уже не обязательно сходятся монотонно. Монотонность можно восстановить, если брать пересечение после каждого шага.

Замечания. Утверждения, аналогичные тем, которые мы доказали для метода релаксации, справедливы и для симметрического метода релаксации (SR)

$$x^{(k+1/2)} = (1 - \omega) x^{(k)} + \omega \{\mathcal{L} x^{(k+1/2)} + \mathcal{U} x^{(k)} + b\},$$

$$x^{(k+1)} = (1 - \omega) x^{(k+1/2)} + \omega \{\mathcal{L} x^{(k+1/2)} + \mathcal{U} x^{(k+1)} + b\}, \quad k \geqslant 0,$$

который при $\omega = 1$ сводится к симметрическому короткошаговому методу (SS), описанному в п.3.2. Необходимым и достаточным условием сходимости этого метода к единственной неподвижной точке при произвольном начальном интервальном векторе является

$$\rho((\mathcal{I}_p - \omega|\mathcal{U}|)^{-1}(|1-\omega|\mathcal{I}_p + \omega|\mathcal{L}|)(\mathcal{I}_p - \omega|\mathcal{L}|)^{-1} \times (|1-\omega|\mathcal{I}_p + \omega|\mathcal{U}|)) < 1.$$

Если $\rho(|\mathcal{A}|) < 1$, то предыдущее условие выполняется при

$$0 < \omega < 2/(1 + \rho(|\mathcal{A}|)).$$

Доказательство можно провести так же, как для метода релаксации.

Если $\mathcal{A}$ — точечная матрица, то по аналогии с рассуждением из теоремы 2 можно показать, что асимптотический фактор сходимости, введенный в определении 1, равен

$$\tilde{\alpha}_{SR} = \rho((\mathcal{I}_p - \omega|\mathcal{U}_p|)^{-1}(|1-\omega|\mathcal{I}_p + \omega|\mathcal{L}_p|) \times (\mathcal{I}_p - \omega|\mathcal{L}_p|)^{-1}(|1-\omega|\mathcal{I}_p + \omega|\mathcal{U}_p|)),$$

и по аналогии с теоремой 3 имеем

$$\tilde{\alpha}_{SS} \leqslant \tilde{\alpha}_{SR}$$

для $0 < \omega < 2/(1 + \rho(|\mathcal{A}_p|))$.

Покажем, что имеет место $\tilde{\alpha}_{SS} \leqslant \tilde{\alpha}_S$. Матрица

$$(\mathcal{I}_p - |\mathcal{L}_p|)^{-1}|\mathcal{U}_p|$$

неотрицательна и всегда приводима, так как ее первый столбец состоит из нулей. Если добавить положительную матрицу $\Delta\mathcal{U}_p$ (вообще говоря, не являющуюся верхней треугольной), то можно сделать матрицу $(\mathcal{I}_p - |\mathcal{L}_p|)^{-1}(|\mathcal{U}_p| + \Delta\mathcal{U}_p)$ неприводимой. Используя теорему Перрона и Фробениуса, мы получаем

$$(\mathcal{I}_p - |\mathcal{L}_p|)^{-1}(|\mathcal{U}_p| + \Delta\mathcal{U}_p)x_p = \lambda x_p,$$

где λ — спектральный радиус матрицы из левой части и вектор x_p имеет только положительные компоненты. Простое преобразование (корректное при $0 < \lambda < 1$ и достаточно малой $\Delta\mathcal{U}_p$) дает

$$x_p = (\mathcal{I}_p - (1/\lambda)(|\mathcal{U}_p| + \Delta\mathcal{U}_p))^{-1}|\mathcal{L}_p|x_p \geqslant (\mathcal{I}_p - (|\mathcal{U}_p| + \Delta\mathcal{U}_p))^{-1}|\mathcal{L}_p|x_p.$$

Окончательно получаем

$$(\mathcal{I}_p - (|\mathcal{U}_p| + \Delta\mathcal{U}_p))^{-1}|\mathcal{L}_p|(\mathcal{I}_p - |\mathcal{L}_p|)^{-1}(|\mathcal{U}_p| + \Delta\mathcal{U}_p)x_p \leqslant \lambda x_p.$$

Это неравенство верно и при $X = 0$. Применяя теперь известную теорему, получаем, что спектральный радиус матрицы из левой части не превосходит λ. Так как это верно для всех матриц $\Delta\mathcal{U}_p$, делающих матрицу $(\mathcal{I}_p - |\mathcal{L}_p|)^{-1}(|\mathcal{U}_p| + \Delta\mathcal{U}_p)$ неприводимой, то мы получаем нужное утверждение, устремляя $\Delta\mathcal{U}_p$ к нулю, так как собственные числа непрерывно зависят от элементов матрицы.

3.4.Оптимальность симметрического короткошагового метода со взятием пересечения на каждом шаге

В этом микромодуле предполагается, что все интервальные матрицы взяты из пространства $M_{nn}(R(\mathbb{C}))$, а все интервальные векторы — из $V_n(R(\mathbb{C}))$.

Мы собираемся теперь исследовать некоторые модификации полношагового, короткошагового и симметрического корочкошагового методов. Если полношаговый метод

$$x^{(k+1)} = \mathcal{A}x^{(k)} + b$$

имеет начальный вектор $x^{(0)}$ для которого $x^* \subseteq x^{(0)}$, то из монотонности включениясследует, что

$$x^* = \mathcal{A}x^* + b \subseteq \mathcal{A}x^{(0)} + b = x^{(1)}.$$

Это показывает, что и $x^{(1)}$ содержит неподвижную точку и вектор $x^{(0)}$, а значит, и их пересечение $x^{(0)} \cap x^{(1)}$. Поэтому естественно продолжать итерацию, используя это новое включение. Это приводит к итерационной процедуре

$$x^{(k+1)} = \{\mathcal{A}x^{(k)} + b\} \cap x^{(k)}, \quad k \geqslant 0,$$

которую мы будем называть полношаговым методом со взятием пересечения на каждом шаге (TI).

Если провести те же рассмотрения для короткошагового метода, то получится итерационная процедура

$$x^{(k+1)} = \{\mathcal{L}x^{(k+1)} + (\mathcal{D} + \mathcal{U})x^{(k)} + b\} \cap x^{(k)}, \quad k \geqslant 0,$$

которую мы назовем короткошаговым методом со взятием пересечения на каждом шаге (SI).

В случае короткошагового метода имеется еще одна возможность:

$$X_i^{(k+1)} = \left\{\sum_{j=1}^{i-1} A_{ij}X_j^{(k+1)} + \sum_{j=i}^{n} A_{ij}X_j^{(k)} + B_i\right\} \cap X_i^{(k)}, \quad 1 \leqslant i \leqslant n, \quad k \geqslant 0.$$

После того как вычислена первая компонента, образуется пересечение со старым приближением, дающее новое приближение. Это новое приближение используется для вычисления нового приближения для второй компоненты и т. д. Эта модификация называется короткошаговым методом со взятием пересечения после каждой компоненты (SIC).

Наконец, для случая, когда все диагональные элементы матрицы $\mathcal{A}$ обращаются в нуль, рассмотрим (снова в предположении $x^* \subseteq x^{(0)}$) итерационную процедуру

$$X_i^{(k+1/2)} = \left\{ \sum_{j=1}^{i-1} A_{ij} X_j^{(k+1/2)} + \sum_{j=i+1}^{n} A_{ij} X_j^{(k)} + B_i \right\} \cap X_i^{(k)}, \quad 1 \leqslant i \leqslant n,$$

$$X_i^{(k+1)} = \left\{ \sum_{j=1}^{i-1} A_{ij} X_j^{(k+1/2)} + \sum_{j=i+1}^{n} A_{ij} X_j^{(k+1)} + B_i \right\} \cap X_i^{(k+1/2)},$$

$$1 \leqslant i \leqslant n, \quad k \geqslant 0.$$

Мы будем называть эту процедуру симметрическим коротко-шаговым методом со взятием пересечения после каждой компоненты (SSIC).

Метод (SSIC) можно выполнять таким образом, что он будет требовать на каждом шаге (за исключением самого первого) того же количества интервальных умножений, что и метод (SIC). В обоих случаях требуется $n^2 - n$ умножений (в предположении, что диагональные элементы обращаются в 0). При этом используются следующие соображения. Допустим, что для некоторого $k>0$ известны суммы

$$\sum_{j=i+1}^{n} A_{ij} X_j^{(k)}, \quad 1 \leqslant i \leqslant n-1.$$

Вычисление векторов $x^{(k+1/2)}$ по методу (SSIC) требует $\frac{1}{2}(n^2 - n)$ умножений. Если запоминается n — 1 сумма

$$\sum_{j=1}^{i-1} A_{ij} X_j^{(k+1/2)}, \quad 2 \leqslant i \leqslant n,$$

то вычисление $x^{(k+1)}$ из $x^{(k+1/2)}$ требует еще $\frac{1}{2}(n^2 - n)$ умножений.

Таким образом, всего для вычисления приближения $x^{(k+1)}$, исходя из $x^{(k)}$ нам требуется $n^2 - n$ умножений, как и в методе (SIC). Если при вычислении $x^{(k+1)}$ из $x^{(k)}$ запоминаются суммы

$$\sum_{j=i+1}^{n} A_{ij} X_j^{(k+1)},$$

то все эти рассуждения проходят для индекса, увеличенного на 1.

Следующее утверждение показывает, что итерационные методы (TI), (SI), (SIC) и (SSIC) сходятся к неподвижной точке x^*.

Теорема 1. *Пусть* $\mathcal{A}$ *—интервальная матрица и* $\rho(|\mathcal{A}|) < 1$. *Если* x^* *— неподвижная точка уравнения* $x = \mathcal{A}x + b$ *и* $x^{(0)} \supseteq x^*$, *то итерационные методы* (TI), (SI), (SIC) *и* (SSIC) *сходятся к* x^*.

Доказательство. Докажем теорему для метода (TI). Так как мы берем пересечения, полученная последовательность приближений $\{x^{(k)}\}_{k=0}^{\infty}$ удовлетворяет условию

$$x^{(0)} \supseteq x^{(1)} \supseteq \ldots \supseteq x^{(k)} \supseteq x^{(k+1)} \supseteq \ldots .$$

В силу следствия 8 из п.2.3 эта последовательность сходится к некоторому пределу $\tilde{x}^*$, и этот предел удовлетворяет условию $x^* \subseteq \tilde{x}^*$. Операция взятия пересечения также непрерывна (если пересечение непусто), поэтому при $k \to \infty$ мы имеем

$$\tilde{x}^* = \{\mathcal{A}\tilde{x}^* + b\} \cap \tilde{x}^*,$$

откуда следует, что

$$\mathcal{A}\tilde{x}^* + b \supseteq \tilde{x}^*.$$

Мы рассматриваем полношаговый метод

$$y^{(k+1)} = \mathcal{A}y^{(k)} + b, \quad k \geqslant 0,$$

где $y^{(0)} = \tilde{x}^*$. Из сказанного следует, что

$$y^{(1)} = \mathcal{A}y^{(0)} + b = \mathcal{A}\tilde{x}^* + b \supseteq \tilde{x}^* \supseteq x^*,$$
$$y^{(2)} = \mathcal{A}y^{(1)} + b \supseteq \mathcal{A}\tilde{x}^* + b \supseteq \tilde{x}^* \supseteq x^*$$

и вообще

$$y^{(k+1)} = \mathcal{A}y^{(k)} + b \supseteq \mathcal{A}\tilde{x}^* + b \supseteq \tilde{x}^* \supseteq x^*.$$

Последовательность $\{y^{(k)}\}_{k=0}^{\infty}$, вычисленная с помощью рассматриваемого итерационного метода, сходится к x^* в силу теоремы 1 из п.3.2. Последнее из доказанных включений дает при $k \to \infty$ соотношение

$$x^* \supseteq \tilde{x}^* \supseteq x^*,$$

т. е. $\tilde{x}^* = x^*$. Для остальных методов доказательство аналогично.

Сравним методы (T), (S), (TI), (SI), (SIC) и (SSIC) по скорости сходимости для случая, когда итерации начинаются с интервального вектора, содержащего неподвижную точку x^*.

В первой теореме метод (T) сравнивается с (TI), а мегод (S) - с (SI).

Теорема 2. *Пусть последовательности* $\{x^{(k)}\}_{k=0}^{\infty}$ *и* $\{\tilde{x}^{(k)}\}_{k=0}^{\infty}$ *вычислены согласно итерационным методам* (T) *и* (TI) *в предположении, что* $x^{(0)} \supseteq \tilde{x}^{(0)} \supseteq x^*$. *Тогда имеет место*

$$x^{(k)} \supseteq \tilde{x}^{(k)} \supseteq x^* \quad \textit{для всех} \quad k \geqslant 0.$$

Такое же утверждение справедливо для последовательностей, вычисленных согласно итерационным методам (S) *и* (SI).

Доказательство. Докажем теорему для методов (T) и (TI). Соотношение $x^{(k)} \supseteq x^*$, $k \geqslant 0$, уже было доказано в предположении $\tilde{x}^{(0)} \supseteq x^*$ при получении формул для итерационного метода (TI). Допустим, что для некоторого $k \geqslant 0$ уже доказано, что

$$x^{(k)} \supseteq \tilde{x}^{(k)}.$$

Для $k = 0$ это верно ввиду нашего исходного допущения. Используя монотонность включения, получаем

$$x^{(k+1)} = \mathcal{A}x^{(k)} + \mathcal{b} \supseteq \mathcal{A}\tilde{x}^{(k)} + \mathcal{b} \supseteq \{\mathcal{A}\tilde{x}^{(k)} + \mathcal{b}\} \cap \tilde{x}^{(k)} = \tilde{x}^{(k+1)},$$

что завершает доказательство по индукции.

Доказательство для методов (S) и (SI) можно провести аналогичным образом.

Теорема 3. *Пусть последовательности* $\{x^{(k)}\}_{k=0}^{\infty}$ *и* $\{\tilde{x}^{(k)}\}_{k=0}^{\infty}$ *вычислены согласно итерационным методам* (TI) *и* (SIC) *в предположении, что* $x^{(0)} \supseteq \tilde{x}^{(0)} \supseteq x^*$. *Тогда имеет место*

$$x^{(k)} \supseteq \tilde{x}^{(k)} \supseteq x^* \quad \textit{для всех} \quad k \geqslant 0.$$

Такое же утверждение справедливо для последовательностей, вычисленных согласно итерационным методам (S) *и* (SIC).

Доказательство. Нам нужно доказать лишь соотношение $x^{(k)} \supseteq \tilde{x}^{(k)}$. Мы проведем доказательство для последовательностей, вычисленных согласно итерационным методам (TI) и (SIC). Допустим, что для некоторого $k \geqslant 0$ верно

$$x^{(k)} \supseteq \tilde{x}^{(k)}.$$

Для $k = 0$ это верно в силу нашего исходного допущения.

Тогда в обозначениях

$$x^{(k)} = (X_i^{(k)}), \quad \tilde{x}^{(k)} = (\tilde{X}_i^{(k)}), \quad \mathcal{A} = (A_{ij}), \quad \mathcal{b} = (B_i)$$

мы имеем

$$X_i^{(k+1)} = \left(\sum_{j=1}^{n} A_{ij} X_j^{(k)} + B_i\right) \cap X_i^{(k)},$$

$$\tilde{X}_i^{(k+1)} = \left(\sum_{j=1}^{n} A_{ij} \tilde{X}_j^{(k)} + B_i\right) \cap \tilde{X}_i^{(k)}.$$

Из $\tilde{X}_i^{(k)} \subseteq X_i^{(k)}$, $1 \leqslant i \leqslant n$, и монотонности включения следует, что

$$\sum_{j=1}^{n} A_{1j}\tilde{X}_j^{(k)} + B_1 \subseteq \sum_{j=1}^{n} A_{1j}X_j^{(k)} + B_1,$$

а потому

$$\tilde{X}_1^{(k+1)} \subseteq X_1^{(k+1)}.$$

Ввиду $\tilde{X}_1^{(k+1)} \subseteq \tilde{X}_1^{(k)} \subseteq X_1^{(k)}$ отсюда следует, что

$$A_{21}X_1^{(k+1)} + \sum_{j=2}^{n} A_{2j}\tilde{X}_j^{(k)} + B_2 \subseteq \sum_{j=1}^{n} A_{2j}X_j^{(k)} + B_2,$$

а ввиду

$$X_2^{(k+1)} = \left(\sum_{j=1}^{n} A_{2j}X_j^{(k)} + B_2\right) \cap X_2^{(k)},$$

$$\tilde{X}_2^{(k+1)} = \left(A_{21}\tilde{X}_1^{(k+1)} + \sum_{j=2}^{n} A_{2j}\tilde{X}_j^{(k)} + B_2\right) \cap \tilde{X}_2^{(k)}$$

получаем

$$\tilde{X}_2^{(k+1)} \subseteq X_2^{(k+1)}.$$

Таким же образом мы показываем, что $\tilde{X}_i^{(k+1)} \subseteq X_i^{(k+1)}$, $3 \leqslant i \leqslant n$, т. е. $\tilde{x}^{(k+1)} \subseteq x^{(k+1)}$, что завершает доказательство по индукции. Доказательство для последовательностей, вычисленных согласно итерационным методам (SI) и (SIC), можно провести аналогичным образом.

Теорема 4. *Пусть последовательности* $\{z^{(k)}\}_{k=0}^{\infty}$ *и* $\{x^{(k)}\}_{k=0}^{\infty}$ *вычислены согласно итерационным методам* (SIC) *и* (SSIC) *в предположении* $z^{(0)} \supseteq x^{(0)} \supseteq x^*$. *Тогда имеет место*

$$z^{(k)} \supseteq x^{(k)} \supseteq x^* \quad \textit{для всех} \quad k \geqslant 0.$$

Доказательство. Допустим, что для некоторого $k \geq 0$ верно $z^{(k)} \supseteq x^{(k)} \supseteq x^*$. Для $k = 0$ это верно в силу нашего исходного допущения. Из формул для (SSIC) и первой формулы для (SIC) получаем с помощью математической индукции по индексам компонент, что

$$z^{(k+1)} \supseteq x^{(k+1/2)} \supseteq x^*.$$

С помощью второй формулы для (SSIC) получаем, еще раз применяя индукцию по индексам компонент, что

$$x^{(k+1/2)} \supseteq x^{(k+1)} \supseteq x^*.$$

Сочетание этих включений дает

$$z^{(k+1)} \supseteq x^{(k+1)} \supseteq x^*.$$

Используя доказанные выше утверждения, мы можем теперь указать оптимальный метод. Пусть (M) обозначает любой метод из множества

$$\{(\mathrm{T}), (\mathrm{S}), (\mathrm{TI}), (\mathrm{SI}), (\mathrm{SIC}), (\mathrm{SSIC})\}.$$

Мы допускаем в качестве начального вектора любой интервальный вектор $x^{(0)}$, такой что $x^* \subseteq x^{(0)}$, где x^* — неподвижная точка, т. е. решение уравнения $x = \mathcal{A}x + b$. Введем частичный порядок на рассматриваемом множестве итерационных методов, полагая $(M) \leqslant (\tilde{M})$, если $x^{(k)} \subseteq \tilde{x}^{(k)}$ для всех $k \geqslant 0$. Здесь $\{x^{(k)}\}_{k=0}^{\infty}$ и $\{\tilde{x}^{(k)}\}_{k=0}^{\infty}$ обозначают последовательности, вычисленные согласно методам (M) и ($\tilde{M}$). Из теоремы 2 имеем

$$(\mathrm{TI}) \leqslant (\mathrm{T}) \quad \text{и} \quad (\mathrm{SI}) \leqslant (\mathrm{S}).$$

Аналогично из теоремы 3 имеем

$$(\mathrm{SIC}) \leqslant (\mathrm{TI}) \quad \text{и} \quad (\mathrm{SIC}) \leqslant (\mathrm{SI})$$

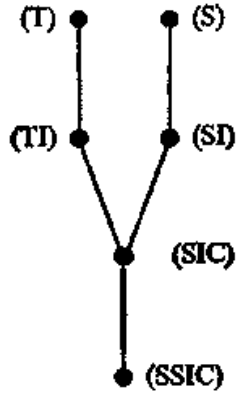
Наконец, из теоремы 4 имеем, что

$$(\mathrm{SSIC}) \leqslant (\mathrm{M})$$

для любого из рассматриваемых итерационных методов (M). Объединяем эти результаты в следующее утверждение.

Теорема 5. *Пусть* $\mathcal{A}$ *— интервальная матрица, такая что* $\rho(|\mathcal{A}|) < 1$, *и* b *— интервальный вектор. Если начать вычисление по одному из методов* (1), (S), (TI), (SI), (SIC) *и* (SSIC) *с вектора* $x^{(0)}$, *удовлетворяющего соотношению* $x^{(0)} \supseteq x^* = \mathcal{A}x^* + b$, *то метод* (SSIC) *всегда даст наименьшую (в смысле включения) локализацию последовательности для* x^*.

Следующая диаграмма наглядно выражает содержание теоремы 5:



Чтобы проиллюстрировать теорему 5, были просчитаны различные примеры на ЭВМ (где мантисса содержит 8 десятичных цифр).

Для каждого примера приводим начальный вектор $x^{(0)} = \left(X_i^{(0)}\right)$ и число итераций k^*, после которого процедура стабилизируется. Примеры показывают, что метод (SIC) требует примерно на 25% больше шагов итерации, чем метод (SSIC).

В первых двух примерах и исходные данные, и неподвижные точки — невырожденные интервалы. В этом случае приводим также и вектор $x^{(k^*)}$. В остальных примерах приводим только наибольшую ширину компоненты вектора $x^{(k)}$, т. е. величину $d^{(k)} = \max\limits_{1 \leqslant i \leqslant n} \left\{d\left(X_i^{(k)}\right)\right\}$. Все примеры были приведены к виду $x = \mathcal{A}x + b$ таким образом, что диагональные элементы матрицы $\mathcal{A}$ равны нулю.

I. Пример.

$$x^{(0)} = \begin{pmatrix} [0.9, & 1.2] \\ [0.4, & 0.7] \\ [0, & 0.2] \\ [-0.4, & -0.1] \end{pmatrix},$$

$$(\mathrm{SIC}): k^* = 37, \quad (\mathrm{SSIC}): k^* = 31,$$

$$x^{(k^*)} = \begin{pmatrix} [1.0328601, & 1.0597579] \\ [0.55075440, & 0.57481398] \\ [0.099483623, & 0.12251379] \\ [-0.24354582, & -0.21269841] \end{pmatrix}.$$

II. Пример.

$$x^{(0)} = \begin{pmatrix} [0.8, & 1.0] \\ [0.65, & 0.85] \\ [0.55, & 0.7] \end{pmatrix}, \qquad x^{(k^*)} = \begin{pmatrix} [0.89636817, & 0.89647991] \\ [0.76595755, & 0.76520225] \\ [0.61424734, & 0.61452184] \end{pmatrix},$$

$$(\mathrm{SIC}): k^* = 18, \quad (\mathrm{SSIC}): k^* = 15.$$

III. Пример.

$$x^{(0)} = \left(X_i^{(0)}\right), \quad \text{где} \quad X_i^{(0)} = [0, 1], \quad i = 1(1)8.$$

$$(\mathrm{SIC}): k^* = 27, \quad (\mathrm{SSIC}): k^* = 22.$$

$d^{(k)}$	k				
	0	5	10	15	20
(SIC)	1	4.0×10^{-2}	8.4×10^{-4}	1.8×10^{-5}	3.8×10^{-7}
(SSIC)	1	7.0×10^{-3}	5.4×10^{-5}	4.1×10^{-7}	3.7×10^{-9}

IV. Пример.

$$x^{(0)} = (X_i^{(0)}), \quad \text{где} \quad X_i^{(0)} = [0,\ 0.5], \quad 1 \leqslant i \leqslant 4.$$

$$(\mathrm{SIC}): k^* = 56, \quad (\mathrm{SSIC}): k^* = 44.$$

$d^{(k)}$	k					
	0	5	10	20	30	40
(SIC)	5.0×10^{-1}	1.1×10^{-1}	2.9×10^{-2}	6.7×10^{-4}	2.7×10^{-5}	1.5×10^{-6}
(SSIC)	5.0×10^{-1}	3.7×10^{-2}	4.7×10^{-3}	7.4×10^{-5}	1.2×10^{-6}	1.8×10^{-8}

V. Пример.

$$x^{(0)} = (X_i^{(0)}), \quad \text{где} \quad X_i^{(0)} = [-0.036016,\ 0.674056], \quad 1 \leqslant i \leqslant 3.$$

$$(\mathrm{SIC}): k^* = 52, \quad (\mathrm{SSIC}): k^* = 42.$$

$d^{(k)}$	k				
	0	5	10	20	40
(SIC)	7.1×10^{-1}	1.2×10^{-1}	2.0×10^{-2}	5.3×10^{-4}	3.7×10^{-7}
(SSIC)	7.1×10^{-1}	5.5×10^{-2}	5.7×10^{-3}	6.2×10^{-5}	3.7×10^{-9}

VI. Пример. Возьмем в примере V

$$X_i^{(0)} = [0.059459,\ 0.643243], \quad 1 \leqslant i \leqslant 3.$$

$$(\mathrm{SIC}): k^* = 51, \quad (\mathrm{SSIC}): k^* = 41$$

$d^{(k)}$	k				
	0	5	10	20	40
(SIC)	5.8×10^{-1}	1.0×10^{-1}	1.7×10^{-2}	4.3×10^{-4}	3.0×10^{-7}
(SSIC)	5.8×10^{-1}	4.5×10^{-2}	4.7×10^{-3}	5.1×10^{-5}	3.7×10^{-9}

Использование локализующих множеств важно при реализации итерационных вычислений на ЭВМ. Если в этом случае мы начинаем вычисления с вектора, представимого в машине и содержащего неподвижную точку, т. е. решение уравнения $x^* = \mathcal{A}x^* + \mathcal{C}$, то все следующие приближения снова содержат эту неподвижную точку. Так как вычисление новых приближений искажается погрешностями округления, мы можем в действительности в какой-то момент «потерять» неподвижную точку: некоторый вновь вычисленный интервал уже не будет содержать ее. Если все операции выполняются в машинной интервальной арифметике, то свойство содержать

неподвижную точку не будет потеряно. Если мы применяем метод, где берутся пересечения, то получается последовательность

$$\tilde{x}^{(0)} \supseteq \tilde{x}^{(1)} \supseteq \ldots \tilde{x}^{(k-1)} \supseteq \tilde{x}^{(k)} = \tilde{x}^{(k+1)} = \ldots.$$

Последовательность приближений, вычисленных на машине, стабилизируется, начиная с некоторого номера k^*. Это следует из того факта, что на цифровой машине представимо лишь конечное количество вещественных чисел.

Покажем, что для методов (TI), (SI), (SIC) и (SSIC) после конечного числа шагов не нужно брать пересечений. Сформулируем и докажем такую теорему для метода итераций (TI).

Теорема 6. *Пусть $\mathcal{A}$ — интервальная матрица, для которой $\rho(|\mathcal{A}|) < 1$. Пусть вектор $x^{(0)} = (X_i^{(0)}) = (\lfloor i(X_i^{(0)}),\ s(X_i^{(0)}) \rfloor)$, выбран таким образом, что для неподвижной точки $x^* = (\lfloor i(X_i^*),\ s(X_i^*) \rfloor)$, т. е. решения уравнения $x = \mathcal{A}x + \mathcal{b}$ выполнено включение $x^{(0)} \supset x^*$, которое вводится соотношениями $i(X_i^{(0)}) < i(X_i^*) \leqslant s(X_i^*) < s(X_i^{(0)}),\ i = 1, 2, \ldots, n.$ Тогда существует $\tilde{k} \geqslant 0$, такое что при всех $k \geqslant \tilde{k}$ для итерационного метода*

$$x^{(k+1)} = (\mathcal{A}x^{(k)} + \mathcal{b}) \cap x^{(k)}, \qquad k \geqslant 0,$$

выполнено равенство

$$x^{(k+1)} = (\mathcal{A}x^{(k)} + \mathcal{b}) \cap x^{(k)} = \mathcal{A}x^{(k)} + \mathcal{b},$$

т. е. верно включение

$$\mathcal{A}x^{(k)} + \mathcal{b} \subseteq x^{(k)}.$$

Доказательство. Мы ограничимся случаем, когда все элементы матрицы $\mathcal{A}$ и векторов $\mathcal{b}$, $x^{(0)}$ принадлежат $I(\mathbb{R})$. Случай, когда разрешены элементы из $R(\mathbb{C})$, может быть рассмотрен аналогичным образом. Сначала мы покажем, что из включения $x^{(0)} \supset x^*$ следует, что не может выполняться соотношение

$$x^{(0)} \subseteq \mathcal{A}x^{(0)} + \mathcal{b}.$$

Действительно, если бы оно выполнялось, то из формул

$$z^{(k+1)} = \mathcal{A}z^{(k)} + \mathcal{b}, \qquad k \geqslant 0,$$

определяющих наш метод итераций, следовало бы при $z^{(0)}=x^{(0)}$, что

$$z^{(1)} = \mathcal{A}z^{(0)} + \mathcal{b} = \mathcal{A}x^{(0)} + \mathcal{b} \supseteq x^{(0)} \supset x^*,$$
$$z^{(2)} = \mathcal{A}z^{(1)} + \mathcal{b} \supseteq \mathcal{A}x^{(0)} + \mathcal{b} \supseteq x^{(0)} \supset x^*$$

и вообще

$$z^{(k+1)} = \mathcal{A}z^{(k)} + \mathcal{b} \supseteq \mathcal{A}x^{(0)} + \mathcal{b} \supseteq x^{(0)} \supset x^*, \quad k \geqslant 0.$$

Так как $\rho(|\mathcal{A}|) < 1$, мы имели бы тогда

$$\lim_{k\to\infty} z^{(k)} = z^*,$$

где $z^* = \mathcal{A}z^* + \mathcal{b}$.

Из последнего соотношения следует, что $z^* \supseteq x^{(0)} \supset x^*$. Это противоречит единственности неподвижной точки, т. е. решения уравнения $x = \mathcal{A}x + \mathcal{b}$. Теперь полагаем

$$x^{(k)} = (X_i^{(k)}), \quad \mathcal{A} = (A_{ij}), \quad \mathcal{b} = (B_i), \quad y^{(k)} = (Y_i^{(k)}),$$

где

$$Y_i^{(k+1)} = \sum_{j=1}^{n} A_{ij}X_j^{(k)} + B_i, \quad k \geqslant 0, \quad 1 \leqslant i \leqslant n.$$

Из только что установленного факта следует, что найдется номер i, $1 \leqslant i \leqslant n$, такой что имеет место в точности одна из следующих двух возможностей:

(a) $Y_i^{(1)} \subset X_i^{(0)}$, т. е. $X_i^{(1)} = Y_i^{(1)} \cap X_i^{(0)} = Y_i^{(1)}$;

(b) $Y_i^{(1)} \not\subset X_i^{(0)}$ и $X_i^{(1)} = Y_i^{(1)} \cap X_i^{(0)} \subset X_i^{(0)}$.

В случае (a) мы получаем ввиду $x^{(0)} \supseteq x^{(1)}$ и монотонности включения, что

$$Y_i^{(2)} = \sum_{j=1}^{n} A_{ij}X_j^{(1)} + B_i \subseteq \sum_{j=1}^{n} A_{ij}X_j^{(0)} + B_i = Y_i^{(1)} = X_i^{(1)},$$
$$X_i^{(2)} = Y_i^{(2)} \cap X_i^{(1)} = Y_i^{(2)},$$

а в общем случае ввиду $x^{(k)} \supseteq x^{(k+1)}$ методом математической индукции получаем

$$Y_i^{(k+1)} = \sum_{j=1}^{n} A_{ij}X_j^{(k)} + B_i \subseteq \sum_{j=1}^{n} A_{ij}X_j^{(k-1)} + B_i = Y_i^{(k)} = X_i^{(k)},$$
$$X_i^{(k+1)} = Y_i^{(k+1)} \cap X_i^{(k)} = Y_i^{(k+1)}.$$

(b). Полагая $i(A) = a_1, s(A) = a_2$ для $A = [a_1, a_2] \in I(\mathbb{R})$, мы можем, не умаляя общности, считать, что

$$i(Y_i^{(1)}) < i(X_i^{(0)}) \leqslant s(Y_i^{(1)}) < s(X_i^{(0)}),$$

т. е.

$$X_i^{(1)} = [i(X_i^{(0)}),\ s(Y_i^{(1)})].$$

(Возможен еще случай

$$i\left(X_i^{(0)}\right) < i\left(Y_i^{(1)}\right) \leqslant s\left(X_i^{(0)}\right) < s\left(Y_i^{(1)}\right),$$

т. е.

$$X_i^{(1)} = \left[i\left(Y_i^{(1)}\right),\ s\left(X_i^{(0)}\right)\right],$$

но он рассматривается аналогично.) Так как $x^{(0)} \supseteq x^{(1)}$, мы имеем

$$Y_i^{(2)} = \sum_{j=1}^{n} A_{ij} X_j^{(1)} + B_i \subseteq \sum_{j=1}^{n} A_{ij} X_j^{(0)} + B_i = Y_i^{(1)},$$

т. е.

$$i\ \left(Y_i^{(1)}\right) \leqslant i\left(Y_i^{(2)}\right),\ s\left(Y_i^{(2)}\right) \leqslant s\left(Y_i^{(1)}\right) = s\left(X_i^{(1)}\right),$$

а также

$$X_i^{(2)} = \left[\max\left\{i\left(Y_i^{(2)}\right),\ i\left(X_i^{(0)}\right)\right\},\ s\left(Y_i^{(2)}\right)\right].$$

Так как $x^{(k)} \supseteq x^{(k+1)}$, то можно показать методом математической индукции, что

$$i\left(Y_i^{(k)}\right) \leqslant i\left(Y_i^{(k+1)}\right),\ s\left(Y_i^{(k+1)}\right) \leqslant s\left(Y_i^{(k)}\right) = s\left(X_i^{(k)}\right), \qquad k \geqslant 1,$$

т. е.

$$X_i^{(k+1)} = \left[\max\left\{i\left(Y_i^{(k+1)}\right),\ i\left(X_i^{(0)}\right)\right\},\ s\left(Y_i^{(k+1)}\right)\right].$$

По предположению мы имеем $i\left(X^{(0)}\right) < i\left(X_i^*\right)$, а по теореме 1 $\lim\limits_{k\to\infty} i\left(X_i^{(k)}\right) = i\left(X_i^*\right)$. Поэтому найдется $k_0 \geq 1$, такое что имеет место

$$\max\left\{i\left(Y_i^{(k_0+1)}\right),\ i\left(X_i^{(0)}\right)\right\} = i\left(Y_i^{(k_0+1)}\right)$$

или

$$X_i^{(k_0+1)} = \left[i\left(Y_i^{(k_0+1)}\right),\ s\left(Y_i^{(k_0+1)}\right)\right],$$

т. е.

$$X_i^{(k_0+1)} = Y_i^{(k_0+1)} \cap X_i^{(k_0)} = Y_i^{(k_0+1)}.$$

Метод математической индукции позволяет теперь установить, что

$$X_i^{(k_0+\nu)} = Y_i^{(k_0+\nu)} \cap X_i^{(k_0+\nu-1)} = Y_i^{(k_0+\nu)}, \qquad \nu \geqslant 1,$$

так как $x^{(k)} \supseteq x^{(k+1)}$. Ввиду соотношений

$$\lim_{k\to\infty} x^{(k)} = x^* \text{ и } x^{(0)} \supset x$$

получаем, что для любого i, $1 \leqslant i \leqslant n$, не удовлетворяющего ни одному из условий (a), (b), хотя бы одно из этих условий выполнится после нескольких следующих шагов итерации. Наконец, получаем

$$x^{(k+1)} = (\mathcal{A}x^{(k)} + \mathcal{b}) \cap x^{(k)} = \mathcal{A}x^{(k)} + \mathcal{b}, \qquad k \geqslant \tilde{k} \geqslant 0.$$

Замечания.

Теоремы этого раздела без всяких изменений переносятся на соответствующие итерационные методы нахождения неподвижной точки, т. е. решения нелинейного уравнения

$$x = \mathcal{f}_{\mathrm{p}}(x),$$

где

$$\mathcal{f}_{\mathrm{p}}(x_{\mathrm{p}}) = \left(f_i\left(x_1, \ldots, x_n; a_{i1}, \ldots, a_{im_i}\right)\right)$$

есть $\mathscr{P}_{\mathrm{p}}$-сжатие.

Самое существенное для доказательства этих теорем свойство — монотонность включения.

3.5. О применимости метода Гаусса к системам уравнений с интервальными коэффициентами

Пусть $\mathcal{A}$ — интервальная матрица, $\mathcal{b}$ — интервальный вектор. Будем предполагать, что обращение $\mathcal{A}_{\mathrm{p}}^{-1}$ существует для всех $\mathcal{A}_{\mathrm{p}} \in \mathcal{A}$. Мы хотим найти множество

$$\Omega = \{x_{\mathrm{p}} \mid \mathcal{A}_{\mathrm{p}} x_{\mathrm{p}} = \mathcal{b}_{\mathrm{p}},\ \mathcal{A}_{\mathrm{p}} \in \mathcal{A},\ \mathcal{b}_{\mathrm{p}} \in \mathcal{b}\}.$$

Это множество в общем случае не имеет простого описания. Поэтому мы ограничимся его локализацией с помощью интервального вектора. Очевидный способ нахождения такого интервального вектора — применение непосредственного обобщения метода Гаусса на системы с интервальными коэффициентами. Иными словами, пусть нам дана таблица коэффициентов

$$\begin{array}{ccc} A_{11} & \ldots & A_{1n} \quad B_1 \\ \cdot & & \cdot \quad \cdot \\ \cdot & & \cdot \quad \cdot \\ \cdot & & \cdot \quad \cdot \\ A_{n1} & \ldots & A_{nn} \quad B_n. \end{array}$$

Применяя формулы

$$A'_{1j} = A_{1j}, \qquad 1 \leqslant j \leqslant n, \qquad B'_1 = B_1,$$
$$A'_{ij} = A_{ij} - A_{1j}(A_{i1}/A_{11}), \qquad 2 \leqslant i,\ j \leqslant n,$$
$$B'_i = B_i - B_1(A_{i1}/A_{11}), \qquad 2 \leqslant i \leqslant n,$$
$$A'_{i1} = 0, \qquad 2 \leqslant i \leqslant n,$$

в предположении $0 \notin A_{11}$, мы вычисляем новую таблицу коэффициентов

$$\begin{matrix} A'_{11} & A'_{12} & \dots & A'_{1n} & B'_1 \\ 0 & A'_{22} & \dots & A'_{2n} & B'_2 \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ 0 & A'_{n2} & \dots & A'_{nn} & B'_n. \end{matrix}$$

Покажем теперь, что имеет место

$$\{x_p \mid \mathscr{A}_p x_p = \mathscr{b}_p,\ \mathscr{A}_p \in \mathscr{A},\ \mathscr{b}_p \in \mathscr{b}\} \subseteq \{y_p \mid \mathscr{A}'_p y_p = \mathscr{b}'_p, \mathscr{A}'_p \in \mathscr{A}',\ \mathscr{b}'_p \in \mathscr{b}'\}.$$

Допустим, что $\mathscr{A}_p = \mathscr{A}$ и $\mathscr{b}_p \in \mathscr{b}$, и рассмотрим систему линейных уравнений

$$\mathscr{A}_p x_p = \mathscr{b}_p.$$

Строим матрицу $\mathscr{A}'_p = (a'_{ij})$ и вектор $\mathscr{b}'_p = (b'_i)$, где

$$a'_{1j} = a_{1j}, \qquad 1 \leqslant j \leqslant n, \qquad b'_1 = b_1,$$

и

$$a'_{ij} = a_{ij} - a_{1j}(a_{i1}/a_{11}), \qquad 2 \leqslant i,\ j \leqslant n,$$
$$b'_i = b_i - b_1(a_{i1}/a_{11}), \qquad 2 \leqslant i \leqslant n,$$
$$a'_{i1} = 0, \qquad 2 \leqslant i \leqslant n.$$

Из теории линейных уравнений хорошо известно, что система уравнений $\mathscr{A}'_p y_p = \mathscr{b}'_p$ имеет те же решения, что и $\mathscr{A}_p x_p = \mathscr{b}_p$. Из монотонности включения следует $\mathscr{A}'_p \in \mathscr{A}'$ и $\mathscr{b}'_p \in \mathscr{b}'$, что и доказывает наше утверждение. Если описанный выше шаг проведен n—1 раз, то исходная таблица коэффициентов превращается в верхнюю треугольную матрицу

$$\begin{matrix} \tilde{A}_{11} & \tilde{A}_{12} & \dots & \tilde{A}_{1n} & \tilde{B}_1 \\ & \tilde{A}_{22} & & \vdots & \vdots \\ & & \ddots & \vdots & \vdots \\ & & & \tilde{A}_{nn} & \tilde{B}_n \end{matrix},$$

для которой имеет место

$$\{x_\mathrm{p} \mid \mathscr{A}_\mathrm{p} x_\mathrm{p} = b_\mathrm{p},\ \mathscr{A}_\mathrm{p} \in \mathscr{A},\ b_\mathrm{p} \in b\} \subseteq \{\tilde{x}_\mathrm{p} \mid \tilde{\mathscr{A}}_\mathrm{p} \tilde{x}_\mathrm{p} = \tilde{b}_\mathrm{p}, \tilde{\mathscr{A}}_\mathrm{p} \in \tilde{\mathscr{A}},\ \tilde{b}_\mathrm{p} \in \tilde{b}\}.$$

Используя формулы

$$X_n = \tilde{B}_n / \tilde{A}_{nn},$$

$$X_i = \left(\tilde{B}_i - \sum_{j=i+1}^{n} \tilde{A}_{ij} X_j\right) / \tilde{A}_{ii}, \quad 1 \leqslant i \leqslant n-1,$$

получаем тогда интервальный вектор $x = (X_i)$, удовлетворяющий условию

$$\{x_\mathrm{p} \mid \mathscr{A}_\mathrm{p} x_\mathrm{p} = b_\mathrm{p},\ \mathscr{A}_\mathrm{p} \in \mathscr{A},\ b_\mathrm{p} \in b\} \subseteq x.$$

В частности, если $\mathscr{A}_\mathrm{p} = (a_{ij})$— невырожденная точечная матрица, то метод Гаусса применим, когда в правой части стоит произвольный интервальный вектор. При этом в процессе исключения по Гауссу может потребоваться перестановка столбцов. Это эквивалентно умножению матрицы $\mathscr{A}_\mathrm{p}$ слева на матрицу перестановки перед началом процесса исключения.

Теперь определим отображение

$$g_\mathrm{p}\colon\ M_{nn}(\mathbb{C}) \times V_n(\mathbb{C}) \to V_n(\mathbb{C})$$

для невырожденной матрицы $\mathscr{A}_\mathrm{p}$ и точечного вектора b_p. Это отображение представляет собой применение метода Гаусса к системе линейных уравнений

$$\mathscr{A}_\mathrm{p} x_\mathrm{p} = b_\mathrm{p},$$

дающее результат

$$x_\mathrm{p} = g_\mathrm{p}(\mathscr{A}_\mathrm{p},\ b_\mathrm{p}).$$

Отображение g_p единственно, но, как обычно, для g_p имеются различные выражения. Например, мы имеем $\mathscr{A}_\mathrm{p}^{-1} b_\mathrm{p} = g_\mathrm{p}(\mathscr{A}_\mathrm{p},\ b_\mathrm{p})$. Кроме того, метод Гаусса дает различные выражения для g_p в зависимости от выбора главных элементов.

Следующие свойства не зависят от выбора главных элементов. Интервальное выражение для g_p обозначается через $g_\mathrm{p}(\mathscr{A},\ b)$. Поэтому интервальный вектор x, получаемый после выполнения

описанного выше метода Гаусса, можно задать равенством $x = g_p(\mathcal{A}, b)$.

Имеем следующие свойства:

$$\mathcal{A}, \mathcal{B} \in M_{nn}(I(\mathbb{C})), \quad a, b \in V_n(I(\mathbb{C})), \quad \mathcal{A} \subseteq \mathcal{B}, \quad a \subseteq b. \tag{1}$$

Отсюда следует, что

$$g_p(\mathcal{A}, a) \subseteq g_p(\mathcal{B}, b).$$

$$\mathcal{A}_p \in M_{nn}(\mathbb{C}), \quad b = u + v \in V_n(I(\mathbb{C})); \tag{2}$$

Отсюда следует, что

$$g_p(\mathcal{A}_p, b) = g_p(\mathcal{A}_p, u) + g_p(\mathcal{A}_p, v).$$

$$\mathcal{A}_p \in M_{nn}(\mathbb{R}), \quad b \in V_n(I(\mathbb{R})). \tag{3}$$

Отсюда следует, что

$$\mathcal{A}_p^{-1} b \subseteq g_p(\mathcal{A}_p, b).$$

$$\mathcal{A}_p \in M_{nn}(\mathbb{C}), \quad a, b \in V_n(I(\mathbb{C})), \quad d(a) \leqslant \alpha d(b) \quad \text{для некоторого } \alpha \geqslant 0. \tag{4}$$

Поэтому для ширины имеет место

$$d(g_p(\mathcal{A}_p, a)) \leqslant \alpha d(g_p(\mathcal{A}_p, b)).$$

По поводу доказательства этих свойств заметим, что (1) сразу следует из монотонности включения, а (2) — из соотношения $a(B + C) = aB + aC$ и формул, определяющих метод Гаусса. Чтобы доказать (3), используем следующий факт.

Если имеются два рациональных выражения f_1 и f_2 для одной и той же функции $f: \mathbb{R} \to \mathbb{R}$, причем f_1 содержит переменную x ровно один раз, а f_2 содержит эту переменную m раз, то для вычислений f_1 и f_2 в интервальной арифметике имеет место $f_1(X) \subseteq f_2(X)$. Аналогичное утверждение верно и для функций от нескольких переменных. Рассмотрим теперь i-е, $1 \leqslant i \leqslant n$, компоненты векторов $\mathcal{A}_p^{-1} b$ и $g_p(\mathcal{A}_p, b)$. Для точечных векторов b_p имеет место $(\mathcal{A}_p^{-1} b_p)_i = (g_p(\mathcal{A}_p, b_p))_i$. Из формул, определяющих метод Гаусса, видно, что компоненты вектора b_p могут входить несколько раз в выражение $(g_p(\mathcal{A}_p, b_p))_i$. Так как они входят всего один раз в $(\mathcal{A}_p^{-1} b_p)_i$, мы получаем (3).

Наконец, (4) получается путем использования формул, описывающих метод Гаусса, правил (10 п.1.2), (14 п.1.2), (12 п.1.5) и (16 п.1.5), а также предположения $d(a_p) \leqslant \alpha d(b_p)$.

Формулы, описывающие метод Гаусса, применимы только к таким интервальным матрицам, для которых условие $0 \notin A_{ii}$ выполняется на всех шагах приведения к верхней треугольной форме. Если это не так при некотором i, $1 \leqslant i \leqslant n-1$, т. е. $0 \in A_{ii}$, то все еще возможно, что подходящая перестановка столбцов позволит избежать соотношения $0 \in A_{ii}$. Однако это не всегда возможно, даже если предположить, что сначала $\mathscr{A}_{\mathrm{p}}^{-1}$ существует для всех $\mathscr{A}_{\mathrm{p}} \in \mathscr{A}$. Причина заключается в следующем.

Очевидно, что в предположении существования $\mathscr{A}_{\mathrm{p}}^{-1}$ для всех $\mathscr{A}_{\mathrm{p}} \in \mathscr{A}$ мы всегда можем выполнить первый шаг метода Гаусса. Если бы мы не смогли сделать этот шаг из-за того, что все элементы первого столбца содержат нуль, то исходная интервальная матрица $\mathscr{A}$ содержала бы вырожденную точечную матрицу $\mathscr{A}_{\mathrm{p}}$, а это противоречит нашему предположению. Рассмотрим теперь матрицу $\mathscr{A}^{(1)} = (A_{ij}^{(1)})$ размерности (n—1)×((n—1), для которой $A_{ij}^{(1)} = A'_{ij}$, $2 \leqslant i, j \leqslant n$. Пусть $\mathscr{U}$ обозначает интервальную матрицу

$$\mathscr{U} = \begin{pmatrix} -A_{21}/A_{11} & 1 & & 0 \\ -A_{31}/A_{11} & 0 & 1 & \\ \vdots & & & \ddots \\ -A_{n1}/A_{11} & 0 & \cdots & 0 \quad 1 \end{pmatrix},$$

размерности (n— 1)×n, а $\mathscr{V}$—интервальную матрицу

$$\mathscr{V} = \begin{pmatrix} A_{12} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{n2} & \cdots & A_{nn} \end{pmatrix}.$$

размерности n× (n—1). Мы имеем $\mathscr{A}^{(1)} = \mathscr{U}\mathscr{V}$. Аналогичным образом положим

$$\mathscr{R}_{\mathrm{p}} = \begin{pmatrix} -a_{21}/a_{11} & 1 & & 0 \\ -a_{31}/a_{11} & 0 & 1 & \\ \vdots & & & \ddots \\ -a_{n1}/a_{11} & 0 & \cdots & 0 \quad 1 \end{pmatrix},$$

$$\mathscr{S}_{\mathrm{p}} = \begin{pmatrix} a_{12} & \cdots & a_{1n} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ a_{n2} & \cdots & a_{nn} \end{pmatrix},$$

где $a_{ij} \in A_{ij}$, $1 \leqslant i, j \leqslant n$. Мы получаем, что матрица $\mathscr{Z}_p = \mathscr{R}_p \mathscr{S}_p$ невырожденная, так как $\mathscr{A}_p \in \mathscr{A}$ невырожденная. Поэтому в первом столбце матрицы $\mathscr{Z}_p$ найдется хотя бы один ненулевой элемент. Из монотонности включения следует, что

$$\{\mathscr{Z}_p = \mathscr{R}_p \mathscr{S}_p \mid \mathscr{A}_p \in \mathscr{A}\} \subseteq \{\mathscr{A}_p^{(1)} \mid \mathscr{A}_p^{(1)} \in \mathscr{A}^{(1)} = \mathscr{U}\mathscr{V}\},$$

и в общем случае здесь нет равенства, как показывает простой пример из п. 2.3. Поэтому нет гарантии, что в первом столбце матрицы $\mathscr{A}^{(1)}$ будет хотя бы один элемент, не содержащий нуля.

Те же рассуждения проходят и для следующих шагов метода. Покажем это с помощью следующего простого примера.

Пример. Рассмотрим интервальную матрицу

$$\mathscr{A} = \begin{pmatrix} [1, 5] + i[-1, 1] & 1 \\ 25 & [-1, 1] + i[-1, 1] \end{pmatrix}.$$

Покажем сначала, что $\mathscr{A}$ не содержит вырожденной точечной матрицы. Так как

$$\det \mathscr{A}_p = a_{11}a_{22} - a_{21}a_{12} \in A_{11}A_{22} - A_{21}A_{12},$$

получаем для любой точечной матрицы $\mathscr{A}_p \in \mathscr{A}$, что

$$\det \mathscr{A}_p \in [-31, -19] + i[-6, 6].$$

Из того что интервал в правой части не содержит нуля, следует, что $\mathscr{A}_p$ невырожденная.

Теперь шаг исключения согласно методу Гаусса порождает (в силу определения 3 п.1.4 и следующего за ним замечания) интервал

$$A'_{22} = \left[-126, \frac{1}{26}\right] + i[-26, 26].$$

Поэтому невозможно продолжить применение метода Гаусса. Даже если мы отделим в матрице $\mathscr{A}$ вещественную и мнимую части, нам все равно не удастся решить получившуюся систему из 4 интервальных уравнений с 4 неизвестными с помощью метода Гаусса, хотя исходная интервальная матрица снова не содержит вырожденных точечных матриц.

Покажем теперь, что метод Гаусса всегда можно выполнить при $n \leq 2$, если коэффициенты — вещественные интервалы и все точечные матрицы $\mathscr{A}_p \in \mathscr{A}$ невырожденные.

Теорема 1. *Пусть* $1 \leq n \leq 2$ *интервальная матрица* $\mathscr{A}_p = (A_{ij})$ *размерности* $n \times n$ *не содержит невырожденных матриц* $\mathscr{A}_p$. *Тогда можно выполнить метод Гаусса.*

Доказательство. При $n = 1$ наше предположение означает, что $0 \notin A_{11}$, и в этом случае наша теорема доказана для $\mathscr{A} = (A_{11})$. При $n = 2$ хотя бы один из интервалов A_{11}, A_{21} не

содержит нуля. Если бы это было не так, то существовала бы вырожденная матрица $\mathscr{A}_p \in \mathscr{A}$, что противоречит условию теоремы. Не умаляя общности, считаем, что $0 \notin A_{11}$; если это не так, переставим строки матрицы $\mathscr{A}$. Теперь метод Гаусса дает

$$A'_{22} = A_{22} - (1/A_{11})\, A_{21} A_{12}.$$

Мы можем рассматривать A'_{22} как оценивание рациональной функции a'_{22} от четырех переменных a_{11}, a_{21}, a_{12} и a_{22} в интервальной арифметике, определяемое формулой

$$a'_{22}(a_{11}, a_{12}, a_{21}, a_{22}) = a_{22} - (1/a_{11})\, a_{21} a_{12}.$$

По условию теоремы мы имеем для любой $\mathscr{A}_p \in \mathscr{A}$ соотношение

$$\det(\mathscr{A}_p) = a_{11}a_{22} - a_{21}a_{12} \neq 0,$$

откуда

$$a'_{22}(a_{11}, a_{12}, a_{21}, a_{22}) = (1/a_{11}) \det(\mathscr{A}_p) \neq 0.$$

Приведенное интервальное оценивание даст точную локализацию, если заменить a_{11} на A_{11}, a_{12} на A_{12}, a_2^1 на A_2^1 и a_{22} на A_{22}, так как каждая переменная входит в выражение для a'_{22} всего один раз. Поэтому мы имеем $0 \notin A'_{22}$, что и означает возможность выполнения метода Гаусса.

Данное выше доказательство не обобщается на случай $n \geq 3$. Даже для $n = 2$ теорема будет неверна, если элементы интервальных матриц берутся из $R(C)$ или из $K(C)$.

Рассмотрим теперь один конкретный класс интервальных матриц, для которого метод Гаусса всегда может быть выполнен. В этом классе можно даже не переставлять строки или столбцы. В дальнейшем мы ограничимся системами уравнений, где элементы матрицы коэффициентов и правой части принадлежат множеству вещественных интервалов или комплексных круговых интервалов. Системы уравнений, элементами которых являются комплексные прямоугольные интервалы, можно свести к первому из упомянутых случаев, отделяя вещественную и мнимую части.

Чтобы объединить приводимые ниже доказательства для вещественных интервалов и комплексных круговых интервалов, мы заметим, что вещественный интервал вида $A = [a_1, a_2]$ можно представить в виде

$$A = [a - r, a + r],$$

где

$$a = \frac{1}{2}(a_1 + a_2), \qquad r = \frac{1}{2}\, d(A) = \frac{1}{2}(a_2 - a_1).$$

Здесь a — центр интервала, а r — радиус, т. е. половина ширины. Мы вводим обозначение

$$A = [a - r,\ a + r] =: \langle a,\ r\rangle,$$

чтобы подчеркнуть аналогию с комплексными круговыми интервалами. Арифметические операции на вещественных интервалах также можно определить через центр и ширину. Пусть $A = \langle a, r\rangle$, $B = \langle b,\ s\rangle$. Тогда сложение и вычитание интервалов A и B можно записать в виде

$$A \pm B = \langle a \pm b,\ r + s\rangle.$$

Формально это соответствует сложению и вычитанию комплексных круговых интервалов. Для умножения нам нужно только равенство

$$[-r,\ r][-s,\ s] = \langle 0,\ r\rangle\langle 0,\ s\rangle = \langle 0,\ rs\rangle.$$

Если мы заметим теперь, что при $0 \notin A = [a_1,\ a_2]$ имеет место представление

$$\frac{1}{A} = \left[\frac{1}{a+r},\ \frac{1}{a-r}\right] = \left[\frac{a}{a^2-r^2} - \frac{r}{a^2-r^2},\ \frac{a}{a^2-r^2} + \frac{r}{a^2-r^2}\right],$$

то получим

$$\frac{1}{A} = \left\langle \frac{a}{a^2-r^2},\ \frac{r}{a^2-r^2}\right\rangle.$$

Формально это соответствует обращению круговою комплексного интервала. Для представления вещественных интервалов в виде $A = \langle a, r\rangle$ верно также

$$|A| = \max\{|a_1|, |a_2|\} = |a| + r.$$

Кроме того, имеем

$$0 \notin A \Leftrightarrow |a| - r > 0.$$

Наконец, имеем

$$A = \langle a,\ r\rangle \subseteq \langle 0,\ |A|\rangle = \langle 0,\ |a| + r\rangle.$$

Докажем сначала следующее утверждение.

Лемма 2. *Пусть*

$$A = \langle a, r_1\rangle,\ B = \langle b, r_2\rangle,\ C = \langle c, r_3\rangle \text{ и } D = \langle d, r_4\rangle$$

— вещественные интервалы или круговые комплексные интервалы, причем $0 \notin D$. *Тогда для*

$$Z = \langle z,\ r_5\rangle = A - (1/D)BC$$

справедливо неравенство

$$|a| - r_1 - |B||C|/(|d| - r_4) \leqslant |z| - r_5.$$

Доказательство. Из монотонности включения следует, что

$$Z = \langle z,\ r_5\rangle = A - BC\frac{1}{D} \subseteq A - \langle 0,\ |B|\rangle\langle 0,\ |C|\rangle\left\langle \frac{\bar{d}}{d\bar{d} - r_4^2},\ \frac{r_4}{d\bar{d} - r_4^2}\right\rangle$$

$$= \langle a,\ r_1\rangle - \left\langle 0,\ |B||C|\frac{|d|}{|d\bar{d} - r_4^2|} + |B||C|\frac{r_4}{d\bar{d} - r_4^2}\right\rangle$$

$$= \langle a,\ r_1\rangle - \left\langle 0,\ |B||C|\frac{1}{|d| - r_4}\right\rangle$$

$$= \left\langle a,\ r_1 + |B||C|\frac{1}{|d| - r_4}\right\rangle =: \langle a,\ r_6\rangle.$$

Отсюда и из $Z \subseteq \langle a,\ r_6\rangle$ следует, что

$$|a| - |z| \leqslant |a - z| \leqslant r_6 - r_5$$

или

$$|z| - r_5 \geqslant |a| - r_1 - |B||C|(1/(|d| - r_4)).$$

Чтобы сформулировать следующее утверждение, нам нужно понятие M-матрицы. Мы применим здесь эквивалентное определение. Вещественная матрица $\mathscr{B}_p = (b_{ij})$ называется M-матрицей, если выполнены условия

$$b_{ij} \leqslant 0, \quad i \neq j, \tag{1}$$

$$\mathscr{B}_p^{-1} \geqslant O_p. \tag{2}$$

В (2) подразумевается покомпонентный частичный порядок. Условие (2) можно заменить следующим условием:

(2') Существует вещественный вектор $u_p = (u_i)$, такой что $u_i > 0,\ 1 \leqslant i \leqslant n$ и $\mathscr{B}_p u_p > o_p$.

Этот факт, а также то обстоятельство, что диагональные элементы M-матрицы положительны, используется в следующем утверждении.

Теорема 3. *Пусть* $\mathscr{A} = (A_{ij})$ *— интервальная матрица, причем* $A_{ij} = \langle a_{ij},\ r_{ij}\rangle,\ 1 \leqslant i,\ j \leqslant n,$ *и пусть*

$$\mathscr{B}_p = (b_{ij})$$

— вещественная матрица, определенная соотношением

$$b_{ij} = \begin{cases} |a_{ii}| - r_{ii}, & i = j, \\ -|A_{ij}| & \text{в противном случае.} \end{cases}$$

Если $\mathscr{B}_p$ *является M-матрицей, то для матрицы* $\mathscr{A}_p$ *можно выполнить алгоритм Гаусса без перестановки строк или столбцов.*

Доказательство. Предположение, что $\mathscr{B}_p$ является M-матрицей, означает существование вектора $u_p = (u_i)$ с положительными элементами, такого что $\mathscr{B}_p u_p > o_p$, т. е. верно

$$(|a_{ii}|-r_{ii})u_i > \sum_{j=1,\ j\neq i}^{n} |A_{ij}|u_j, \qquad 1\leqslant i\leqslant n.$$

Ввиду неравенства $|a_{11}|-r_{11}>0$ выполнено условие $0\notin A_{11}$ и применимы формулы из первой части этого микромодуля. Теперь мы покажем, что условия теоремы выполнены и для интервальной матрицы $\tilde{A}'=(\tilde{A}'_{ij})$ размерности $(n-1)\times(n-1)$, где

$$\tilde{A}'_{ij}=A'_{ij}=\langle a'_{ij},\ r'_{ij}\rangle,\ 2\leqslant i,\ j\leqslant n.$$

Это позволит немедленно завершить доказательство теоремы с помощью математической индукции.

Для $i\geq 2$ имеет место

$$\sum_{j=2,\ j\neq i}^{n} |A'_{ij}|u_j = \sum_{j=2,\ j\neq i}^{n} \left|A_{ij}-A_{1j}\frac{A_{i1}}{A_{11}}\right|u_j$$

$$\leqslant \sum_{j=2,\ j\neq i}^{n} |A_{ij}|u_j + |A_{i1}|\left|\frac{1}{A_{11}}\right| \sum_{j=2,\ j\neq i}^{n} |A_{1j}|u_j.$$

С помощью неравенства

$$\sum_{j=2,\ j\neq i}^{n} |A_{1j}|u_j < (|a_{11}|-r_{11})u_1 - |A_{i1}|u_i,$$

справедливого в силу условия теоремы, можно получить оценки

$$\sum_{j=2,\ j\neq i}^{n} |A'_{ij}|u_j \leqslant \sum_{j=2,\ j\neq i}^{n} |A_{ij}|u_j$$

$$+|A_{i1}|\frac{1}{|a_{11}|-r_{11}}\{(|a_{11}|-r_{11})u_1-|A_{1i}|u_i\}$$

$$=\sum_{j=1,\ j\neq i}^{n} |A_{ij}|u_j - \frac{|A_{i1}||A_{1i}|}{|a_{11}|-r_{11}}u_i$$

$$<u_i\left(|a_{ii}|-r_{ii}-\frac{|A_{i1}||A_{1i}|}{|a_{11}|-r_{11}}\right)\leqslant(|a'_{ii}|-r'_{ii})u_i,$$

где последнее неравенство получается по лемме 2. Это завершает доказательство.

Следующее определение вводит важный класс интервальных матриц, удовлетворяющих условиям теоремы 3.

Определение 4. Говорят, что интервальная матрица $\mathcal{A}=(A_{ij})$, компоненты которой $A_{ij}=\langle a_{ij},\ r_{ij}\rangle$ являются вещественными интервалами или круговыми комплексными интервалами, имеет сильно доминирующую диагональ (или что ее диагональ сильно доминирует), если

$$|a_{ii}| - r_{ii} > \sum_{j=1,\, j\neq i}^{n} |A_{ij}|, \quad 1 \leqslant i \leqslant n.$$

Очевидно, что элементы сильно доминирующей диагонали не содержат нулей и что для любой точечной матрицы $\hat{\mathcal{A}}_p = (\hat{a}_{ij}) \in \mathcal{A}$ выполнено соотношение

$$|\hat{a}_{ii}| > \sum_{j=1,\, j\neq i}^{n} |\hat{a}_{ij}|, \quad 1 \leqslant i \leqslant n.$$

Поэтому любая точечная матрица $\hat{\mathcal{A}}_p \in \mathcal{A}$ имеет сильно доминирующую диагональ в обычном смысле и потому невырождена.

Для матрицы с сильно доминирующей диагональю можно выполнить условия теоремы 3, если взять вектор $u_p = (u_i)$, такой что $1 \leq i \leq n$. Мы доказали следующее утверждение.

Следствие 5. *Пусть интервальная матрица $\mathcal{A}$ имеет сильно доминипующую диагональ Тогда метод Гаусса можно выполнить для $\mathcal{A}$ без перестановки строк или столбцов.*

Требование строгого доминирования диагонали можно ослабить, сохранив применимость метода Гаусса, если данная интервальная матрица имеет вид

$$\mathcal{A} = \begin{pmatrix} A_1 & C_1 & & 0 \\ B_2 & A_2 & C_2 & \\ & \ddots & \ddots & \ddots \\ 0 & & B_n & A_n \end{pmatrix};$$

т. е. является трехдиагональной интервальной матрицей. Мы предположим еще, что $C_i \neq 0$, $1 \leqslant i \leqslant n-1$ и $B_i \neq 0$, $2 \leqslant i \leqslant n$, так как в противном случае задача распадается на меньшие задачи, для которых эти условия выполнены.

Теорема 6. *Пусть $\mathcal{A}$ — трехдиагональная интервальная матрица, такая что*

$$A_i = \langle a_i, r_i \rangle, \qquad 1 \leqslant i \leqslant n,$$
$$B_i = \langle b_i, s_i \rangle \neq 0, \qquad 2 \leqslant i \leqslant n,$$
$$C_i = \langle c_i, t_i \rangle \neq 0, \qquad 1 \leqslant i \leqslant n-1.$$

Предположим далее, что

$$|a_1| - r_1 > |C_1|,$$
$$|a_i| - r_i \geqslant |B_i| + |C_i|, \quad 2 \leqslant i \leqslant n-1,$$
$$|a_n| - r_n > |B_n|.$$

Тогда метод Гаусса может быть выполнен без перестановки строк или столбцов.

Замечание. В случае трехдиагональной матрицы $\mathscr{A}$ условия из определения 4 выполнены только для первой и последней строк.

Доказательство теоремы 6. Первый шаг метода Гаусса состоит в порождении трехдиагональной матрицы $\mathscr{A}'$, для которой

$$\begin{aligned}
&A'_i = A_i, \quad C'_i = C_1, \\
&B'_2 = 0, \quad B'_i = B_i, && 3 \leqslant i \leqslant n, \\
&A'_2 = A_2 - C_1 B_2 (1/A_i), \quad A'_i = A_i, && 3 \leqslant i \leqslant n, \\
&C'_i = C_i, && 2 \leqslant i \leqslant n-1.
\end{aligned}$$

Покажем, что в матрице $\mathscr{A}'$ сильный критерий суммы по строкам выполнен не только для первой, но и для второй строки, т. е. верно

$$|a'_2| - r'_2 > |C_2| = |C'_2|.$$

Имеем

$$\begin{aligned}
A'_2 &= A_2 - C_1 (B_2/A_1) \subseteq A_2 - |1/A_1| \langle 0, |C_1| \rangle \langle 0, |B_2| \rangle \\
&= \langle a_2, r_2 + |C_1||B_2|/(|a_1| - r_1) \rangle,
\end{aligned}$$

т. е.

$$|a'_2| - r'_2 \geqslant |a_2| - \left(r_2 + \frac{|C_1||B_2|}{|a_1| - r_1} \right).$$

Так как

$$|a_1| - r_1 > |C_1| > 0, \quad |B_2| > 0$$

и

$$-|L_2| - r_2 + |a_2| \geqslant |C_2|,$$

получаем, что

$$\begin{aligned}
|a'_2| - r'_2 &\geqslant |a_2| - \left(r_2 + \frac{|C_1||B_2|}{|a_1| - r_1} \right) \\
&> -|B_2| - r_2 + |a_2| \geqslant |C_2| = |C'_2|.
\end{aligned}$$

После n—1 шага такого типа приходим к матрице $\hat{\mathscr{A}}$, имеющей ненулевые элементы только для главной диагонали и супердиагонали (расположенной непосредственно над главной), причем никакой элемент главной диагонали не содержит нуля. Третье предположение $|a_n| - r_n > |B_n|$ применяется на (n—1)-м шаге. Заметим, что доказательство теоремы 6 можно несколько сократить. Из условий теоремы 6 следует, что матрица $\mathscr{B}_P$, введенная в теореме 3, имеет неприводимо сильно доминирующую главную диагональ, а

потому является *M*-матрицей. Это означает, что наше утверждение — просто частный случай теоремы 3.

Рассмотрим теперь систему, имеющую более подходящий для итерации вид

$$x = \mathscr{C}x + c,$$

где $\mathscr{C} = (C_{ij})$, $C_{ij} = \langle c_{ij}, r_{ij} \rangle$, $1 \leqslant i, j \leqslant n$ — вещественная интервальная матрица и c — интервальный вектор. Согласно теореме 1, из микромодуля 31, итерационный метод

$$x^{(k+1)} = \mathscr{C}x^{(k)} + c, \quad k = 0, 1, 2, \ldots$$

сходится для любого начального интервального вектора $x^{(0)}$ к единственной неподвижной точке, т. е. решению уравнения $x = \mathscr{C}x + x$, тогда и только тогда, когда спектральный радиус матрицы $|\mathscr{C}| = (|C_{ij}|)$ меньше единицы. Мы хотим показать, что при этом условия теоремы 3 всегда выполнены для матрицы $\mathscr{A} = \mathscr{I}_p - \mathscr{C} = (A_{ij})$, где $\mathscr{I}_p$ — единичная матрица.

Имеем

$$A_{ij} = \begin{cases} \langle 1 - c_{ii}, r_{ii} \rangle, & i = j, \\ -C_{ij} & \text{в противном случае.} \end{cases}$$

Из $\rho(|\mathscr{C}|) < 1$ следует, что $|C_{ii}| = |c_{ii}| + r_{ii} < 1$, $1 \leqslant i \leqslant n$.

Матрица $\mathscr{B}_p = (b_{ij})$, введенная в теореме 3 имеет для $\mathscr{A} = \mathscr{I}_p - \mathscr{C}$ элементы

$$b_{ij} = \begin{cases} |1 - c_{ii}| - r_{ii}, & i = j, \\ -|C_{ij}| & \text{в противном случае.} \end{cases}$$

Мы рассмотрим теперь вещественную матрицу $\mathscr{B}_{p1} = \mathscr{I}_p - |\mathscr{C}|$. Ввиду $\rho(|\mathscr{C}|) < 1$ обратная матрица $\mathscr{B}_{p1}^{-1}$ существует в силу известной теоремы, причем $\mathscr{B}_{p1}^{-1} \geqslant O_p$. Рассмотрим теперь представление матрицы $\mathscr{B}_{p1}$ в виде

$$\mathscr{B}_{p1} = \mathscr{M}_{p1} - \mathscr{N}_{p1},$$

где

$$\mathscr{M}_{p1} = \operatorname{diag}(1 - |C_{ii}|), \quad \mathscr{N}_{p1} = -(\mathscr{B}_{p1} - \mathscr{M}_{p1}) = \mathscr{M}_{p1} - \mathscr{B}_{p1}.$$

Имеем $\mathscr{M}_{p1}^{-1} \geqslant O_p$, $\mathscr{N}_{p1} \geqslant O_p$. Отсюда в силу известной теоремы следует, что $\rho(\mathscr{M}_{p1}^{-1}\mathscr{N}_{p1}) < 1$.

Рассмотрим, наконец, представление матрицы $\mathscr{B}_p$ в виде

$$\mathscr{B}_p = \mathscr{M}_p - \mathscr{N}_p,$$

где

$$\mathcal{M}_p = \operatorname{diag}(|1 - c_{ii}| - r_{ii}), \quad \mathcal{N}_p = -(\mathcal{B}_p - \mathcal{M}_p) = \mathcal{M}_p - \mathcal{B}_p.$$

Имеем

$$|1 - c_{ii}| - r_{ii} \geqslant 1 - |c_{ii}| - r_{ii} = 1 - |C_{ii}| > 0, \quad 1 \leqslant i \leqslant n,$$

откуда

$$\mathcal{M}_p \geqslant \mathcal{M}_{p1}, \text{ т. е. } \mathcal{M}_{p1}^{-1} \geqslant \mathcal{M}_p^{-1}.$$

Отсюда и из $\mathcal{N}_p = \mathcal{N}_{p1}$ следует, что $\mathcal{M}_p^{-1}\mathcal{N}_p \leqslant \mathcal{M}_{p1}^{-1}\mathcal{N}_{p1}$.

Теперь теорема Перрона-Фробениуса дает

$$\rho(\mathcal{M}_p^{-1}\mathcal{N}_p) \leqslant \rho(\mathcal{M}_{p1}^{-1}, \mathcal{N}_{p1}), \text{ т. е. } \rho(\mathcal{M}_p^{-1}\mathcal{N}_p) < 1.$$

Применяя известную теорему, мы получаем, наконец, что $\mathcal{B}_p$ является *M*-матрицей.

Замечания. Утверждения о применимости метода Гаусса (теорема 3) можно найти в литературе. Доказательство теоремы 3 — это непосредственное обобщение доказательства таких же утверждений для точечных матриц. Пусть матрица $\mathcal{A}^{\dot{T}}$— транспонированная для интервальной матрицы $\mathcal{A}$. Мы имеем утверждение, соответствующее следствию 5. Если $\mathcal{A}^T$ имеет сильно доминирующую диагональ (в смысле определения 4), то метод Гаусса может быть выполнен для интервальной матрицы $\mathcal{A}$ без перестановок строк. Доказательство получается ссылкой на теорему 3, так как введенная там матрица $\mathcal{B}_p$ снова оказывается *M*-матрицей. Аналогичное утверждение справедливо для трехдиагональной матрицы $\mathcal{A}$, если для $\mathcal{A}^T$ выполнены условия теоремы 6.

Только что доказанные утверждения будут далее использованы для решения систем нелинейных точечных уравнений.

Вопрос о применимости метода Гаусса не был пока что достаточно исследован удовлетворительным образом. В литературе было показано, что этот метод применим для частного класса уравнений.

3.6. Метод и процедура Хансена

1. Метод Хансена

Если в системе линейных интервальных уравнений диагональ не является сильно доминирующей (определение 4 из п.3.5), то ее можно решать с помощью преобразования, предложенного Хансеном. Цель

этого преобразования — сделать данную систему интервальных уравнений системой с сильно доминирующей диагональю. Пусть дана интервальная матрица $\mathcal{A} = (A_{ij})$, где $A_{ij} = \langle a_{ij}, r_{ij} \rangle$ — элементы из *I*(R) или *K(*C*)*. Будем предполагать, что существуют обращения всех точечных матриц $\mathcal{A}_p \in \mathcal{A}$. Берем обращение точечной матрицы $m(\mathcal{A}) := (a_{ij})$ и с помощью $m(\mathcal{A})^{-1}$ строим интервальную матрицу

$$\tilde{\mathcal{A}} = m(\mathcal{A})^{-1}\mathcal{A}$$

и интервальный вектор

$$\tilde{b} = m(\mathcal{A})^{-1} b.$$

Имеем

$$\{x_p \mid \mathcal{A}_p x_p = b_p \mathcal{A}_p \in \mathcal{A},\ b_p \in b\} \subseteq \{y_p \mid \mathcal{A}_p y_p = \tilde{b}_p \tilde{\mathcal{A}}_p \subset \tilde{\mathcal{A}},\ \tilde{b}_p \in \tilde{b}\}.$$

Чтобы показать это, предположим, что x_p принадлежит множеству из левой части, т. е

$$\mathcal{A}_p x_p = b_p, \text{ где } \mathcal{A}_p \in \mathcal{A},\ b_p \in b.$$

Тогда

$$m(\mathcal{A})^{-1}\mathcal{A}_p x_p = m(\mathcal{A})^{-1} b_p$$

и наше утверждение следует из

$$m(\mathcal{A})^{-1}\mathcal{A}_p \in \tilde{\mathcal{A}}, \quad m(\mathcal{A})^{-1} b_p \in \tilde{b}.$$

Идея рассматриваемого преобразования состоит в том, что если элементы матрицы $\mathcal{A}$ имеют не слишком большую ширину, то диагональ матрицы $\tilde{\mathcal{A}}$ будет сильно доминирующей. Тогда можно применить метод Гаусса. В пределе мы имеем $d(\mathcal{A}) = \mathcal{O}_p$, т. е. $\tilde{\mathcal{A}} = \mathcal{I}_p$, так что в этом случае матрица $\tilde{\mathcal{A}}$ наверняка имеет сильно доминирующую диагональ. Если же ширина компонент матрицы $\mathcal{A}$ невелика, то $\tilde{\mathcal{A}}$ несильно отличается от $\mathcal{I}_p$.

Ясно, что сильное доминирование диагонали для матрицы $\tilde{\mathcal{A}} = m(\mathcal{A})^{-1}\mathcal{A}$ зависит не только от ширины компонент матрицы $\mathcal{A}$. Действительно, если мы представим компоненты этой интервальной матрицы (которые могут быть круговыми комплексными или вещественными интервалами) в виде

$$A_{ij} = \langle a_{ij}, r_{ij} \rangle, \quad 1 \leqslant i,\ j \leqslant n,$$

то, вводя еще матрицу

$$\mathscr{D} = (D_{ij}),\ D_{ij} = \langle 0,\ r_{ij} \rangle,\ 1 \leqslant i,\ j \leqslant n,$$

получим, что

$$\begin{aligned} \tilde{\mathscr{A}} = m(\mathscr{A})^{-1}\mathscr{A} &= m(\mathscr{A})^{-1}(m(\mathscr{A}) + \mathscr{D}) \\ &= \mathscr{I}_p + m(\mathscr{A})^{-1}\mathscr{D} = \mathscr{I}_p + |m(\mathscr{A})^{-1}|\mathscr{D} \\ &= \mathscr{I}_p + \mathscr{H}, \end{aligned}$$

где

$$\mathscr{H} = |m(\mathscr{A})^{-1}|\mathscr{D}.$$

Так как

$$\begin{aligned} \|\,|\mathscr{H}|\,\| &\leqslant \|\,|m(\mathscr{A})^{-1}|\,\| \cdot \|\,|\mathscr{D}|\,\| = \frac{1}{2}\|\,|m(\mathscr{A})^{-1}|\,\|\,\|d(\mathscr{A})\| \\ &\leqslant \frac{1}{2}\|\,|m(A)^{-1}|\,\| \cdot \|d(\mathscr{A})\|/\|m(\mathscr{A})\|, \end{aligned}$$

то мы видим, что при данной точечной матрице $m(\mathscr{A})$ диагональ матрицы $\mathscr{A}$ будет сильно доминировать с тем большей вероятностью, чем меньше число

$$\hat{\varkappa} = \|\,|m(\mathscr{A})^{-1}|\,\|\,\|m(\mathscr{A})\|$$

Если $m(\mathscr{A})$ — вещественная точечная матрица и мы используем монотонную матричную норму, то

$$\tilde{\varkappa} = \|m(\mathscr{A})^{-1}\|\,\|m(\mathscr{A})\|,$$

где $\tilde{\varkappa}$ —хорошо известная величина, обусловленность матрицы $m(\mathscr{A})$. Поэтому применимость метода Хансена к вещественной интервальной матрице зависит не только от величины $d(\mathscr{A})$, но (даже более существенным образом) и от обусловленности матрицы $m(\mathscr{A})$.

В предыдущем разделе мы описали метод преобразования множества линейных интервальных уравнений к верхней треугольной форме. Другие методы, разработанные в теории вещественных систем линейных уравнений, также могут быть обобщены на линейные системы интервальных уравнений. Мы упомянем метод Гаусса — Жордана, которым данная матрица приводится к диагональному виду. После этого вычисление решения исходной системы требует еще n добавочных делений. Подробное описание этого варианта алгоритма Гаусса можно найти в литературе. Тем же методом, что и в теореме 3 из п.3.5, можно показать, чго этот метод всегда применим к данной системе линейных интервальных уравнений, если ее матрица имеет сильно доминирующую диагональ.

Пусть дана вещественная интервальная матрица $\mathcal{A} = (A_{ij})$. Будем предполагать, что обратная матрица $\mathcal{A}_p^{-1}$ существует для любой $\mathcal{A}_p \in \mathcal{A}$. Пусть далее $\mathcal{b} = (B_i)$ — вещественный интервальный вектор.

Запишем в виде

$$\ell = (L_1,\ L_2,\ \ldots,\ L_n)^T$$

вещественный интервальный вектор с компонентами L_i, $1 \leqslant i \leqslant n$, которые получаются из множества

$$\mathfrak{L} = \{x_p \mid \mathcal{A}_p x_p = \mathcal{b}_p,\ \mathcal{A}_p \in \mathcal{A},\ \mathcal{b}_p \in \mathcal{b}\}$$

проектированием на соответствующие оси координат. Иными словами,

$$L_i = L_i(\mathcal{A},\ \mathcal{b}) = \{l_i \mid (l_1,\ \ldots,\ l_i,\ \ldots,\ l_n)^T \in \mathfrak{L}\},\ 1 \leqslant i \leqslant n.$$

ℓ — интервальный вектор наименьшей ширины, содержащий множество $\mathfrak{L}$.

Мы хотим теперь исследовать вопрос о том, насколько хорошо интервальные векторы, вычисляемые по методу Хансена, аппроксимируют вектор ℓ. Для «решения» преобразованной системы линейных уравнений мы используем тогда метод Гаусса — Жордана. Будет показано, что разность между шириной полученного вектора и шириной вектора ℓ стремится к нулю при стремлении к нулю ширины векторов $\mathcal{A}$ и $\mathcal{b}$. Это показывает, что метод Хансена дает вектор, достаточно близкий к вектору ℓ, если ширина исходных данных не слишком велика. Мы снова представляем вещественный интервал $A = [a_1, a_2]$ с помощью его середины $a = \frac{1}{2}(a_1 + a_2)$ и полуширины $r = \frac{1}{2} d(A) = \frac{1}{2}(a_2 - a_1)$:

$$A = \langle a,\ r \rangle.$$

Лемма 1. *Пусть* $\mathcal{A}_p x_p = \mathcal{b}_p$, *где* $\mathcal{A}_p = (a_{ij})$ — *вещественная невырожденная матрица, имеющая обратную* $\mathcal{A}_p^{-1} = \mathcal{B}_p = (b_{ij})$, *и пусть* $x_p = (x_i)$, $\mathcal{b}_p = (b_i)$ — *точечные векторы. Пусть далее* $\mathcal{A} = (A_{ij})$ — *интервальная матрица с элементами* $A_{ij} = \langle a_{ij},\ r_{ij} \rangle$, $1 \leqslant i,\ j \leqslant n$, *а* $\mathcal{b} = (B_i)$ — *интервальный вектор с элементами* $B_i = \langle b_i,\ r_i \rangle$, $1 \leqslant i \leqslant n$. *Тогда для k-й компоненты* L_k *введенного выше интервального вектора* $\ell = (L_i)$ *справедливо соотношение*

$$\frac{1}{2} d(L_k) = \sum_{i=1}^{n} \sum_{l=1}^{n} | b_{ki} x_j r_{lj} | + \sum_{i=1}^{n} | b_{ki} r_i | + O(d^2).$$

Здесь

$$d = \max \left\{ \max_{1 \leqslant l,\ i \leqslant n} \{r_{ij}\},\ \max_{1 \leqslant i \leqslant n} \{r_i\} \right\}.$$

Запись $O(d^2)$ *обозначает любую вещественную функцию f от d для которой*

$$| f/d^2 | \leqslant \gamma \quad при \quad d \leqslant d_0,$$

где $\gamma \geqslant 0,\ d_0 > 0$ *— константы.*

Доказательство В силу правила Крамера множество L_k является образом $(n^2 + n)$-мерного гиперкуба при отображении

$$x_k = x_k(\mathscr{A}_p,\ b_p).$$

Из теоремы о среднем значении мы получаем, что

$$x_k(\mathscr{A}_p,\ \hat{b}_p) = x_k(\mathscr{A}_p,\ b_p) + \sum_{i=1}^{n} \sum_{l=1}^{n} \frac{\partial x_k}{\partial a_{ij}} (\hat{a}_{ij} - a_{ij}) + \sum_{i=1}^{n} \frac{\partial x_k}{\partial b_i} (\hat{b}_i - b_i) + \frac{1}{2} x_k''(u_p + t(v_p - u_p))(v_p - u_p) \times (v_p - u_p).$$

Здесь $t \in (0,\ 1)$, через x_k'' обозначен гессиан отображения x_k, а u_p, v_p — это векторы из $V_{n^2+n}(\mathbb{R})$, причем значениями компонент вектора u_p (соответственно вектора v_p) являются элементы матрицы $\mathscr{A}_p$ и вектор b_p (соответственно элементы матрицы $\hat{\mathscr{A}}_p$ и вектор $\hat{b}_p$). Если теперь продифференцировать n уравнений

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad 1 \leqslant i \leqslant n$$

по a_{ij}, то, используя соотношение

$$\mathscr{A}_p \frac{\partial}{\partial a_{ij}} x_p = -x_j e_p^i,$$

мы получим уравнение

$$\frac{\partial}{\partial a_{ij}} x_p = \left(\frac{\partial x_1}{\partial a_{ij}},\ \frac{\partial x_2}{\partial a_{ij}},\ \ldots,\ \frac{\partial x_n}{\partial a_{ij}} \right)^{\mathrm{T}},$$

где e_p^i есть i-й единичный вектор. Отсюда мы получаем

$$\frac{\partial}{\partial a_{ij}} x_p = -x_j \mathscr{A}_p^{-1} e_p^i \text{ или } \frac{\partial x_k}{\partial a_{ij}} = -b_{ki} x_j.$$

Из равенства

$$x_p = \mathscr{B}_p \mathscr{b}_p$$

мы получаем

$$\frac{\partial x_k}{\partial b_l} = b_{kl}.$$

Если использовать эти формулы для производных в теореме о среднем значении, то мы получим

$$x_k(\hat{\mathscr{A}}_p,\ \hat{\mathscr{b}}_p) \in x_k(\mathscr{A}_p,\ \mathscr{b}_p) + \sum_{i=1}^{n}\sum_{j=1}^{n} |\, b_{kl} x_j \,| \langle 0,\ r_{ij} \rangle + \sum_{i=1}^{n} |\, b_{kl} \,| \langle 0,\ r_i \rangle + \ldots,$$

так как

$$\mathscr{A}_p \in \mathscr{A}, \quad \mathscr{b}_p \in \mathscr{b}.$$

Иными словами,

$$\frac{1}{2} d(L_k) = \sum_{i=1}^{n}\sum_{j=1}^{n} |\, b_{kl} x_j r_{ij} \,| + \sum_{i=1}^{n} |\, b_{kl} r_i \,| + O(d^2).$$

Лемма 2. *Пусть $\mathscr{A}_p = (a_{ij})$ — вещественная невырожденная матрица, для которой*

$$\mathscr{A}_p^{-1} = \mathscr{B}_p = (b_{ij}),$$

и пусть — $\mathscr{b}_p = (b_i)$ вещественный вектор. Пусть далее $\mathscr{A} = (A_{ij})$ — вещественная интервальная матрица с элементами $A_{ij} = \langle a_{ij},\ r_{ij} \rangle$, $1 \leqslant i,\ j \leqslant n$, и $\mathscr{b} = (B_i)$ — вещественный интервальный вектор с элементами $B_i = \langle b_i,\ r_i \rangle$, $1 \leqslant i \leqslant n$. Запишем в виде $\tilde{\mathscr{A}} = (A_{ij})$ интервальную матрицу $\mathscr{A}_p^{-1}\mathscr{A} = \mathscr{B}_p\mathscr{A}$, а в виде $\tilde{\mathscr{b}} = (B_i)$ — интервальный вектор $\mathscr{A}_p^{-1}\mathscr{b} = \mathscr{B}_p\mathscr{b}$. Предположим, что все точечные матрицы, принадлежащие $\tilde{\mathscr{A}}$, невырожденные. Тогда интервальный вектор

$$\tilde{\tilde{l}} = (\tilde{L}_1,\ \tilde{L}_2,\ \ldots,\ \tilde{L}_n)^T,$$

где

$$L_i = L_i(\tilde{\mathscr{A}},\ \tilde{\mathscr{b}}) = \{\tilde{l}_i \,|\, (\tilde{l}_1,\ \ldots,\ \tilde{l}_i,\ \ldots,\ \tilde{l}_n)^T \in \mathfrak{L}\}, \quad 1 \leqslant i \leqslant n,$$

$$\mathfrak{L} = \{\tilde{x}_p \,|\, \mathscr{A}_p = \tilde{\mathscr{b}}_p,\ \mathscr{A}_p \in \tilde{\mathscr{A}},\ \tilde{\mathscr{b}}_p \in \tilde{\mathscr{b}}\},$$

удовлетворяет соотношению

$$d(\tilde{L}_k) = d(L_k) + O(d^2), \quad 1 \leqslant k \leqslant n.$$

Доказательство. Пусть $\mathscr{A}_p x_p = \mathscr{b}_p$. Из леммы 1 следует, что

$$\frac{1}{2} d(L_k) = \sum_{i=1}^{n} \sum_{j=1}^{n} |b_{ki} x_j r_{ij}| + \sum_{i=1}^{n} |b_{ki} r_i| + O(d^2).$$

Элементы интервальной матрицы $\tilde{\mathscr{A}} = (\tilde{A}_{ij})$ имеют вид

$$\tilde{A}_{ij} = \left\langle \delta_{ij}, \sum_{m=1}^{n} |b_{im} r_{im}| \right\rangle, \quad 1 \leqslant i, j \leqslant n,$$

где δ_{ij} — символ Кронекера. Элементы интервального вектора $\tilde{b} = (\tilde{B}_i)$ имеют вид

$$\tilde{B}_i = \left\langle x_i, \sum_{m=1}^{n} |b_{im} r_m| \right\rangle, \quad 1 \leqslant i \leqslant n.$$

Применяя лемму 1 еще раз, мы получаем

$$\begin{aligned}
\frac{1}{2} d(\tilde{L}_k) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \delta_{ki} x_j \left(\sum_{m=1}^{n} |b_{im} r_{im}| \right) \right| \\
&\quad + \sum_{i=1}^{n} \left| \delta_{ki} \left(\sum_{m=1}^{n} |b_{im} r_m| \right) \right| + O(\tilde{d}^2) \\
&= \sum_{j=1}^{n} \left| x_j \left(\sum_{m=1}^{n} |b_{km} r_{mj}| \right) \right| + \sum_{m=1}^{n} |b_{km} r_m| + O(\tilde{d}^2) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} |b_{ki} x_j r_{ij}| + \sum_{i=1}^{n} |b_{ki} r_i| + O(\tilde{d}^2),
\end{aligned}$$

где

$$\tilde{d} = \max \left\{ \max_{1 \leqslant i, j \leqslant n} \{\tilde{r}_{ij}\}, \max_{1 \leqslant i \leqslant n} \{\tilde{r}_i\} \right\}.$$

Из формул для элементов матрицы $\tilde{\mathscr{A}}$ и вектора $\tilde{b}$ мы сразу усматриваем, что

$$\tilde{d} = O(d).$$

Сравнение с выражением для $\frac{1}{2} d(L_k)$ дает

$$d(\tilde{L}_k) = d(L_k) + O(d^2),$$

что и доказывает лемму.

Теперь метод Гаусса — Жордана за конечное число шагов порождает по данной интервальной матрице $\mathscr{A}$ некоторую диагональную матрицу. В результате каждого шага в новой матрице появляется хотя бы один новый нуль вне диагонали. Предположим, что даны интервальная матрица $\mathscr{H} = (H_{ij})$ и интервальный вектор $h = (H_i)$, такие что

$$H_{ij} = \langle h_{ij},\ e_{ij}\rangle, \quad 1 \leqslant i,\ j \leqslant n,$$
$$H_i = \langle h_i,\ e_i\rangle, \quad 1 \leqslant i \leqslant n,$$

где $h_{ij} = \delta_{ij}$, $h_i = x_i$
и

$$\max\left\{\max_{1 \leqslant i,\ j \leqslant n}\{e_{ij}\},\ \max_{1 \leqslant i \leqslant n}\{e_i\}\right\} = O(d).$$

Пусть, кроме того, $\mathscr{A} = (A_{ij})$, $\mathscr{b} = (b_{ij})$, где

$$A_{ij} = \langle a_{ij},\ r_{ij}\rangle, \quad 1 \leqslant i,\ j \leqslant n,$$
$$B_i = \langle b_i,\ r_i\rangle, \quad 1 \leqslant i \leqslant n.$$

Положим по определению

$$d = \max\left\{\max_{1 \leqslant i,\ j \leqslant n}\{r_{ij}\},\ \max_{1 \leqslant i \leqslant n}\{r_i\}\right\}.$$

Теперь мы можем доказать по индукции, что наши предположения о матрице $\mathscr{H}$ и векторе h верны на любом шаге алгоритма. Как показано в лемме 2, эти предположения справедливы для матрицы $\mathscr{H} := \mathscr{A} = \mathscr{A}_0^{-1}\mathscr{A}$ и интервального вектора $h := \mathscr{b} = \mathscr{A}_0^{-1}\mathscr{b}$. Кроме того, из леммы 2 следует, что

$$d(\tilde{L}_k) = d(L_k(\mathscr{H},\ h)) = d(L_k) + O(d^2), \quad 1 \leqslant k \leqslant n.$$

Чтобы получить нуль для пары $(r,\ s)$ индексов, (где $r \neq s$) по матрице $\mathscr{H}'$ и вектору h' строятся матрица $\mathscr{H}$ и вектор h согласно следующим формулам:

$$H'_{ij} = H_{ij}, \qquad i \neq r,$$
$$H'_{rj} = H_{rj} - H_{sj}H_{rs}/H_{ss}, \quad j \neq s,$$
$$H'_{rs} = 0,$$
$$H'_i = H_i, \qquad i \neq r,$$
$$H_r = H_r - H_sH_{rs}/H_{ss}.$$

Ввиду $h_{ij} = \delta_{ij}$ отсюда следует, что

$$\begin{aligned} H'_{rj} &= \langle h_{rj},\ e_{rj}\rangle - \langle h_{sj},\ e_{sj}\rangle\langle h_{rs},\ e_{rs}\rangle/\langle h_{ss},\ e_{ss}\rangle \\ &= \langle \delta_{rj},\ e_{rj}\rangle - \langle 0,\ e_{sj}\rangle\langle 0,\ e_{rs}\rangle/\langle h_{ss},\ e_{ss}\rangle \\ &= \langle \delta_{rj},\ e_{rj}\rangle - \langle 0,\ e_{sj}e_{rs}\rangle(1/|1(-e_{ss}))| \\ &= \langle \delta_{rj},\ e_{rj} + [e_{sj}e_{rs}/(1 - e_{ss})]\rangle \end{aligned}$$

и

$$\begin{aligned} H'_r &= \langle x_r,\ e\rangle - \langle x_s,\ e_s\rangle\langle 0,\ e_{rs}\rangle/\langle h_{ss},\ e_{ss}\rangle \\ &= \langle x_r,\ e_r + ((|x_s|e_{rs} + e_{rs}e_s)/(1 - e_{ss}))\rangle. \end{aligned}$$

Поэтому представление

$$H'_{ij} = \langle \delta_{ij},\ e'_{ij} \rangle, \quad H'_i = \langle x_i,\ e'_i \rangle$$

справедливо также для матрицы $\mathcal{H}' = (H'_{ij})$ и вектора $k' = (H'_i)$.

Из леммы 1 следует, что

$$\frac{1}{2} d(L_k(\mathcal{H},\ k)) = \sum_{i=1}^{n} \sum_{j=1}^{n} |\delta_{ki} x_j e_{ij}| + \sum_{i=1}^{n} |\delta_{ki} e_i| + O(d^2)$$

$$= \sum_{j=1}^{n} |x_j e_{kj}| + e_k + O(d^2).$$

Аналогично мы получаем

$$\frac{1}{2} d(L_k(\mathcal{H}',\ k')) = \sum_{j=1}^{n} |x_j e'_{kj}| + e'_k + O(d^2).$$

Для $k \neq r$ имеем $e'_{kj} = e_{kj}$ и $e'_k = e_k$. Из допущения

$$d(L_k(\mathcal{H},\ k)) = d(L_k) + O(d^2)$$

получаем поэтому

$$d(L_k(\mathcal{H}',\ k')) = d(L_k) + O(d^2), \quad k \neq r.$$

Для $k = r$ имеем

$$\sum_{j=1}^{n} |x_k e'_{kj}| = \sum_{j=1,\ j \neq s}^{n} |x_j e_{kj}| + O(d^2)$$

и

$$e'_k = e_k + |x_s e_{ks}| + O(d^2),$$

откуда

$$\frac{1}{2} d(L_k(\mathcal{H}',\ k')) = \sum_{j=1}^{n} |x_j e_{kj}| + e_k + O(d^2),$$

т. е.

$$d(L_k(\mathcal{H}',\ k')) = d(L_k) + O(d^2) \text{ для } k = r.$$

Этим завершается доказательство по индукции. Решение диагональной системы уравнений не увеличивает ширину элементов, поэтому соотношение

$$d(L_k(\widehat{\mathcal{H}},\ \widehat{k})) = d(L_k(\mathcal{A},\ b)) + O(d^2)$$

справедливо для заключительной диагональной матрицы $\mathcal{H}$ и соответствующего интервального вектора k.

2. Процедура Купермана и Хансена

Пусть $\mathscr{A}$ — интервальная матрица, такая, что $\mathscr{A}_p$ — невырожденная для всех $\mathscr{A}_p \in \mathscr{A}$, и пусть

$$\mathfrak{L} = \{x_p \mid \mathscr{A}_p x_p = \ell_p,\ \mathscr{A}_p \in \mathscr{A},\ \ell_p = \ell\}$$

— множество всех решений для данного интервального вектора ℓ Даже простые примеры показывают, что метод Хансена, описанный в предыдущем пункте, вычисляет в общем случае лишь некоторое подмножество векгора ℓ, имеющего компоненты

$$L_k = \{l_k \mid (l_1, \ \ldots, \ l_k, \ \ldots, \ l_n)^T \in \mathfrak{L}\}.$$

Куперман описал неинтервальную процедуру, которая в некоторых случаях дает лучшую локализацию множества $\mathfrak{L}$. Впоследствии Хансен обобщил эту процедуру до интервального метода.

Рассмотрим множество линейных уравнений

$$\mathscr{A}_p x_p = \ell_p, \quad x_p = (x_i),$$

где $\mathscr{A}_p$ — неособенная точечная матрица и ℓ_p — вещественный вектор. Если частная производная неотрицательна

$$\frac{\partial x_k}{\partial a_{ij}} \geqslant 0,$$

то x_k — неубывающая функция от a_{ij}. Если теперь a_{ij} может изменяться в вещественном интервале

$$A_{ij} = [a_{ij}^1, \ a_{ij}^2],$$

то величина x_k, рассматриваемая как функция от a_{ij} на интервале $A_{ij} = [a_{ij}^1, \ a_{ij}^2]$, принимает свое наименьшее значение при $a_{ij} = a_{ij}^1$, а наибольшее значение при $a_{ij} = a_{ij}^2$. То же самое можно сказать и о зависимости компонент x_k от ℓ_i.

Пусть теперь $\mathscr{A} = (A_{ij})$ — данная вещественная интервальная матрица и $\ell = (B_i)$ — данный интервальный вектор. Чтобы локализовать интервал

$$L_k = [l_k^1, \ l_k^2], \quad 1 \leqslant k \leqslant n,$$

поступаем следующим образом.

Начинаем с вещественной интервальной матрицы $\mathscr{A} = (A_{ij})$,

$A_{ij} = [a_{ij}^1, \ a_{ij}^2], \quad 1 \leqslant i, \ j \leqslant n,$ и вещественного интервального вектора $\ell = (B_i), \ B_i = [b_i^1, \ b_i^2], \ 1 \leqslant i \leqslant n$. Затем строим интервальные матрицы

$$\check{\mathscr{A}} = (\check{A}_{ij}) \ \text{и} \ \hat{\mathscr{A}} = (\hat{A}_{ij})$$

и интервальные векторы

$$\tilde{\mathscr{b}} = (\tilde{B}_l) \text{ и } \hat{\mathscr{b}} = (\hat{B}_l)$$

согласно следующим правилам:

$$\tilde{A}_{ij} = \begin{cases} [a_{ij}^1, a_{ij}^1], & \text{если } \partial x_k/\partial a_{ij} \geqslant 0 \text{ для всех } \mathscr{A}_p \in \mathscr{A} \text{ и } \mathscr{b}_p \in \mathscr{b}, \\ [a_{ij}^2, a_{ij}^2], & \text{если } \partial x_k/\partial a_{ij} \leqslant 0 \text{ для всех } \mathscr{A}_p \in \mathscr{A} \text{ и } \mathscr{b}_p \in \mathscr{b}, \\ A_{ij} & \text{в противном случае.} \end{cases}$$

$$\hat{A}_{ij} = \begin{cases} [a_{ij}^2, a_{ij}^2], & \text{если } \partial x_k/\partial a_{ij} \geqslant 0 \text{ для всех } \mathscr{A}_p \in \mathscr{A} \text{ и } \mathscr{b}_p \in \mathscr{b}, \\ [a_{ij}^1, a_{ij}^1], & \text{если } \partial x_k/\partial a_{ij} \leqslant 0 \text{ для всех } \mathscr{A}_p \in \mathscr{A} \text{ и } \mathscr{b}_p \in \mathscr{b}, \\ A_{ij} & \text{в противном случае.} \end{cases}$$

$$\tilde{B} = \begin{cases} [b_i^1, b_i^1], & \text{если } \partial x_k/\partial b_i \geqslant 0 \text{ для всех } \mathscr{A}_p \in \mathscr{A} \text{ и } \mathscr{b}_p \in \mathscr{b}, \\ [b_i^2, b_i^2], & \text{если } \partial x_k/\partial b_i \leqslant 0 \text{ для всех } \mathscr{A}_p \in \mathscr{A} \text{ и } \mathscr{b}_p \in \mathscr{b}, \\ B_i & \text{в противном случае.} \end{cases}$$

$$\hat{B}_i = \begin{cases} [b_i^2, b_i^2], & \text{если } \partial x_k/\partial b_i \geqslant 0 \text{ для всех } \mathscr{A}_p \in \mathscr{A} \text{ и } \mathscr{b}_p \in \mathscr{b}, \\ [b_i^1, b_i^1], & \text{если } \partial x_k/\partial b_i \leqslant 0 \text{ для всех } \mathscr{A}_p \in \mathscr{A} \text{ и } \mathscr{b}_p \in \mathscr{b}, \\ B_i & \text{в противном случае.} \end{cases}$$

Теперь вычисляем локализующие интервалы

$$X_k = [x_k^1, x_k^2] \quad \text{и} \quad Y_k = [y_k^1, y_k^2]$$

для интервалов

$$L_k(\tilde{\mathscr{A}}, \tilde{\mathscr{b}}) = [\tilde{l}_k^1, \tilde{l}_k^2] \quad \text{и} \quad L_k(\hat{\mathscr{A}}, \hat{\mathscr{b}}) = [\hat{l}_k^1, \hat{l}_k^2],$$

используя, например, метод Хансена, описанный в предыдущем пункте. Предыдущие рассуждения показывают, что для $L_k(\mathscr{A}, \mathscr{b}) = [l_k^1, l_k^2]$ имеет место

$$l_k^1 \geqslant \tilde{l}_k^1 \text{ и } l_k^2 \leqslant \hat{l}_k^2.$$

Отсюда следует, что верно

$$[\hat{l}_k^1, \hat{l}_k^2] \supseteq L_k(\mathscr{A}, \mathscr{b}).$$

Поэтому имеем также

$$[x_k^1, y_k^2] \supseteq [\tilde{l}_k^1, \hat{l}_k^2] \supseteq L_k(\mathscr{A}, \mathscr{b}).$$

Для построения интервальных матриц $\tilde{\mathscr{A}}$, $\hat{\mathscr{A}}$ и интервальных векторов $\tilde{\mathscr{b}}$, $\hat{\mathscr{b}}$ при фиксированном k нужны частные производные $\partial x_k/\partial a_{ij}$ и $\partial x_k/\partial b_i$. Формулы для них уже были получены в предыдущем пункте: полагая $\mathscr{A}_p^{-1} = (\bar{a}_{ij})$, имеем

$$\frac{\partial x_k}{\partial a_{ij}} = -\bar{a}_{ki} x_j, \quad \frac{\partial x_k}{\partial b_i} = \bar{a}_{ki}.$$

Чтобы выяснить распределение знаков этих производных, мы вычисляем, используя, например, метод Хансена, интервальный

вектор $\bar{\ell} = (\bar{L}_i)$, содержащий множество $\mathfrak{L}$, и интервальную матрицу $\mathscr{A} = (\bar{A}_{ij})$, содержащую обращения всех $\mathscr{A}_p \in \mathscr{A}$. Если теперь нижняя граница интервала $\bar{A}_{ki}\bar{L}_i$ неотрицательна для некоторого фиксированного k, то $\partial x_k/\partial a_{ij} \leqslant 0$ для всех $\mathscr{A}_p \in \mathscr{A}$ и $\ell_p \in \ell$. Аналогично, если верхняя граница интервала $\bar{A}_{ki}\bar{L}_i$ неположительна, то $\partial x_k/\partial a_{ij} \geqslant 0$. Соответствующее утверждение верно и для $\partial x_k/\partial b_i$. Если $0 \in \bar{A}_{ki}\bar{L}_i$ (соответственно $0 \in \bar{A}_{ki}$), то мы все еще можем иметь $\partial x_k/\partial a_{ij} \geqslant 0$ или ≤ 0 (соответственно $\partial x_k/\partial b_i \geqslant 0$ или ≤ 0), так как ℓ только содержит множество $\mathfrak{L}$, а $\bar{\mathscr{A}}$ только локализует множество обращений матриц $\mathscr{A}_p \in \mathscr{A}$. При построении матриц $\tilde{\mathscr{A}}$, $\hat{\mathscr{A}}$ и векторов $\tilde{\ell}$, $\hat{\ell}$ элементы матрицы $\mathscr{A}$ и вектора ℓ преобразуются в этом случае так, как будто $\partial x_k/\partial a_{ij}$ (соответственно $\partial x_k/\partial b_i$) меняет знак, потому, что вычисленные значения не позволяют принять иное решение. Этот метод дает, вообще говоря, гораздо лучшую локализацию, чем метод Хансена из п.1. Его недостаток —большой объем вычислений. В общем случае приходится не только вычислять интервальную матрицу $\bar{\mathscr{A}}$, содержащую обращения всех матриц $\mathscr{A}_p \in \mathscr{A}$, но и решать две системы интервальных уравнений для каждой компоненты $L_k(\mathscr{A}, \ell)$.

Замечания. Метод Купермана и Хансена применим и к итерационным методам. Можно искать метод, основанный на решении $2n$ систем уравнений методом Гаусса. Однако использование метода Хансена всегда дает лучшую локализацию.

3.7. Итерационные методы для локализации обратной матрицы и разложения на треугольные

Пусть даны невырожденная матрица $\mathscr{A}_p \in M_{nn}(\mathbb{R})$ размерности $n \times n$ и интервальная матрица $\mathscr{X}^{(0)} \in M_{nn}(I(\mathbb{R}))$, такая, что $\mathscr{X}^{(0)}$ локализует матрицу, обратную к $\mathscr{A}_p$, т. е. $\mathscr{A}_p^{-1} \in \mathscr{X}^{(0)}$.

Рассмотрим здесь процедуры, которые итерационно улучшают локализующую матрицу $\mathscr{X}^{(0)}$.

При этом будем использовать преобразование m, отображающее множество интервальных матриц размерности $n \times n$ во множестве вещественных точечных матриц размерности $n \times n$. Оно переводит

каждую интервальную матрицу в такую, элементами которой являются середины соответствующих элементов исходной матрицы. Иными словами, в обозначениях

$$i(X) = x_1,\ s(X) = x_2 \quad X = [x_1,\ x_2] \in I(\mathbb{R})$$

вводим отображение.

$$m:\ M_{nn}(I(\mathbb{R})) \to M_{nn}(\mathbb{R}) \tag{1}$$

равенствами $m(\mathscr{X}) = \frac{1}{2}(i(X_{ij}) + s(X_{ij}))$.

Это срединное отображение интервальных матриц очевидным образом непрерывно. Оно обладает следующими свойствами:

$$m(\mathscr{X} \pm \mathscr{Y}) = m(\mathscr{X}) \pm m(\mathscr{Y}), \quad \mathscr{X},\ \mathscr{Y} \in M_{nn}(I(\mathbb{R})), \tag{2}$$

$$m(\mathscr{B}_p\mathscr{X}) = \mathscr{B}_p m(\mathscr{X}),\ m(\mathscr{X}\mathscr{B}_p) = m(\mathscr{X})\mathscr{B}_p, \tag{3}$$

$$\mathscr{B}_p \in M_{nn}(\mathbb{R}),\ \mathscr{X} \in M_{nn}(I(\mathbb{R})),$$

$$m(\mathscr{B}_p) = \mathscr{B}_p, \quad \mathscr{B}_p \in M_{nn}(\mathbb{R}). \tag{4}$$

Вот краткое доказательство соотношения (3):

$$m(\mathscr{B}_p\mathscr{X}) = m\left(\sum_{k=1}^{n} b_{lk}X_{kj}\right) = \left(\sum_{k=1}^{n} m(b_{lk}[i(X_{kj}),\ s(X_{kj})])\right)$$

$$= \left(\sum_{k=1}^{n} b_{lk}\frac{1}{2}(i(X_{kj}) + s(X_{kj}))\right) = \left(\sum_{k=1}^{n} b_{lk}m(X_{kj})\right)$$

$$= \mathscr{B}_p m(\mathscr{X}).$$

Утверждения (2) и (4) могут быть доказаны аналогично. Теперь сформулируем первый метод, позволяющий вычислять последовательность локализаций для обратной матрицы $\mathscr{A}_p^{-1}$. Пусть $r \geq 1$ — фиксированное натуральное число.

Для произвольной точечной матрицы $\mathscr{B}_p$ положим

$$\mathscr{B}_p^{(0)} = \mathscr{I}_p \text{ (где } \mathscr{I}_p - \text{единичная матрица).}$$

Рассмотрим итерационную процедуру

$$\mathscr{X}^{(k+1)} = m(\mathscr{X}^{(k)}) \sum_{\nu=0}^{r-2} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^\nu + \mathscr{X}^{(k)}(\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1}, \quad k \geq 0. \tag{5}$$

В случае $r = 2$ получаем формулу

$$\mathscr{X}^{(k+1)} = m(\mathscr{X}^{(k)}) + \mathscr{X}^{(k)}(\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)})), \quad k \geq 0,$$

которую можно считать интервальным вариантом метода Шульца для вычисления обратной матрицы.

Свойства итерационного метода (5) собраны в следующем утверждении.

Теорема 1. *Пусть* $\mathscr{A}_p$ *— невырожденная матрица размерности* $n \times n$ *и* $\mathscr{X}^{(0)}$ *— интервальная матрица той же размерности, такая что* $\mathscr{A}_p^{-1} \in \mathscr{X}^{(0)}$. *Пусть последовательность* $\{\mathscr{X}^{(k)}\}_{k=0}^{\infty}$ *интервальных матриц вычисляется по формулам* (5). *Тогда*

каждое приближение $\mathscr{X}^{(k)}$, $k \geqslant 0$, *содержит* $\mathscr{A}_p^{-1}$; (6)

последовательность $\{\mathscr{X}^{(k)}\}_{k=0}^{\infty}$ *сходится к* $\mathscr{A}_p^{-1}$ *тогда и* (7)

только тогда, когда спектральный радиус $\rho(\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)}))$ *меньше* 1;

для матричной нормы $\|\cdot\|$ *последовательность* $\{d(\mathscr{X}^k)\}_{k=0}^{\infty}$ (8)

удовлетворяет условию $\|d(\mathscr{X}^{(k+1)})\| \leqslant \gamma \|d(\mathscr{X}^{(k)})\|^r$, $\gamma \geqslant 0$, *т. е. R-порядок метода* (5) *удовлетворяет неравенству* $O_R((5), \mathscr{A}^{-1}) \geqslant r$ *(см. приложение А, теорема* 2).

Доказательство. (6): Для произвольной матрицы $m(\mathscr{X}^{(k)}) \in M_{nn}(\mathbb{R})$ легко проверить соотношение

$$m(\mathscr{X}^{(k)}) \sum_{\nu=0}^{r-2} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{\nu} = \mathscr{A}_p^{-1} - \mathscr{A}_p^{-1} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1}.$$

Для $k = 0$ утверждение (6) верно в силу условия теоремы.

Допустим теперь, что $\mathscr{A}_p^{-1} \in \mathscr{X}^{(k)}$. Используя только что полученное равенство и соотношения (10' из п.2.3), получаем

$$\begin{aligned}\mathscr{A}_p^{-1} &= m(\mathscr{X}^{(k)}) \sum_{\nu=0}^{r-2} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{\nu} + \mathscr{A}^{-1} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1} \\ &\in m(\mathscr{X}^{(k)}) \sum_{\nu=0}^{r-2} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{\nu} + \mathscr{X}^{(k)} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1} \\ &= \mathscr{X}^{(k+1)}.\end{aligned}$$

Этим завершается доказательство соотношения (6) методом математической индукции.

(7): Используя равенства (2)—(4) для срединного отображения, участвующего в формулах (5), получаем для последовательности $\{m(\mathscr{X}^{(k)})\}_{k=0}^{\infty}$ следующую рекуррентную формулу:

$$m(\mathscr{X}^{(k+1)}) = m(\mathscr{X}^{(k)}) \sum_{\nu=0}^{r-1} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{\nu}.$$

Это — обобщение итерационной процедуры Шульца. Умножая обе части этого равенства на $\mathscr{A}_p$, получаем

$$\mathscr{A}_p m(\mathscr{X}^{(k+1)}) = (\mathscr{I}_p - (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))) \sum_{\nu=0}^{r-1} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^\nu$$
$$= \mathscr{I}_p - (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^r$$

или

$$\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k+1)}) = (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^r$$
$$= (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)}))^{r^{(k+1)}}.$$

Отсюда следует, что

$$\lim_{k\to\infty} m(\mathscr{X}^{(k)}) = \mathscr{A}_p^{-1} \Leftrightarrow \lim_{k\to\infty} (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)}))^k = \mathscr{O}_p$$
$$\Leftrightarrow \rho(\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})) < 1.$$

Теперь мы покажем, что последовательность $\{\mathscr{X}^{(k)}\}_{k=0}^{\infty}$ сходится к $\mathscr{A}_p^{-1}$ тогда и только тогда, когда последовательность $\{m(\mathscr{X}^{(k)})\}_{k=0}^{\infty}$ срединных матриц сходится к $\mathscr{A}_p^{-1}$. Действительно, рассмотрим последовательность $\{d(\mathscr{X}^{(k)})\}_{k=0}^{\infty}$, которая в силу (12 из п.2.3) и (19 из п.2.3) удовлетворяет рекуррентному соотношению

$$d(\mathscr{X}^{(k+1)}) = d(\mathscr{X}^{(k)}) \left| (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1} \right|.$$

Если мы имеем теперь $\lim_{k\to\infty} m(\mathscr{X}^{(k)}) = \mathscr{A}_p^{-1}$, то из последнего соотношения следует, что

$$\lim_{k\to\infty} d(\mathscr{X}^{(k)}) = \mathscr{O}_p.$$

С другой стороны, из непрерывности отображенияm и равенства (4) сразу получается, что

$$\lim_{k\to\infty} \mathscr{X}^{(k)} = \mathscr{A}_p^{-1}$$

влечет за собой $\lim_{k\to\infty} m(\mathscr{X}^{(k)}) = \mathscr{A}_p^{-1}$.

Так как выше уже было показано, что условие

$$\rho(\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})) < 1$$

необходимо и достаточно для сходимости последовательности $\{m(\mathscr{X}^{(k)})\}_{k=0}^{\infty}$, мы получаем (7).

(8): Имеем

$$\begin{aligned} d\left(\mathscr{X}^{(k+1)}\right) &= d\left(\mathscr{X}^{(k)}\right)\left|\left(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(k)}\right)\right)^{r-1}\right| \\ &= d\left(\mathscr{X}^{(k)}\right)\left|\left(\mathscr{A}_{\mathrm{p}}\mathscr{A}_{\mathrm{p}}^{-1} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(k)}\right)\right)^{r-1}\right| \\ &\leqslant d\left(\mathscr{X}^{(k)}\right)\left(\left|\mathscr{A}_{\mathrm{p}}\right|\left|\mathscr{A}_{\mathrm{p}}^{-1} - m\left(\mathscr{X}^{(k)}\right)\right|\right)^{r-1} \\ &\leqslant d\left(\mathscr{X}^{(k)}\right) 2^{-(r-1)}\left(\left|\mathscr{A}_{\mathrm{p}}\right| d\left(\mathscr{X}^{(k)}\right)\right)^{r-1}. \end{aligned}$$

Мы используем монотонную и мультипликативную матричную норму $\|\cdot\|'$, поэтому из только что доказанного соотношения следует

$$\left\| d\left(\mathscr{X}^{(k+1)}\right)\right\|' \leqslant 2^{-(r-1)}\left\|\mathscr{A}_{\mathrm{p}}\right\|^{(r-1)}\left\| d\left(\mathscr{X}^{(k)}\right)\right\|'^{r}.$$

Неравенство

$$\|\mathscr{B}_{\mathrm{p}}\|\gamma_1 \leqslant \|\mathscr{B}_{\mathrm{p}}\|' \leqslant \gamma_2\|\mathscr{B}_{\mathrm{p}}\|, \quad \gamma_1 > 0,\ \gamma_2 > 0,$$

верно для любой матричной нормы $\|\cdot\|$. Из этого неравенства следует

$$\left\| d\left(\mathscr{X}^{(k+1)}\right)\right\|\gamma_1 \leqslant 2^{-(r-1)}\gamma_2^{-(r-1)}\left\|\mathscr{A}_{\mathrm{p}}\right\|^{r-1}\gamma_2^{n}\left\| d\left(\mathscr{X}^{(k)}\right)\right\|^{r},$$

что и доказывает (8).

Из доказательства видно, что сходимость имеет место для произвольной матрицы $\mathscr{X}^{(0)}$, не обязательно содержащей $\mathscr{A}_{\mathrm{p}}^{-1}$.

В этом случае, однако, последовательные приближения не обязаны содержать $\mathscr{A}_{\mathrm{p}}^{-1}$. Отметим, что критерий (7) зависел не от всей локализующей матрицы $\mathscr{X}_{\mathrm{p}}^{(0)}$, а только от ее срединной матрицы $m\left(\mathscr{X}^{(0)}\right)$. При этом ширина $a\left(\mathscr{X}^{(0)}\right)$ может быть произвольной. Это значит, что имея подходящую аппроксимацию $m\left(\mathscr{X}^{(0)}\right)$ матрицы $\mathscr{A}_{\mathrm{p}}^{-1}$, удовлетворяющую неравенству $\rho\left(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(0)}\right)\right) < 1$, с помощью определенных оценок понорме всегда можно построить интервальную матрицу $\mathscr{X}^{(0)}$, такую что $\mathscr{A}_{\mathrm{p}}^{-1} \in \mathscr{X}^{(0)}$. Тогда последовательные приближения, полученные согласно (5), сходятся к $\mathscr{A}_{\mathrm{p}}^{-1}$ по теореме 1.

Так как последовательные приближения из (5) всегда содержат $\mathscr{A}_{\mathrm{p}}^{-1}$ в силу (6), кажется естественным брать пересечение следующего приближения с предыдущим и продолжать итерационный процесс с этим новым потенциально улучшенным приближением. Это приводит к следующей итерационной процедуре:

$$\begin{cases} \mathscr{Y}^{(k+1)} = m\left(\mathscr{X}^{(k)}\right)\sum\limits_{\nu=0}^{r-2}\left(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(k)}\right)\right)^{\nu} \\ \qquad + \mathscr{X}^{(k)}\left(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m\left(\mathscr{X}^{(k)}\right)\right)^{r-1}, \\ \mathscr{X}^{(k+1)} = \mathscr{Y}^{(k+1)} \cap \mathscr{X}^{(k)}, \quad k \geqslant 0. \end{cases} \tag{9}$$

Применяя эту итерационную процедуру, получаем монотонную последовательность $\mathscr{X}^{(0)} \supseteq \mathscr{X}^{(1)} \supseteq \mathscr{X}^{(2)} \supseteq \ldots$ локализаций для матрицы $\mathscr{A}_{\mathrm{p}}^{-1}$. Следующий численный пример показывает, что в этом случае критерий (7), вообще говоря, не достаточен для сходимости.

Возьмем $r = 2$ и положим

$$\mathscr{A}_{\mathrm{p}} = \begin{pmatrix} 0.4 & 0.6 \\ -0.6 & 0.4 \end{pmatrix}, \quad \mathscr{X}^{(0)} = \begin{pmatrix} [-2, 4] & [-3, 3] \\ [-3, 3] & [-2, 4] \end{pmatrix},$$

откуда следует, что $m(\mathscr{X}^{(0)}) = \mathscr{I}_{\mathrm{p}}$. Мы получаем

$$\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(0)}) = \begin{pmatrix} 0.6 & -0.6 \\ 0.6 & 0.6 \end{pmatrix},$$

откуда

$$\rho(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(0)})) < 1.$$

Поэтому процедура (5) с этим начальным приближением сходится к $\mathscr{A}_{\mathrm{p}}^{-1}$. Используя (9), получаем

$$\mathscr{Y}^{(1)} = m(\mathscr{X}^{(0)}) + \mathscr{X}^{(0)}(\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(0)})) = \begin{pmatrix} [-2, 5.2] & [-4, 2.3] \\ [-3, 4.2] & [-2, 5.2] \end{pmatrix},$$

откуда следует, что $\mathscr{X}^{(1)} = \mathscr{X}^{(0)}$. Поэтому последовательность приближений, вычисленная по формулам (9), не сходится к $\mathscr{A}_{\mathrm{p}}^{-1}$ в противоположность последовательности, вычисленной по формулам (5). Условие сходимости для итерации (9) содержится в следующей теореме.

Теорема 2. *Пусть* $\mathscr{A}_{\mathrm{p}}$ *— невырожденная матрица размерности* $n \times n$, *а* $\mathscr{X}^{(0)}$ *— интервальная матрица размерности* $n \times n$, *такая что* $\mathscr{A}_{\mathrm{p}}^{-1} \in \mathscr{X}^{(0)}$. *Тогда*

каждое приближение $\mathscr{X}^{(k)}$, $k \geqslant 0$, *содержит* $\mathscr{A}_{\mathrm{p}}^{-1}$; (6')

если неравенство $\rho(|\mathscr{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} \mathscr{X}_{\mathrm{p}}|) < 1$ *выполнено для всех* $\mathscr{X}_{\mathrm{p}} \in \mathscr{X}^{(0)}$, (10)

то последовательность $\{\mathscr{X}^{k}\}_{k=0}^{\infty}$ *сходится к* $\mathscr{A}_{\mathrm{p}}^{-1}$;

последовательность $\{d(\mathscr{X}^{(k)})\}_{k=0}^{\infty}$ *может быть следующим образом ограничена в матричной норме* $\|\cdot\|$: (8')

$$\|d(\mathscr{X}^{(k+1)})\| \leqslant \gamma' \|d(\mathscr{X}^{(k)})\|^{r}, \quad \gamma \geqslant 0,$$

т. е. R-порядок итерационной процедуры (9) *удовлетворяет неравенству* $O_R((9), \mathscr{A}_{\mathrm{p}}^{-1}) \geqslant r$ *(см. приложение* А, *теорема* 2).

Доказательство. (6'): Как и в доказательстве утверждения (6), мы устанавливаем сначала, что $\mathscr{A}_{\mathrm{p}}^{-1} \in \mathscr{Y}^{(k+1)}$, откуда ввиду

$\mathscr{A}_{\mathrm{p}}^{-1} \in \mathscr{X}^{(k)}$ немедленно следует $\mathscr{A}_{\mathrm{p}}^{-1} \in \mathscr{X}^{(k+1)}$.

(10): В силу следствия 8 из п.2.3 последовательные приближения

$$\mathscr{X}^{(0)} \supseteq \mathscr{X}^{(1)} \supseteq \mathscr{X}^{(2)} \supseteq \ldots$$

всегда сходятся к некоторой интервальной матрице $\mathscr{X}$. Теперь покажем, что в условиях нашей теоремы выполнено равен-ство $d(\mathscr{X}) = \mathcal{O}_{\mathrm{p}}$. Положив

$$\mathscr{Y} = m(\mathscr{X}) \sum_{\nu=0}^{r-2} (\mathcal{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}))^{\nu} + \mathscr{X} (\mathcal{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}))^{r-1},$$

получаем $\mathscr{X} = (X_{ij} \cap Y_{ij}) \subseteq \mathscr{Y}$ в силу (9). Используя (11 из п.2.3), получим $d(\mathscr{X}) \leqslant d(\mathscr{Y})$. Для $d(\mathscr{X})$ имеем из (9) соотношение

$$d(\mathscr{X}) |\mathcal{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X})|^{r-1} \geqslant d(\mathscr{X}) |(\mathcal{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}))^{r-1}| = d(\mathscr{Y}) \geqslant d(\mathscr{X}),$$

откуда следует, что

$$d(\mathscr{X})(\mathcal{I}_{\mathrm{p}} - |\mathcal{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X})|^{r-1}) \leqslant \mathcal{O}_{\mathrm{p}}.$$

Из условия $\rho(|\mathcal{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X})|) < 1$ следует существование матрицы $(\mathcal{I}_{\mathrm{p}} - |\mathcal{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X})|^{r-1})^{-1}$. Эта обратная матрица также неотрицательна. Отсюда следует, что $d(\mathscr{X}) \leqslant \mathcal{O}_{\mathrm{p}}$, т. е. $d(\mathscr{X}) = \mathcal{O}_{\mathrm{p}}$. Ввиду (6') мы имеем поэтому $\mathscr{X} = \mathscr{A}_{\mathrm{p}}^{-1}$.

(8'): Как и в доказательстве утверждения (8), мы показываем сначала, что для монотонной и мультипликативной матричной нормы $\|\cdot\|'$ имеет место неравенство

$$\|d(\mathscr{Y}^{(k+1)})\|' \leqslant \gamma \|d(\mathscr{X}^{(k)})\|'^{r}.$$

Отсюда с помощью (11 из п.2.3), монотонности нормы $\|\cdot\|'$ и включения $\mathscr{X}^{(k+1)} \subseteq \mathscr{Y}^{(k+1)}$ следует неравенство

$$\|d(\mathscr{X}^{k+1})\|' \leqslant \|d(\mathscr{Y}^{(k+1)})\|' \leqslant \gamma \|d(\mathscr{X}^{(k)})\|'^{r}.$$

Так же как и в доказательстве утверждения (8), мы используем теперь теорему об эквивалентности норм для доказательства утверждения (8).

В отличие от критерия (7) условие сходимости (10) зависит от ширины матрицы $\mathscr{X}^{(0)}$, локализующей $\mathscr{A}_{\mathrm{p}}^{-1}$. Эту зависимость легко охарактеризовать формулами. Если, например, матрица $\mathscr{X}^{(0)}$ удовлетворяет для монотонной мультипликативной нормы $\|\cdot\|$ неравенству $\|\mathcal{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(0)})\| < 1$, то условие

$$\|d(\mathscr{X}^{(0)})\| < 2(1 - \|\mathcal{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} m(\mathscr{X}^{(0)})\|)/\|\mathscr{A}_{\mathrm{p}}\| \qquad (11)$$

достаточно для того, чтобы $\|\mathcal{I}_{\mathrm{p}} - \mathscr{A}_{\mathrm{p}} \mathscr{X}\| < 1$ было верно для

всех $\mathscr{X}_p \in \mathscr{X}^{(0)}$. Рассмотрим теперь вкратце вопрос о нахождении подходящей интервальной матрицы $\mathscr{X}^{(0)}$. Допустим, что $\mathscr{A}_p$ можно представить в виде

$$\mathscr{A}_p = \mathscr{T}_p - \mathscr{B}_p, \text{ где } \|\mathscr{B}_p\| < 1.$$

При $m(\mathscr{X}^{(0)}) := \mathscr{T}_p$ мы имеем

$$\|\mathscr{T}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})\| = \|\mathscr{B}_p\| < 1,$$

так что последовательность (5) сходится в силу критерия (7) для любой интервальной матрицы $\mathscr{X}^{(0)}$, для которой $m(\mathscr{X}^{(0)}) = \mathscr{T}_p$ Чтобы обеспечить соотношение $\mathscr{A}_p^{-1} \in \mathscr{X}^{(1)}$, рассмотрим равенство

$$\mathscr{A}_p \mathscr{X}_p = (\mathscr{T}_p - \mathscr{B}_p)\mathscr{X}_p = \mathscr{T}_p$$

или

$$\mathscr{X}_p = \mathscr{B}_p \mathscr{X}_p + \mathscr{T}_p.$$

Из него следует в силу мультипликативности матричной нормы $\|\cdot\|$, что

$$\|\mathscr{X}_p\| \leqslant a := 1/(1 - \|\mathscr{B}_p\|).$$

Если теперь мы используем норму, задаваемую суммами по столбцам или суммами по строкам, то получим

$$-a \leqslant x_{ij} \leqslant a, \quad 1 \leqslant i, \ j \leqslant n,$$

для элементов матрицы $\mathscr{X}_p = (x_{ij})$. Для матрицы $\mathscr{X}^{(0)}$ с элементами

$$X_{ij}^{(0)} = \begin{cases} [-a, a] & \text{для } i \neq j, \\ [-a, 2+a] & \text{для } i = j \end{cases}$$

имеем $\mathscr{A}_p^{-1} \in \mathscr{X}^{(0)}$ и $m(\mathscr{X}^{(0)}) = \mathscr{T}_p$. Поэтому итерационный метод (5) сходится к $\mathscr{A}_p^{-1}$ в силу теоремы 1. Если теперь неравенство (11) выполняется после некоторого шага итерации, то процесс можно продолжать дальше по формулам (9).

При практическом выполнении алгоритма (5) встречающиеся выражения вычисляются по аналогии со схемой Горнера.

Это дает формулу

$$\mathscr{X}^{(k+1)} = \ldots (\mathscr{X}^{(k)} \mathscr{T}_p^{(k)} + m(\mathscr{X}^{(k)}))\mathscr{T}_p^{(k)} + m(\mathscr{X}^{(k)})\mathscr{T}_p^{(k)} \ldots)\mathscr{T}_p^{(k)} + m(\mathscr{X}^{(k)}), \tag{5'}$$

где

$$\mathscr{T}_p^{(k)} = \mathscr{T}_p - \mathscr{A}_p m(\mathscr{X}_p^{(k)}).$$

Так как умножение матриц становится неассоциативным при появлении интервальных матриц, мы имеем в общем случае

$$\mathscr{X}^{(k)}(\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)}))^{r-1} \neq (\ldots(\mathscr{X}^{(k)}(\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)})))\ldots) \times (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)})).$$

Даже при одной и той же начальной матрице формулы (5) и (5') порождают в общем случае разные последовательности. Однако теорема 1 все же верна и для итераций (5'). Рассмотрим теперь объем вычислений, нужных на каждом шаге (5').

Если $\mathscr{A}_p$ — матрица размерности $n \times n$, то (5') требует на каждом шаге

$$rn^3 \text{ умножений и } rn^3 - n^2 + n \text{ сложений.}$$

Даже те члены в (5'), которые, как $\mathscr{T}_p^{(k)}$, не содержат ничего интервального, приходится вычислять в интервальной арифметике, чтобы обеспечить локализацию матрицы $\mathscr{A}_p^{-1}$. Если пренебречь более низкими степенями n, то мы увидим, что объем вычислений по алгоритму (5') пропорционален r.

Теперь мы хотим оценить число k шагов итерации, которые требуются, чтобы, исходя из данного $\mathscr{X}^{(0)}$, достичь величины

$$\|d(\mathscr{X}^{(k)})\|,$$

меньшей, чем заранее предписанная погрешность.

Так же, как в доказательстве теоремы 1, мы получаем для (5') следующее соотношение:

$$d(\mathscr{X}^{(k+1)} = d(\mathscr{X}^{(k)}) | (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(k)})) |^{r-1} \leqslant d(\mathscr{X}^{(k)}) | (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})) |^{r^k(r-1)}$$

т. е.

$$d(\mathscr{X}^{(k+1)}) \leqslant d(\mathscr{X}^{(0)}) \prod_{\nu=0}^{k} | (\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})) |^{r^\nu(r-1)}.$$

Если по-прежнему мы используем монотонную и мультипликативную матричную норму $\|\cdot\|$ и допустим, что $\|\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})\| < 1$, то получим, что

$$\|d(\mathscr{X}^{(k+1)})\| \leqslant \|d(\mathscr{X}^{(0)})\| \left(\prod_{\nu=0}^{k} \|\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})\|^{r^\nu} \right)^{r-1} = \|d(\mathscr{X}^{(0)})\| \|\mathscr{I}_p - \mathscr{A}_p m(\mathscr{X}^{(0)})\|^{r^{k+1}-1}.$$

Это выражение позволяет нам оценить $\bar{k}$ при сделанных предположениях

Исходя из соотношения

$$\| d(\mathscr{X}^{(k)}) \| \leqslant \| d(\mathscr{X}^{(0)}) \| \| \mathscr{I}_p - \mathscr{A}_p m(\mathscr{Y}^{(0)}) \|^{k-1},$$

мы определим, при каком значении r итерационный метод требует наименьшего объема вычислений для достижения заданной точности для $\| d(\mathscr{X}^{(k)}) \|$. Согласно предшествующим рассуждениям, этот объем вычислений можно считать пропорциональным величине r. Пусть теперь даны $r^{(1)}>1$ и $r^{(2)}>1$, причем $r^{(1)} \neq r^{(2)}$. После $p^{(1)}$ (соответственно $p^{(2)}$) шагов итерации (5') со значением $r = r^{(1)}$ (соответственно $r = r^{(2)}$) при одном и том же начальном значении $\mathscr{X}^{(0)}$ мы выполним один и тот же объем вычислений. Иными словами, имеем

$$r^{(1)} p^{(1)} = r^{(2)} p^{(2)}.$$

Точность, достигнутую при использовании этих методов, можно оценить величиной $(r^{(1)})^{p^{(1)}} - 1$ (соответственно $(r^{(2)})^{p^{(2)}} - 1$).

Мы требуем от «оптимальной» итерационной процедуры $r = r^{(1)}$ чтобы для всех других значений $r^{(2)}$ и количества шагов $p^{(1)}$, $p^{(2)}$ мы имели бы

$$(r^{(1)})^{p^{(1)}} > (r^{(2)})^{p^{(2)}},$$

что ввиду $p^{(2)} = r^{(1)} p^{(1)} / r^{(2)}$ эквивалентно неравенству

$$(r^{(1)})^{1/r^{(1)}} > (r^{(2)})^{1/r^{(2)}}.$$

Так как функция $x^{1/x}$ для начальных x имеет максимум при x=3, получаем, что итерация (5') оптимальна в описанном смысле при этом значении.

Заметим еще, что метод (5) можно применять и для комплексных матриц, используя арифметику в $R(\mathbb{C})$. При этом будет верна теорема 1. Мы упомянем в этой связи более общие исследования. Рассмотрим теперь методы монотонной локализации обратной матрицы, обладающие свойствами, похожими на свойства метода (9). Основные вычисления в них вообще не используют интервальной арифметики. Верхняя и нижняя границы вычисляются по отдельным формулам. Этот метод, однако, применим лишь в случае, когда $\mathscr{A}_p^{-1} \geqslant \mathscr{O}_p$.

Итак, пусть $\mathscr{A}_p$ —невырожденная матрица и $r \geq 2$ — натуральное число. Рассмотрим итерационную процедуру

$$\begin{cases} \mathscr{X}_p^{(k+1)} = \mathscr{X}_p^{(k)} + (\mathscr{I}_p - \mathscr{X}_p^{(k)} \mathscr{A}_p) \sum\limits_{v=0}^{r-2} (\mathscr{I}_p - \mathscr{X}_p^{(k)} \mathscr{A}_p)^v \mathscr{X}_p^{(k)}, \\ \mathscr{Y}_p^{(k+1)} = \mathscr{Y}_p^{(k)} + (\mathscr{I}_p - \mathscr{Y}_p^{(k)} \mathscr{A}_p) \sum\limits_{v=0}^{r-2} (\mathscr{I}_p - \mathscr{X}_p^{(k)} \mathscr{A}_p)^v \mathscr{X}_p^{(k)}, \\ k \geqslant 0. \end{cases} \tag{12}$$

с заданными $\mathscr{X}_p^{(0)}$, $\mathscr{Y}_p^{(0)}$.

Исполняя эту процедуру, мы получим две последовательности точечных матриц, для которых верно следующее утверждение.

Теорема 3. *Пусть* $\mathcal{A}_p$ *— невырожденная матрица размерности* $n \times n$, *причем* $\mathcal{A}_p^{-1} \geqslant \mathcal{O}_p$. *Пусть далее* $\mathcal{X}_p^{(0)}$, $\mathcal{Y}_p^{(0)}$ *— две матрицы размерности* $n \times n$, *для которых верно*

$$\mathcal{X}_p^{(0)} \geqslant \mathcal{O}_p \text{ и } \mathcal{X}_p^{(0)} \mathcal{A}_p \leqslant \mathcal{I}_p \leqslant \mathcal{Y}_p^{(0)} \mathcal{A}_p.$$

Пусть последовательности $\{\mathcal{X}_p^{(k)}\}_{k=0}^{\infty}$ и $\{\mathcal{Y}_p^{(k)}\}_{k=0}^{\infty}$ *вычислены по формулам* (12). *Тогда верны следующие утверждения:*

$$\mathcal{O}_p \leqslant \mathcal{X}_p^{(0)} \leqslant \ldots \leqslant \mathcal{X}_p^{(k)} \leqslant \mathcal{X}_p^{(k+1)} \leqslant \ldots \leqslant \mathcal{A}_p^{-1} \leqslant \ldots \leqslant \mathcal{Y}_p^{(k+1)} \leqslant \mathcal{Y}_p^{(k)} \leqslant \ldots \leqslant \mathcal{Y}_p^{(0)}. \quad (13)$$

(7') *Обе последовательности*

$$\{\mathcal{X}_p^{(k)}\}_{k=0}^{\infty}, \quad \{\mathcal{Y}^{(k)}\}_{k=0}^{\infty}$$

сходятся к $\mathcal{A}_p^{-1}$ *тогда и только тогда, когда спектральный радиус* $\rho(\mathcal{I}_p - \mathcal{X}_p^{(0)} \mathcal{A}_p)$ *меньше* 1.

Если процедура сходится, то величины

$$d^{(k)} = \| \mathcal{Y}_p^{(k)} - \mathcal{X}_p^{(k)} \|$$

удовлетворяют соотношению (14) $d^{(k+1)} \leqslant \gamma (d^{(k)})^r, \quad \gamma \geqslant 0.$

Поэтому, если понимать (12) *как метод итераций для вычисления интервальных матриц* $([x_{ij}^{(k)}, y_{ij}^{(k)}])$, *то верно*

$$O_R((12), \mathcal{A}_p^{-1}) \geqslant r$$

(см. приложение А, *теорема* 2).

Доказательство. (13): Докажем соотношение

$$\mathcal{O}_p \leqslant \mathcal{X}_p^{(0)} \leqslant \ldots \leqslant \mathcal{X}_p^{(k-1)} \leqslant \mathcal{X}_p^{(k)},$$
$$\mathcal{Y}_p^{(k)} \leqslant \mathcal{Y}_p^{(k-1)} \leqslant \ldots \leqslant \mathcal{Y}_p^{(0)},$$
$$\mathcal{X}_p^{(k)} \mathcal{A}_p \leqslant \mathcal{I}_p \leqslant \mathcal{Y}_p^{(k)} \mathcal{A}_p$$

математической индукцией по $k \geq 0$. Эти неравенства выполнены для $k = 0$ по условию теоремы. Из

$$\mathcal{I}_p - \mathcal{Y}_p^{(k)} \mathcal{A}_p \leqslant \mathcal{O}_p \leqslant \mathcal{I}_p - \mathcal{X}_p^{(k)} \mathcal{A}_p, \quad \mathcal{X}_p^{(k)} \geqslant \mathcal{O}_p$$

следует, что

$$(\mathcal{I}_p - \mathcal{Y}_p^{(k)}\mathcal{A}_p)\sum_{\nu=0}^{r-2}(\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^\nu \mathcal{X}_p^{(k)}$$
$$\leqslant \mathcal{O}_p \leqslant (\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)\sum_{\nu=0}^{r-2}(\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^\nu \mathcal{X}_p^{(k)},$$

т. е

$$\mathcal{X}_p^{(k)} \leqslant \mathcal{X}_p^{(k+1)}, \quad \mathcal{Y}_p^{(k+1)} \leqslant \mathcal{Y}_p^{(k)}.$$

Из

$$\mathcal{I}_p - \mathcal{X}_p^{(k+1)}\mathcal{A}_p = \mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p - (\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)$$
$$\times \sum_{\nu=0}^{r-2}(\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^\nu \mathcal{X}_p^{(k)}\mathcal{A}_p = (\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^r \geqslant \mathcal{O}_p$$

и

$$\mathcal{Y}_p^{(k+1)}\mathcal{A}_p - \mathcal{I}_p = \mathcal{Y}_p^{(k)}\mathcal{A}_p - \mathcal{I}_p - (\mathcal{Y}_p^{(k)}\mathcal{A}_p - \mathcal{I}_p)$$
$$\times \sum_{\nu=0}^{r-2}(\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^\nu \mathcal{X}_p^{(k)}\mathcal{A}_p = (\mathcal{Y}_p^{(k)}\mathcal{A}_p - \mathcal{I}_p)(\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^{r-1} \geqslant \mathcal{O}_p$$

следует, что

$$\mathcal{X}_p^{(k+1)}\mathcal{A}_p \leqslant \mathcal{I}_p \leqslant \mathcal{Y}_p^{(k+1)}\mathcal{A}_p.$$

Ввиду $\mathcal{A}_p^{-1} \geqslant \mathcal{O}_p$ получаем

$$\mathcal{X}_p^{(k+1)} \leqslant \mathcal{A}_p^{-1} \leqslant \mathcal{Y}_p^{(k+1)}.$$

(7'): Используя соотношения

$$\mathcal{I}_p - \mathcal{X}_p^{(k+1)}\mathcal{A}_p = (\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^r,$$
$$\mathcal{Y}_p^{(k+1)}\mathcal{A}_p - \mathcal{I}_p = (\mathcal{Y}_p^{(k)}\mathcal{A}_p - \mathcal{I}_p)(\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p)^{r-1},$$

установленные при доказательстве неравенства (13), можем показать по индукции, что

$$\mathcal{I}_p - \mathcal{X}_p^{(k)}\mathcal{A}_p = (\mathcal{I}_p - \mathcal{X}_p^{(0)}\mathcal{A}_p)^{r^k},$$
$$\mathcal{Y}_p^{(k)}\mathcal{A}_p - \mathcal{I}_p = (\mathcal{Y}_p^{(0)}\mathcal{A}_p - \mathcal{I}_p)(\mathcal{I}_p - \mathcal{X}_p^{(0)}\mathcal{A}_p)^{r^k-1},$$

откуда и следует нужное утверждение.

(14): Снова используя соотношения, установленные при доказательстве (13), получаем

$$\|\mathcal{A}_p^{-1} - \mathcal{X}_p^{(k+1)}\| \leqslant \|\mathcal{A}_p\|^{r-1}\|\mathcal{A}_p^{-1} - \mathcal{X}_p^{(k)}\|^r,$$
$$\|\mathcal{Y}_p^{(k+1)} - \mathcal{A}_p^{-1}\| \leqslant \|\mathcal{A}_p\|^{r-1}\|\mathcal{Y}_p^{(k)} - \mathcal{A}_p^{-1}\|\,\|\mathcal{A}_p^{-1} - \mathcal{X}_p^{(k)}\|^{r-1}.$$

С помощью монотонной матричной нормы эту оценку можно продолжить следующим образом:

$$\|\mathcal{Y}_p^{(k+1)} - \mathcal{X}_p^{(k+1)}\| \leqslant \|\mathcal{A}_p^{-1} - \mathcal{X}_p^{(k+1)}\| + \|\mathcal{Y}_p^{(k+1)} - \mathcal{A}_p^{-1}\| \leqslant 2\|\mathcal{A}_p\|^{r-1}\|\mathcal{Y}_p^{(k)} - \mathcal{X}_p^{(k)}\|^r,$$

т. е. $d^{(k+1)} \leqslant \gamma (d^{(k)})^r$, $\gamma = 2\|\mathcal{A}_p\|^{r-1}$. Тепень нужное соотношение следует из теоремы 2 приложения А.

В методе (12) интервальные операции не используются. Несмотря на это, он порождает, подобно методу (9), монотонную последовательность границ для матрицы $\mathcal{A}_p^{-1}$. Был дан также необходимый и достаточный критерий сходимости, аналогичный критерию для метода (5). Однако применимость метода (12) ограничивается требованием $\mathcal{A}_p^{-1} \geqslant \mathcal{O}_p$.

Для определенных классов матриц можно указать вполне общие начальные значения $\mathcal{X}_p^{(0)}$, $\mathcal{Y}_p^{(0)}$, при которых (12) всегда будет сходиться. Если матрица $\mathcal{A}_p = (a_{ij})$ удовлетворяет условиям

$$a_{ii} > 0, \quad 1 \leqslant i \leqslant n,$$

$$a_{ij} \leqslant 0, \quad 1 \leqslant i, j \leqslant n, \quad i \neq j,$$

$$\sum_{i=1}^{n} a_{ij} > 0, \quad 1 \leqslant j \leqslant n,$$

то $\mathcal{A}_p$ является *M*-матрицей. Начальные матрицы $\mathcal{X}_p^{(0)} = (x_{ij})$ и $\mathcal{Y}_p^{(0)} = (y_{ij})$, такие что

$$x_{ij} = \begin{cases} 1/a_{ii} & \text{для } i = j \\ 0 & \text{в противном случае,} \end{cases} \quad \text{и} \quad y_{ij} = 1 \Big/ \sum_{\nu=1}^{n} a_{\nu i}, \quad 1 \leqslant i, j \leqslant n,$$

удовлетворяют условиям теоремы 3.

Рассмотрим теперь невырожденную вещественную матрицу $\mathcal{A}_p = (a_{ij})$ размерности $n \times n$, строки которой были переставлены, чтобы стало возможным разложение

$$\mathcal{A}_p = (\mathcal{I}_p + \mathcal{L}_p^*)\,\mathcal{U}_p^*, \tag{15}$$

где $\mathcal{I}_p$ — единичная матрица, $\mathcal{L}_p^*$ — строго нижняя треугольная матрица, $\mathcal{U}_p^*$ — верхняя треугольная матрица. Как известно, это разложение можно найти с помощью метода Гаусса. Мы хотим описать итерационную структуру, постепенно улучшающую границы, между которыми заключены элементы матриц $\mathcal{L}_p^*$ и $\mathcal{U}_p^*$. Такие границы, включающие все ошибки округления, можно вычислить, если использовать при исполнении метода Гаусса для точечной матрицы $\mathcal{A}_p$ машинную интервальную арифметику с округлением наружу. Допустим поэтому, что $\mathcal{L}^{(0)}$ — строго нижняя треугольная

интервальная матрица и $\mathcal{U}^{(0)}$ — верхняя треугольная матрица, для которых

$$\mathcal{L}_{\mathrm{p}}^{*} \in \mathcal{L}^{(0)}, \quad \mathcal{U}_{\mathrm{p}}^{*} \in \mathcal{U}^{(0)}. \tag{16}$$

Предположим сначала, что $\mathcal{L}_{\mathrm{p}}$ и $\mathcal{U}_{\mathrm{p}}$ — произвольные, но фиксированные треугольные матрицы, причем $\mathcal{L}_{\mathrm{p}}$ — строго нижняя, а $\mathcal{U}_{\mathrm{p}}$ — верхняя. Тогда мы имеем из (15)

$$\mathcal{R}_{\mathrm{p}} = (r_{ik}) := \mathcal{A}_{\mathrm{p}} - (\mathcal{I}_{\mathrm{p}} + \mathcal{L}_{\mathrm{p}})\,\mathcal{U}_{\mathrm{p}} = (\mathcal{I}_{\mathrm{p}} + \mathcal{L}_{\mathrm{p}}^{*})\,(\mathcal{U}_{\mathrm{p}}^{*} - \mathcal{U}_{\mathrm{p}}) + (\mathcal{L}_{\mathrm{p}}^{*} - \mathcal{L}_{\mathrm{p}})\,\mathcal{U}_{\mathrm{p}}.$$

Будем считать матрицы $\mathcal{U}_{\mathrm{p}}^{*} - \mathcal{U}_{\mathrm{p}}$ и $\mathcal{L}_{\mathrm{p}}^{*} - \mathcal{L}_{\mathrm{p}}$ неизвестными, а множитель $\mathcal{I}_{\mathrm{p}} + \mathcal{L}_{\mathrm{p}}^{*}$ перед $\mathcal{U}_{\mathrm{p}}^{*} - \mathcal{U}_{\mathrm{p}}$ известным и будем затем решать это уравнение путем поочередного нахождения сначала строки матрицы $\mathcal{U}_{\mathrm{p}}^{*} - \mathcal{U}_{\mathrm{p}}$, а затем столбца матрицы $\mathcal{L}_{\mathrm{p}}^{*} - \mathcal{L}_{\mathrm{p}}$. В предложении $u_{ij} = 0$ для

$$1 \leqslant i \leqslant n$$

получаем тогда

$$\begin{cases} u_{ik}^{*} - u_{ik} = r_{ik} - \sum\limits_{j=1}^{i-1} l_{ij}^{*}\,(u_{jk}^{*} - u_{jk}) - \sum\limits_{j=1}^{i-1} (l_{ij}^{*} - l_{ij})\,u_{jk}, \quad i \leqslant k \leqslant n, \\ l_{ki}^{*} - l_{ki} = \dfrac{1}{u_{ii}} \left\{ r_{ki} - \sum\limits_{j=1}^{i} l_{kj}^{*}\,(u_{ji}^{*} - u_{ji}) - \sum\limits_{j=1}^{i-1} (l_{kj}^{*} - l_{kj})\,u_{ji} \right\}, \quad i < k \leqslant n, \\ 1 \leqslant i \leqslant n. \end{cases}$$

Полагая

$$\left.\begin{aligned} r_{ik} &= a_{ik} - \sum_{j=1}^{i-1} l_{ij} u_{jk}, - u_{ik}, \quad i \leqslant k \leqslant n, \\ r_{ki} &= a_{ki} - \sum_{j=1}^{i} l_{kj} u_{ji}, \qquad\qquad i+1 \leqslant k \leqslant n, \end{aligned}\right\} \; 1 \leqslant i \leqslant n,$$

получаем тогда соотношения

$$\begin{cases} u_{ik}^{*} = a_{ik} - \sum\limits_{j=1}^{i-1} l_{ij}^{*} u_{jk} - \sum\limits_{j=1}^{i-1} l_{ij}^{*}\,(u_{jk}^{*} - u_{jk}), \quad i \leqslant k \leqslant n, \\ l_{ki}^{*} = \dfrac{1}{u_{ii}} \left\{ a_{ki} - \sum\limits_{j=1}^{i-1} l_{kj}^{*} u_{ji} - \sum\limits_{j=1}^{i} l_{kj}^{*}\,(u_{ji}^{*} - u_{ji}) \right\}, \quad i < k < n, \\ 1 \leqslant i \leqslant n. \end{cases}$$

Выделенные жирным шрифтом выражения происходят из множителя $\mathcal{I}_p + \mathcal{L}_p^*$, который мы предположили известным. Если воспользоваться предположением, что $\mathcal{L}_p^* \in \mathcal{L}^{(0)}$ для этих величин и выбрать $\mathcal{U}_p := \mathcal{U}_p^0 \in \mathcal{U}^{(0)}$, где $u_{ii}^{(0)} \neq 0$, $1 \leqslant i \leqslant n$, то получим для элементов матриц $\mathcal{L}_p^*$ и $\mathcal{U}_p^*$ следующие локализующие интервалы:

$$\begin{cases} u_{ik}^* \in U_{ik}^{(1)} = \left\{ a_{ik} - \sum\limits_{j=1}^{i-1} L_{ij}^{(1)} u_{jk}^{(0)} - \sum\limits_{j=1}^{i-1} L_{ij}^{(0)} \left(U_{jk}^{(1)} - u_{jk}^{(0)} \right) \right\} \cap U_{ik}^{(0)}, \\ \qquad\qquad 1 \leqslant k \leqslant n, \\ l_{ki}^* \in L_{ki}^{(1)} = \left\{ \frac{1}{u_{ii}^{(0)}} \left(a_{ki} - \sum\limits_{j=1}^{i-1} L_{kj}^{(1)} u_{ji}^{(0)} - \sum\limits_{j=1}^{i} L_{kj}^{(0)} \left(U_{ji}^{(1)} - u_{ji}^{(0)} \right) \right) \right\} \\ \qquad\qquad \cap L_{ki}^{(0)}, \quad i < k \leqslant n, \\ 1 \leqslant i \leqslant n. \end{cases}$$

Систематическое повторение этой процедуры приводит к следующему итерационному методу:

$$\begin{cases} U_{ik}^{(m+1)} = \left\{ a_{ik} - \sum\limits_{j=1}^{i-1} L_{ij}^{(m+1)} u_{jk}^{(m)} - \sum\limits_{j=1}^{i-1} L_{ij}^{(m)} \left(U_{jk}^{(m+1)} - u_{jk}^{(m)} \right) \right\} \cap U_{ik}^{(m)}, \\ \qquad\qquad i \leqslant k \leqslant n, \\ L_{ki}^{(m+1)} = \left\{ \frac{1}{u_{ii}^{(m)}} \left(a_{ki} - \sum\limits_{j=1}^{i-1} L_{kj}^{(m+1)} u_{ji}^{(m)} - \sum\limits_{j=1}^{i} L_{kj}^{(m)} \left(U_{ji}^{(m+1)} - u_{ji}^{(m)} \right) \right) \right\} \\ \qquad\qquad \cap L_{ki}^{(m)}, \quad i < k \leqslant n, \\ 1 \leqslant i \leqslant n, \\ m \geqslant 0. \end{cases}$$

С помощью рассуждений, аналогичных проведенным при описании первого шага этого метода, можно показать, что в общем случае верно следующее утверждение.

Теорема 4. *Пусть матрица* $\mathcal{A}_p$ *имеет разложение* $\mathcal{A}_p = (\mathcal{I}_p + \mathcal{L}_p^*)\,\mathcal{U}_p^*$. *Пусть* $\mathcal{L}_p^* \in \mathcal{L}^{(0)}$, $\mathcal{U}_p^* \in \mathcal{U}^{(0)}$. *Если* $\mathcal{L}_p^{(m)} \in \mathcal{L}^{(m)}$, $\mathcal{U}_p^{(m)} \in \mathcal{U}^{(m)}$, *то для всех* $m \geq 0$ *верно*

$$\mathcal{L}_p^* \in \mathcal{L}^{(m+1)}, \quad \mathcal{U}_p^* \in \mathcal{U}^{(m+1)}. \tag{18}$$

Мы отметим, что метод (17) может быть выполнен при начальных интервалах произвольной, но конечной ширины, если выполнено предположение (16). Единственное деление встречается при

вычислении $L_{kl}^{(m+1)}$. Так как $0 \neq u_{ll}^{*} \in U_{ll}^{(m)}$, мы всегда можем выбрать $u_{ll}^{(m)} \in U_{ll}^{(m)}$ так, чтобы $u_{ll}^{(m)} \neq 0$.

Покажем теперь, что метод (17) при отсутствии ошибок округления дает точное треугольное разложение за конечное число шагов.

Теорема 5. *В условиях предыдущей теоремы метод* (17) *вычисляет точное разложение матрицы* $\mathcal{A}_0$ *размерности* $n \times n$ *самое большее за* $2n-1$ *шагов.*

Доказательство. При $m = 0$ мы получаем из (17) для i=l, что

$$U_{1k}^{(1)} = a_{1k} \cap U_{1k}^{(0)} = a_{1k} = u_{1k}^{*}, \quad 1 \leqslant k \leqslant n.$$

Это значит, что для произвольных начальных матриц, удовлетворяющих условиям локализации (16), матрица $\mathcal{U}^{(1)}$ имеет в первой строке значения, нужные для треугольного разложения. Матрица $\mathcal{U}^{(2)}$ имеет нужные значения в первой строке точно так же, как $\mathcal{U}^{(1)}$. Поэтому для первою столбца матрицы $\mathcal{L}^{(2)}$ мы получаем из (17) при m= 1, i= 1, что

$$L_{k1}^{(2)} = \left(\frac{a_{k1}}{u_{11}}\right) \cap L_{k1}^{(1)} = \frac{a_{k1}}{a_{11}} = l_{k1}^{*}, \quad 2 \leqslant k \leqslant n.$$

Поэтому после второго шага итерации нужные значения имеют и первый столбец матрицы $\mathcal{L}^{(2)}$, и первая строка матрицы $\mathcal{U}^{(2)}$. Покажем теперь, что если первые i строк матрицы $\mathcal{U}^{(m)}$ и первые i столбцов матрицы $\mathcal{L}^{(m)}$ уже имеют нужные значения, то не менее чем i+1 строк матрицы $\mathcal{U}^{(m+1)}$ (и i столбцов матрицы $\mathcal{L}^{(m+1)}$) имеют нужные значения. Предыдущее рассуждение показывает, что это верно при $i = 0$. Для i>0 применим математическую индукцию.

Заметим прежде всего, что $\mathcal{L}^{(m+1)}$ и $\mathcal{U}^{(m+1)}$ имеют те же элементы, что $\mathcal{L}^{(m)}$ и $\mathcal{U}^{(m)}$ в первых i строках и столбцах. Таким образом эти элементы все еще имеют нужные значения. Это непосредственно следует из (17). Из (17) мы получаем также следующее соотношение для $(i + 1)$-й строки матрицы $\mathcal{U}^{(m+1)}$:

$$U_{i+1,k}^{(m+1)} = \left\{a_{i+1,k} - \sum_{j=1}^{i-1} l_{i+1,j}^{*} u_{jk}^{*}\right\} \cap U_{i+1,k}^{(m)} = u_{i+1,k}^{*}, \quad i+1 \leqslant k \leqslant n.$$

Чтобы завершить доказательство по индукции, мы должны показать, что если для некоторого $m \geq 0$ первые i строк матрицы $\mathcal{U}^{(m)}$ и первые i—1 столбцов матрицы $\mathcal{L}^{(m)}$ имеют нужные значения, то $\mathcal{L}^{(m+1)}$ имеют нужные значения в первых столбцах (а $\mathcal{U}^{(m+1)}$ — в первых i строках).

Это было доказано выше для i= 1. Для $i > 1$ заметим, что в силу (17)

матрица $\mathcal{U}^{(m+1)}$ имеет нужные значения по крайней мере в тех же строках, что и $\mathcal{U}^{(m)}$, а матрица $\mathcal{L}^{(m+1)}$ — по крайней мере в тех же строках, что $\mathcal{L}^{(m)}$. Тогда для i-го столбца матрицы $\mathcal{L}^{(m+1)}$ получаем из (17), что

$$L_{ki}^{(m+1)} = \left\{ \frac{1}{u_{ii}^{*}} \left(a_{ki} - \sum_{j=1}^{i-1} l_{kj}^{*} u_{ji}^{*} \right) \right\} \cap L_{ki}^{(m)} = l_{ki}^{*}, \quad i+1 \leqslant k \leqslant n.$$

Таким образом, самое большее через $2n$— 1 шаг мы получим точное решение.

Следующая теорема показывает, что данный метод обладает так называемым «квадратичным» свойством сходимости, хорошо знакомым по методу Ньютона — Рафсона.

Теорема 6. *Пусть*

$$d^{(m)} := \max_{1 \leqslant i,j \leqslant n} \{\max \{d(L_{ij}^{(m)}),\ d(U_{ij}^{(m)})\}\}.$$

Тогда для метода (17) *имеет место*

$$d^{(m+1)} \leqslant \alpha (d^{(m)})^2,$$

где α *— неотрицательное вещественное число, не зависящее от m, т. е. на каждом шаге ширина интервала примерно возводится в квадрат.*

Доказательство (методом математической индукции). Из (17) мы получаем

$$\left\{ \begin{aligned} d(U_{ik}^{(m+1)}) &\leqslant \sum_{j=1}^{i-1} d(L_{ij}^{(m+1)}) |u_{jk}^{(m)}| + \sum_{j=1}^{i-1} d(L_{ij}^{(m)}) |U_{jk}^{(m+1)} - u_{jk}^{(m)}| \\ &\quad + \sum_{j=1}^{i-1} |L_{ij}^{(m)}| d(U_{jk}^{(m+1)}), \quad i \leqslant k \leqslant n, \\ d(L_{ki}^{(m+1)}) &\leqslant \frac{1}{|u_{ii}^{(m)}|} \left\{ \sum_{j=1}^{i-1} d(L_{kj}^{(m+1)}) |u_{ji}^{(m)}| \right. \\ &\quad + \sum_{j=1}^{i} d(L_{kj}^{(m)}) |U_{ji}^{(m+1)} - u_{ji}^{(m)}| \\ &\quad \left. + \sum_{j=1}^{i} |L_{kj}^{(m)}| d(U_{ji}^{(m+1)}) \right\}, \quad i < k \leqslant n, \\ 1 &\leqslant i \leqslant n. \end{aligned} \right. \tag{17'}$$

Положим теперь при $0 \notin U_{ii}^{(0)}, \quad 1 \leqslant i \leqslant n,$

$$\begin{cases} \alpha_{ik} = \begin{cases} \sum\limits_{l=1}^{i-1} (\beta_{il} |U_{lk}^{(0)}| + 1 + |L_{il}^{(0)}| \alpha_{lk}), & i \leqslant k \leqslant n, \\ 0 & \text{в противном случае,} \end{cases} \\ \beta_{ki} = \begin{cases} \left| \frac{1}{U_{ii}^{(0)}} \right| \left\{ \sum\limits_{l=1}^{i-1} \beta_{kl} |U_{li}^{(0)}| + \sum\limits_{l=1}^{i} (1 + |L_{kl}^{(0)}| \alpha_{li}) \right\}, & i < k \leqslant n, \\ 0 & \text{в противном случае,} \end{cases} \end{cases} \tag{17''}$$

и наконец

$$\alpha = \max_{1 \leqslant i,\ k \leqslant n} \{\max \{\alpha_{ik}, \beta_{ki}\}\}.$$

Используя определение (17''), мы немедленно получаем из (17') при $i = 1$, что

$$l\left(U_k^{(m+1)}\right) \leqslant \alpha_{1k} \left(d^{(m)}\right)^2, \quad 1 \leqslant k \leqslant n,$$

и

$$d\left(L_{k1}^{(m+1)}\right) \leqslant \beta_{k1} \left(d^{(m)}\right)^2, \quad 1 < k \leqslant n.$$

Допустим теперь, что для первых $i \neq 1$ строк и столбцов имеет место при $(1 \leqslant l \leqslant i-1)$

$$\begin{cases} d\left(U_{lk}^{(m+1)}\right) \leqslant \alpha_{lk} \left(d^{(m)}\right)^2, & l \leqslant k \leqslant n, \\ d\left(L_{kl}^{(m+1)}\right) \leqslant \beta_{kl} \left(d^{(m)}\right)^2, & l \leqslant k \leqslant n. \end{cases} \tag{17'''}$$

Это очевидно при $i = 1$. Теперь мы получаем из (17') и (17''), что

$$d\left(U_{ik}^{(m+1)}\right) \leqslant \sum_{l=1}^{i-1} \beta_{il} |U_{lk}^{(0)}| \left(d^{(m)}\right)^2 + \sum_{l=1}^{i-1} \left(d^{(m)}\right)^2 + \sum_{l=1}^{i-1} |L_{il}^{(0)}| \alpha_{lk} \left(d^{(m)}\right)^2$$
$$= \alpha_{ik} \left(d^{(m)}\right)^2, \quad i \leqslant k \leqslant n.$$

Аналогично

$$d\left(L_{ki}^{(m+1)}\right) \leqslant \left| \frac{1}{U_{ii}^{(0)}} \right| \left\{ \sum_{l=1}^{i-1} \beta_{kl} |U_{li}^{(0)}| \left(d^{(m)}\right)^2 + \sum_{l=1}^{i} \left(d^{(m)}\right)^2 \right.$$
$$\left. + \sum_{l=1}^{i} |L_{kl}^{(0)}| \alpha_{li} \left(d^{(m)}\right)^2 \right\} = \beta_{ki} \left(d^{(m)}\right)^2, \quad i < k \leqslant n.$$

Из этих соотношений следует доказываемое утверждение

$$d^{(m+1)} \leqslant \alpha \left(d^{(m)}\right)^2.$$

Если исполнять (17) на вычислительной машине, применяя машинную интервальную арифметику, то в противоположность теореме 5 мы, вообще говоря, не получим за конечное число шагов точного треугольного разложения матрицы $\mathscr{A}_p$. Рассмотрим, какой

окончательной точности здесь можно достичь. В этих рассуждениях примем те же допущения, которые привели нас к формулам (4. 15a) и (4. 15b), а тем самым и к (4. 22), (4.23). Последние две формулы были использованы при доказательстве формул (4. 24) и (4.25). Теперь мы применим эти две формулы к (17). Это дает ширину вычисленных элементов

$$\begin{cases} d\left(\bar{U}_{ik}^{(m+1)}\right) \leqslant d\left(U_{ik}^{(m+1)}\right) + 2\varepsilon \sum_{j=1}^{2i-2} \left|\bar{S}_j\right| + 2\varepsilon\left(3 + 3\varepsilon + \varepsilon^2\right) \\ \qquad \times \sum_{j=1}^{i-1} \left(\left|L_{ij}^{(m)}\right| \left|\bar{U}_{jk}^{(m+1)} - u_{jk}^{(m)}\right| + \left|\bar{L}_{ij}^{(m+1)}\right| \left|u_{jk}^{(m)}\right|\right), \\ \qquad\qquad i \leqslant k \leqslant n, \\ d\left(\bar{L}_{ki}^{(m+1)}\right) \leqslant d\left(L_{ki}^{(m+1)}\right) + \left|\frac{1}{u_{ii}^{(m)}}\right| \left\{ 2\varepsilon \sum_{j=1}^{2i-1} \left|\bar{T}_j\right| + 2\varepsilon_{2i} \left|\bar{T}_{2i}\right| \right. \\ \qquad + 2\varepsilon\left(3 + 3\varepsilon + \varepsilon^2\right) \left(\sum_{j=1}^{i-1} \left|\bar{L}_{kj}^{(m+1)}\right| \left|u_{ji}^{(m)}\right| \right. \\ \left. \left. \qquad + \sum_{j=1}^{i} \left|L_{kj}^{(m)}\right| \left|\bar{U}_{ji}^{(m+1)} - u_{ji}^{(m)}\right| \right) \right\}, \quad i < k \leqslant n, \\ 1 \leqslant i \leqslant n. \end{cases}$$

В этих неравенствах S_j и $\bar{T}_j$, обозначают фактические результаты промежуточных вычислений. Эти неравенства можно интерпретировать следующим образом. Допустим, что все элементы матрицы $\mathscr{L}_p$ не превосходят 1 по абсолютной величине. Это предположение выполняется хотя бы приближенно, если строки матрицы $\mathscr{A}_p$ упорядочены таким образом, что не приходится переставлять строки в процессе исключения по Гауссу с выбором главных элементов по столбцам. Если ширина интервалов $L_{ij}^{(m)}$ не слишком велика, то $\left|L_{ij}^{(m)}\right|$ не намного больше 1, а в силу того, что в (17) берутся пересечения, это верно и для $\left|L_{ij}^{(m+1)}\right|$. При тех же предположениях те же рассуждения показывают, что $\left|\bar{U}_{jk}^{(m+1)}\right| - \left|u_{jk}^{(m)}\right|$ и $\left|U_{ji}^{(m+1)}\right| - \left|u_{ji}^{(m)}\right|$ малы. Поэтому делаем вывод, что при малой ширине элементов на m-м шаге разность между $d\left(\bar{U}_{ik}^{(m+1)}\right)$ и $d\left(U_{ik}^{(m+1)}\right)$ существенно зависит от $\left|u_{jk}^{(m)}\right|$ и от величины промежуточных результатов. То же верно и для разности между

$d\left(\overline{L}_{ki}^{(m+1)}\right)$ и $d\left(L_{ki}^{(m+1)}\right)$, если добавить еще, что малые величины $\left|u_{ii}^{(m)}\right|$ могут ухудшить эту разность.

Так как в силу теоремы 6 первые члены приведенных выше неравенств приближенно равны квадратам таких же членов на предыдущем шаге, мы получим небольшую ширину, если выполнены следующие условия:

(a) Элементы матрицы $\mathcal{U}_{\mathrm{p}}^{*}$ по абсолютной величине не намного больше единицы.

(b) Диагональные элементы матрицы $\mathcal{U}_{\mathrm{p}}^{*}$ по абсолютной величине не намного меньше единицы.

(c) Элементы матрицы $\mathcal{L}_{\mathrm{p}}^{*}$ по абсолютной величине не больше единицы.

(d) Вычисленные промежуточные результаты в (17) не слишком велики по абсолютной величине.

V. Шкалы

Шкала — это знаковая система, для которой задано отображение, ставящее в соответствие реальным объектам тот или иной количественный элемент шкалы.

Формально **шкалой** называют кортеж, <X, φ, Y>, где X — реальный объект, φ — отображение, Y — знаковая система.

Зна́ковая систе́ма — система из знаков и отношений между ними, основное понятие семиотики. Состоит из однообразно интерпретируемых и трактуемых сообщений или сигналов, которыми можно обмениваться в процессе общения. Знаковые системы структурируют процесс общения, придавая ему предсказуемость.

Понятие знаковой системы близко понятию языка: иногда они используются взаимозаменяемо, однако понятие языка несёт коннотации, связанные с естественными чуловеческими языками, и обычно является менее общим. Другим примером знаковой системы является метаязык — **язык для описания языков.**

Различные типы измерительных шкал широко используются в теоретической и практической человеческой деятельности, в науке и технике — в том числе во многих гуманитарных научных областях, таких как экономика, психометрия, социология и др.

Приведем альтернативное определение понятия «шкала»

Шкала́ (лат. *scala* — лестница) — часть показывающего устройства средства измерений, представляющая собой упорядоченный ряд отметок вместе со связанной с ними нумерацией или техническая отметка на шкале измерительного прибора. Шкалы могут располагаться по окружности, дуге или прямой линии. Показания отсчитываются невооружённым глазом при расстояниях между делениями до 0,7 мм, при меньших — при помощи лупы или микроскопа, для долевой оценки делений применяют дополнительные шкалы — нониусы.

Следует заметить, что термин «шкала» в метрологической практике имеет, по крайней мере, два различных значения. Во-первых, шкалой или, точнее, шкалой измерений (шкалой физической величины) называют принятый по соглашению порядок определения и обозначения всевозможных проявлений (значений) конкретного свойства (величины). Во-вторых, шкалой называют отсчётные устройства аналоговых средств измерений, это значение используется в данной статье.

Круговую шкалу часов, курвиметров и некоторых других приборов называют циферблатом.

1. Выбор измеряемых критериев.

Одной из важнейших задач при оценке значений выбранных критериев объектов является задача измерения рассматриваемых критериев по выбранным шкалам.

В работе рассматриваются вопросы измерения выбранных критериев объектов. Будем считать измерением процесс присвоения чисел критериям (признакам) распознаваемых объектов (вещей, предметов или событий) согласно некоторой системе правил.

Для измерения критериев будем выделять три свойства чисел: тождество, ранговый порядок и аддитивность. Эти свойства будем задавать аксиомами A_1-A_9.

Тождество обусловлевается тремя аксиомами:

A_1. Либо $x=y$, либо $x \neq y$.

A_2. Если $x=y$, то $y = x$.

A_3. Если $x=y$ и $y = z$, то $x=z$.

Свойства рангового порядка обусловливаются двумя аксиомами:

A_4. Если $x>y$, то $y < x$.

A_5. Если $x>y$, и $y >z$, то $x>z$.

Свойство аддитивности обусловливается следующими аксиомами:

A_6. Если $x=y$ и $z>0$, то $x + z>y$.

A_7. $x +z=z+x$.

A_8. Если $x=y$ и $z = p$, то $x+z=y+p$.

A_9. $(x +y) +z=x+ (y+z)$.

Приведенные аксиомы позволяют нам выделить четыре уровня измерения. Перечислим эти уровни измерения в порядке их усиления: шкалы наименований, шкалы порядка, шкалы интервалов и шкалы отношений. Чем выше уровень шкалы, тем больше математических и статистических операции можно выполнять, над полученными при измерении числами.

Приведем ряд определений, которые будут использоваться при построении моделей объектов и процессов ИИ.

Определение 1. Множесто Y выходов процесса P вместе с заданным на нем множеством отношений R_{Y^i}, $i \in I$, $I = \{1, 2, \ldots, n\}$, будем называть системой с отношениями и обозначать (рис. 1)

$$\bar{Y} = (Y, R_{y^i}), \quad i \in I, I = \{1, 2, \ldots, n\}, \quad Y = \{y_1, y_2, \ldots, y_k\}.$$

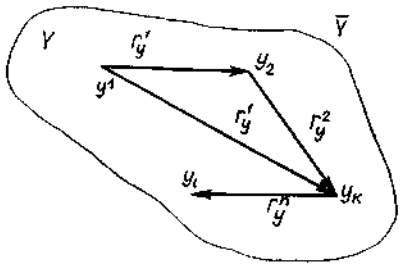


Рис. 1. Пример системы с отношениями.

Для целей построения моделей объектов и процессов будем рассматривать конечные множества.

Определение 2. Отношением эквивалентности на множестве $Y=\{y_1,\ y_2,\ \ldots,\ y_k\}$ будем называть бинарное отношение r_y^1, если оно обладает следующими свойствами (рис. 2):

1) рефлексивности, т. е.

$$y_i r_y y_i \forall_i \in I,\ I=\{1, 2,\ \ldots,\ k\};$$

2) симметричности, т. е.

$$y_a r_y^1 y_k \leftrightarrow (y_a r_y y_k,\ y_k r_y^1 y_a);$$
$$\forall y_a,\ y_k \in Y;$$

3) транзитивности, т. е.

$$y_a r_y^1 y_k \wedge y_k r_y^1 y_m \rightarrow y_a r_y^1 y_m .$$

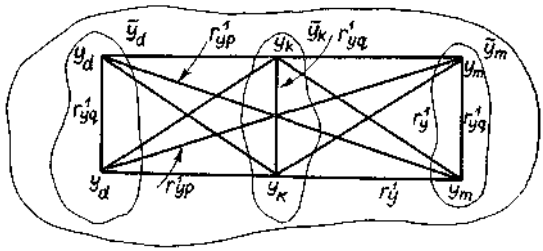


Рис. 2. Пример отношения эквивалентности.

Определение 3. Гомоморфизмом системы с отношениями Y на систему такого же типа $\bar{N}=(N_Y,\ R_{N^i},\ i\in I,\ I=\{1,\ 2,\ \ldots,\ n\})$ будем называть такое отображение γ, при котором для всех $i\in I$ и $(y_1, \ldots, y_{k_i})\in Y^{k_i}$ будет иметь место

$$R_{Y^i}(y_1, y_2,\ \ldots,\ y_{k_i}) = R_{N^i}(\gamma(y_1), \gamma(y_2),\ \ldots,\ \gamma(y_{k_i}))\ .$$

Отображение γ будет гомоморфизмом системы $\overline{Y}$ в $\bar{N}$ тогда и только тогда, когда для всех $i\in I$ имеет место (рис. 3)

$$R_{Y^i}=\gamma_{k_i}^{-1}(R_{N^i})\ .$$

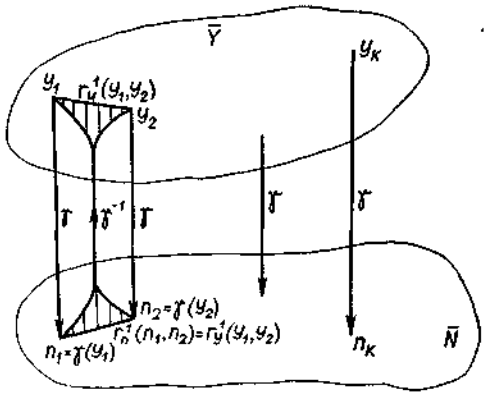


Рис. 3. Гомоморфизм системы с отношениями.

2. Типы шкал

На сегодняшний день различают четыре основных типа шкал измерений: номинальная, порядковая, интервальная и относительная. Каждый тип шкалы обладает определенными признаками, которые рассматриваются ниже; сейчас же рассмотрим какую роль играет техника измерений в процессе классификации.

Часто при классификации исследователь не имеет возможности численно измерить исследуемый параметр. Например, отношение человека к чему-либо, степень его предпочтения и т.д. Способы измерения в данном случае отличаются от традиционных способов. Измерением в данном случае будет считаться любой способ приписывания числовых значений символам, которые отражают качественные характеристики объектов. При этом должны существовать устойчивые взаимосвязи между символами и качествами, которые они отражают. Иными словами, для осуществления кластеризации объекта с качественными характеристиками необходимо использовать приемы техники шкалирования.

В процессе использования техники шкал традиционно выделяют ряд стадий, качество выполнения которых оказывает непосредственное влияние на результат выделения кластеров. На первом этапе необходимо дать четкое определение тому, что собираются измерять.

Далее следует указать, как измерение будет осуществлено на практике или что/кто конкретно подлежит измерению. После чего выбирают тип шкалы измерения, который предопределяет метод сбора информации. Любые измерения связаны с ошибками, но поскольку измерение в данном случае имеет специфику, то исследователь может самостоятельно оценить некоторые случайные отклонения исследуемого параметра и исключить его из кластера. Традиционно объекты наблюдения могут быть представлены в следующих типах шкал.

Шкалы измерений принято классифицировать по типам измеряемых данных, которые определяют допустимые для данной шкалы математические преобразования, а также типы отношений, отображаемых соответствующей шкалой. Современная классификация шкал была предложена в 1946 году Стэнли Смитом Стивенсом.

Шкала наименований (номинальная, классификационная)
Используется для измерения значений качественных признаков. Значением такого признака является наименование класса эквивалентности, к которому принадлежит рассматриваемый объект. Примерами значений качественных признаков являются названия государств, цвета, марки автомобилей и т. п. Такие признаки удовлетворяют аксиомам тождества:

- Либо $A = B$, либо $A \neq B$;
- Если $A = B$, то $B = A$;
- Если $A = B$ и $B = C$, то $A = C$.

При большом числе классов используют иерархические шкалы наименований. Наиболее известными примерами таких шкал являются шкалы, используемые для классификации животных и растений.
С величинами, измеряемыми в шкале наименований, можно выполнять только одну операцию — проверку их совпадения или несовпадения. По результатам такой проверки можно дополнительно вычислять частоты заполнения (вероятности) для различных классов, которые могут использоваться для применения различных методов статистического анализа[5] — критерия согласия Хи-квадрат, критерия Крамера для проверки гипотезы о связи качественных признаков и др.

Логическая основа шкал определена нами выше в аксиомах A_1—A_3. Построить шкалу наименований — это значит использовать число как название.

Свойство тождества чисел заключается в том, что два объекта либо тождественны, либо различны, а также, что объекты, равные одному и тому же объекту, равны между собой. Аксиома 2 гласит, что отношение равенства симметрично.

Свойство тождества чисел будем использовать в теории ИИ для порождения бесконечного набора (множества) различных названий. Названием может быть и порядковый номер. Можно нумеровать технические чертежи, показатели состояния процесса, действия и т. д. Это не будет значить ничего иного, кроме того, что каждый отдельный объект или предмет должен иметь свое обозначение. Шкалы наименований будем использовать для распознавания различных объектов.

Шкалы наименований по существу качественны, однако они могут допускать не которые статистические операции. Можно сосчитать число представителей каждого класса объектов, определить их частоту и найти наиболее многочисленный класс. Отношения элементов шкалы наименований показаны на рис. 4.

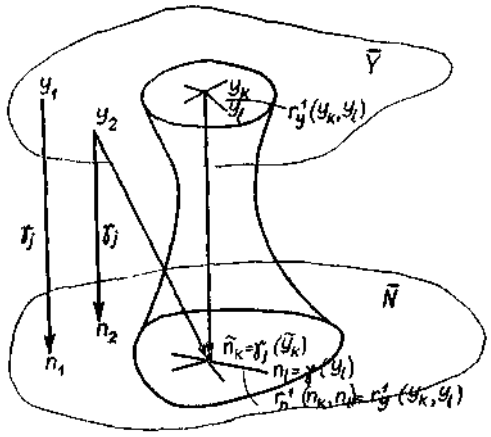


Рис. 4. Отношения в шкале наименований.

Приведем еще одно определение.

Определение 4. Взаимно-однозначное отображение системы (объекта) (Y, r_y^1) в систему (N, r_n^1), где r^1_y— отношение

эквивалентности, а r^1_n — отношение равенства на множестве натуральных чисел *N*, будем называть шкалой наименований.

На рис. 4 система с отношениями $\bar{Y} = (Y, r_y^1)$ на множестве *Y* имеет только одно отношение — отношение эквивалентности r^i_y, которое отображается гомоморфно в отношение r^1_n.

Примеры:
Для обозначения в номинальной шкале могут быть использованы:
- слова естественного языка (например, географические названия, собственные имена людей и т. д.);
- произвольные символы (гербы и флаги государств, эмблемы родов войск, всевозможные значки и т. д.);
- номера (регистрационные номера автомобилей, официальных документов, номера на майках спортсменов);
- их различные комбинации (например, почтовые адреса, экслибрисы личных библиотек, печати и пр.).

Однако необходимость классификации возникает и в тех случаях, когда классифицируемые состояния образуют непрерывное множество (или континуум). Задача сводится к предыдущей, если все множество разбить на конечное число подмножеств, искусственно образуя тем самым классы эквивалентности; тогда принадлежность состояния к какому-либо классу снова можно регистрировать в шкале наименований. Однако условность введенных классов (не их шкальных обозначений, а самих классов) рано или поздно проявится на практике.

Примеры:
1. Например, возникают трудности точного перевода с одного языка на другой при описании цветовых оттенков: в английском языке голубой, лазоревый и синий цвета не различаются.
2. Названия болезней также образуют шкалу наименований. Психиатр, ставя больному диагноз «шизофрения», «паранойя», «маниакальная депрессия» или «психоневроз», ис-пользует номинальную шкалу; и все же иногда врачи не зря вспоминают, что «нужно лечить больного, а не болезнь»: название болезни лишь обозначает класс, внутри которого на самом деле имеются различия, так как эквивалентность внутри класса носит условный характер.

Необходимо понимать, что обозначения классов — это только символы, даже если для этого использованы номера. С этими номерами нельзя обращаться как с числами — это только цифры.

Пример. Если у одного спортсмена на спине номер 1, а другого — 2, то никаких других выводов, кроме того, что это разные участники соревнований, делать нельзя: например, нельзя сказать, что «второй в

два раза лучше».
При обработке экспериментальных данных, зафиксированных в номинальной шкале, непосредственно с самими данными можно выполнять только операцию проверки их совпадения или несовпадения.

Порядковая шкала (или ранговая)

Строится на отношении тождества и порядка. Субъекты в данной шкале ранжированы. Но не все объекты можно подчинить отношению порядка. Например, нельзя сказать, что больше круг или треугольник, но можно выделить в этих объектах общее свойство-площадь, и таким образом становится легче установить порядковые отношения. Для данной шкалы допустимо монотонное преобразование. Такая шкала груба, потому что не учитывает разность между субъектами шкалы. Пример такой шкалы: балльные оценки успеваемости (неудовлетворительно, удовлетворительно, хорошо, отлично), шкала Мооса.

Рассмотрим усиление шкалы наименований. Первое усиление шкалы наименований возможно только в том случае, когда мы осуществляем сравнивание двух распознаваемых объектов по одному общему признаку. Например, если объекты охарактеризованы по объему, то можно установить, какой из них имеет больший объем по сравнению с другим. Согласно аксиоме A_4, упорядочивающее отношение r^2 асимметрично, т. е. $y_k r_{yi}^2 y_m$ обозначает, что объект y_k по какому-то признаку i превосходит объект y_m. В случае простого порядка отношение $y_m r^2 y_k$ не имеет места, т. е. $y_m \overline{r}^2 y_k$.

Если сравнивать таким образом каждую пару объектов в соответствующем перечне и если каждая тройка объектов обнаруживает транзитивность (см. аксиому A_5), то можно построить шкалу простого порядка. Например, можно пронумеровать объемы по их величине, присваивая большему объему больший номер.

Таким образом, объекты, отображенные на шкалу простого порядка, должны быть сравнимы и транзитивны по общему признаку.

В схемах простого порядка каждый элемент должен иметь более высокий или более низкий ранг, чем всякий другой элемент: **частота появления любого подмножества равна единице**. Во многих случаях будем допускать равную оценку появления классов, чтобы не проводить различий за границами наблюдений. При этом частота появления подмножества эмпирических объектов может быть больше

единицы. Элементы, измеренные по такой шкале, образуют слабый порядок.

Логическое основание слабого порядка заключается в двух отношениях: **антисимметрии и транзитивности**. Примером антисимметрии служит отношение «≥» для действительных чисел. В словесной форме отношение антисимметрии будем выражать высказываниями вида: *«y_k* по меньшей мере так же хорошо, как и *y_d»*. Для слабого порядка аксиомы A_4 и A_5 можно записать так:

A_4'. Либо $x \geq y$, либо $y \leq x$.

A_5'. Если $x \geq y$ и $y \geq z$, то $x \geq z$.

Если при этом ни $x \geq y$, ни $y \leq x$, то будем говорить говорить, что x и y несравнимы. Если одновременно $x \geq y$ и $y \geq x$, то получается рефлексивное отношение $x = y$. Отметим, что несравнимость не всегда то же самое, что безразличие. В случае, когда некоторые элементы перечня несравнимы по упорядочивающему отношению, а остальное подмножество допускает сравнение, то будем говорить, что имеет место частичный порядок.

Аксиомы упорядочения допускают те же статистические операции, что и аксиомы тождества, т. е. получение **частот и мод**. Ранговый порядок позволяет вычислять **медианы, центили и коэффициенты ранговой корреляции**, оценивающие степень близости упорядочений в сравниваемых множествах.

Арифметические и другие статистические операции, кроме перечисленных выше, исключаются, так как распознаваемые объекты на шкалах порядка не обязательно располагаются равномерно.

С помощью операций измерения $c_1, c_2, \ldots, c_j$ можно получить на множестве чисел N множество шкал порядка, гомоморфно отображающих систему с отношениями

$$\bar{Y} = (Y, r_y^1, r_y^2)$$

в систему с отношениями

$$\bar{N} = (N, r_n^1, r_n^2),$$

где r_n^2— отношение слабого порядка (рис. 5).

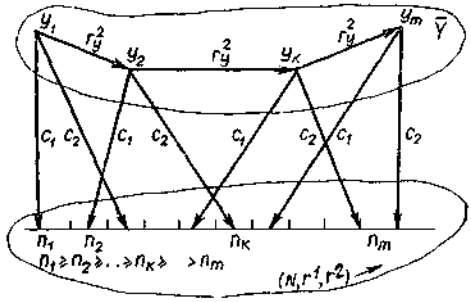


Рис. 5. Отношения в шкале порядка.

Приведем следующее определение

Определение 5. Шкалу $П: Y \rightarrow N$ будем называть шкалой порядка, если она единственна с точностью до монотонности возрастающих непрерывных отображений множества $\gamma(Y)$ в множество N.

Измерение в шкале порядка может применяться, например, в следующих ситуациях:

- когда необходимо упорядочить объекты во времени или пространстве. Это ситуация, когда интересуются не сравнением степени выраженности какого-либо их качества, а лишь взаимным пространственным или временным расположением этих объектов;
- когда нужно упорядочить объекты в соответствии с каким-либо качеством, но при этом не требуется производить его точное измерение;
- когда какое-либо качество в принципе измеримо, но в настоящий момент не может быть измерено по причинам практического или теоретического характера.

Типовые порядковые шкалы

Обозначив такие классы символами и установив между этими символами отношения порядка, мы получим шкалу простого порядка: A → B → C → D → E → F.

Примеры:

Нумерация очередности, неимение знания, призовые места в конкурсе, социально-экономический статус («низший класс», «средний класс», «высший класс»).

Разновидностью шкалы простого порядка являются оппозиционные шкалы. Они образуются из пар антонимов (например, сильный-

слабый), стоящих на разных концах шкалы, где за середину берется позиция, соответствующая среднему значению наблюдаемой сущности. Как пра-вило, остальные позиции никак не шкалируются.
Иногда оказывается, что не каждую пару классов можно упорядочить по предпочтению: неко-торые пары считаются равными — одновременно $A \geq B$ и $B \leq A$, т. е. $A = B$.
Шкала, соответствующая такому случаю, называется шкалой слабого порядка.
Иная ситуация возникает, когда имеются пары классов, несравнимые между собой, т. е. ни $A \geq B$, ни $B \leq A$. В таком случае говорят о шкале частичного порядка. Шкалы частичного порядка часто возникают в социологических исследованиях субъективных предпочтений. Например, при изучении покупательского спроса субъект часто не в состоянии оценить, какой именно из двух разнородных товаров ему больше нравится (например, клетчатые носки или фруктовые консервы, велосипед или магнитофон и т. д.); затрудняется человек и упорядочить по предпочтению любимые занятия (чтение литературы, плавание, вкусная еда, слушание музыки).

Характерной особенностью порядковых шкал является то, что отношение порядка ничего не говорит о дистанции между сравниваемыми классами. Поэтому порядковые экспериментальные данные, даже если они изображены цифрами, нельзя рассматривать как числа. Например, нельзя вычислять выборочное среднее порядковых измерений.
Пример. Рассматривается испытание умственных способностей, при котором измеряется время, затрачиваемое испытуемым на решение тестовой задачи. В таких экспериментах время хотя и измеряется в числовой шкале, но как мера интеллекта принадлежит порядковой шкале.
Порядковые шкалы определяются только для заданного набора сравниваемых объектов, у этих шкал нет общепринятого, а тем более абсолютного стандарта.
Примеры:
1. При определенных условиях правомерно выражение «первый в мире, второй в Европе» — просто чемпион мира занял второе место на европейских соревнованиях.
2. Само расположение шкал является примером порядковой шкалы.

Модифицированные порядковые шкалы
Опыт работы с сильными числовыми шкалами и желание уменьшить относительность порядковых шкал, придать им хотя бы внешнюю

независимость от измеряемых величин побуждают исследователей к различным модификациям, придающим порядковым шкалам некоторое (чаще всего кажущееся) усиление. Кроме того, многие величины, измеряемые в порядковых (принципиально дискретных) шкалах, имеют действительный или мыслимый непрерывный характер, что порождает попытки модификации (усиления) таких шкал. При этом иногда с полученными данными начинают обращаться как с числами, что приводит к ошибкам, неправильным выводам и решениям.

Примеры:

1. В 1811 г. немецкий минералог Ф. Моос предложил установить стандартную шкалу твердости, постулируя только десять ее градаций. За эталоны приняты следующие минералы с возрастающей твердостью: 1 — тальк; 2 — гипс; 3 — кальций, 4 — флюорит, 5 — апа-тит, б — ортоклаз, 7 — кварц, 8 — топаз, 9 — корунд, 10 — алмаз. Из двух минералов тверже тот, который оставляет на другом царапины или вмятины при достаточно силь-ном соприкосновении. Однако номера градаций алмаза и апатита не дают основания утверждать, что алмаз в два раза тверже апатита.
2. В 1806 г. английский гидрограф и картограф адмирал Ф. Бофорт предложил балльную шкалу силы ветра, определяя ее по характеру волнения моря: 0 — штиль (безветрие), 4 — умеренный ветер, 6 — сильный ветер, 10 шторм (буря), 12 — ураган.
3. В 1935 г. американский сейсмолог Ч. Рихтер предложил 12-балльную шкалу для оценки энергии сейсмических волн в зависимости от последствий прохождения их по данной территории. Затем он развил метод оценки силы землетрясения в эпицентре по его магнитуде (условная величина, характеризующая общую энергию упругих колебаний, вызванных землетрясением или взрывами) на поверхности земли и глубине очага.

Интервальная шкала (она же Шкала разностей)

Здесь происходит сравнение с эталоном. Построение такой шкалы позволяет большую часть свойств существующих числовых систем приписывать числам, полученным на основе субъективных оценок. Например, построение шкалы интервалов для реакций. Для данной шкалы допустимым является линейное преобразование. Это позволяет

приводить результаты тестирования к общим шкалам и осуществлять, таким образом сравнение показателей. Пример: шкала Цельсия.

Начало отсчёта произвольно, единица измерения задана.
Допустимые преобразования — сдвиги. Пример: измерение времени.
Распознаваемые объекты распознаются по значения их признаков. Эти значания определяются, как правило, в определенных (заданных) интервалах. Сами интервалы задаются шкалами (на шкалах) интервалов.
Свойства шкал интервалов будем определять аксиомами A_1—A_5 в случае, если упорядоченное множество представляет собой множество дейсвительных чисел. При этом оказывается возможным упорядочить интервалы между точками шкалы порядка. Иногда шкалы интервалов будем называть дважды упорядоченными шкалами, т е.

$$\gamma(y_k) - \gamma(y_m) = \gamma(y_l) - \gamma(y_n) = \ldots =$$

$$= y_k - y_m = y_l - y_n = \ldots = n_k - n_m = n_l - n_n = \ldots$$

Шкалы интервалов, к сожалению, не обладают таким важным свойством аддитивности, которое определено аксиомами A_6—A_9. Это значит, что на шкалах интервалов нельзя применять ни одно из основных арифметических действий, поскольку вычитание, умножение и деление есть частные случаи сложения.

В необходимых случаях на шкалах интервалов будем выбирать произвольный нуль, что позволит рассматривать разности как абсолютные величины, обладающие свойством аддитивности. Календарное время и высота над уровнем моря суть шкалы интервалов, но с ними обычно обращаются как со шкалами отношений, так как общепринято соглашение о нуле (уровень моря и т. п.).

На шкалах интервалов можно выполнять те же операции, что и на шкалах наименований и порядка, а кроме того, процедуры вычисления **математического ожидания, стандартного отклонения, коэффициента асимметрии и смешанных моментов**. Не имеет смысла только одна операция: **определение отношения стандартного отклонения к математическому ожиданию, так как эта величина зависит от положения нулевой точки.**

На рис. 6 символом $r_{\Delta y}{}^2$ обозначено отношение порядка на множестве различий по некоторому качеству *y,* что соответствует случаю, когда

$$r_{\Delta n}{}^2(n_k, n_m) = r_{\Delta n}(n_l, n_n),$$

т. е. имеет место

$$r^1(r_{\Delta n}{}^2(n_k, n_m), r_{\Delta n}{}^2(n_l, n_m)).$$

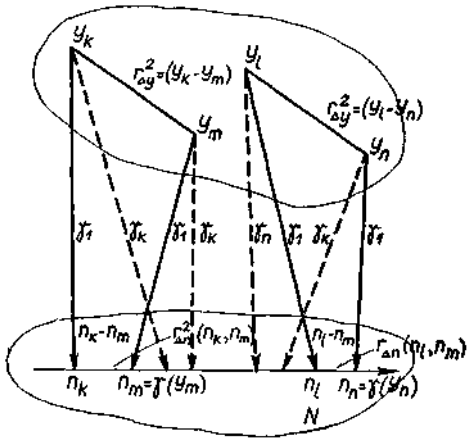


Рис 6. Отношения в шкале интервалов.

В шкале интервалов присутствуют упорядоченность и интервальность, но нет нулевой точки. Шкалы могут иметь произвольные начала отсчета, а связь между показаниями в таких шкалах является линейной:
y = ax + b,
где a > 0; — ∞ < b < ∞.
Для этой шкалы справедливо следующее свойство:

$$\frac{x_1 - x_2}{x_3 - x_4} = \frac{y_1 - y_2}{y_3 - y_4} = const$$

Примеры:
1. Температура, время, высота местности — величины, которые по физической природе либо не имеют абсолютного нуля, либо допускают свободу выбора в установлении начала отсчета.
2. Часто можно услышать фразу: «Высота … над уровнем моря». Какого моря? Ведь уровень морей и океанов разный, да и меняется со временем. В России высоты точек земной поверхности отсчитывают от среднемноголетнего Уровня Балтийского моря в районе Кронштадта.
В этой шкале только интервалы имеют смысл настоящих чисел и только над интервалами следует выполнять арифметические операции. Если произвести арифметические операции над самими отсчетами по шкале, забыв об их относительности, то имеется риск получить

бессмыс-ленные результаты.
Пример. Нельзя сказать, что температура воды увеличилась в два раза при ее нагреве от 9 до 18° по шкале Цельсия, поскольку для того, кто привык пользоваться шкалой Фаренгейта, это будет звучать весьма странно, так как в этой шкале температура воды в том же опыте изменится от 37 до 42°.

Частным случаем интервальных шкал являются **шкалы разностей:** циклические (периодические) шкалы, шкалы, инвариантные к сдвигу. В такой шкале значение не изменяется при любом числе сдвигов.
$y = x + nb$,
$n = 0, 1, 2,\ldots$
Постоянная b называется периодом шкалы.
Примеры. В таких шкалах измеряется направление из одной точки (шкала компаса, роза ветров и т. д.), время суток (циферблат часов), фаза колебания (в градусах или радианах).
Однако соглашение о хотя и произвольном, но едином для нас начале отсчета шкалы позволяет использовать показания в этой шкале как числа, применять к нему арифметические действия (до тех пор пока кто-нибудь не забудет об условности нуля, например при переходе на летнее время или обратно).

Абсолютная шкала (она же Шкала отношений)

Это интервальная шкала, в которой присутствует дополнительное свойство — естественное и однозначное присутствие нулевой точки. Пример: число людей в аудитории. В шкале отношений действует отношение «во столько-то раз больше». Это единственная из четырёх шкал имеющая абсолютный ноль. Нулевая точка характеризует отсутствие измеряемого качества. Данная шкала допускает преобразование подобия (умножение на константу). Определение нулевой точки — сложная задача для психологических исследований, накладывающая ограничение на использование данной шкалы. С помощью таких шкал могут быть измерены масса, длина, сила, стоимость (цена). Пример: шкала Кельвина (температур, отсчитанных от абсолютного нуля, с выбранной по соглашению специалистов единицей измерения — Кельвин).

Для описания свойств предложенных выше шкал введем шкалу отношений.

Рассматриваемая ниже шкала отношений обладает всеми свойствами описанных шкал, а также важным свойством аддитивности, определяемым аксиомами A_6—A_9. Изменение шкалы не

изменяет отношения результатов одного измерения к другому. Другими словами, на шкале отношений величина y подвергается преобразованию

$$y=cx,$$

где c — любой ненулевой скаляр.

Нуль шкалы отношений естествен. Вес, объем, давление, скорость и другие физические величины измеряются по шкалам отношений. На шкалах отношений можно выполнять все арифметические операции.

Введем следующие определения.

Определение 6. Группа Q_p положительных линейных преобразований из N на N состоит из всех преобразований вида

$$q_{\alpha,\beta}: y \to \alpha y + \beta,$$

где $\alpha \in N^+$, $\beta \in N$ (N^+— множество всех положительных чисел).

Группу

$$Q=\{ q_{\alpha,0}: \alpha \in N^+\}$$

будем называется группой растяжений, а

$$Q_S=\{ q_{1,\beta}: \beta \in N\}$$

— группой сдвигов.

Определение 7. Шкалу будем называть шкалой интервалов, если она единственна с точностью до положительных линейных преобразований (растяжений).

Пример шкалы отношений приведен на рис. 7.

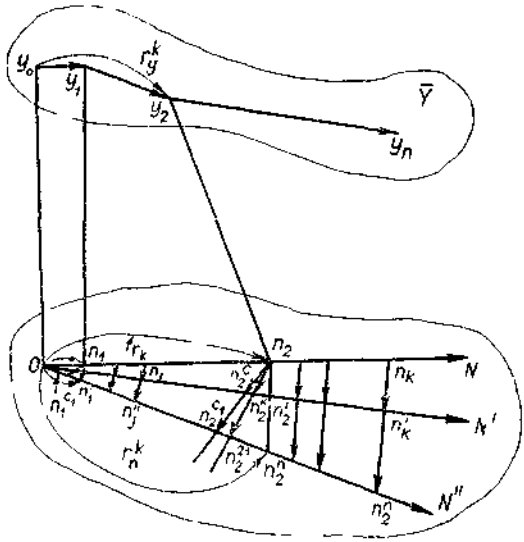


Рис. 7. Пример шкалы отношений.

Таким образом, важно отметить, что рассмотренные шкалы позволяют отобразить на множество чисел определенные отношения на множестве свойств и характеристик распознаваемых объектов. В этой связи представляется возможность упорядочивать множество распознаваемых объектов по множеству показателей.

Абсолютная шкала

Абсолютная (метрическая) шкала имеет и абсолютный нуль (b = 0), и абсолютную единицу (a = 1). В качестве шкальных значений при измерении количества объектов используются натуральные числа, когда объекты представлены целыми единицами, и действительные числа, если кроме целых единиц присутствуют и части объектов. Именно такими качествами обладает числовая ось, которую естественно называть абсолютной шкалой.
Важной особенностью абсолютной шкалы по сравнению со всеми остальными является отвлеченность (безразмерность) и абсолютность ее единицы. Указанная особенность позволяет производить над показаниями абсолютной шкалы такие операции, которые недопустимы для показаний других шкал, — употреблять эти показания в качестве показателя степени и аргумента логарифма.
Примеры:
1. Абсолютные шкалы применяются, например, для измерения количества объектов, предметов, событий, решений и т. п.
2. Примером абсолютной шкалы также является шкала температур по Кельвину.
Числовая ось используется как измерительная шкала в явной форме при счете предметов, а как вспомогательное средство присутствует во всех остальных шкалах.

Из рассмотренных шкал первые две являются неметрическими, а остальные — метрическими.

Шкалирование

Шкалирование представляет собой отображение какого-либо свойства объекта или явления в числовом множестве.
Можно сказать, что чем сильнее шкала, в которой производятся измерения, тем больше сведений об изучаемом объекте, явлении,

процессе дают измерения. Поэтому так естественно стремление каждого исследователя провести измерения в возможно более сильной шкале. Однако важно иметь в виду, что выбор шкалы измерения должен ориентироваться на объективные отношения, которым подчинена наблюдаемая величина, и лучше всего производить измерения в той шкале, которая максимально согласована с этими отношениями. Можно измерять и в шкале более слабой, чем согласованная (это приведет к потере части полезной информации), но применять более сильную шкалу опасно: полученные данные на самом деле не будут иметь той силы, на которую ориентируется их обработка. Иногда же исследователи усиливают шкалы; типичный случай — «оцифровка» качественных шкал: классам в номинальной или порядковой шкале присваиваются номера, с которыми дальше «работают» как с числами. Если в этой обработке не выходят за пределы допустимых преобразований, то «оцифровка» — это просто перекодировка в более удобную (например, для ЭВМ) форму. Однако применение других операций сопряжено с заблуждениями, ошибками, так как свойства, навязываемые подобным образом, на самом деле не имеют ме-ста.

По мере развития соответствующей области знания тип шкалы может меняться.

Пример. Температура сначала измерялась по порядковой шкале (холоднее — теплее), затем — по интервальным шкалам (Цельсия, Фаренгейта, Реомюра), а после открытия абсолютного нуля температур — по абсолютной шкале (Кельвина).

Основные положения

1. В основе любого наблюдения и анализа лежат измерения, которые представляют собой алгоритмические операции: данному наблюдаемому состоянию объекта ставится в соответствие определенное обозначение: число, номер или символ. Множество таких обозначений, используемых для регистрации состояний наблюдаемого объекта, называется измерительной шкалой.
2. В зависимости от допустимых операций на измерительных шкалах их различают по их силе.
3. Самой слабой шкалой является номинальная шкала, представляющая собой конечный набор обозначений для никак не связанных между собой состояний (свойств) объекта.
4. Следующей по силе считается порядковая шкала, дающая возможность в каком-то отноше-нии сравнивать разные классы наблюдаемых состояний объекта, выстраивая их в определенном порядке. Различают шкалы простого, слабого и частичного порядка.

Численные значения порядковых шкал не должны вводить в заблуждение относительно допустимости математических операций над ними.
5. Еще более сильная шкала — шкала интервалов, в которой кроме упорядочивания обозначений, можно оценить интервал между ними и выполнять математические действия над этими интервалами. Разновидностью шкалы интервалов является шкала разностей или циклическая.
6. Следующей по силе идет шкала отношений. Измерения в такой шкале являются «полноправными» числами, с ними можно выполнять любые арифметические действия (правда, при условии однотипности единиц измерения).
7. И, наконец, самая сильная шкала — абсолютная, с которой можно выполнять любые математические действия без каких-либо ограничений.
8. Отображение какого-либо свойства объекта или явления в числовом множестве называется шкалированием. Чем сильнее шкала, в которой производятся измерения, тем больше сведений об изучаемом объекте, явлении, процессе дают измерения. Однако применять более сильную шкалу опасно: полученные данные на самом деле не будут иметь той силы, на которую ориентируется их обработка. Лучше всего производить измерения в той шкале, которая максимально согласована с объективными отношениями, которым подчинена наблюдаемая величина. Можно измерять и в шкале, более слабой, чем согласованная, но это приведет к потере части полезной информации.

3. Элементы, виды, свойства шкал

Элементы шкалы

- Отметка шкалы — знак на шкале (чёрточка, зубец, точка и т.д.), соответствующий некоторому значению физической величины.
- Числовая отметка шкалы — отметка шкалы, у которой проставлено число.
- Нулевая отметка — отметка шкалы, соответствующая нулевому значению измеряемой величины.
- Деление шкалы — промежуток между двумя соседними отметками шкалы.
- Длина деления шкалы — расстояние между осями (или центрами) двух соседних отметок шкалы, измеренное вдоль

воображаемой линии, проходящей через середины самых коротких отметок шкалы.
- Цена деления шкалы — разность значений величины, соответствующих двум соседним отметкам шкалы.
- Длина шкалы — длина линии, проходящей через центры всех самых коротких отметок шкалы и ограниченной начальной и конечной отметками. Линия может быть реальной или воображаемой, кривой или прямой.

Интервал деления шкалы (деление шкалы) — расстояние между осями симметрии двух рядом лежащих штрихов (выражается в линейных или в угловых единицах)

Виды шкал средств измерений

- **Односторонняя шкала** — шкала с нулевой отметкой, расположенной в начале или в конце шкалы
- **Двусторонняя шкала** — шкала с нулевой отметкой, расположенной между начальной и конечной отметками. Различают симметричные (начальная и конечная отметки соответствуют одинаковым значениям измеряемой величины) и несимметричные двусторонние шкалы (начальной и конечной отметкам соответствуют разные значения).

Свойства шкал

- Начальное значение шкалы — наименьшее значение измеряемой величины, которое может быть отсчитано по шкале средства измерений. Во многих случаях шкала начинается с нулевой отметки, однако могут быть и другие значения — например, у медицинского термометра это 34,3 °C.
- Конечное значение шкалы — наибольшее значение измеряемой величины, которое может быть отсчитано по шкале средства измерений.
- Характер шкалы — функциональная зависимость $\boldsymbol{a = f(x)}$ между линейным (или угловым) расстоянием **a** какой-либо отметки от начальной отметки шкалы, выраженным в долях всей длины шкалы, и значением **x** измеряемой величины, соответствующим этой отметке:

- Равномерная шкала — шкала, отметки на которой нанесены равномерно.
- Неравномерная шкала — шкала, отметки на которой нанесены неравномерно.
- Логарифмическая или гиперболическая шкала — шкала с сужающимися делениями, характеризуемыми тем, что отметка, соответствующая полусумме начального и конечного значений, расположена между 65 и 100 % длины шкалы. Следует заметить, что выражение «логарифмическая шкала» используется и по отношению к другому значению понятия «шкала» (см.: Шкала физической величины, Логарифмический масштаб).
- Степенная шкала — шкала с расширяющимися или сужающимися делениями, но не подпадающая под определение логарифмической (гиперболической) шкалы.

Свойства шкал согласно классификации Стэнли Смита Стивенса

С вопросом о типе шкалы непосредственно связана проблема адекватности методов математической обработки результатов измерения. В общем случае адекватными являются те статистики, которые инвариантны относительно допустимых преобразований используемой шкалы измерений.

Типы шкал и их свойства согласно классификации Стэнли Смита Стивенса				
	Номинальная шкала	**Порядковая** шкала	**Интервальная** шкала	шкала **Отношений**
Логические / математические операции × ÷	X	X	X	*
+ -	X	X	*	*

 >	X	*	*	*
= ≠	*	*	*	*
Примеры: *Дихотомические и недихотомические* переменные	*Дихотомические:* Пол (мужской vs. женский) *Недихотомические:* Национальность (американец/китаец/ и т.д)	*Дихотомические:* Состояние здоровья (здоровый vs. больной), Красота (красивый vs.уродливый) *Недихотомические:* Мнение ('полностью согласен'/ 'скорее согласен'/ 'скорее несогласен'/ 'полностью несогласен')	Дата (с 1457 до н. э. до 2013 н.э) Широта (от +90° до −90°) Температура (от 10 °C до 20 °C)	Возраст (от 0 до 99 лет)
Мера центральной тенденции	Мода	Медиана	Среднее арифметическое	Среднее геометрическое
Метрическая или нет	Неметрическая (качественная)	Неметрическая (качественная)	Метрическая (количественная)	Метрическая (количественная)

Критика типологии Стивенса

Анализируя различные типы шкал Ф. Н. Ильясов приходит к выводу, что номинальная и интервальная шкала являются исследовательскими артефактами.

Хотя типология Стивенса все еще широко применима, она до сих пор является объектом критики теоретиков, в частности в случае с номинальной и порядковой шкалой.

Основные моменты критики шкал Стивенсона:

- Сведéние выбора только к тем статистическим методам, которые «демонстрируют инвариантность, подходящую для данного типа шкалы», представляется опасным для анализа данных практикой.
- Его таксономия слишком строга, чтобы ее возможно было применять для реальных данных.
- Стивенсовские ограничения часто ведут к понижению уровня данных через их преобразование в ранги и последующее ненужное обращение к непараметрическим методам.

Лорд критиковал аргументы Стивенса, показав, что выбор допустимых статистических тестов для некоторого набора данных не зависит от проблем репрезентации или единственности, а зависит от осмысленности.

Бейкер, Хардик и Петринович, а также Боргатта и Борнштедт подчеркнули тот факт, что следование Стивенсовским ограничениям часто заставляет исследователей прибегать к ранговому упорядочению данных и тем самым отказываться от использования параметрических тестов. К сожалению, их аргументация носила скорее ad hoc характер и завершалась предложением использовать стандартные параметрические процедуры вместо того, чтобы связываться с проблемой робастности.

Гуттман в более общем смысле доказывал, что статистическая интерпретация данных зависит от того, какой вопрос обращен к данным и какое доказательство мы готовы принять в ответ на этот

вопрос. Он определил это доказательство в терминах функции потерь, выбранной для проверки качества модели.

Джон Тьюки также критиковал стивенсовские ограничения как опасные для хорошего статистического анализа. Подобно Лорду и Гуттману, Тьюки отметил важность смысла данных при определении и шкалы, и подходящего способа анализа. Поскольку шкальные типы Стивенса абсолютны, в ситуации когда, например, данные нельзя считать полностью интервальными, их следует понизить в ранге до ординальных.

Даже сам Стивенс оговаривался, замечая: «Фактически большая часть шкал, широко и эффективно применяемых психологами, – это шкалы порядка. Обычные статистики, включая средние и стандартные отклонения, при строгом подходе не должны использоваться при работе с этим шкалами, однако такому неправомочному использованию может быть дано известное прагматическое оправдание: во многих случаях оно приводит к плодотворным результатам»

Дункан (1986) возразил против употребления слова “измерение” в описании номинальной шкалы, но Стивенс (1975) после дал собственное определения **“измерения” которое звучит, как “приписывание признака по какому-либо правилу.** Единственное правило, которое не может быть использовано для этих целей - случайность приписывания”. Однако, так называемое **“номинальное измерение” включает оценочное суждение исследователя, а возможные трансформации этого измерения бесконечны.** Это одно из замечаний, сделанных Лордом в 1953 году в сатиристической статье *On the Statistical Treatment of Football Numbers.*

Использование "среднего" в качестве меры центральной тенденции для порядкового типа по-прежнему спорно среди тех, кто принимает типологию Стивенса. Не смотря на это, многие ученые, занимающиеся поведенческими исследованиями, используют среднее для порядковых данных. Обычно это оправдывают тем, что порядковый тип в поведенческих науках находится где-то между истинным порядковым и интервальным типами. Хотя разница интервалов между двумя порядковыми разрядами не является постоянной, она зачастую имеет тот же порядок.

К примеру, применение измерительных моделей в образовательном контексте показывает, что общие оценки имеют довольно линейную зависимость с измерениями в пределах диапазона оценки. Таким образом, некоторые утверждают, что пока разница интервалов между порядковыми разрядами не очень большая, статистические данные интервальных шкал (к примеру "средняя") может иметь значимый результат для порядковых шкал. Программное обеспечение для статистического анализа (например SPSS) требует от пользователя указание соответствующего класса измерений для каждой переменной. Это гарантирует, что непреднамеренные ошибки пользователя не приведут к бессмысленному анализу (пример: анализ корреляции с номинальной переменной).

Терстоун добился прогресса в разработке обоснования для получения интервального типа, основанного на законе сравнительного суждения. Общим применением закона является аналитический процесс иерархии. Геогр Раш (англ.) достиг дальнейшего прогресса, разработав вероятностную модель Rasch model (англ.), которая дает теоретическую основу и обоснование для получения интервальных измерений из подсчета наблюдений (например общее количество баллов по оценкам).

Несмотря на всю критику, в широком диапазоне ситуаций опыт показывает, что применение запрещенных статистик к данным приводит к научно значимым результатам, важным при принятии решений и ценным для дальнейших исследований.

4. Типологии шкал

Существуют иные типологии, отличные от Стивенса. К примеру: Mostller Mosteller и Tukey (1977), Nelder (1990) создали описание непрерывного отсчета, непрерывных отношений и категориальных моделях данных. См. также: Chrisman (1998), van den Berg (1991).

Типология Мостеллера и Тьюки (1977)

Mosteller and Tukey заметили, что 4 уровня недостаточно и предложили следующее деление:[13]

1. Имена

2. Оценочные суждения (e.g. новичок, второкурсник etc.)
3. Оценки ограниченные 0 и 1
4. Счетные (положительные целые числа)
5. Натуральные (положительные вещественные числа)
6. Сбалансированные (любые вещественные числа)

Например, проценты (вариант фракций в терминах Мостлера-Тьюки) не подходят к теории Стивенса, так как не существует полностью допустимых трансформаций.[7]

Типология Крисмана (1998)

Николас Крисман предложил расширенный поиск уровней измерения для учета разных измерений, которые не обязательно соответствуют традиционным представлениям уровней измерения. Измерения, связанные с диапазоном и повторением (к примеру радиальные градусы по кругу, часы и тд), градуированные категории членства и другие типа измерений, не соответствуют оригинальной работе Стивена, приводящие к внедрению 6-ти новых уровней измерения к существующим 10-ти:

1. Номинальная
2. Градуированное членство
3. Порядковая
4. Интервальная
5. Интервальная логарифмическая
6. Экстенсивное отношение
7. Циклическое отношение
8. Производное отношение
9. Счетная
10. Абсолютная

Расширенные уровни измерений редко используются вне академической географии.

Типы шкал и “операционная теория измерения” Стивенса

Теория типов шкал это своеобразная "интеллектуальная служанка" операционной теории измерения Стивенса, которая стала окончательной в психологии и поведенческих науках, несмотря на

критику Мичелла за противоречивость с измерениями в естественных науках (Michell, 1999). На самом деле, операционная теория измерения была реакцией на выводы комитета, созданного British Association for the Advancement of Science (англ.) в 1932 для изучения возможности подлинных научных измерений в психологических и поведенческих науках. Этот комитет, который стал известен как "Комитет Фергюсона", опубликовал окончательный отчет (Ferguson, et al., 1940, p. 245), в котором шкала Стивенса сон (Stevens & Davis, 1938) была объектом критики.

> *…любо закон имеющий целью выразить количественное отношение между интенсивностью ощущения и интенсивностью стимула не только ложный, но и фактически не имеющий смысла до тех пор, пока смысл не обретет понятие сложения, примененное к ощущению.*

Значит, если шкала сонов Стивенса действительно измеряет интенсивность ощущений аудитории, должно быть произведено доказательство того, что эти ощущения являются количественными атрибутами. Необходимым доказательством было присутствие “аддитивных структур” – концепт, разработанный немецким математиком Отто Холдером (Hölder, 1901). В условиях доминации физика и теоретика измерений Нормана Роберта Кампбелла (англ.) в обсуждении фергюсонского комитета, было постановлено, что измерения в социальных науках невозможны из-за отсутствия операции конкатенации. Впоследствии это решение было признано неверным после разработки теории совместных измерений Дебрю, а также независимо Люсом и Тьюки. Однако Стивенс хотел не введения дополнительных экспериментов для обнаружения аддитивных структур, а признания решения фергюсонского комитета полностью недействительным путем предложения новой теории измерений.

> *Перефразируя Н.Р. Кампбела (Final Report, p.340), можно сказать что измерение, в самом широком смысле, определяется как присваивание чисел объектам и события согласно некоторому правилу (Стивенс, 1946, p.677).*

Огромное влияние на Стивенса оказали идеи другого гарвардского академика, лауреата нобелевской премии, физика Перси Бриджмена (1927), чью доктрину “Операционизм” Стивенс использовал для определения термина “измерение”. К примеру, в определении

Стивенса используется рулетка, которая определяет длину (объект измерения) как измеримую (следовательно количественную). Критики операционализма возражают, что он смешивает отношения между двумя объектами или событиями для свойств одного из объектов или событий (Hardcastle, 1995; Michell, 1999; Moyer, 1981a,b; Rogers, 1989).

Канадский исследователь измерительных теорий William Rozeboom (1966) был одним из первых критиков, резко высказавшихся против теории типов шкал Стивенса

Тип переменной зависит от контекста

Еще одна проблема может заключаться в том, что одна и та же переменная может иметь разные типы шкал в зависимости от способа её измерения и целей анализа. Например, цвет волос обычно считается номинальной переменной, так как не имеет определенного порядка.[15] Тем не менее, расположить цвета в определенном порядке возможно несколькими способами, в том числе и по оттенкам, с помощью колориметрии.

Использование в психометрии

Используя различные шкалы, можно производить различные психологические измерения[16]. Самые первые методы психологических измерений были разработаны в психофизике. Основной задачей психофизиков являлось то, каким образом определить, как соотносятся физические параметры стимуляции и соответствующие им субъективные оценки ощущений. Зная эту связь, можно понять, какое ощущение соответствует тому или иному признаку. Психофизическая функция устанавливает связь между числовым значением шкалы физического измерения стимула и числовым значением психологической или субъективной реакцией на этот стимул.

Некоторые распространённые шкалы

- Температурные шкалы разных стран и времён (Цельсия, Фаренгейта, Кельвина и др.)

- Шкала Рихтера
- Шкала Бофорта скорости ветра
- Шкала Мооса — шкала твёрдости минералов
- Цветовая палитра, Атлас цветов

5. Схемы оценки решений в ИИ

Множество оценок Z выхода Y процесса *P* или состояний объекта состоит из величин, измеренных в различных шкалах. Это не позволяет без дополнительных преобразований выполнять адекватные арифметические и другие операции без искажения информации, заложенной при первичных измерениях.

Одной из эффективных процедур, позволяющих устранить разнородность шкал и разный масштаб измерения показателей *Z*, является нормализация. Пусть множество показателей *Z* конечно и фиксированно. Возможность добавления хотя бы одного элемента к Z не рассматривается. Тогда общая схема процедуры нормализации может быть представлена следующим выражением:

$$\overline{Z}_i = \frac{z_i - z_i^{\min}}{z_i^{\max} - z_i^{\min}},$$

где $\overline{Z}_i$ — нормированное значение *i*-го показателя *z*; z_1 — нормируемое значение *i*-го показателя *z*; $z_i^{\min}$ — минимальное значение *i*-го показателя *z*; $z_i^{\max}$ — максимальное значение *i*-го показателя *z*. При этом наибольшее значение *z* для *i*-го показателя отобразится на множестве чисел в единицу, а наименьшее — в нуль, независимо от масштаба и типа шкалы измерения.

На рис. 8 показано, как состояния процесса P_1 и P_2 с помощью операций измерения $c_1, c_2, \ldots, c_k$ отображаются на шкалы $N_1, N_2, \ldots, N_k$ для соответствующих показателей $z_1, z_2, \ldots, z_k$.

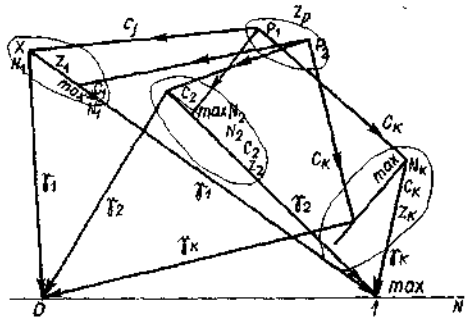


Рис. 8. Пример измерения состояния процесса .

С помощью операции нормализации γ разнородные шкалы отображаются на числовой интервал [0, 1] множества действительных чисел $\overline{N}$.

В результате операции нормализации для *m* возможных альтернатив A_1, A_2,…,A_m получается *k* числовых оценок z_1, z_2,…,z_k, соответствующих каждой оцениваемой альтернативе. Оценка z_{ij} соответствует *i*-й альтернативе по *j*-му показателю (табл. 1).

Таблица 1

Матрица оценок альтернатив

Альтернативы	Показатели			
	z_1	z_2 . . .	z_j . . .	z_k
a_1	z_{11}	z_{12} . . .	z_{1j} . .	z_{1k}
a_2	z_{21}	z_{22} . .	z_{2j} . .	z_{2k}
. . .	. . .	. . . . .	. . . . .	. . .
a_i	z_{i1}	z_{i2} . . .	z_{ij} . .	z_{ik}
. . .	. .	. . . . .	. . . . .	. . .
a_m	z_{m1}	z_{m2} . . .	z_{mj} . . .	z_{mk}

Предлагается несколько критериев выбора числовых оценок степения распознавания объектов для принятия оптимального решения для альтернатив, характеризующихся матрицей оценок, аналогичных приведенным в табл. 1. Рассмотрим некоторые из них.

Максимальный критерий Вальда, или принцип гарантированного результата, в качестве оптимальной выбирает такую альтернативу *A,* при которой максимальное значение нормированного показателя максимально:

$$W = \max_i \min_j \overline{z}_{ij}.$$

Если ЛРО действует согласно этому критерию, то при равном приоритете всех показателей выбирается такая альтернатива, для которой «наихудший» показатель является «наилучшим» по сравнению с другими. Пользуясь критерием Вальда, ЛРО обеспечивает для себя наилучшие характеристики в случае пессимистической оценки степения распознаваемого объектов и больше внимания уделяет слабым показателям и неуспехам, чем достоинствам и сильным сторонам характеризуемого процесса распознавания.

Рассмотрим пример оценки решений по критерию Вальда. В табл. 2 приведены оценки альтернатив по трем показателям.

Таблица 2

Пример для оценки альтернатив по трем показателям

Альтернативы	Показатели		
	z_1	z_2	z_3
a_1	1,0	0,3	0,6
a_2	0	0,2	0,4
a_3	0,6	0,5	0,4
a_4	0,8	1,0	0,2

На первом этапе находится минимальное значение показателя для каждой строки и выписываются в столбец значения, соответствующие $\min \overline{z}_{ij}$. На втором этапе определяется максимальный элемент в столбце $\min \overline{z}_{ij}$. Для первой альтернативы это значение показателя z_2, равное 0,3. Для второй альтернативы a_2 минимальное значение, равное нулю, будет у первого показателя z_1. Для альтернативы a_3 на втором этапе замечаем, что минимальное нормированное значение показателя z_3, равное 0,4, является наибольшим по сравнению с другими альтернативами. Согласно критерию Вальда, a_3 определяется как оптимальная альтернатива, которая лучше альтернативы a_1. В свою очередь, альтернатива a_1 лучше альтернативы a_4. Наконец, альтернатива a_4 лучше, чем альтернатива a_2 (табл. 3).

Таблица 3

Пример оценки альтернатив по критерию Вальда

Альтернативы	Показатели			Критерий Вальда		
	z_1	z_2	z_3	$\min \overline{z}_{ij}$	$\max \min \overline{z}_{ij}$	Ранг
a_1	1,0	0,3	0,6	0,3	—	2
a_2	0	0,2	0,4	0	—	4
a_3	0,6	0,5	0,4	0,4	0,4	1
a_4	0,8	1,0	0,2	0,2	—	3

Рассмотрим второй критерий - критерий Гурвица.

Критерий Гурвица, иначе называемый критерием пессимизма-оптимизма, рекомендует при оценке альтернатив не руководствоваться ни крайним пессимизмом (всегда рассчитывать на худшее стечение обстоятельств и обращать внимание только на недостатки), ни крайним легкомысленным оптимизмом (считать что все обойдется лучшим образом, и обращать внимание только на успехи и достоинства альтернативы). Критерий Гурвица имеет вид

$$H = \max_i \{\alpha \min_j \overline{z}_{ij} + (1-\alpha) \max_i \overline{z}_{ij}\}. \qquad (1)$$

где α — коэффициент, значение которого заключено между нулем и единицей. Анализ выражения (1) показывает, что при α=1 критерий Гурвица превращается в пессимистический критерий Вальда, а при α=0 — в критерий «крайнего оптимизма». Последний рекомендует выбирать альтернативу, для которой существует такая наилучшая оценка показателя, что ее «не могут превзойти характеристики других альтернатив». При 0< α <1 получаем что-то среднее между крайним пессимизмом и крайним оптимизмом. Можно заметить, что коэффициент α обозначает как бы степень пессимизма ЛРО. Этот коэффициент назначается ЛРО из субъективных соображений и получает значение тем ближе к единице, чем больше ЛРО хочет «подстраховаться» в неприятных и опасных ситуациях.

Для альтернатив, заданных табл. 2, при α=0,5 в соответствии с критерием Гурвица принимается, что в равной степени рассматриваются как «сильные», так и «слабые» стороны альтернатив. При этом вычисленный по формуле критерий Гурвица принимает наибольшее значение, равное 0,65, для альтернативы a_1:

$$H(a_1) = \max\{0,5 \min z_{1j} + (1-0,5) \max z_{1j}\} =$$

$$\max_1\{0,5 \cdot 0,3 + 0,5 \cdot 1\} = 0,65.$$

Значения критерия Гурвица и ранги для других альтернатив приведены в табл. 4.

При сравнении результатов оценки альтернатив по критериям Гурвица и Вальда отметим различие в рангах для альтернатив a_1 и a_3. По критерию Вальда альтернатива a_3 предпочтительнее, чем альтернатива a_1. По критерию Гурвица уже не только альтернатива a_1, но и a_4 лучше, чем a_3.

Таблица 4

Пример оценки альтернатив по критерию Гурвица

Альтернативы	Показатели			Критерий Гурвица при $\alpha=0,5$						
	z_1	z_2	z_3	$\min z_{ij}$	$\max z_{ij}$	$\alpha\min z_{ij}$	$(1-\alpha)\times\max z_{ij}$	H	$\max H$	Ранг
a_1	1	0,3	0,6	0,3	1	0,15	0,5	0,65	0,65	1
a_2	0	0,2	0,4	0	0,4	0	0,2	0,20	—	4
a_3	0,6	0,5	0,4	0,4	0,6	0,20	0,3	0,50	—	3
a_4	0,8	1	0,2	0,2	1	0,10	0,5	0,60	—	2

Рассмотрим очередной критерий – критерий Лапласа.

По критерию Лапласа учитываются не только «лучшие» и «худшие» значения показателей оцениваемых альтернатив, но и все остальные, промежуточные значения. При этом оценка каждой альтернативы приводится как бы к среднему значению, обобщающему значения всех показателей для каждой альтернативы. Это среднее значение альтернативы будем вычислять по формуле

$$S_j = \frac{1}{n}\sum_{j=1}^{n} z_{ij},$$

где n — общее число показателей для i-й альтернативы.

В табл. 5 приведены оценки для альтернатив, вычисленные по критерию Лапласа.

Наибольшее значение критерия Лапласа ($S\approx0,67$) получает альтернатива a_4:

$$S_4 = \frac{1}{n}\sum_{j=1}^{n} z_{ij} = \frac{1}{3}(0,8+1+0,2) \approx 0,67.$$

Таблица 5

Пример оценки альтернатив по критерию Лапласа

Альтернативы	Показатели			Критерий Лапласа		
	z_1	z_2	z_3	Σz_{ij}	S_i	Ранг
a_1	1,0	0,3	0,6	1,9	0,63	2
a_2	0	0,2	0,4	0,6	0,20	4
a_3	0,6	0,5	0,4	1,5	0,50	3
a_4	0,8	1,0	0,2	2,0	0,67	1

Отметим при этом, что альтернатива a_4 лучше a_1, которая, в свою очередь, лучше, чем a_3. Таким образом, для каждого критерия выбора получается своя лучшая альтернатива, оказывающаяся не лучшей по другим критериям (табл. 6).

Таблица 6

Пример сравнения критериев оценки альтернатив

Альтернативы	Показатели			Ранги по критериям				
	z_1	z_2	z_3	Вальда V	Гурвица H	Лапласа S	ЛРО	
							Случай 1	Случай 2
a_1	1	0,3	0,6	2	1	2	2,4	1,2
a_2	0	0,2	0,4	4	4	4	2,4	4
a_3	0,6	0,5	0,4	1	3	3	1	3
a_4	0,8	1	0,2	3	2	1	2,4	1,2

На практике может иметь место случай, когда ЛРО предъявляется множество альтернатив и становится известен результат выбора лучшей. Для примера, приведенного в табл. 6, будем считать, что установлено только три вышеуказанных критерия оценки альтернатив, которыми может пользоваться ЛРО, но заведомо неизвестно, каким именно критерием оно пользовалось.

Если ЛРО выбрало одну альтернативу a_3, а остальные альтернативы остались неупорядоченными (случай 1 в табл. 6), то альтернативе a_3 присвоен ранг 1, а всем остальным — 2, 3 и 4, т. е. средний ранг равен 3. В первом приближении можно утверждать, что критерий, по которому ЛРО выбрало альтернативу, ближе к критерию Вальда, чем к другим.

Когда ЛРО выбрало как наилучшие и равные между собой алыерпативы a_1 и a_4, а затем определило, что a_3 лучше, чем a_2 (случаи 2 в табл. 6), то в первом приближении можно утверждать, что при определении схемы оценки решений ЛРО пользовалось скорее критериями Гурвица и Лапласа, чем Вальда.

Задача оценки адекватности поведения ЛРО по упорядочению множества альтернатив и схем оценки степени распознавания объектов реализуется с помощью методов ранговой корреляции. Это позволяет алгоритмическим путем определить схему оценки степени распознавания объектов для конкретной стуации распознавания объектов и конкретного ЛРО, используя методы векторной оптимизации.

Приложение 1.

Примеры решения задач

Задача 1. Определить доверительный интервал, в который с вероятностью *P* попадают значения измеряемой величины при условии, что:
а) результаты измерений распределены равномерно;
б) результаты измерений распределены по треугольному закону;
в) результаты измерений распределены по трапецеидальному закону.
Решение.
а) Плотность вероятности равномерного распределения зададим в виде:

$$P(x)=\begin{cases}\frac{1}{2}a & при\ \ x\in[m-a,m+a]\\ 0 & при\ \ x\notin[m-a,m+a]\end{cases},$$

где *m* - центр распределения. Можно было бы положить *m=0,* что не влияет на конечный результат, но мы рассмотрим общий случай. Установим связь между доверительной вероятностью *P* и доверительным интервалом Δ. Для этого запишем выражение для вероятности попадания случайной величины *x* в интервал $[m-\Delta,\ m+\Delta]$, где Δ - текущее значение интервала; $\Delta \le a$. Имеем по определению вероятности (см. ф. (2.5)):

$$P\{x\in[m-\Delta,m+\Delta]\}=\int_{m-\Delta}^{m+\Delta}p(x)\,dx\ .$$

Подставляя выражение для *p*(x) и преводя вычисления, получим:

$$P=\frac{\Delta}{a}\Rightarrow\Delta=P\cdot a\ ,$$

т.е. доверительный интервал линейно зависит от доверительной вероятности. В частности, при *P=1;* Δ*=a,* т.е. интервал равен области задания распределения в которой $p(x)\neq 0$. Отметим, что на полученный результат не влияет положение центра распределения, о чем говорилось выше.

б) Вероятность попадания случайной величины в интервал (-Δ,Δ) равна:

$$P\{-\Delta \le x \le \Delta\} = \int_{-\Delta}^{\Delta} p(x)\,dx = F(\Delta) - F(-\Delta), \quad (1)$$

где $p(x)$ - плотность треугольного распределения.

Для удобства вычисления будем считать, что распределение симметрично относительно 0 и определено в интервале $[-2a,\ 2a]$, что не влияет на конечный итог.

Воспользуемся формулой (2.52) для интегральной функции распределения и получим:

$$P(-\Delta \le x \le \Delta) = \frac{-\Delta^2}{8a^2} + \frac{\Delta}{2a} + \frac{1}{2} - \left(\frac{\Delta^2}{8a^2} - \frac{\Delta}{2a} + \frac{1}{2}\right) = -\frac{\Delta^2}{4a^2} + \frac{\Delta}{a} = \frac{\Delta}{a}\left(1 - \frac{\Delta}{4a}\right). \quad (2)$$

Из (2) следует, что вероятность квадратично зависит от величины интервала. В частности, при $\Delta = 2a$ имеем $p=1$, что очевидно.

в) В этом случае воспользуемся соотношением (2.56) для интегральной функции распределения, что дает:

$$P\{-\Delta \le x \le \Delta\} = 1 - \frac{(a+b-\Delta)^2}{8ab} - \frac{(a+b-\Delta)^2}{8ab} = 1 - \frac{(a+b-\Delta)^2}{4ab} \quad (3)$$

т. е. вероятность квадратично зависит от интервала.

Соотношение (3) справедливо, если левая граница интервала $[-\Delta,\ \Delta]$ заключена между $-(a+b)$ и $-(a-b)$, а правая между $(a-b)$ и $(a+b)$. Если же искомый интервал заключен между $-(a-b)$ и $(a-b)$, то имеем:

$$P\{-\Delta \le x \le \Delta\} = \frac{a+\Delta}{2a} - \frac{a-\Delta}{2a} = \frac{\Delta}{a} \quad (4)$$

т.е. результат совпадает с полученным выше для равномерного распределения, что не удивительно, так как в интервале $[-(a-b);\ (a-b)]$ трапецеидальное распределение совпадает с равномерным.

Задача 2. Даны результаты десяти измерений электрического сопротивления R_i Ом, резистора. Найти оценку истинного значения и ошибки измерения при следующих условиях:

а) результаты измерений распределены нормально;

б) результаты измерений распределены равномерно;

в) распределение неизвестно.

124; 122; 122; 121; 125; 125; 119; 123; 124; 121.

Решение.

а) распределение нормальное.
За оценку истинного значения принимается выборочное среднее:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

и так как n=10, то $\overline{x}_n = \frac{1}{10}\sum_{i=1}^{10} x_i = 122{,}60.$

Выборочная дисперсия равна:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2 = \frac{34{,}4}{9} = 3{,}82.$$

Дисперсия выборочного среднего равна:

$$S_{\overline{x}}^2 = \frac{S_n^2}{n} = \frac{3{,}82}{10} = 0{,}382.$$

СКО выборочного среднего:

$$S_{\overline{x}} = \sqrt{S_{\overline{x}}^2} = 0{,}618.$$

Определим доверительный интервал. Выберем вероятность, например, P=0,99 и по таблицам распределения Стьюдента при $k=n-1=9$ найдем: t=3,250. Тогда $\Delta = t \cdot S_{\overline{x}} = 3{,}250 \cdot 0{,}618 = 2{,}009 \approx 2{,}01.$
Результат измерения запишем в виде:

$$x = x \pm \Delta = 122{,}60 \pm 2{,}01.$$

б) распределение равномерное.
За оценку истинного значения принимается центр размаха:

$$x_R = \frac{x_{\min} + x_{\max}}{2} = \frac{119+125}{2} = 122.$$

Определим доверительный интервал (см. решение задачи 1), считая, чтораспределение сосредоточено в интервале $\left[x_R - a, x_R + a\right]$, где

$$a = \left(x_{\max} - x_{\min}\right)/2:$$

$$\Delta = \frac{P\left(x_{\max} - x_{\min}\right)}{2} = \frac{0{,}99\left(125 - 119\right)}{2} = 2{,}97.$$

Результат запишем в виде:

$$x = 122{,}00 \pm 2{,}97.$$

в) распределение неизвестно.

Для расчета доверительного интервала используем неравенство Чебышева:

$$P\left\{\left|x - M\left[x\right]\right| \le \varepsilon\right\} \ge 1 - \frac{\sigma_x^2}{\varepsilon^2}.$$

Отсюда при $P=0{,}99$ найдем:

$$\varepsilon^2 = \frac{\sigma_x^2}{0{,}01} = 100\sigma^2 \quad \Rightarrow \quad \varepsilon = 10\sigma_x.$$

Результат записывается в виде:

$$x = M[x] \pm 10\sigma_x$$

Оценки $M[x]$, σ_x в данном случае неизвестны. Полученные результаты показывают, что с уменьшением априорной информации о законе распределения, доверительный интервал возрастает.

Задача 3. При аттестации измерительного канала информационно-измерительной системы (ИИС) проводилось 11 измерений быстродействия, в мсек, канала. Нормативное значение быстродействия составляет m_0 =16 мсек. Требуется выяснить равно, занижено или завышено быстродействие канала по сравнению с нормативным. Доверительную вероятность принять равной P = 0,95.

17,9; 20,5; 18,3; 17,2; 14,0; 18,9; 18,3; 19,8; 19,6; 17,8; 22,1

Решение.

1) Сформулируем гипотезу:

$$H_0: \overline{x} = m_0; \quad \overline{H_0}: \overline{x} > m_0.$$

Критерий имеет вид:

$$T = \frac{\overline{x} - m_0}{S_n/\sqrt{n}}.$$

Проведем расчеты (n = 11):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{11} \cdot 204{,}4 = 18{,}582;$$

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \sqrt{\frac{1}{10} \cdot 42{,}724} = 2{,}067,$$

$$\hat{T} = \frac{18{,}58 - 16}{2{,}067/\sqrt{11}} = 2{,}94.$$

Определим критическую область. В данном случае выбирается односторонняя критическая область $K: \mid t \mid \geq \varepsilon.$

Определим ε по таблице распределения Стьюдента из соотношения

$$P\left\{\left|\hat{T}\right| < \varepsilon\right\} = 1 - \alpha.$$

При вероятности P=0,95 и числе степеней свободы $k=n$-1=10, получим ε =2,228.

Так как $\hat{T} > \varepsilon,$ то гипотеза H_0 отклоняется, т. е. мы установили, что измеренное значение быстродействия больше нормативного.

Задача 4. На выходе усилителя переменного тока проводилось две серии измерений. В первой серии было проведено 5 измерений электрического напряжения и получена оценка выборочной дисперсии $S_5^2 = 100 \ \ \text{мкВ}^2$, во второй серии - 9 измерений, и оценка дисперсии составила $S_9^2 = 80 \ \ \text{мкВ}^2.$ Среднее значение напряжения в обеих сериях $\bar{x} = 0.$ Требуется выяснить свидетельствуют ли данные измерения о том, что значение дисперсии в первой серии превышает значение дисперсии во второй серии на уровне значимости α = 0,1.

Решение. Задача решается в соответствии с рекомендациями §2.5. Сначала сформулируем гипотезу:

$$H_0 : S_{\bar{x}}^{2(1)} = S_{\bar{x}}^{2(2)}; \overline{H_0} : S_{\bar{x}}^{2(1)} > S_{\bar{x}}^{2(2)}.$$

Определим критерий в виде:

$$F = T = \frac{S_5^{\,2}}{S_9^{\,2}}.$$

При условии справедливости гипотезы H_0 критерий имеет F-распределение Фишера. Выберем одностороннюю критическую область $t > \varepsilon$. Тогда, если значение $\hat{T}$ попадает в критическую область, то гипотеза отвергается, и принимается в противном случае. Значение ε находится из соотношения $P\left\{\hat{T} < \varepsilon\right\} = 1 - \alpha$ по таблицам распределения Фишера. Для этого определяют критическое значение критерия $\varepsilon(P,\ k_1,\ k_2)$, где α—уровень значимости:

$k_1 = n_1 - 1;\ k_2 = n_2 - 1.$ Проведем расчеты. В нашем случае $T = 100/80 = 1{,}25;$ выберем α=0,05. По таблице определим

$\varepsilon(0{,}95;4;8) = 3{,}84$. Так как $\hat{T} = 1{,}25 < \varepsilon = 3{,}84,$ то гипотеза принимается, т. е обе серии принадлежат генеральной совокупности с одной и той же дисперсией.

Задача 5. Проводилось исследование передаточных характеристик аналогового преобразователя в стационарном режиме. Для этого на вход подавался сигнал *x,* а на выходе регистрировался сигнал *y.* Из теоретических соображений известно, что зависимость имеет вид полинома не выше 2-ой степени. Требуется определить вид функции преобразования *y=f(x),* оценить отклонение от линейности и максимальную ошибку преобразования в диапазоне изменения измеряемой величины. Рассмотреть два случая:

а) ошибка измерения неизвестна;

б) ошибка составляет 10%.

Результаты измерений представлены в табл.1.

Таблица 1

x	1	2	3	4	5	6	7	8	9
y/x	10,1	10,9	12,2	13,1	13,9	15,3	16,3	16,7	18

Решение.

Обозначим $y/x \equiv z$. Задача состоит в нахождении зависимости $z=f(x)$. Воспользуемся результатами §2.3. Начнем поиск зависимости с наименьшей степени полинома.

1. Модель I: $z_I = a_0$

а) Составим систему нормальных уравнений. В данном случае имеем одно уравнение.

$$a_0[x^0] = [zx^0] = [z] \Rightarrow a_0 = \frac{[z]}{[x_0]} = \frac{\sum_{i=1}^{N} z_i}{N}.$$

б) Найдем оценку параметра a_0:

$$\hat{a}_0 = \frac{\sum_{i=1}^{9} z_i}{9} = \frac{126{,}5}{9} = 14{,}06.$$

Модель I имеет вид: $z_I = \hat{a}_0 = 14{,}06$.

в) Найдем дисперсию оценки параметра $\hat{a}_0$ непосредственным вычислением, т. к. система вырожденная (состоит из одного уравнения).

$$S^2_{\hat{a}_0} = \frac{1}{N} S^2_{0I},$$

где $S^2_{0I} = \frac{1}{N-m} \sum_{i=1}^{N} \left(z_i - \hat{z}_i\right)^2$ - дисперсия ошибки эксперимента.

Полагая $N = 9$; $m = 1$, найдем:

$$S^2_{0I} = \frac{1}{9-1} \sum_{i=1}^{9} \left(z_i - \hat{z}_i\right)^2 = 7{,}39,$$

$$S^2_{\hat{a}_0} = \frac{7{,}39}{9} = 0{,}82 \Rightarrow S_{\hat{a}_0} = 0{,}90.$$

2. Усложним модель, увеличив степень полинома.

Модель II:

$$z_{II} = a_0 + a_1 x$$

а) Запишем систему нормальных уравнений:

$$Na_0 + a_1[x] = [z]$$
$$a_0[x] + a_1[x^2] = [zx]$$

б) Найдем оценки параметров,

$$\hat{a}_0 = \frac{\Delta_0}{\Delta},$$

где $\Delta = \begin{vmatrix} N & [x] \\ [x] & [x^2] \end{vmatrix} = N[x^2] - [x][x];$

$$\Delta_0 = \begin{vmatrix} [z] & [x] \\ [zx] & [x^2] \end{vmatrix} = [z][x^2] - [x][zx],$$

где

$$[x] = \sum_{i=1}^{N} x_i = \sum_{i=1}^{9} x_i = 45; \quad [x^2] = \sum_{i=1}^{N} x_i = 285;$$

$$[z] = \sum_{i=1}^{N} z_i = 126{,}5; \quad [zx] = \sum_{i=1}^{N} x_i z_i = 691{,}9.$$

Получим:

$\Delta = 9 \cdot 285 - 45 \cdot 45 = 2565 - 2025 = 540;$

$\Delta_0 = 126{,}5 \cdot 285 - 45 \cdot 691{,}9 = 36052{,}5 - 31135{,}5 = 4917$

$$\hat{a}_0 = \frac{4917}{540} = 9{,}106$$

$$\hat{a}_1 = \frac{\Delta_1}{\Delta};$$

где $\Delta_1 = \begin{vmatrix} N & [z] \\ [x] & [zx] \end{vmatrix} = \begin{vmatrix} 9 & 126{,}5 \\ 45 & 691{,}9 \end{vmatrix} = 6227{,}1 - 5692{,}5 = 534{,}6;$

$$\hat{a}_1 = \frac{\Delta_1}{\Delta} = \frac{534{,}6}{540} = 0{,}99.$$

в) Выпишем модель II:

$$z_{II} = 9{,}106 + 0{,}99x$$

г) Найдем дисперсию оценок параметров:

$$S^2_{\hat{a}_0} = \frac{\left[x^2\right]}{\Delta} \cdot S^2_{0\text{II}} \; ; \; S^2_{\hat{a}_1} = \frac{N}{\Delta} \cdot S^2_{0\text{II}},$$

где $S^2_{0\text{II}} = \frac{1}{N-2} \sum_{i=1}^{N} \left[z_i - \left(\hat{a}_0 + \hat{a}_1 x_i\right)\right]^2$

Расчеты дают:

$$S^2_{0\text{II}} = 0{,}4524 \Rightarrow S_{0\text{II}} = 0{,}67$$

$$S^2_{\hat{a}_0} = \frac{285}{540} \cdot 0{,}4524 = 0{,}239 \Rightarrow S_{\hat{a}_0} = 0{,}49$$

$$S_{\hat{a}_1} = \frac{9}{540} \cdot 0{,}4524 = 0{,}0075 \Rightarrow S_{\hat{a}_1} = 0{,}086$$

Определим доверительные интервалы для оценок параметров. Зададим значение доверительной вероятности P=0,95. По таблицам распределения Стьюдента найдем:
для модели I при

$$k = 9 - 1 = 8: t_{0,95;8} = 2{,}306; \; \Delta_{\hat{a}_0} = t_{0,95;8} \cdot S_{\hat{a}_0} = 2{,}08;$$

для модели II при

$$k = 9 - 2 = 7: t_{0,95;7} = 2{,}365; \; \Delta_{\hat{a}_0} = t_{0,95;7} \cdot S_{\hat{a}_0} = 1{,}16;$$

$$\Delta_{\hat{a}_1} = t_{0,95;7} \cdot S_{\hat{a}_1} = 0{,}2.$$

Дисперсия и доверительный интервал для функции в случае первой модели совпадают с таковыми же для оценки параметра $\hat{a}_0$. В случае второй модели:

$$S^2_{\hat{Z}(x)} = S^2_{\hat{a}_0} + x^2 S^2_{\hat{a}_1}; \; \Delta_{\hat{Z}(x)} = t_{0,95;7} \cdot S_{\hat{Z}(x)} = \sqrt{\Delta^2_{\hat{a}_0} + x^2 \Delta^2_{\hat{a}_1}},$$

т. е. дисперсия и квадрат доверительного интервала растут квадратично с увеличением значения входного сигнала x.
3. Результаты расчетов сведем в табл. 2,3.
Таблица 2

$\hat{a}_0 / S_{\hat{a}_0}$	$\hat{a}_0 / S_{\hat{a}_0}$	$\hat{a}_1 / S_{\hat{a}_1}$
(модель I)	(модель II)	
$14{,}06 / 0{,}90$ $\Delta_{\hat{a}_0} = 2{,}08$	$9{,}106 / 0{,}49$ $\Delta_{\hat{a}_0} = 1{,}16$	$0{,}99 / 0{,}086$ $\Delta_{\hat{a}_1} = 0{,}2$

Таблица 3

x_i	z_i	$\hat{z}_{i\text{I}} \pm \Delta_{\hat{z}_{i\text{I}}}$	$\hat{z}_{i\text{II}} \pm \Delta_{\hat{z}_{i\text{II}}}$	$\lvert\hat{z}_{i\text{I}} - \hat{z}_{i\text{II}}\rvert$	$\hat{z}_{i\text{II}} - z_i$
1	10,1	14,06±2,08	10,096±1,82	3,964	–0,004
2	10,9	14,06±2,08	11,086±1,90	2,974	+0,186
3	12,2	14,06±2,08	12,076±2,03	1,984	–0,124
4	13,1	14,06±2,08	13,066±2,20	0,994	–0,034
5	13,9	14,06±2,08	14,056±2,39	0,004	+0,156
6	15,3	14,06±2,08	15,046±2,60	0,986	–0,254
7	16,3	14,06±2,08	16,036±2,84	1,976	–0,264
8	16,7	14,06±2,08	17,026±3,10	2,966	+0,326
9	18	14,06±2,08	18,016±3,36	3,956	+0,016

4. Оценка ма ксимальной погрешности преобразования (см. 6-ой столбец табл. 3).

$$\sigma_{np} = \frac{\max\left|z_i - \hat{z}_{i\text{II}}\right|}{\hat{z}_{i\text{II}}} \cdot 100\%,$$

$$\sigma_{np} = \frac{0{,}326}{17{,}026} \cdot 100\% \approx 1{,}9\%$$

5. Оценка максимального отклонения от линейности (см. 5-й столбец табл.3).

$$\sigma_{нл} = \frac{\max\left|z_{i\text{I}} - z_{i\text{II}}\right|}{z_{i\text{II}}} \cdot 100\%$$

$$\sigma_{нл} = \frac{3{,}964}{10{,}096} \cdot 100\% \approx 39\%$$

6. Проверка адекватности модели.

а) дисперсия ошибки эксперимента неизвестна. Составляем критерий.

$$F = \frac{S_{0\text{I}}^{\,2}}{S_{0\text{II}}^{\,2}} = \frac{7{,}39}{0{,}4524} = 16{,}34$$

По таблице распределения Фишера находим критическое значение критерия (выбирается односторонняя критическая область):

$F_{кр}(P, k_1, k_2); P = 0{,}95; k_1 = N - m_1 = 8; k_2 = N - m_2 = 7,$

что дает

$$F_{кр}(0{,}95;\ 8;\ 7) = 3{,}73$$

Так как $F > F_{кр}$, то принимается модель II с двумя параметрами.

б) Известна дисперсия ошибки эксперимента:

$$S_0^2 = (0{,}1)^2 = 0{,}01.$$

Составляем отношение $F = \dfrac{S_{0\text{II}}^2}{S_0^2}$ (достаточно проверить отношение только для модели II): $F = \dfrac{0{,}4524}{0{,}01} = 45{,}24.$

По таблице распределения Фишера определяем критическое значение критерия $F_{кр}(P,\ k_1,\ k_2);\ P = 0{,}95;\ k_1 = 7;\ k_2 = \infty$

(Число степеней свободы для S_0^2: $k_2 = \infty$, так как это теоретическая оценка, которой формально соответствует бесконечное число измерений):

$$F_{кр}(0{,}95;\ 7;\ \infty) = 2{,}01$$

Так как $F > F_{кр}$, то гипотеза о равенстве дисперсий отвергается. Поскольку модель адекватна, то можно сделать вывод, что априорное значение дисперсии ошибки эксперимента занижено.

Приложение 2

Совместная обработка количественных и качественных данных

Задача совместной обработки количественных и качественных данных возникает при оценке интегральных свойств, зависящих от ряда факторов (например, оценка качества, диагностирование и т.п.). Сформируем задачу в следующем виде. Пусть имеется свойство (характеристика состояния), определяемое по *n* критериям (факторам), каждый из которых представлен двумя оценками: числовой (количественной) *a,* выражающей объективную информацию, полученную измерением, и словесной (качественной) *b,* отражающей субъективное мнение экспертов. Например, для фактора «температура»: $a=37{,}3°C;$ *b*=«повышенная температура».

Требуется определить общую (интегральную) оценку данного свойства с учетом как количественной, так и качественной информации. Рассмотрим несколько случаев.

1. Предположим, что оценки являются статистическими и находятся опросом *m* экспертов. Тогда возможный способ решения задачи состоит в переводе всех оценок в порядковую (ранговую) шкалу. Обозначим: *j* - номер критерия $(1 \le j \le n);$ *i* - номер эксперта $(1 \le i \le m);$ a_{ij} - количественная оценка критерия *j* для эксперта *i*, приведенная к десятибалльной шкале (1-10); b_{ij} — качественная оценка критерия *j* для эксперта *i* по шкале (1-10); r_j- коэффициент корреляции между a_{ij} и b_{ij}; p_{ij} - относительный вес *j*-го критерия для эксперта *i*. Для агрегирования данной информации можно использовать одну из сверток, например, аддитивную или по наихудшему критерию. При использовании аддитивной свертки расчеты выполняются по следующей схеме:

$$\overline{b}_j = \frac{1}{m}\sum_{i=1}^{m} p_{ij} b_{ij}$$

$$\overline{a}_j = \left(\frac{1}{m}\sum_{i=1}^{m} p_{ij}\right)\left(\frac{1}{m}\sum_{i=1}^{m} a_{ij}\right) = \overline{p}_{ij} \cdot \overline{a}_{ij}$$

Общая оценка равна:

$$\hat{c} = \frac{1}{n}\sum_{j=1}^{n} c_j ,$$

где $c_j = \frac{1}{2} r_j \left(\bar{a}_j + \bar{b}_j\right);$ r_j — коэффициент ранговой корреляции Спирмена, определяемый из соотношения:

$$r_j = 1 - \frac{6\sum_{i=1}^{m}\left(\hat{a}_{ij} - \hat{b}_{ij}\right)^2}{m\left(m^2 - 1\right)},$$

где $\hat{a}_{ij}, \hat{b}_{ij}$ - ранговые оценки факторау для *i*-го эксперта.

Если ранжировки содержат совпадающие ранги, то выражение для коэффициента корреляции усложняется, так как должно учитывать число повторений рангов в ранжировках. Степень согласия между экспертами проверяется дополнительно по коэффициенту конкордации Кендалла.

Наряду с коэффициентом *r* могут использоваться и другие коэффициенты связи: коэффициент Юла *(Q),* коэффициент коллигации (Y), коэффициент абсолютной связи *(V)* .

$$Q = \frac{\alpha\delta - \beta\gamma}{\alpha\delta + \beta\gamma} = \frac{n\Delta}{\alpha\delta + \beta\gamma};$$

$$\Delta = \frac{\alpha\delta - \beta\gamma}{m};$$

где α- число оценок (a_j , b_j); β- число оценок $(a_j, \text{не} - b_j);$ $\gamma - (\text{не} - a_j, b_j);$ δ - (не - a_j , не - b_j).

Коэффициенты *Q* и *Y* эквиваленты друг другу и связаны соотношением:

$$Q = \frac{2Y}{1 + Y^2}.$$

Коэффициент *Q* равен нулю, если оценки a_{ij}, b_{ij} (объективная и субъективная) независимы, и принимает значение +1 в случае полной связанности (все оценки a_{ij} одновременно являются b_{ij} либо наоборот), а значение -1 в случае отрицательной связанности (все оценки a_{ij} не являются b_{ij}).

Коэффициент абсолютной связи определяется соотношением:

$$V = \frac{(\alpha\delta - \beta\gamma)}{\{(\alpha+\beta)(\alpha+\gamma)(\beta+\delta)(\gamma+\delta)\}^{\frac{1}{2}}}.$$

Он равен нулю, когда Δ=0, и принимает значение +1 только, когда все a_{ij} одновременно являются b_{ij} и все b_{ij} одновременно являются a_{ij}. При использовании свертки по наихудшему критерию:

$$\overline{b}_j = \min_i p_{ij} b_{ij}; \quad \overline{a}_j = \min_i p_{ij} \cdot \min_i a_{ij}$$

$$\hat{c}_L = \min_j c_j; \quad c_j = \frac{1}{2} r_j \left(\hat{a}_j + \hat{b}_j\right), \quad \hat{c}_u = \max_j c_j,$$

где $\hat{c}_L, \hat{c}_u$ - нижняя и верхняя граница соответственно.

В этом случае имеем интервальную оценку:

$$\hat{c} = \left[\hat{c}_L, \hat{c}_u\right],$$

причем нижняя граница соответствует стратегии пессимизма, а верхняя - оптимизма.

2. Рассмотрим случай, когда количественные оценки представлены в виде интервалов, а качественные - в виде (возможно усредненных по экспертам) словесных оценок, например: a_t=(37,2±0,1)°*C*; b_t =«повышенная температура». Применим для построения общей оценки нечеткие модели, тогда a_t и b_t имеют вид нечеткого числа и нечеткого интервала соответственно (рис.1).

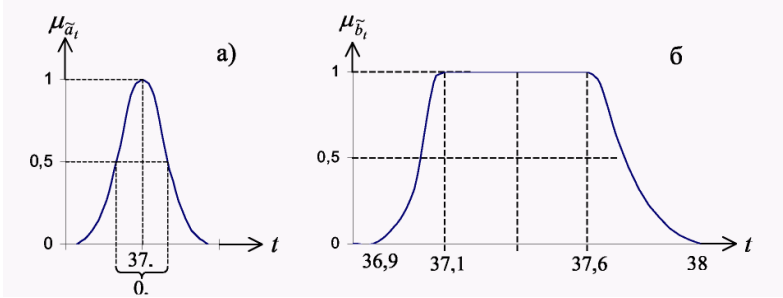


Рис. 1. Нечеткое число (а) и нечеткий интервал (б).

Для нечеткого числа граничные значения получаются как центральное значение плюс (минус) размах, т. е. 37,2±0,2. Для нечеткого

интервала граничные значения могут быть несимметричны относительно центра интервала.

Определим индекс (степень) согласования оценок $a_j b_j$ в виде :

$$\alpha_{jab} = F\left(\alpha\left(\tilde{a}_{j\alpha} R \tilde{b}_{j\alpha}\right) = \varnothing\right),$$

где F - свертка определяемая выбранной стратегией принятия решения, R - отношение согласования. Например, если R задается операцией пересечения, то $F = 1 - \inf a$ (мягкая стратегия) либо $F = \sup \alpha$ (жесткая стратегия); $\tilde{a}_{j\alpha}, \tilde{b}_{j\alpha}$ - α-срезы соответствующих нечетких множеств. В частности, если R задается операцией типа min, а $F = \sup \alpha$, то имеем:

$$\alpha_{jab} = \sup_{\alpha} \min\left(\mu_{\tilde{a}_{j\alpha}}(t), \mu_{\tilde{b}_{j\alpha}}(t)\right).$$

Достоверность оценки α_{jab} определяется соотношением :

$$\alpha_{jab} > \nu_R$$

либо

$$\alpha_{jab} > \frac{\nu_R}{2},$$

где V_R - индекс нечеткости множества, индуцированного отношением R.

$$\nu_R = 2 \sup_{\alpha}\left(R_\alpha \cap \overline{R}_\alpha \neq \varnothing\right) = 2 \sup_{\alpha} \min\left(\mu_{R_\alpha}, 1 - \mu_{R_\alpha}\right)$$

Если R задается операцией пересечения со сверткой типа min, то имеем:

$$\mu_{R_\alpha} = \min\left(\mu_{\tilde{a}_{j\alpha}}, \mu_{\tilde{b}_{j\alpha}}\right)$$

Общая оценка определяется выражением:

$$\hat{c} = \left[\hat{c}_L, \hat{c}_u\right], \quad \hat{c}_L = \min_j \alpha_{jab}, \quad \hat{c}_u = \operatorname{sum}_j \alpha_{jab}, \quad \operatorname{sum}(\alpha, \beta) = \alpha + \beta - \alpha\beta,$$

где $\hat{c}_L, \hat{c}_u$ - нижняя и верхняя граница интервала соответственно.

Нижняя граница соответствует противоречивым факторам, а верхняя - взаимодополнительным.

3. Рассмотрим теперь случай, когда количественная и качественная оценки относятся к разным критериям (факторам) например, $a_t = (39{,}1 \pm 0{,}1)°C$; b_t=«повышенное содержание

лейкоцитов». Применим нечеткие модели. В этом случае оценки a_j , b_j суммируются непосредственно, и мы имеем:

$$\hat{c} = [\hat{c}_L, \hat{c}_u], \quad \hat{c}_L = \min_j \mu_{\tilde{a}_j}(\tilde{b}_j), \quad \hat{c}_u = \operatorname{sum}_j \mu_{\tilde{a}_j}(\tilde{b}_j).$$

При определении типа операции свертки (min, sum, max и т.д.) следует учитывать дополнительную априорную информацию о семантике взаимосвязи оценок $\tilde{a}_j(\tilde{b}_j)$, $\tilde{a}_k(\tilde{b}_k)$ и интегральной оценки, и в зависимости от характера взаимосвязи того или иного фактора с другими выбирается тип свертки. Для обоснованного выбора свертки кроме индекса α_{jab} целесообразно также рассчитывать и другие индексы, в частности, индекс согласования $\lceil \tilde{a}_j$ с $\tilde{b}_j$:

$$\alpha_{j\bar{a}b} = F(\alpha(\bar{\tilde{a}}_{j\alpha} R \tilde{b}_{j\alpha}) \neq \varnothing);$$

индекссогласования $\tilde{a}_j$ с $\lceil \tilde{b}_j$:

$$\alpha_{ja\bar{b}} = 1 - F(\alpha((\tilde{a}_{j\alpha} R \tilde{b}_{j\alpha}) \cup (\tilde{a}_{j\alpha} R(\bar{\tilde{a}}_{j\alpha} R \tilde{b}_{j\alpha})) \neq \varnothing));$$

индекс согласования $\tilde{a}_j$ с $\bar{\tilde{a}}_j$ по фактору $\tilde{b}_j$:

$$\alpha_{ja\bar{a}} = F(\alpha(\tilde{a}_{j\alpha} R \tilde{b}_{j\alpha}) R(\bar{\tilde{a}}_{j\alpha} R \tilde{b}_{j\alpha}) \neq \varnothing).$$

В зависимости от соотношения четырех индексов уточняется вид операции свертки.

4. Рассмотрим решение задачи при тех же условиях, что и в п. 3, используя нечеткую логику. Пусть *щ* набор количественных оценок, b_j - качественных, причем a_i и b_j представлены в виде нечетких интервалов, так что a_i принимает значения в нечетком множестве $\tilde{A}_i$, b_j - в нечетком множестве $\tilde{B}_j$. Связь между исходными оценками a_i, b_j и искомой оценкой c можно выразить в виде набора правил:

Если $(\overset{*}{i} a_i$ есть $\tilde{A}_i) \to c$ есть $\tilde{C}_1$

Если $(\overset{*}{j} b_j$ есть $\tilde{B}_j) \to c$ есть $\tilde{C}_2$

где $\tilde{C}_1, \tilde{C}_2$ - нечеткие множества, в которых принимает значение оценка *c,* * - связка «и» («или»), → - операция импликации. Агрегирование количественных и качественных оценок проводится раздельно. Перепишем набор правил, вводя функции принадлежности:

$$\mu_{\tilde{C}_1} = \sup_a\left(\mu_{R\to}\left(\tilde{A},\tilde{C}_1\right) * \mu_{\tilde{A}}(a)\right);\ \ \mu_{\tilde{A}}(a) = \underset{i}{*}\mu_{\tilde{A}_i}(a_i);$$

$$\mu_{\tilde{C}_2} = \sup_b\left(\mu_{R\to}\left(\tilde{B},\tilde{C}_2\right) * \mu_{\tilde{B}}(b)\right);\ \ \mu_{\tilde{B}}(b) = \underset{j}{*}\mu_{\tilde{B}_j}(b_j),$$

где * - операция min, max или sum.

Итоговая оценка получается расчетом индекса согласования оценок $\tilde{C}_1$ и $\tilde{C}_2$, например в виде (см. выше):

$$\hat{c} = \alpha_{C_1,C_2} = \sup_{\alpha}\min\left(\mu_{\tilde{C}_1\alpha}, \mu_{\tilde{C}_2\alpha}\right)$$

Достоверность оценки проверяется сравнением с индексом нечеткости отношения согласования нечетких множеств $\tilde{C}_1$ и $\tilde{C}_2$.

Предложенный подход целсообразно применять при обработке разнородной информации. При этом ошибка измерения отдельных характеристик возрастает из-за перехода к порядковой (нечеткой) шкале, однако ошибка модели (предсказания, вывода, концепции) уменьшается за счет того, что отдельные части информации согласуются друг с другом.

Приложение 3

Таблицы наиболее часто используемых распределений

3.1. Интегральная функция нормированного нормального распределения

Значения *z* для различных Φ(z)

$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z
0,0005	–3,2905	0,35	–0,3853	0,80	+0,8416
0,005	–2,575	0,40	–0,2533	0,85	+1,0364
0,01	–2,3267	0,45	–0,1257	0,90	+1,2816
0,05	–1,6449	0,50	–0,0000	0,95	+1,6449
0,10	–1,2816	0,55	+0,1257	0,99	+2,3267
0,15	–1,0364	0,60	+0,2533	0,995	+2,575
0,20	–0,8416	0,65	+0,3853	0,9995	+3,2905
0,25	–0,6745	0,70	+0,5244		
0,30	–0,5244	0,75	+0,6745		

3.2. Распределение Стьюдента

$$P\left\{|t| < t_p\right\} = 2\int_0^{t_p} S(t;k)\,dt$$

k	P			
	0,90	0,95	0,98	0,99
1	6,314	12,706	31,821	63,657
2	2,920	4,303	6,965	9,925
3	2,353	3,182	4,541	5,841
4	2,132	2,776	3,747	4,604
5	2,015	2,571	3,365	4,032
6	1,943	2,447	3,143	3,707
7	1,895	2,365	2,998	3,499
8	1,860	2,306	2,896	3,355
9	1,833	2,262	2,821	3,250
10	1,812	2,228	2,764	3,169
11	1,796	2,201	2,718	3,106
12	1,782	2,179	2,681	3,055
13	1,771	2,160	2,650	3,012
14	1,761	2,145	2,624	2,977
15	1,753	2,131	2,602	2,947
16	1,746	2,120	2,583	2,921
17	1,740	2,110	2,567	2,898
18	1,734	2,101	2,552	2,878
19	1,729	2,093	2,539	2,861
20	1,725	2,086	2,528	2,845
21	1,721	2,080	2,518	2,831
22	1,717	2,074	2,508	2,819
23	1,714	2,069	2,500	2,807
24	1,711	2,064	2,492	2,707
25	1,708	2,060	2,485	2,787
26	1,706	2,056	2,479	2,779
27	1,703	2,052	2,473	2,771
28	1,701	2,048	2,467	2,763
29	1,699	2,045	2,462	2,756
30	1,697	2,042	2,457	2,750

k	P			
	0,90	0,95	0,98	0,99
∞	1,64485	1,95996	2,32634	2,57852

3.3. Интегральная функция χ^2 - распределение Пирсона

Значения $\chi^2_{k;P}$ для различных k и P

k	P			
	0,90	0,95	0,98	0,99
1	2,706	3,841	5,412	6,635
2	4,605	5,991	7,824	9,210
3	6,251	7,815	9,837	11,345
4	7,779	9,488	11,668	13,277
5	9,236	11,070	13,388	15,086
6	10,645	12,592	15,033	16,812
7	12,017	14,067	16,622	18,475
8	13,362	15,507	18,168	20,090
9	14,684	16,919	19,679	21,666
10	15,987	18,307	21,161	23,209
11	17,275	19,675	22,618	24,725
12	18,549	21,026	24,054	26,217
13	19,812	22,362	25,472	27,688
14	21,064	23,685	26,873	29,141
15	22,307	24,996	28,259	30,578
16	23,542	26,296	29,633	32,000
17	24,769	27,587	30,995	33,409
18	25,989	28,869	32,346	34,805
19	27,204	30,144	33,687	36,191
20	28,412	31,410	35,020	37,566
21	29,615	32,671	36,343	38,932
22	30,813	33,924	37,659	40,289
23	32,007	35,172	38,968	41,638
24	33,196	36,415	40,270	42,980
25	34,382	37,652	41,566	44,314
26	35,563	38,885	42,856	45,642
27	36,741	40,113	44,140	46,963
28	37,916	41,337	45,419	48,278
29	39,087	42,557	46,693	49,588

k	P			
	0,90	0,95	0,98	0,99
30	40,256	43,773	47,962	50,892

3.4. F^-распределение Фишера

Значение F_{k1k2} для различных доверительных вероятностей α

k_2	α	k_1																			
		1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	50	60	100	∞
1	0,75	5,83	7,50	8,20	8,58	8,82	8,98	9,10	9,19	9,26	9,32	9,36	9,41	9,49	9,58	9,63	9,67	9,71	9,74	9,76	9,78	9,85
	0,90	39,9	49,5	53,6	55,8	57,2	58,2	58,9	59,4	59,9	60,2	60,5	60,7	61,2	61,7	62,0	62,3	62,5	62,7	62,8	63,0	63,3
	0,95	161	200	216	225	230	234	237	239	241	242	243	244	246	248	249	250	251	252	252	253	254
2	0,75	2,57	3,00	3,15	3,23	3,28	3,31	3,34	3,35	3,37	3,38	3,39	3,39	3,41	3,43	3,44	3,44	3,45	3,45	3,46	3,47	3,48
	0,90	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,40	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,47	9,48	9,49
	0,95	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5	19,5
	0,99	98,5	99,0	99,2	99,2	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,5	99,5	99,5	99,5	99,5	99,5	99,5
3	0,75	2,02	2,28	2,36	2,39	2,41	2,42	2,43	2,44	2,44	2,44	2,45	2,45	2,46	2,46	2,46	2,47	2,47	2,47	2,47	2,47	2,47
	0,90	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,15	5,14	5,13
	0,95	10,1	9,55	9,28	9,28	9,10	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,70	8,66	8,64	8,62	8,59	8,58	8,57	8,55	8,53
	0,99	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,1	27,1	26,9	26,7	26,6	26,5	26,4	26,4	26,3	26,2	26,1
4	0,75	1,81	2,00	2,05	2,06	2,07	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08

Приложение 4

Конспект лекций по теории измерений

Основные понятия.

Расчёт приборов и систем производится на основе математических моделей их функционирования. Очевидно, что любые математические модели строятся исходя из допущений. Эти допущения позволяют упростить модель, но увеличивают вероятность появления непрогнозируемой погрешности.

Таким образом, основная проблема расчета точности возникае в связи с наличием факторов, оказывающих влияние на результа измерения, но не охватываемых рабочей моделью исследуемог измерительного средства.

В результате расчетная точность прибора и фактическ полученная, после его разработки могут существенно различаться Отметим, что погрешность от некоторых влияющих факторов можн уменьшить.

Основная задача теории измерений это выработк рекомендаций по выполнению измерений и их обработке обеспечивающих минимальную погрешность при наличи неконтролируемых возмущений.

Кроме того, теория измерений исследует вопрос планирования, эксперимента, во время которого определяется каки методы целесообразно использовать? Сколько нужно измерений?

Теория измерений, как и любая другая теория, имеет основны положения (постулаты) и понятия.

К числу понятий в первую очередь следует отнести поняти физической величины и ее единицы.

Введение понятия физических величин и установление и единиц является необходимой предпосылкой измерений. Однако всяко измерение всегда выполняется применительно к конкретному объекту И общее определение измеряемой физической величины необходим

конкретизировать, учитывая свойства данного объекта и цел измерения. Так, по существу, вводится и определяется истинно значение измеряемой величины.

Первый постулат теории измерений звучит так:

- *Существует истинное значение измеряемой величины;*

Второй постулат

- *Истинное значение измеряемой величины отыскат невозможно.*

Третий постулат

- *Истинное значение измеряемой величины постоянно.*

1. Классификация измерений.

В соответствии с РМГ29-99, введенным вместо ГОСТ 16263 (метрология, основные термины и определения) измерения физических величин делятся на:

- Равноточные измерения. Это ряд измерений какой-либ величины, выполненных одинаковыми по точности средствам измерений в одних и тех же условиях с одинаковой тщательностью Отметим, что прежде чем обрабатывать ряд измерений, необходим убедиться в том, что все измерения этого ряда являются равноточными

- Неравноточные измерения. Это ряд измерений какой-либ величины, выполненных различающимися по точности средствам измерений и (или) в разных условиях. Отметим, что неравноточны измерения обрабатывают с учетом веса отдельных измерений входящих в ряд.

- Однократное. Это измерение, выполненное один раз.

- Многократное. Это измерение физической величины одного того же размера, результат которого получен из нескольких следующи друг за другом измерений, т.е. состоящего из ряда однократны измерений.

- Статическое измерение. Измерение физической величины принимаемой в соответствии с конкретной измерительной задачей з неизменную на протяжении времени измерения.

- Динамическое измерение. Измерение изменяющейся п размеру физической величины.

- Абсолютное измерение. Измерение, основанное на прямы измерениях одной или нескольких основных величин и (или использовании значений физических констант. Например, измерени силы, падающего на землю тела F=mg основано на измерении основно величины –массы и использовании физической постоянной g.

- Относительное измерение. Измерение отношения величины одноименной величине, играющей роль единицы, или измерени изменения величины по отношению к одноименной величин принимаемой за исходную.

- Прямое измерение. Измерение, при котором искомо значение физической величины получают непосредственно. Отметим что термин прямое возник в противоположность термину косвенное.

- Косвенное измерение. Определение искомого значени физической величины на основании результатов прямых измерени других физических величин, функционально связанных с искомо величиной.

- Совокупное измерения. Проводимые одновременн измерения нескольких одноименных величин, при которых искомы значения величин определяются путем решения системы уравнений получаемых при измерениях этих величин в различных сочетаниях Отметим, что для определения значений искомых величин числ уравнений должно быть не меньше числа величин.

- Совместные измерения. Проводимые одновременно измерения двух или нескольких не одноименных величин для определения зависимости между ними.

2. Методы измерений.

2.1 Термины и определения в соответствие с РМГ 29-99

Принцип измерений - физическое явление и эффект, положенное в основу измерений.

Метод измерений – приём или совокупность приёмов сравнения измеряемой физической величины с её единицей в соответствии с реализованным принципом измерений. (Примечание – метод измерений обычно обусловлен устройством средств измерений)

РМГ даёт определение следующих методов измерений:

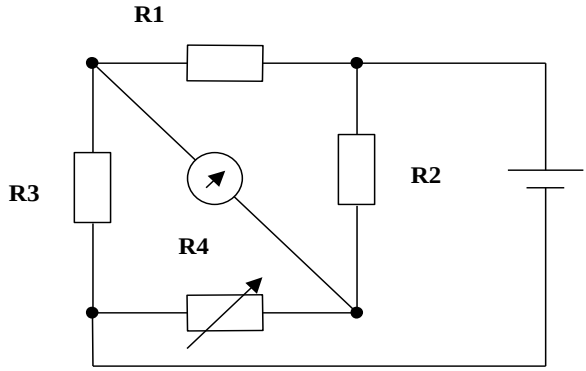
1. *Метод непосредственной оценки – метод измерений, при котором значение величины определяют непосредственно по показывающему средству измерений.*
2. *Метод сравнения с мерой (метод сравнения) – метод измерений, в котором измеряемую величину сравнивают с величиной, воспроизводимой мерой. (Метод Борда)*
3. *Нулевой метод измерений – метод сравнений с мерой, в котором результирующий эффект воздействия измеряемой величины и меры на прибор сравнения доводят до нуля.*
4. *Метод измерения замещением (метод замещения). Метод сравнения с мерой, в котором измеряемую величину замещают мерой с известным значением величины.*
5. *Метод измерений дополнением (метод дополнения). Метод сравнения с мерой, в котором значение измеряемой величины дополняется мерой этой же величины с таким расчётом, чтобы на прибор сравнения воздействовала их сумма, равная заранее заданному значению.*
6. *Дифференциальный метод измерений (дифференциальный метод). Метод измерений , при котором измеряемая величина сравнивается с однородной величиной, имеющей известное значение, незначительно отличающиеся от значения измеряемой величины, и при котором измеряется разность между этими двумя величинами.*
7. *Контактный метод измерений - метод измерений, основанный на том, что чувствительный элемент прибора приводится в контакт с объектом измерений.*

8. *Бесконтактный метод измерений – метод, основанный на том, что чувствительный элемент средства измерений не приводится в контакт с объектом измерений.*

В методе непосредственной оценки показания используемого измерительного прибора полностью определяют результат измерения.

В методе сравнения с мерой результат измерения полностью определяется значением меры, при этом показания прибора используются только для оценки разности между мерой и применяемой физической величиной. Метод сравнения с мерой определяет методы измерений замещением, нулевой метод и метод измерений дополнением. Нулевой метод основан на том, что влияние неизвестной величины на измерительную систему устраняется путём компенсирующего воздействия, величина которого известна.

Примером использования нулевого метода является мостовой измеритель сопротивления.



В том случае, когда R3=R2=Constant, значение R1 неизвестно, а значение R4 может подбираться и однозначно определено, то с помощью такого прибора можно измерить

сопротивление, помещаемое вместо R1. Отличительной чертой мостовой схемы является то, что состояние равновесия не зависит от источника питания. Помимо использования метода для измерения электрических сопротивлений, метод может быть полезен для измерения других физических величин, например гидравлических сопротивлений, температуры с помощью терморезисторов и т. д.

Если в методе измерения дополнением неизвестная и известная величина одновременно участвуют в каждом измерении, то в методе измерений замещением они участвуют по отдельности одна вслед за другой.

В методе измерения замещением сначала по отклонению стрелки или по показанию индикатора измерительной системы определяется неизвестное значение измеряемой величины. Затем неизвестная величина заменяется известной и регулируемой величиной, которая подстраивается таким образом, чтобы получился точно такой же результат измерения. Показания измерительной системы в этом случае играют лишь промежуточную роль. Поэтому характеристики измерительной системы не должны влиять на результат измерения. Важным является лишь стабильность системы во времени и её разрешающая способность.

При калибровке измерительной системы применяется, по существу, метод измерений замещением. Сначала система калибруется по известной величине. Затем можно точно измерить неизвестную величину, если значение совпадает с одной из откалиброванных точек. Метод измерения замещением часто используют также в качестве простого средства установления «равенства», когда точность используемой измерительной системы не имеет значения.

Дифференциальный метод рассмотрим на примере следующей задачи:

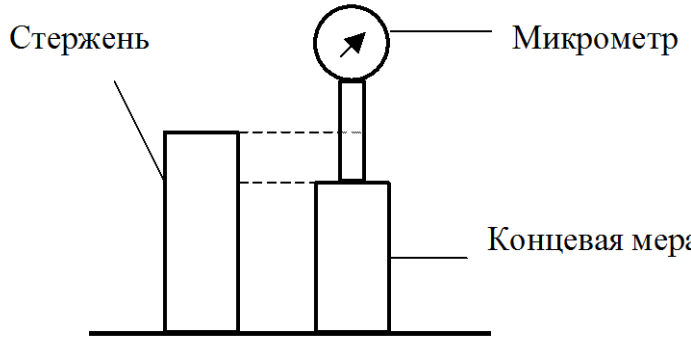
Необходимо измерить длину стержня примерно равную 101мм., с погрешностью $1 \bullet 10^{-4}$ мм. Если мы применим метод непосредственной оценки, то необходимо

воспользоваться измерительным инструментом, например штангенциркулем с нониусом, дающим погрешность в $1 \bullet 10^{-4}$ мм.

Точность существующих штангенциркулей не позволяет этого сделать, поэтому применяется дифференциальный метод.

Применяется концевая мера, представляющая собой мерный брусок, имеющий необходимую длину. Применим меру 100,000мм. $\pm 10^{-5}$. Различие по длине примерно равное 1мм. Измерим с помощью измерительной головки. Точность такой головки равна 0,1мкм.

Таким образом, точность измерений обеспечивается.



Следует отметить, что в РМГ 29-99 даны определения основных методов измерений, при этом рекомендации не оговаривают общее количество этих методов и их назначения.

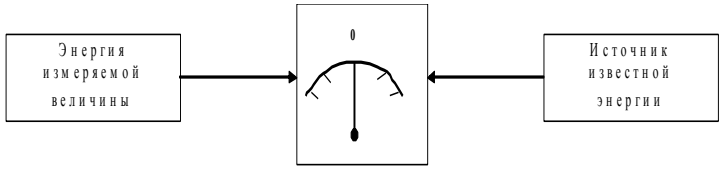
2.2 Методы измерения не включённые в РМГ 29-99

В связи с этим разберём другие методы, которые, однако, часто встречаются в практике измерений и фигурируют в американской классификации методов измерений.

Компенсационный метод – это такой метод измерения, при котором влияние неизвестной величины на измерительную систему устраняется путём компенсации его влияния известной величиной. Иными словами измерение неизвестной величины проводится путём компенсации её влияния на прибор, при этом величина компенсирующего воздействия известна. Особенностью этого метода является то, что когда действие неизвестной величины полностью скомпенсировано, никакая энергия не перекачивается в источник неизвестной величины и ни какая энергия не потребляется от него. Источник неизвестной величины не нагружается входом измерительной системы. Степень компенсации можно определять с помощью нуль органа.

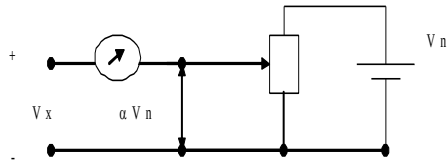
Чтобы скомпенсировать разность электрических потенциалов Vx применяется схема представленная на рисунке . в этой схеме от источника питания Vn снимается напряжения αVn. Если подстроить потенциометр так, чтобы αVn = Vx индикатор примет нулевое положение. В том случае если потенциометр отградуирован (положение движка соответствует, какому либо напряжению) и значение Vn постоянно, то эта схема становится измерительной схемой прибора.

Для компенсационного метода характерно использование двух источников энергии, таких как Vn и Vx, недостаток заключается в том, что если значение Vx или Vn не постоянны, то компенсация не возможна.



Нуль орган

Измерение постоянного напряжения:



Метод аналогий.

В этом методе используется модель объекта, от которой мы хотим получить измерительную информацию. Измерения, выполненные на модели, обеспечивают нас сведениями о неизвестном объекте в той мере, в какой модель соответствует объекту в наиболее существенных моментах. Этим методом аналогий пользуются чаще всего в тех случаях, когда измерения на самом объекте невозможны, в случаях, когда проектируется сам объект и необходимо подобрать его свойства соответствующие известным результатам измерений, в случае, когда объект при измерениях может быть уничтожен и т. д.

Один класс используемых моделей это математические модели. В этом случае модель описывается теме же самыми математическими соотношениями, что и действительный объект. Например, механические весы с плечами разной длины можно рассматривать как модель электрической мостовой схемы (моста Уитсона). Пусть длины плеч равны L_1 и L_2, массы грузов на чашки весов – m_1 и m_2, а ускорение силы тяжести – g, условие равновесия будет $m_1 L_1 g = m_2 L_2 g$.

Здесь величина g играет роль источника энергии, а стрелка весов служит нуль индикатором.

Другой класс моделей образуют масштабные модели, представляющие собой линейно увеличенные или уменьшенные копии измеряемого объекта. Этот тип моделей используется при исследовании динамических характеристик различных процессов, например акустики больших залов.

Третий класс состоит из моделей, являющихся результатом нелинейного масштабирования. Увеличение или уменьшение производят таким образом, что в модель переносят без искажения только определенные свойства объекта.

Примерами таких моделей являются испытания в аэродинамических трубах и опытовых бассейнах.

Математическое моделирование во многих случаях связано с экспериментальным исследованием некоторых характеристик, от которых зависит работа математической модели.

Поэтому обычно нелинейное и линейное масштабирование способствует улучшению качества математической модели.

Метод повторений.

Согласно этому методу производиться несколько измерений одной и той же неизвестной величины, причем процедура измерений каждый раз выбирается другой. Например, самые фундаментальные физические константы измерены несколькими различными способами. Это позволяет предотвратить возможность проявления одних и тех же систематических погрешностей, характерных для того или иного типа измерений. Другие методы измерений будут приводить к близким результатам, но погрешности измерений окажутся взаимно независимыми и это является свидетельством надежности результатов измерений.

Метод перечисления.

Этот метод заключается в определении отношения двух величин (известной и неизвестной) путем подсчета. Подсчитывать можно только объекты, структуры и события. Физические величины заданной физической размерности должны быть измерены. При измерениях появляются погрешности.

При подсчете их нет (в предположении, что со счета не сбились).

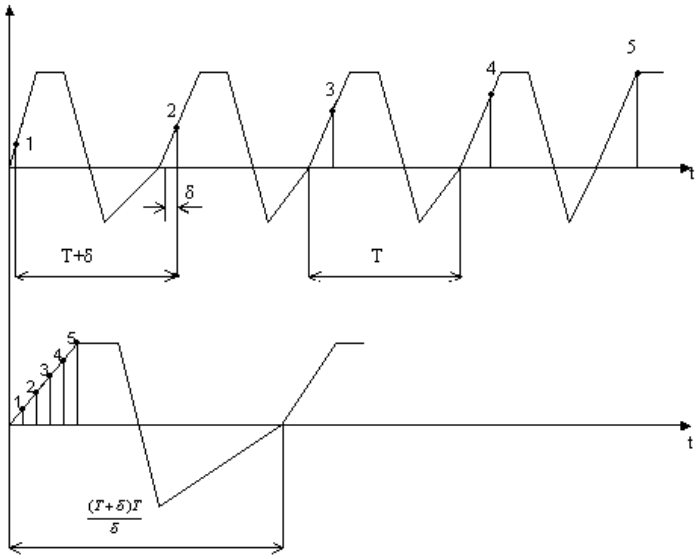
Метод перечисления применяется, например, при измерениях частоты. Частота периодического сигнала измеряется путем простого подсчета числа периодов за единицу времени.

Метод перечисления важен также в аналого-цифровом преобразовании. Однако иногда легче перейти от перечисления к измерению. Например, для определения числа шурупов в упаковке проще их взвесить, чем пересчитать.

2.3 Некоторые методы, определяющие стратегию измерений

Метод когерентных отсчетов.

Использование этого метода позволяет обрабатыват измерительный сигнал с шириной спектра F, значительно большей, че ширина полосы B измерительной системы, при условии, что сигна является периодическим. Беря отсчёты значений измеряемого сигнала интервалом, немного превосходящим n периодов сигнала (n – цело число), можно запомнить форму сигнала и получить верно представление о нем. Если интервал между отсчетами обозначит $nT+\delta$, где T – период измеряемого сигнала, то период восстановленног (по этим отсчетам) сигнала будет равен $\frac{T(nT+\delta)}{\delta}$. Это означает, что измеренном сигнале произошло уменьшение частоты в $\frac{\delta}{nT+\delta}$ раз.



На рисунке интервал между отсчетами выбран $T+\delta$ (так, чт n=1). На нижнем графике восстановленный по отсчетам сигнал. Есл $\delta<<T$, то сигнал можно обрабатывать измерительной системой полосой $B<<fo$. Из рисунка видно, что число пропускаемых периодо n и отношение T/δ (число отсчетов на период восстановленног сигнала) выбираются таким образом, чтобы частотный спект восстановленного сигнала, представляющего собой огибающу пиковых значений был уже полосы пропускания измерительно системы, применяемой для обработки исходного сигнала, из которог берутся отсчеты.

Такой способ измерений реализуется при стробоскопических измерениях.

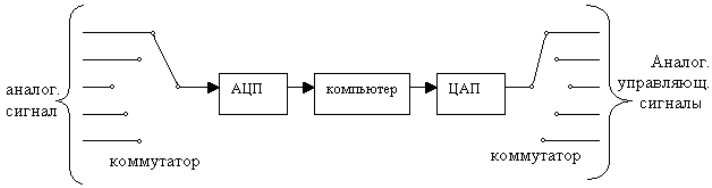
Метод случайных отсчетов.

Очевидно, что при когерентном взятии отсчетов требуется выполнение определенных условий, гарантирующих, что отсчеты с номерами 1,2,3 и т.д. будут производиться точно в нужные моменты времени. Однако бывают случаи, когда нас интересует только информация о величине, а не о форме сигнала. В этом случае отсчеты можно брать в произвольные моменты времени. Таким образом, можно определить, например, среднеквадратическое значение сигнала с широкополосным спектром. При этом сигнал не обязательно должен быть периодическим.

В общем случае можно утверждать, что случайный характе отсчетов не оказывает влияния на статистические параметры относящиеся к величине сигнала (среднее, среднеквадратическо отношение и т.д.). Иногда случайное взятие отсчетов осуществляют определенной частотой, никак не связанной с наблюдаемым сигналом Однако, при этом, может наступить корреляция между измеряемы сигналом и процедурой взятия отсчетов и поэтому информация величине сигнала, получаемая из отсчетов может содержать ошибки Чтобы избежать этих ошибок, частоту отсчетов качают в определенны пределах. При случайном взятии отсчетов полоса пропускани "*B*" измерительной системы также может быть меньше ширин спектра *F* измеряемого сигнала.

Метод мультиплексирования.

Применение этого способа позволяет одновременно (при частотном мультиплексировании) или последовательно (при временно мультиплексировании) обрабатывать несколько процессов. Целесообразно использовать этот способ в том случае, когда полоса "*B*" измерительной системы много больше частотного спектра *F* измеряемых сигналов, при этом измерительных сигналов несколько. Н рис. в качестве иллюстрации показано как можно использовать временное мультиплексирование в системе контроля параметров и управления процессом ректификации нефти. С помощью аналоговых датчиков контролируются различные параметры «крекинг-процесса» такие как температура, давление, объем и т.д.



При частотном мультиплексировании узкополосный измеряемый сигнал сдвигается в другой диапазон частот. Это производится таким образом, чтобы спектры нескольких преобразуемых измеряемых сигналов занимали соседние интервалы частот без перекрытия. На выходе измерительной системы необходим демодулятор, чтобы восстановить каждый сигнал в его полосе частот. Этот способ часто используют в телеметрии и телефонии.

3. Общие вопросы теории погрешностей

3.1 Виды погрешностей и особенности терминологии в соответствие с РМГ

В курсе метрологии вы изучали классификацию погрешностей поэтому останавливаться на ней мы не будем. Остановимся подробнее на терминологических особенностях и соответствующем смысловом наполнении этих терминов в соответствии с РМГ 29-99.

3.1.1 Погрешность средств измерений и погрешность результата измерений.

В соответствии с РМГ погрешность измерения - это отклонени результата измерения от истинного (действительного) значения измеряемой величины. Погрешность средства измерения – это разност между показанием средства измерений и истинным (действительным) значением измеряемой физической величины. Определение

погрешности средства измерения, даваемое в РМГ представляется не очень удачным. Представляется, что погрешность прибора это его свойство, для описание которого, приходится использовать соответствующие правила.

Поэтому полагать, что воспользовавшись, например, вольтметром класса точности 1,0 т.е. имеющим предел приведённой погрешности , равный 1% , мы получаем и результат измерения с погрешностью 1% - грубейшая ошибка. Исторически часть наименований погрешностей закрепилась за погрешностями средств измерений, другая – за погрешностями результатов измерения, а некоторые применяются по отношению и к тем и к другим. Поэтому, рассматривая эти термины, следует обращать внимание на области их применения.

3.1.2 Инструментальная и методическая погрешности. Инструментальная погрешность – составляющая погрешности измерения, обусловленная погрешностью применяемого средства измерений. Погрешность метода измерений (или методическая погрешность) – составляющая систематической погрешности измерений, обусловленная несовершенством принятого метода измерений. Очевидно, что величину инструментальной погрешности можно узнать из паспорта прибора. Оценка значения методической погрешности является более сложной процедурой. Очень часто причиной возникновения методической погрешности является то, что, организуя измерения, нередко вынуждены измерять не ту величину, которая в принципе должна быть измерена, а некоторую другую, близкую, но не равную ей. Этот приём замены позволяет создавать наиболее простые, надёжные и универсальные приборы. Наглядный пример этого – выбор метода построения прибора для измерения запас горючего в баке автомобиля. Ясно, что суммарная энергия, запасённая в топливе, определяется его массой (а не объёмом) и для её измерения нужны весы. Но совмещение топливного бака с весовым механизмом резко усложняет конструкцию. Поэтому разработчик заменяет весы простейшим поплавковым уровнемером, хотя уровень топлива зависит и от наклона бака, и от температуры и лишь весьма приближённо отражает массу топлива. Автомобилисты знают, что летом на жаре

бензин расходуется быстрее, т.к. при той же массе имеет больший объём. Ещё один пример, это измерение напряжения вольтметром. Вследствие шунтирования входным сопротивлением вольтметра того участка цепи, на котором измеряется напряжение, оно оказывается меньшим, чем было до присоединения вольтметра. Поэтому на низкоомных участках цепи эта погрешность ничтожна, а на высокоомных может быть очень значительной. К методическим погрешностям относятся все погрешности, которые могут быть определены и количественно оценены с помощью математической модели измерительной процедуры. Количественная оценка погрешностей и их характеристик при этом выполняется на основе расчётов или имитационного моделирования. Таким образом, отличительной особенностью методических погрешностей является то что они могут быть определены лишь путём создания математических моделей или имитационным моделированием измеряемого объекта и н могут быть найдены сколь угодно тщательным исследованием лишь самого измерительного прибора.

3.1.3 Основная и дополнительная погрешности. Основная погрешность это погрешность средства измерений, применяемого в нормальных условиях. Дополнительная погрешность, это составляющая погрешности, возникающая дополнительно к основной погрешности вследствие отклонения какой – либо из влияющих величин от нормального её значения или вследствие её выхода за пределы нормальной области значений. Очевидно, что любой прибор работает в сложных, изменяющихся во времени условиях. Изменение этих условий в том случае, если они влияют на прибор (например: температуры, влажности, вибрации, напряжения питания и др.) приводит к изменению показаний прибора. Таким образом, интересующий нас единственный фактор, из всего множества влияющих на прибор мы называем измеряемой величиной. При проектировании прибора мы стремимся к тому, чтобы влияние всех остальных факторов на показания прибора было минимальным. При аттестации или градуировке прибора в лабораторных условиях все значения влияющих величин могут поддерживаться в узких пределах. Такие условия, оговорённые в технической документации, называются

нормальными. В соответствие с РМГ под нормальными условиями измерений понимаются условия измерений, характеризуемые совокупностью значений или областей значений влияющих величин, при которых изменением результата измерений пренебрегают вследствие малости. Следует отметить, что дополнительная погрешность нормируется и соотносится с режимами работы прибора. Так авиационные приборы имеют рабочий диапазон при температуре от +60 до – 80 градусов Цельсия. Морские приборы должны работать при влажности 100 % и т.д. Рабочая область значений влияющей величины определяется РМГ как область, в пределах которой нормируют дополнительную погрешность или изменение показаний средства измерений. Нормирование дополнительной погрешности во многих случаях производят путём указания коэффициентов влияния, например $\gamma_{доп} = \psi \Delta\theta$, где ψ - коэффициент влияния, $\Delta\theta$ - отклонение от нормальных условий.

3.1.4 Статические и динамические погрешности. Эти погрешности присущи как средствам, так и методам измерений. Их различают по зависимости от скорости изменения измеряемой величины с течением времени. Статическая погрешность, это погрешность средства измерения, возникающая при измерении физической величины, принимаемой за неизменную. Динамическая погрешность – погрешность средства измерений, возникающая при измерении изменяющейся (в процессе измерений) физической величины. Таким образом, динамические погрешности являются одной из разновидностей дополнительных погрешностей. Вы уже знаете, что они имеют специфические методы нормирования и расчёта.

3.1.5 Систематические и случайные погрешности. РМГ дают следующие определения: систематическая погрешность результата измерения – это составляющая погрешности результата измерений, остающаяся постоянной или закономерно изменяющаяся при повторных измерениях одной и той же физической величины. Случайная погрешность измерения, это составляющая погрешности результата измерения, изменяющаяся случайным образом (по знаку и значению) при повторных измерениях, проведённых с одинаковой тщательностью, одной и той же физической величины. Основной

отличительный признак систематических погрешностей состоит в том, что они могут быть предсказаны и благодаря этому почти полностью устранены. Отличительная черта случайных погрешностей это их непредсказуемость от одного отсчёта к другому. Как правило, они проявляются в виде разброса значений относительно некоторой величины. Следует отметить, что систематическая погрешность определённого средства измерения, как правило, будет отличаться от систематической погрешности другого экземпляра средства измерения такого же типа, вследствие чего для группы однотипных средств измерений систематическая погрешность может иногда рассматриваться как случайная погрешность.

Систематические погрешности, в свою очередь, подразделяются в зависимости от характера измерения на постоянные, прогрессивные, периодические и погрешности, изменяющиеся по сложному закону. Постоянные погрешности – погрешности, которые длительное время сохраняют своё значение, например, в течение времени выполнения всего ряда измерений. Прогрессивные погрешности, в соответствии с РМГ, это непрерывно возрастающие ил убывающие погрешности. Периодические погрешности – погрешности значение которых является периодической функцией времени или перемещения указателя измерительного прибора. Что касается погрешностей, изменяющихся по сложному закону, то они происходят вследствие совместного действия нескольких систематических погрешностей.

3.1.6 Неисключённая систематическая погрешность, погрешность градуировки и вариация показаний. Во многих случаях при калибровке прибора получают ряд точек, по этим точкам проводят плавную среднюю кривую, которую и принимают за характеристику. Систематически наблюдающиеся отклонения от выбранной в качестве характеристики плавной кривой в общем случае определяется как погрешность адекватности выбранной функциональной зависимости фактической характеристике прибора. РМГ определяет неисключённую систематическую погрешность как «составляющую погрешности результата измерений, обусловленную

погрешностями вычисления и введения поправок на влияние систематических погрешностей или систематической погрешностью, поправка, на действие которой не введена вследствие её малости.» Есл в качестве характеристики выбрана прямая линия, то неисключенная систематическая погрешность будет носить характер погрешности от линейности характеристики. В том случае, если неисключённая систематическая погрешность меняет знак, то она может быть охарактеризована как погрешность от гистерезиса или вариации. Такая погрешность во многих случаях обусловлена зоной нечувствительност прибора. Под вариацией показаний измерительного прибора понимается разность его показаний в одной и той же точке диапазона измерений при плавном подходе к этой точке со стороны меньших и больших значений измеряемой величины.

3.2 Термины, позволяющие нормировать погрешности средств измерений.

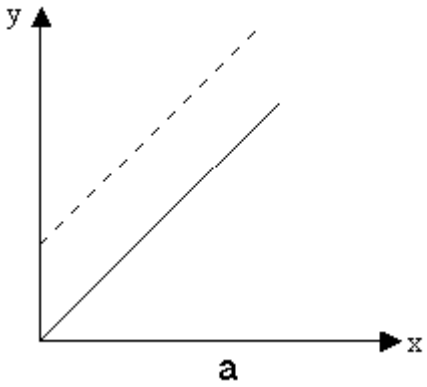
3.2.1 Абсолютная, относительная и приведённая погрешности средств измерений.

Погрешность измерения, выраженная в единицах измеряемой величины, есть абсолютная погрешность. Следует отличать абсолютную погрешность от абсолютного значения погрешности. Абсолютное значение погрешности это значение погрешности без учёта её знака (модуль погрешности). Относительная погрешность, это погрешность измерения, выраженная отношением абсолютной погрешности измерения к действительному или измеренному значени измеряемой величины. Относительная погрешность обычно выражаетс в процентах и рассчитывается по формуле $\delta = \frac{\Delta x}{x} 100\%$, где Δx – абсолютная погрешность измерений; x – действительное или измеренное значение величины. Приведённая погрешность средства

измерения в соответствие с РМГ, есть относительная погрешность, выраженная отношением абсолютной погрешности средства измерени к условно принятому значению величины, постоянному во всём диапазоне измерений или в части диапазона. Таким образом, это отношение абсолютной погрешности к протяжённости диапазона измерения, выраженное в процентах.

3.2.2 Аддитивные и мультипликативные погрешности.

РМГ не даёт определения этих погрешностей, однако, эти термины часто встречаются в специальной литературе. Аддитивные погрешности не зависят от уровня измеряемых сигналов. Погрешность возникает от сдвига статической характеристики прибора, угол наклона, при этом, остается неизменным (рис. а). Мультипликативные погрешности характеризуются постоянным приращением наклона характеристики (рис. б).



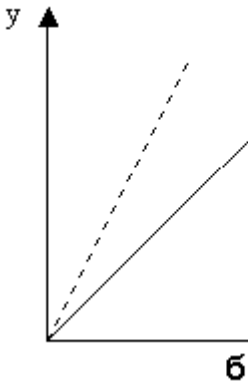


Рис. Аддитивная (а) и мультипликативная (б) погрешности

Эти погрешности применимы как к случайным, так и к систематическим погрешностям. Примером систематической аддитивной погрешности является погрешность от постороннего груза на чашке весов, погрешность от смещения стрелки шкального механизма при её нулевом положении. Примером случайных аддитивных погрешностей является погрешность от наводки

переменной ЭДС на вход прибора, погрешности от тепловых шумов и т.д. Примером мультипликативной погрешности может быть изменени коэффициента усиления усилителя при его прогреве, дрейф нуля интегратора и т.д.

3.3 Методы нормирования погрешностей средств измерений

Для того чтобы ориентироваться в метрологических свойствах конкретного средства измерения, чтобы заранее оценить погрешность, которую внесёт данный прибор в конкретный результат, пользуются так называемыми нормированными значениями погрешности. Под нормированным значением понимаются погрешности, являющиеся предельными для данного типа средств измерений. При этом как систематическая, так и случайная составляющие погрешности отдельных экземпляров приборов одного и того же типа могут различаться, однако в целом для этого типа средств измерений погрешности не превосходят гарантированного значения. Таким образом, нормируются основная и дополнительная погрешности. Именно эти границы основной погрешности, а также коэффициентов влияния и заносятся в паспорт каждого экземпляра прибора. Вся процедура нормирования погрешностей средств измерений, основывается на системе стандартов, обеспечивающих единство измерений.

3.3.1 Класс точности средств измерений

Это есть обобщённая характеристика данного типа средств измерений, как правило, отражающая уровень их точности, выражаема пределами допускаемых основной и дополнительных погрешностей, а также другими характеристиками, влияющими на точность. Основные способы установления допускаемых погрешностей и обозначения классов точности средств измерений установлены ГОСТ 8.401. Основная погрешность средства измерения нормируется четырьмя различными способами. Основное различие в способах нормирования

обусловлено разным соотношением аддитивной и мультипликативной составляющих погрешности тех или иных средств измерения.

При чисто мультипликативной полосе погрешностей средств измерения абсолютная погрешность $\Delta(x)$ возрастает прямо пропорционально текущему значению x измеряемой величины. Поэтому относительная погрешность, т. е. погрешность чувствительности такого преобразователя, $\gamma_S=\Delta(x)/x$ оказывается постоянной величиной при любом значении x и ее удобно использовать для нормирования погрешностей преобразователя и указания его класса точности.

Таким способом нормируются погрешности масштабных преобразователей (делителей напряжения, шунтов, измерительных трансформаторов тока и напряжения и т. п.). Их класс точности указывается в виде значения γ_S выраженного в процентах. Граница относительной погрешности результата измерения $\gamma(x)$ в этом случае постоянна и при любом x просто равна значению γ_S, а абсолютная погрешность результата измерения рассчитывается по формуле $\Delta(x) = \gamma_S(x)$.

Если бы эти соотношения оставались справедливыми для всего диапазона возможных значений измеряемой величины x от 0 до $X_к$ ($X_к$ — предел диапазона измерений), то такие измерительные преобразователи были бы наиболее совершенными, так как они имели бы бесконечно широкий рабочий диапазон, т. е. обеспечивали бы с той же погрешностью измерение сколь угодно малых значений x.

Однако реально таких преобразователей не существует, так как невозможно создать преобразователь, полностью лишенный аддитивных погрешностей. Эти погрешности от шума, дрейфа, трения, наводок, вибраций и т. п. неизбежны в любых типах СИ. Поэтому для реальных СИ, погрешность которых нормируется лишь одним числом — погрешностью чувствительности γ_S, — всегда указываются границы рабочего диапазона, в которых такая оценка остается приближенно справедливой.

При чисто аддитивной полосе погрешностей нормировать абсолютное значение погрешности от нечувствительности (вблизи

нуля) Δ_0 неудобно, так как для многопредельных приборов оно будет различным для каждого поддиапазона, и в паспорте прибора пришлось бы перечислять эти значения для всех поддиапазонов.

Поэтому нормируют не абсолютное Δ_0, а приведенное значени этой погрешности: $\gamma_0=\Delta_0/X_n$, где X_n—так называемое нормирующее значение измеряемой величины. Стандарт 8.401 определяет для приборов с равномерной или степенной шкалой нормирующее значение X_n равным верхнему пределу диапазона измерений, в том случае, если нулевая отметка находится на краю или вне шкалы. Если же нулевая отметка находится посредине шкалы, то X_n равно протяженности диапазона измерений (например, для амперметра со шкалой от —30 до +60 А значение X_n = 60 - (—30) =90 А. Значение приведенной погрешности γ_0 выраженное в процентах, используется для обозначения класса точности таких СИ. Однако полагать, как уже указывалось, что вольтметр класса точности 1,0 обеспечивает во всем диапазоне измерений получение результатов с погрешностью ±1 %, — грубейшая ошибка. В действительности текущее значение относительной погрешности $\gamma(x)=\Delta_0/x$, т. е. растет обратно пропорционально x и изменяется по гиперболе (рис.). Таким образом, относительная погрешность $\gamma(x)$ равна классу точности прибора γ_0 лиш на последней отметке шкалы (при $x = X_к$). При $x =0,1X_к$ она в 10 раз больше γ_0, а при дальнейшем уменьшении x стремится к бесконечност

При уменьшении измеряемой величины x до значения абсолютной погрешности вблизи нуля Δ_0 относительная погрешность результата измерения достигает 100%. Такое значение измеряемой величины, называется ***порогом чувствительности*** СИ. РМГ даёт следующее определение порога чувствительности. Это характеристика средства измерений в виде наименьшего значения изменения физической величины, начиная с которого может осуществляться её измерение данным средством.

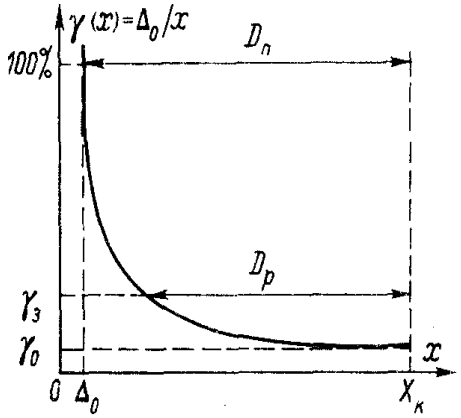


Рис. 1-4

Очевидно, что *диапазон* измеряемых величин $D_п$ (рис) для любого преобразователя ограничивается снизу порогом чувствительности, а сверху — пределом измерений. Так как в области малых значений *x* погрешность измерений очень велика, то *рабочий диапазон* $D_р$ ограничивают снизу таким значением *x,* где относительная погрешность измерений γ(x) не превосходит еще некоторого заранее заданного значения $γ_з$, равного, например, 4, 10 или 20%. Таким образом, рабочий диапазон назначается достаточно произвольно (рис.) и составляет только некоторую часть полного диапазона СИ. В начальной же части шкалы, измерения недопустимы, в чем и заключается отрицательное влияние аддитивной погрешности, не позволяющее использовать один и тот же преобразователь для измерения как больших, так и малых измеряемых величия.

При одновременном присутствии аддитивной и мультипликативной составляющих полоса погрешностей имеет трапецеидальную форму (рис. *а),* а текущее значение абсолютной погрешности *Δ(x)* в функции измеряемой величины *x* описывается соотношением

$$\Delta(x)=\Delta_0+\gamma_s x \qquad (1)$$

где $Δ_0$ — аддитивная, а $γ_s x$ — мультипликативная составляющие абсолютной погрешности.

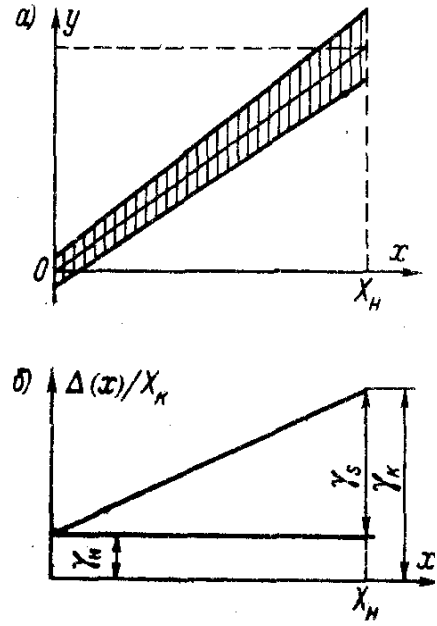


Рис.

Если все члены уравнения (1) разделить на предел измерений *Хк,* то дл приведенного значения погрешности получим

$$\gamma_{пр}(x) = \frac{\Delta(x)}{x_к} = \frac{\Delta_0}{x_к} + \gamma_s \frac{x}{x_к} \qquad (2)$$

Приведенное значение погрешности в начале диапазона (при $x=0$) обозначим через $\Delta_0/x_к=\gamma_н$. Тогда соотношение (2) примет вид

$$\gamma_{пр}(x) = \gamma_н + \gamma_s \frac{x}{x_к}$$

это соотношение отражает график на рис. б.

Таким образом, при наличии у прибора и аддитивной, и мультипликативной составляющих погрешности его приведенная погрешность линейно возрастает от $\gamma_н=\Delta_0/x_к$ в начале диапазона (при $x=0$) до значения $\gamma_к=\gamma_н+\gamma_s$ в конце диапазона (при $x=x_к$).

Относительная погрешность результата измерения, исходя из выражения (1) составляет

$$\gamma(x)=\frac{\Delta(x)}{x}=\frac{\Delta_0}{x}+\gamma_s=\gamma_s+\gamma_н\frac{x_к}{x} \qquad (3)$$

т. е. при $x=x_к$ она будет $\gamma(x)=\gamma_н+\gamma_s=\gamma_к$ а по мере уменьшения x возрастает до бесконечности. Но отличие $\gamma(x)$ от чисто аддитивной погрешности состоит в том, что заметное возрастание $\gamma(x)$ начинается тем позже, чем меньше $\gamma_н$ по сравнению с γ_s

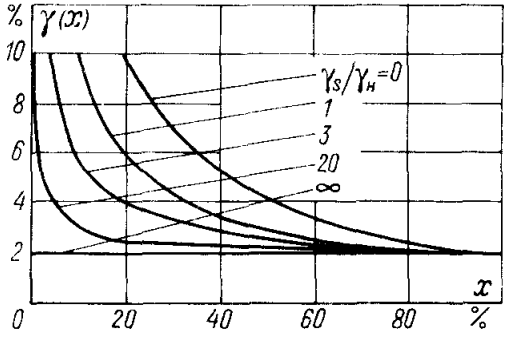


Рис.

Для иллюстрации этого явления на рис. изображены кривы для частного случая $\gamma_н+\gamma_s=\gamma_к=2\%=const$, из которых видно, чт возрастание $\gamma(x)$ происходит при уменьшении x вне зависимости о отношения $\gamma_s/\gamma_н$. Из этих кривых также видно, как расширяется рабочи диапазон СИ по мере увеличения отношения $\gamma_s/\gamma_н$ т. е. уменьшения Δ_0 приближения полосы погрешностей, приведенной на предыдущем ри *а,* к чисто мультипликативной полосе.

Например, если заданное значение погрешности γ

ограничивающее нижнюю границу рабочего диапазона, принять γ_s 4%, то при $\gamma_s/\gamma_н$=0 рабочий диапазон будет двукратным (от 50 до 100% При $\gamma_s/\gamma_н$ = 3 он становится уже пятикратным (от 20 до 100%), а пр $\gamma_s/\gamma_н$ =20—становится двадцатикратным (от 5 до 100%). В последне случае в интервале от 100 до 10% диапазона прибора погрешност результатов измерения почти не изменяется, т. е. большие и малы значения *x* измеряются с одной и той же относительной погрешностью

Форму полосы погрешностей, которая изображена н предыдущем рис., а следовательно, и вытекающие из этого свойств имеют высокоточные потенциометры постоянного тока, цифровы вольтметры и другие высокоточные приборы. Формальны отличительным признаком для них является то, что их класс точност согласно ГОСТ 8.401 обозначается не одним, а двумя числам записываемыми через косую черту, т. е. в виде условной дроби $\gamma_к/\gamma_н$. числителе которой указывается (в процентах) приведенна погрешность $\gamma_к$ в конце диапазона измерений, а в знаменателе – приведенная погрешность $\gamma_н$ в нуле диапазона.

При нормировании погрешностей с помощью формул

Кроме перечисленных разновидностей нормировани погрешностей средств измерений (путем указания классов точности виде γ_s, γ_0, $\gamma_к/\gamma_н$). ГОСТ 8.401 разрешает использовать так называемы специальные формулы нормирования погрешностей. Это необходим для того, чтобы нормировать погрешности средств измерений имеющие более сложный вид, чем тот, который показан н предыдущих рисунках.

К числу таких приборов, например, относятся цифровы частотомеры, погрешность которых зависит не только от измеряемо величины *x,* но и от времени *T,* отводимого для измерения это частоты. Мосты для измерения сопротивлений, отличающиеся тем, чт имеют не только нижний, порог чувствительности, но и верхний преде измерения, ограниченный погрешностью. Особенностью этих приборо является то, что их порог чувствительности, в том числе определяется неопределенностью контактных сопротивлений, а верхний преде измерений ограничен погрешностью при измерении очень больши

сопротивлений из-за приближения измеряемого сопротивления сопротивлению изоляции между зажимами самого моста. В этом случа погрешность результатов измерения описывается трехчленно формулой вида

$$\gamma(x) = \frac{\Delta_0}{x} + \gamma_s + \frac{x}{\Delta_\infty} \qquad (4)$$

где Δ_∞ и Δ_0 — верхний и нижний пороги измеряемых сопротивлени при которых погрешность достигает 100%.

Во всех подобных случаях необходимо внимательно изучат документацию на соответствующий прибор и пользоваться дл вычисления погрешности результата измерения приводимыми в не специальными формулами.

3.3.2 Обозначения классов точности средств измерений.

Согласно ГОСТ 8.401 для указания нормированных значени погрешности чувствительности γ_s приведенной аддитивно погрешности γ_0, приведенных погрешностей в начале $\gamma_н$ и конце диапазона измерений не могут использоваться произвольные числа Выраженные в процентах, они могут иметь значения 6—4—2,5— 1,5— 1,0—0,5—0,2—0,1—0,05—0,02—0,01—0,005—0,002—0,001 и т. д Значение класса точности прибора маркируется на его шкале. Для то чтобы различить, какая из погрешностей обозначена в качестве класс точности, используются следующие условные обозначения.

Если класс точности прибора установлен по значени погрешности чувствительности γ_s, т.е. форма полосы погрешност условно принята чисто мультипликативной, обозначаемое на

шкале значение класса точности обводится кружком. Например 1

обозначает, что $\gamma_s = 1{,}5\%$.

Если же полоса погрешностей принята аддитивной и прибо нормируется приведенной погрешностью нуля γ_0 (таких приборо большинство), то класс точности указывается просто 1,5. На приборах резко неравномерной шкалой, например омметрах, класс точност прибора указывается в долях от длины шкалы и обозначается 1.

Обозначение класса точности в виде, например, 0,02/0,0 указывает, что погрешность прибора нормирована по двучленно формуле с $\gamma_н$ =0,01% и $\gamma_к$ =0,02%.

Таким образом, обозначение класса прибора дает достаточн полную информацию для приближенной оценки погрешносте результатов измерения.

Следует отметить, что хотя ГОСТ 8.401 направлен на то, чтоб нормирование погрешностей СИ производилось единообразно, измерительной практике такого единообразия пока еще нет, так ка используется большое число хороших высокоточных приборо которые были выпущены еще до введения этого стандарта, закупаютс и широко используются приборы иностранного производств нормированные, естественно, не в соответствии с ГОСТ 8.401, и т. д.

Например, погрешность высокоточных потенциометро постоянного тока нормируется чаще всего двучленной формулой, класс точности прибора указывается в виде одного числа - е относительной погрешности чувствительности. В этом случае указани класса точности в виде одного числа γ_s, не является признаком того, чт прибор не имеет аддитивной составляющей погрешности и потребител обязан быть внимательным при расчете погрешностей результато измерения, чтобы не допустить ошибки.

При нормировании погрешностей сложных СИ двучленно формулой (3) ГОСТ 8.401 предусматривает несколько иное е написание. Это производится в том случае, когда текущее значени относительной погрешности $\gamma(x)$ выражается не через значени аддитивной $\gamma_н$ и мультипликативной γ_s составляющих предел допускаемых погрешностей, как в формуле (1), а через указываемые обозначении класса точности приведенные погрешности в начале $\gamma_н$ и конце $\gamma_к$ диапазона измерений. В этом случае, учитывая, что $\gamma_к=\gamma_н+$ соотношение (1) получает вид

$$\gamma(x) = \gamma_к + \gamma_н\left(\frac{X_к}{x} - 1\right) \qquad (5)$$

Практически этим соотношением более удобно пользоватьс для вычисления $\gamma(x)$ по известным x, $X_к$, $\gamma_н$ и $\gamma_к$ чем соотношением (3).

У широкодиапазонных приборов, например мостов дл измерения сопротивлений, в их технической документации вмест указания коэффициентов трехчленной формулы (4) часто приводятс просто диапазоны, в которых погрешность результата измерения н превосходит указанного значения. Например, указывается, чт относительная погрешность не превосходит:

0,5% в диапазоне от 10^2 до 10^4 ом

1% —от 5 до 10^5 Ом

5% — от 0,5 до 10^6 ом

10% —от 0,2 до $2 \cdot 10^6$ Ом

20% —от 0,1 до $4 \cdot 10^6$ Ом.

Как правило, эти данные достаточно точно соответствую трехчленной формуле (4). Поэтому по ним можно определит коэффициенты Δ_0, Δ_∞ и γ_s формулы (4) и использовать ее дл аналитического определения $\gamma(x)$ при любом произвольном значении x.

3.4.Расчёт оценки инструментальной статической погрешности результата измерения по паспортным данным используемого средства измерений.

3.4.1 Вычисление погрешности при различном нормировании класса точности

Результат измерения имеет ценность лишь тогда, когда можно оценить его интервал неопределенности, т. е. степень достоверности. Поэтому согласно ГОСТ 8.011—72 «Показатели точности измерений и формы представления результатов измерений» сообщение о любом результате измерений обязательно должно сопровождаться указанием его погрешности.

Погрешность результата прямого однократного измерения зависит от многих факторов, но, в первую очередь, определяется, естественно, погрешностью используемых средств измерений. Поэтому в первом приближении погрешность результата измерения можно принять равной погрешности, которой в данной точке диапазона измерений

характеризуется используемое средство измерений.

Так как погрешности средств измерений изменяются в диапазоне, то вычисление должно производиться по соответствующим формулам. Вычисляться должна как абсолютная, так и относительная погрешности результата измерения, так как первая из них нужна для округления результата и его правильной записи, а вторая — для однозначной сравнительной характеристики его точности.

Для разных характеристик нормирования погрешностей СИ эти вычисления производятся по-разному, поэтому рассмотрим три характерных случая.

1. Класс точности прибора указан в виде одного числа γ_s заключенного в кружок. Тогда относительная погрешность результата (в процентах) $\gamma(x)=\gamma_s$, а абсолютная его погрешность $\Delta(x)=\gamma_s x/100$.

2. Класс точности прибора указан одним числом γ_0 (без кружка). Тогда абсолютная погрешность результата измерения $\Delta(x)=\gamma_0 X_к$, где $X_к$ — предел измерений, на котором оно производилось, а относительная погрешность измерения (в процентах) находится по формуле

$$\gamma(x) = \frac{\Delta(x)}{x} = \gamma_0 \frac{x_к}{x} \qquad (6)$$

т. е. в этом случае при измерении, кроме отсчета измеряемой величины x, обязательно должен быть зафиксирован и предел измерений $x_к$, иначе впоследствии нельзя будет вычислить погрешность результата.

3. Класс точности прибора указан двумя числами в виде $\gamma_к/\gamma_н$. В этом случае удобнее вычислить относительную погрешность результата по формуле (5), а уже затем найти абсолютную погрешность как $\Delta(x)=\gamma(x)x/100$.

При использовании этих формул полезно помнить, что в формулы для определения $\gamma(x)$ значения γ_s, γ_0, $\gamma_н$ и $\gamma_к$ подставляются в процентах, поэтому и относительная погрешность результата измерения получается также в процентах.

Однако для вычисления абсолютной погрешности $\Delta(x)$ в единицах x значение $\gamma(x)$ (в процентах) необходимо разделить на 100.

3.4.2 Правила округления значений погрешности и результата измерения.

Рассчитывая значения погрешности по формулам (5) и (6), особенно при пользовании калькулятором, значения погрешностей получают с большим числом знаков. Однако исходными данными для расчета являются нормируемые значения погрешности СИ, которые указываются всего с одной или двумя значащими цифрами. Вследствие этого и в окончательном значении рассчитанной погрешности должны быть оставлены только первые одна-две значащие цифры. При этом приходится учитывать следующее. Если полученное число начинается с цифр 1 или 2, то отбрасывание второго знака приводит к очень большой ошибке (до 30—50%), что недопустимо. Если же полученное число начинается, например, с цифры 9, то сохранение второго знака, т. е. указание погрешности, например, 0,94 вместо 0,9, является дезинформацией, так как исходные данные не обеспечивают такой точности.

Исходя из этого, на практике установилось такое правило: если полученное число начинается с цифры, равной или большей $\sqrt{10} \approx 3$, то в нем сохраняется лишь один знак; если же оно начинается с цифр, меньших 3, т. е. с цифр 1 и 2, то в нем сохраняют два знака. В соответствии с этим правилом установлены и нормируемые значения погрешностей средств измерений: в числах 1,5 и 2,5% указываются два знака, но в числах 0,5; 4; 6% указывается лишь один знак.

В итоге можно сформулировать три правила округления рассчитанного значения погрешности и полученного экспериментального результата измерения.

1. Погрешность результата измерения указывается двумя значащими цифрами, если первая из них равна 1 или 2, и одной, — если первая есть 3 и более.

2. Результат измерения округляется до того же десятичного разряда, которым оканчивается округленное значение абсолютной

погрешности.

3. Округление производится лишь в окончательном ответе, а все предварительные вычисления проводят с одним-двумя лишними знаками.

Пример. На вольтметре класса точности 2,5 с пределом измерений 300 В был получен отсчет измеряемого напряжения x =267,5 В.

Расчет погрешности удобнее вести в следующем порядке: необходимо найти абсолютную погрешность, а затем — относительную. Абсолютная погрешность $\Delta(x) = \gamma_0 X_к/100$; при $\gamma_0=2,5\%$ и $X_к = 300\ В$ это дает $\Delta(x) = \dfrac{2,5 \cdot 300}{100} = 7,5\ \ В \approx 8\ В$;

относительная $\gamma(x) = \dfrac{\Delta_0}{x} 100 = \dfrac{7,5}{267,5} 100 = 2,81\% \approx 2,8\%$

Так как первая значащая цифра значения абсолютной погрешности (7,5 В) больше трех, то это значение должно быть округлено по обычным правилам округления до 8 В, но в значении относительной погрешности (2,81%) первая значащая цифра меньше 3, поэтому здесь должны быть сохранены в ответе два десятичных разряда и указано $\gamma(x)$ =2,8%. Полученное значение x=267,5 В должно быть округлено до того же десятичного разряда, которым оканчивается округленное значение абсолютной погрешности, т. е. до целых единиц вольт.

Таким образом, в окончательном ответе должно быть сообщено: Измерение произведено с относительной погрешностью $\gamma(x)$=2,8%. Измеренное напряжение x=(268±8) В или x=268 В ± 8 В.

При этом более наглядно указать пределы интервала неопределенности измеренной величины в виде x=260—276 В или 260 В <$\boldsymbol{x}$< 276 В.

Наряду с изложенными правилами округления значений погрешностей результатов измерения, иногда применяются и другие, некоторые из них изложены в книге Рабинович С.Г. Погрешности измерений. – Л.:Энергия ,1978.

4. Некоторые сведения из теории вероятностей

4.1 Теорема Бернулли

Случайная величина – величина, значение которой изменяется случайным образом при повторении эксперимента. Как правило, можно указать границы, в которых находится случайная величина, а также установить, насколько часто внутри этого интервала интересующая нас случайная величина принимает то или иное значение. Опыт показывает, что в разных случаях некоторые из этих значений появляются чаще, а другие – реже. Совокупность наблюдаемых значений такой величины и частоты появления каждого из этих значений позволяет установить так называемый закон распределения случайной величины, который является её однозначной характеристикой.

Дадим определение понятию вероятность. Вероятность это отношение числа случаев, при котором происходит заданное событие, к общему числу возможных событий. При измерениях физических величин в тех случаях, когда основную роль играют случайные погрешности, все оценки точности измерения можно сделать только с некоторой вероятностью. Действительно, случайные погрешности образуются в результате совокупности ряда мелких не учитываемых причин, каждая из которых вносит незначительный вклад в общую погрешность. Следует считать, что часть из этих погрешностей положительна, часть – отрицательна. Общая погрешность, которая образуется в результате сложения таких элементарных погрешностей, может иметь различные значения, но каждому из них будет соответствовать, в общем случае, разная вероятность.

Формулы сложения, умножения вероятностей, полной вероятности (формула Байеса) оставим на самостоятельную проработку. Для этого рекомендую обратиться к книге

Б.В.Гнеденко, А.Я.Хинчин Элементарное введение в теорию вероятностей. М. Наука 1972. Или любые справочники по математике. Рассмотрим формулу Бернулли.

В случае выполнения взаимно независимых измерений, т.е. измерений при которых вероятность появления того или иного результата в каждом эксперименте не зависит от того, какие результаты наступили или наступят в других экспериментах. В каждом из этих случаев, может получиться (или не получиться) некоторый результат A с вероятностью p не зависящей от номера эксперимента. Если мы имеем k измерений при которых получен результат A, а при остальных n-k этот результат иной, где n – общее число измерений, то вероятность появления k-того измерения с результатом A определится по формуле:

$$P(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Из этой формулы следует важный вывод:
наивероятнейшая доля появлений результата A при
большом числе измерений практически равна вероятности
появления события при отдельном измерении.

Теорема Бернулли, для задач теории измерений может быть сформулирована следующим образом: Если мы производим серию из большого числа n измерений, то с вероятностью близкой к единице, мы можем ожидать, что число k, появлений результата A будет очень близко к своему наивероятнейшему значению, отличаясь от него лишь на незначительную долю общего числа n произведённых измерений.

4.2 Неравенство Чебышёва, закон больших чисел

Очевидно, что знание какой-либо характеристики, определяющей отклонение некой случайной величины от её среднего значения, например, ее среднего квадратического отклонения, позволяет нам создать ориентировочное представление о том, насколько и в дальнейшем следует ожидать

отклонения фактических значений этой величины от ее среднего значения. Однако это замечание само по себе не содержит еще никаких количественных оценок и не дает возможности хотя бы приближенно рассчитать, сколь вероятными могут оказаться большие отклонения. Все это позволяет сделать следующее простое рассуждение, проведенное впервые Чебышевым. Будем исходить из выражения для дисперсии случайной величины *x*

$$D = \sum_{i=1}^{k} (x_i - \bar{x})^2 \, p_i$$

здесь *i* – индекс случайной величины.

Пусть α - любое положительное число; если в этом выражении мы выбросим все члены, в которых $|x_i - \bar{x}| \le \alpha$, и оставим только те, где $|x_i - \bar{x}| > \alpha$, то от этого сумма может только уменьшиться:

$$D \ge \sum_{|x_i - \bar{x}| > \alpha} (x_i - \bar{x})^2 \, p_i$$

Но эта сумма уменьшится еще более, если в каждом члене мы заменим множитель $(x_i - \bar{x})^2$ меньшей величиной α^2;

$$D \ge \alpha^2 \sum_{|x_i - \bar{x}| > \alpha} p_i$$

Сумма, стоящая теперь в правой части, есть сумма вероятностей всех тех значений x_i случайной величины *x*, которые отклоняются от $\bar{x}$ в ту или другую сторону больше, чем на α; по правилу сложения это есть вероятность того, что величина *x* получит какое-либо одно из этих значений. Другими словами, это есть вероятность $P(|x - \bar{x}| > \alpha)$ того, что фактически полученное отклонение окажется больше, чем α; таким образом, мы находим

$$p\left(|x - \bar{x}| > \alpha\right) \le \frac{D_x}{\alpha^2}$$

Полученное соотношение называется *неравенством Чебышёва.* Оно позволяет нам оценить вероятность отклонений, больших чем любое заданное число α, если только известно среднее квадратическое отклонение. Правда, оценка, даваемая неравенством Чебышева, часто оказывается весьма грубой; все же иногда она может быть использована практически, не говоря уже о том, что теоретическое значение ее чрезвычайно велико.

Следствием неравенства Чебышева является закон больших чисел, который можно сформулировать следующим образом: *при очень большом числе случайных явлений средний их результат практически перестаёт быть случайным и может быть предсказан с большой степенью определённости.*

4.3 Нормальный закон распределения

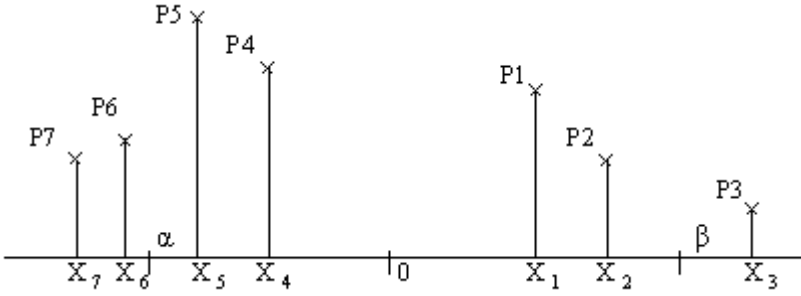
Очевидно, что значительное количество явлений протекает при существенном участии тех или иных случайных величин. Часто до того, как явление, процесс или операция не завершены, все, что мы можем знать об этих случайных величинах,— это их законы распределения, т. е. списки их возможных значений с указанием вероятности каждого из этих значений. Если величина может получать бесчисленное множество различных значений, то предпочтительнее указывать вероятности не отдельных значений ее, а целых участков таких значений (например, вероятность того, что ошибка измерения будет заключена в пределах от —1 *мм* до +1 *мм,* от 0,1 *мм* до 0,25 *мм* и т. д.). Таким образом, необходимо получить возможно точное представление о законе распределения случайной величины. Теоретические исследования показали, что в большом числе встречающихся на практике случаев с достаточным основанием можно ожидать законов распределения определенного типа. Эти законы называются *нормальными законами.* В практике измерений мы имеем дело со случайными погрешностями. В том случае, если все частные погрешности будут независимыми между собой случайными величинами,

причем такими, что каждая из них будет оказывать очень малое влияние на суммарную погрешность, то величина общей погрешности измерений, которую мы хотим исследовать, будет просто суммарным действием всех случайных ошибок, происходящих от отдельных причин. Таким образом интересующая нас ошибка результата измерения является суммой большого числа взаимно независимых случайных величин, и ясно, что подобным же образом дело будет обстоять для большинства случайных погрешностей, с которыми мы. имеем дело на практике.

Теоретические рассуждения показывают, что закон распределения случайной величины, являющейся суммой очень большого числа взаимно независимых случайных величин, уже в силу одного этого, какова бы ни была природа слагаемых, *лишь бы каждое из них было мало по сравнению со всей суммой,* должен быть близок к закону некоторого совершенно определенного типа. Этот тип и есть тип нормальных законов.

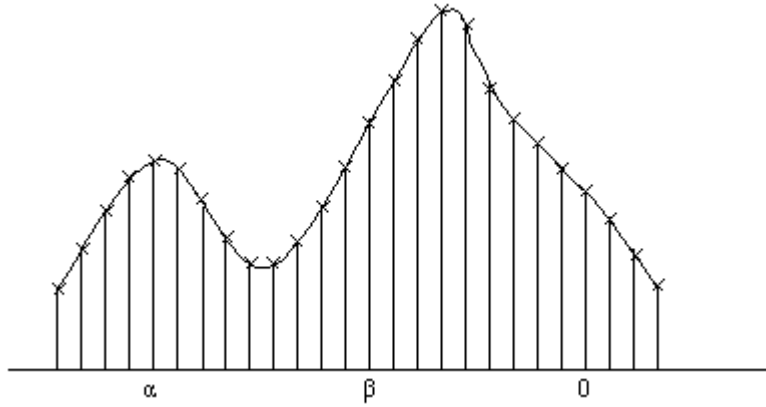
4.3.1 Понятие кривой распределения

Изобразим законы распределения графически, с помощью диаграмм. На горизонтальной прямой откладываются различные возможные значения данной случайной величины, начиная от некоторого начала отсчета О, положительные вправо, отрицательные влево (рис.).

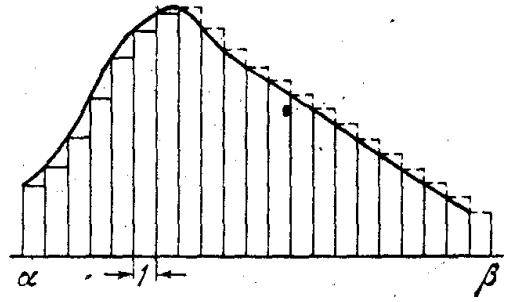


Против, каждого возможного значения откладывают по вертикали вверх вероятность этого значения, абсциссы таких линий называют квантилями. Масштаб в обоих направлениях выбирается таковым, чтобы вся диаграмма имела удобную и легко обозримую форму. Из рисунка видно, что характеризуемая им случайная величина имеет наивероятнейшее значение X_5 (отрицательное) и что по мере удаления возможных значений этой величины от числа X_5, вероятности их довольно быстро убывают. Вероятность того, что величина примет значение, заключенное в каком-либо отрезке (α, β), по правилу сложения равна сумме вероятностей всех возможных значений, лежащих в этом отрезке и геометрически изображается суммой длин вертикальных черточек, расположенных над этим .отрезком; на рис. $P(\alpha<x<\beta)=p_1+p_2+p_4+p_5$. Если же, как это часто бывает на практике, число возможных значений очень велико, то выбирается большой масштаб в горизонтальном направлении, вследствие чего возможные значения располагаются чрезвычайно густо (рис.), так что верхушки проведенных вертикальных черточек сливаются для глаза в одну сплошную кривую линию, которую называют *кривой плотности распределения вероятностей*. Она всегда положительна и подчинена условию:

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$



Здесь вероятность того, что *x* лежит в интервале (α, β) графически изображается суммой длин вертикальных чёрточек, расположенных над отрезком (α, β). Допустим теперь, что расстояние между двумя соседними возможными значениями всегда равно единице; это будет, например, если возможные значения выражаются рядом последовательных целых .чисел, чего практически всегда можно достигнуть, выбирая достаточно мелкую единицу масштаба. Тогда длина каждой вертикальной черточки численно равна площади прямоугольника, высотой которого служит эта черточка, а основанием — равное единице расстояние ее от соседней черточки (рис.). Таким образом, вероятность неравенств, $a<x<\beta$ графически может быть изображена суммой площадей нарисованных на чертеже прямоугольников, расположенных над этим отрезком. Но если возможные значения расположены очень густо, как на предпоследнем рис., то сумма площадей таких прямоугольников практически не будет отличаться от площади криволинейной фигуры, ограниченной снизу отрезком (α, β), сверху — кривой распределения, а с боков — вертикальными черточками, проведенными из точек α и β (рис.).



Таким образом, на графике плотности вероятности вероятность попадания данной случайной величины в любой отрезок просто, и удобно выражается площадью, лежащей над этим отрезком ниже кривой распределения. Если закон распределения данной величины задается такой кривой плотности распределения вероятности, то на ней вовсе не проводят вертикальных отрезков, которые только загромоздили бы собой рисунок; да и самый вопрос о вероятностях отдельных значений здесь теряет свою актуальность; если возможных значений очень много, то вероятности отдельных значений будут, как правило, ничтожно малы и теряют всякий интерес. Таким образом, если случайная величина принимает очень много значений, то нам важно знать вероятности не отдельных этих значений, а вероятности целых отрезков таких значений. Но именно эти вероятности даются площадями на криволинейных диаграммах.

На основании такого подхода вводится понятие квантильных оценок погрешности, т.е. значений погрешности, с заданной доверительной вероятностью $p_д$, как границ интервала неопределённости $\pm\Delta_д$ на протяжении которого встречается $p_д$ процентов всех значений погрешности, а $1- p_д$ процентов общего числа их значений остаются за границами этого интервала. Так как

квантили ограничивающие доверительный интервал погрешности, могут быть выбраны различными, то при сообщении такой оценки должно одновременно указываться значение принятой доверительной вероятности $p_д$.

4.3.2Свойства нормальных кривых распределения

Величина, распределенная по нормальному закону, всегда имеет бесчисленное множество возможных значений; поэтому нормальные законы удобно графически изображать криволинейными диаграммами. На рис. изображено несколько кривых распределения подчинённых нормальному закону.

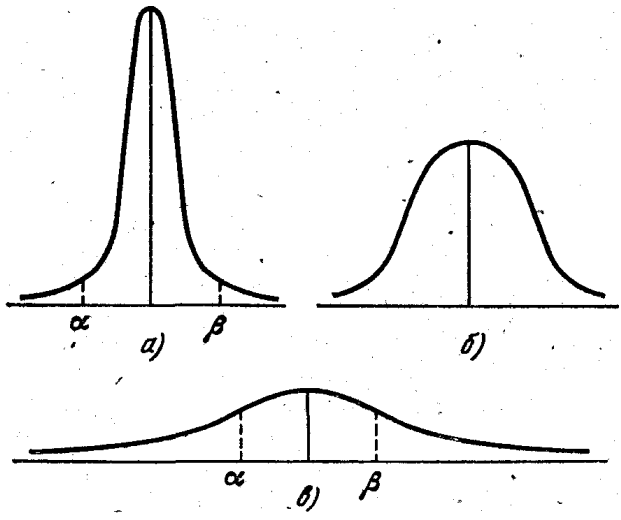


Рис. 15.

1. Все кривые имеют одну наивысшую точку, при удалении от которой вправо или влево они понижаются. Это

означает, что при удалении значений случайной величины от ее наивероятнейшего значения вероятности их постоянно убывают.

2. Все кривые симметричны относительно вертикальной прямой, проведенной через наивысшую точку. Это означает, что все значения, равноудаленные от наивероятнейшего значения, имеют одинаковые вероятности.

3. Все кривые имеют колоколообразную форму, плотность вероятности такого распределения определяется уравнением

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-a)^2}{2\sigma^2} \right)$$

где $\sigma^2 = D$ – дисперсия, a – среднее значение случайной величины Очевидно, что для всякой кривой распределения расположенная под ней площадь равна единице т.е. площадь эта равна вероятности того, что данная случайная величина примет какое бы то ни было из своих значений, т. е. вероятности достоверного события. Отличие отдельных кривых распределения друг от друга состоит в том, что эта суммарная площадь, одна и та же для всех кривых, различным образом распределена между различными участками. Для нормальных законов, как показывают кривые на рис. вопрос в основном заключается в том, какая доля этой суммарной площади сосредоточена над участками, непосредственно примыкающими *к* наивероятнейшему значению, и какая над участками, более удаленными от этого значения. Для закона, изображаемого на рис. *а*, почти вся площадь сосредоточена в непосредственной близости наивероятнейшего значения; это означает, что случайная величина с подавляющей вероятностью и, значит, в подавляющем большинстве случаев принимает значения, близкие к ее наивероятнейшему значению. В связи с тем, что наивероятнейшее значение всегда совпадает со средним значением, мы можем *сказать,* что случайная величина, подчиненная закону *а*, мало рассеяна; в частности, ее дисперсия и среднее квадратическое отклонение малы. Наоборот, в случае, изображенном на рис. *в,* - площадь, сосредоточенная в

непосредственной близости наивероятнейшего значения, составляет лишь небольшую долю суммарной площади. Здесь весьма вероятно, что случайная величина будет получать значения, заметно отклоняющиеся от ее наивероятнейшего значения, она сильно рассеяна и ее дисперсия и среднее квадратическое отклонение велики. Случай б, очевидно, занимает положение, промежуточное между случаями *а* и *в*.

Сформулируем основные свойства нормального распределения

Свойство 1. Если величина x распределена по нормальному закону, то

1. при любых постоянных $c>0$ и d, величина $cx+d$ также распределена по некоторому нормальному закону, и

2. обратно, для любого нормального закона найдется такая (единственная) пара чисел $c>0$,и d, что величина $cx+d$ распределена именно по этому закону. Таким образом, если случайная величина x распределена по нормальному закону, то законы распределения, которым подчиняются величины $cx+d$ при всевозможных значениях постоянных $c>0$ и d,— это все нормальные законы.

Свойство 2. Если случайные величины x и y взаимно независимы и распределены по нормальным законам, то и сумма их $z=x+y$ распределена по некоторому нормальному закону.

Доказательства этих положений приводятся в высшей математике. Указанные свойства позволяют перейти к следующим положениям, имеющим практическое применение.

- Если случайная величина распределена по одному из нормальных законов, то все её особенности могут быть однозначно определены путём задания ее среднего значения и дисперсии. В частности, зная среднее значение и дисперсию такой величины, мы можем вычислить вероятность того, что значение ее будет принадлежать тому или другому произвольно выбранному участку.

- Отношение срединного (вероятного) отклонения к среднему квадратическому отклонению одно и то же для всех нормальных законов.

- В том случае, если x и y – взаимно независимые случайные величины, подчинённые нормальным законам, и $z=x+y$, то тогда вероятные отклонения величин x, y и z связаны между собой выражением $E_z = \sqrt{E_x^2 + E_y^2}$

4.4 Деформация законов распределения при суммировании случайных величин. Центральная предельная теорема.

Особенность законов распределения таких случайных величин, как погрешности приборов и результатов измерений, состоит в их большом разнообразии. Это вызвано тем, что результирующая погрешность прибора или результата измерения складывается из ряда составляющих. Если эти составляющие рассматривать как случайные величины, то суммирование погрешностей сводится к суммированию случайных величин. Но при суммировании случайных величин законы их распределения резко изменяют свою форму.

Закон распределения суммы независимых случайных величин $p(x)=p(x_1+x_2)$, имеющих распределения $p_1(x)$ и $p_2(x)$ называется *композицией* и выражается интегралом свертки. Изменение формы законов распределения при образовании композиции показано на рис.

Так, при суммировании двух равномерно распределенных погрешностей (рис. *а) с* шириной распределений *а* >b результирующая погрешность имеет распределение в форме трапеции с верхним основанием *а—b* и нижним *а + b.* Эту деформацию можно представить себе более наглядно как «размыв» резко ограниченных концов более широкого распределения (шириной *а)* на величину протяженности b менее широкого распределения как это показано штриховыми линиями на рис. *а.*

Композиция двух одинаковых (с шириной а) равномерных распределений является треугольной (так называемое *распределение Симпсона),* так как в этом случае верхнее основание трапеции

обращается в нуль, а нижнее — в *2a*.

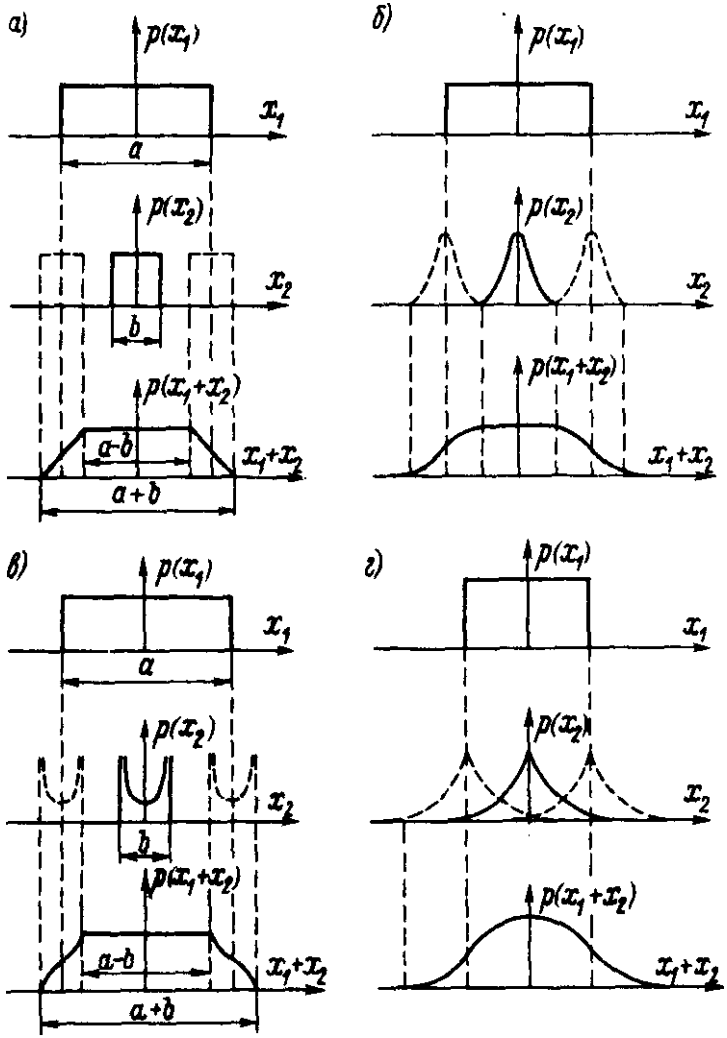
Подобным же образом образуется композиция равномерного и нормального распределений (рис. б), лишь с тем отличием, что подъем и спад по краям результирующего распределения происходит по кривой интегрального закона нормального распределения, аналогично тому, как на рис. *а* он происходил по кривой интегрального закона равномерного распределения (по прямой линии).

Образование композиции равномерного распределения шириной *a* и арксинусоидального распределения шириной b показано на рис. *в.* Композиция представляет собой криволинейную трапецию с верхним основанием *a — b,* нижним *a* + b и спадами по кривым интегрального закона арксинусоидального распределения (функции арксинуса).

Композиция равномерного распределения и распределения Лапласа (двустороннее экспоненциальное распределение на рис. *б)* показана на рис. *г* и имеет длинные, полого спадающие «хвосты» кривой результирующего распределения.

Распределения, показанные на рис. построены без соблюдения относительного масштаба кривых по вертикали. Этот масштаб определяется каждый раз тем, что площадь под любой из кривых плотности должна быть равна единице.

В том случае, если мы имеем дело с композицией n распределений независимых случайных величин, то суммарный закон распределения приводит к закону как угодно близкому к нормальному, причём, чем больше X_n, тем большая степень приближения к нормальному закону. Об этом говорит центральная предельная теорема.

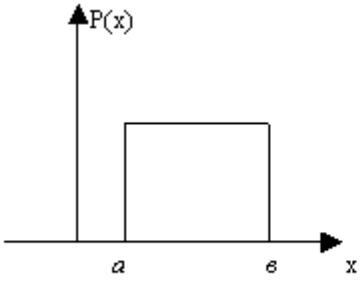


4.5 Другие виды законов распределения

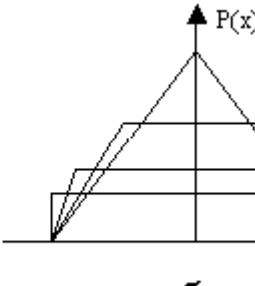
4.5.1 Прямоугольное (равномерное) распределение

Это такое распределение случайной величины, при котором є
плотность вероятности P(x) имеет вид, показанный на рис. *а*. В отличи

от нормального закона распределения непрерывная случайная величин здесь принимает значение только в пределах некоторого конечног интервала. В общем случае равномерное распределение относится классу трапециидальных распределений, которые показаны на рис. Во многих случаях этот вид принимает композиция распределений пр суммировании двух равномерно распределённых случайных величин.



а



б

По прямоугольному закону распределяются случайны составляющие погрешностей измерений, обусловленные сухи трением, погрешности округления отсчётов, погрешности квантовани в цифровых приборах и другие. Выражения для мат. ожидани дисперсии и с.к.о. случайных величин, распределённых по таком закону, имеет вид:

$$M(x) = \frac{b+a}{2} \qquad D(x) = \frac{(b-a)^2}{12} \qquad \sigma(x) = \frac{b-a}{2\sqrt{3}}$$

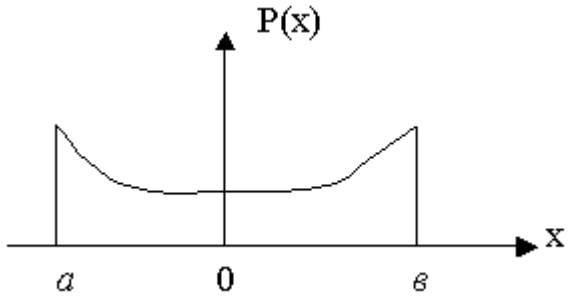
В частном случае равномерного распределения симметричного относительно оси OY с границами *(-a)* и *(+a)*

$$M(x) = 0 \qquad D(x) = \frac{a^2}{\sqrt{3}} \qquad \sigma(x) = \frac{a}{\sqrt[4]{3}}$$

4.5.2 Арксинусоидальные распределения

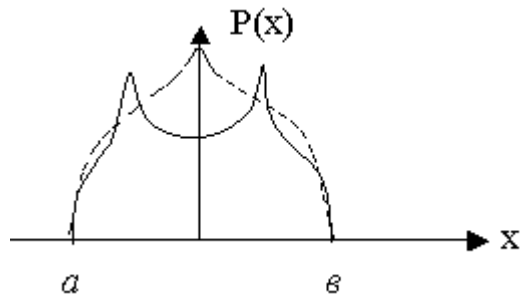
Распределение отсчётов синусоидально изменяющейся во времени величины $x=X_m \sin \omega t$, если моменты этих отсчётов равномерно

распределены во времени называется арксинусоидальным. Эти распределения характерны для погрешностей, возникающих от наводо на линиях связей или на входе прибора, от сетей питания с частотой 50 или 400 Гц. Эта помеха, складываясь с полезным сигналом , создаёт, как правило, аддитивную погрешность и в ряде случаев ограничивает порог чувствительности измерительного устройства. Вид арксинусоидального распределения представлен на рис.



Его плотность описывается выражением $p(x) = \frac{1}{\pi\sqrt{X_m^2 - x^2}}$,

с.к.о. такого распределения $\sigma = \frac{X_m}{\sqrt{2}}$. На практике, однако, напряжение наводки, как правило загрязнено дополнительно присутствием высших гармоник. Распределение суммы двух синусоидально изменяющихся в времени с разными частотами величин является композицией двух арксинусоидальных распределений, в этом случае распределение получает вид, приведённый на рис.



4.5.3 Экспоненциальные распределения

Это широкий класс распределений. В книге П.В.Новицкий, И.А.Зограф Оценка погрешностей результатов измерений Л.Энергоатомиздат 1991., показано, что распределение Лапласа, нормальное и равномерное относятся к разряду экспоненциальных распределений, который может быть описан единой аналитической моделью вида

$$p(x) = \frac{\alpha}{2\lambda\sigma\Gamma(1/\alpha)} \exp\left(-\left| \frac{x - X_ц}{\lambda\sigma} \right|^{\alpha} \right) \qquad (7)$$

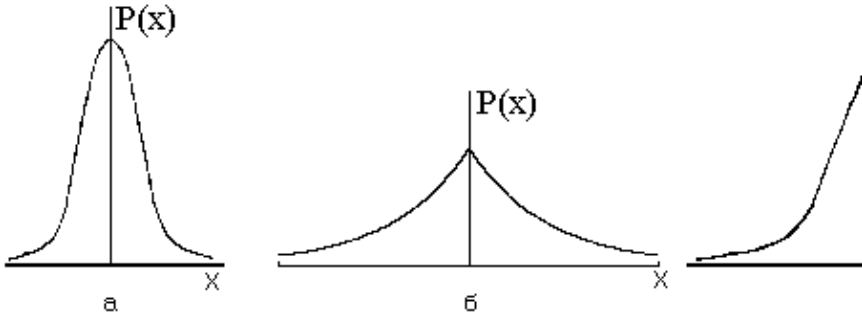
где $\lambda = \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}$, σ - с.к.о., $X_ц$ – координата центра распределения, Γ – гамма функция, α - некоторая характерная для данного распределения постоянная (коэффициент) – его показатель степени.

Для иллюстрации влияния показателя степени α на форму описываемого распределения положим $X_ц=0$, а произведение $\lambda\sigma=1$. Тогда:

$$P(x) = \frac{\alpha}{2\Gamma(1/\alpha)} \exp(-|x|^{\alpha}) = A(\alpha) \exp(-|x|^{\alpha}) \qquad (8)$$

где A(α) – нормирующий множитель распределения, зависящий от α.

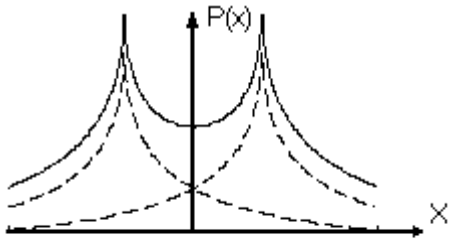
При α<1 аналитические модели (7), (8) описывают распределения с очень пологими спадами, близкие по своим свойствам к распределению Коши рис. *а*. При α=1 они соответствуют распределению Лапласа, рис. *б* а при α=2 – нормальному распределению (Гаусса) рис. *в*, при α>2, она описывает распределение по своим свойствам близкое к трапециидальному и при α→∞ соответствует равномерному распределению. Таким образом, модель (7) очень удобна для описания погрешностей приборов и измерений.

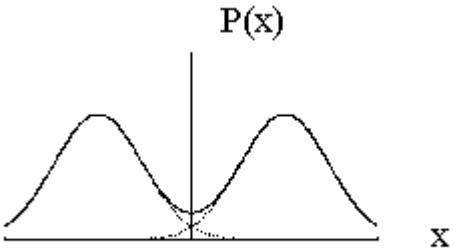


4.5.4 Класс двухмодальных распределений

В практике измерений кривая плотности распределения погрешностей имеет достаточно выраженный максимум, совпадающий с координатой центра распределения, такие распределения называются одномодальными. Трапециидальное распределение за исключением треугольного, не имеют моды, т.е. являются безмодальными. Однако иногда погрешности оказываются распределёнными с кривой плотности , имеющей два симметричных относительно центра максимума. Такие распределения называются симметричными

двухмодальными. В качестве аналитической модели для описания симметричных двухмодальных распределений может использоваться композиция дискретного двузначного распределения и экспоненциальных распределений с произвольным значением показателя степени α. Рис. Подобные кривые обусловлены погрешностями от механического гистерезиса упругих элементов приборов и датчиков.





4.5.5 Семейство законов распределения Стьюдента

Эти законы распределения описывают плотность вероятности значений среднего арифметического, вычисленного по выборке из *n* случайных отсчётов из нормально распределённой генеральной совокупности. Следует отметить, что это не один какой-то закон распределения, а целое семейство законов, так как вид этого распределения зависит от числа *n* отсчётов по которым рассчитывается

среднее значение. В центрированном и нормированном виде семейство распределений Стьюдента описывается выражением

$$p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{x^2}{\nu}\right)^{\frac{\nu+1}{2}}} \qquad (9)$$

где Γ – обозначение гамма функции, ν - так называемое число степеней свободы, зависящее от числа n усредняемых отсчётов: $\nu=n-1$. Для нормированных распределений Стьюдента с числом степеней свободы $\nu>4$ справедливо соотношение:

$$\sigma = \sqrt{\frac{n-1}{n-3}} = \sqrt{\frac{\nu}{\nu-2}} \qquad (10)$$

Как видно из формулы (10), особенность распределений Стьюдента состоит в том, что при $n \le 3$ т.е. при $\nu \le 2$ их с.к.о. становится равным бесконечности, т.е. дисперсионная оценка ширины разброса перестаёт работать. Она одинаково будет равна бесконечности у распределений как с большим, так и с меньшим разбросом. Таким образом, классический аппарат моментов для оценки ширины и формы распределений Стьюдента с малым числом степеней свободы оказывается не работоспособным и их ширина и форма могут быть оценены лишь с использованием доверительных оценок. Этим распределение Стьюдента резко отличается от всех других рассмотренных ранее законов распределений.

4.5.6 Закон распределения Коши

Важен для теории измерений тем, что ему подчиняется, например , распределение отношения двух нормально распределённых центрированных случайных величин. Распределение Коши - это

предельное распределение семейства распределений Стьюдента с минимально возможным числом степеней свободы ν=1. Подставляя в выражение (9) ν=1 получим:

$$p(x) = \frac{\Gamma(1)}{\sqrt{\pi} \cdot \Gamma 0{,}5(1+x^2)} = \frac{1}{\pi(1+x^2)}$$

Кривая плотности вероятности этого распределения была представлена в разделе 4.5.3. Она с первого взгляда кажется похожей н кривую плотности нормального распределения, однако в действительности это совсем не так, ибо её свойства существенно отличны от свойств экспоненциальных распределений. Так дисперсия отсчётов при таком законе распределения вероятностей принципиальн не может быть указана, т.к. определяющий её интеграл расходится. На практике это означает, что оценка дисперсии и с.к.о. определяемые по полученным экспериментальным данным будут неограниченно возрастать, по мере увеличения объёма n этих данных. Естественно, чт использование такой оценки неправомерно. Оценка координаты центра X_u распределения Коши в виде среднего арифметического всех наблюдаемых отсчётов также неправомерна, т.к. её рассеяние, при σ=∞ равно бесконечности т. е. Распределение Коши не имеет даже определённого значения математического ожидания. Однако, если по практически полученным экспериментальным данным, например при измерении активного сопротивления по падению напряжения на нём от шумового тока, построить кривую распределения плотности вероятности, то получим фигуру представленную ранее на рис. ясно показывающую и положение центра распределения на числовой оси, и ширину разброса экспериментальных данных. Таким образом шумовы измерения отнюдь не являются неправомерными. Но классический метод моментов теории вероятностей не способен дать оценку параметров таких распределений. Оценку ширины разброса экспериментальных данных при подобных распределениях, возможно произвести только на основе теории информации.

4.6 Вероятностные оценки ширины распределения

Для оценки величины разброса случайных погрешностей относительно центра, т. е. ширины распределения, на практике используются различные приемы, приводящие к существенно разным результатам. Поэтому целесообразно сопоставить эти приемы и уяснить себе их особенности.

4.6.1 «Предельная», или «максимальная», оценка случайной погрешности.

Она теоретически правомерна только для ограниченных распределений (равномерного, трапецеидального, треугольного, арксинусоидального и т. п.). Для этих распределений действительно существует такое значение $\pm X_m$, которое ограничивает с обеих сторон возможные значения случайной величины. Однако эти распределения являются лишь теоретической идеализацией и реальные распределения погрешностей, строго говоря, им никогда не соответствуют. Кривые плотности реальных распределений погрешностей, за редкими исключениями, не имеют четко выраженных границ. И поэтому указание для них «предельных», или «максимальных», значений неправомерно. На практике такая оценка есть указание наибольшего по модулю отклонения, встретившегося в данном, произвольно ограниченном ряду наблюдений, так как с увеличением объема выборки экспериментальных данных «предельные» значения монотонно возрастают. Предельная погрешность прибора, найденная экспериментально по 100 отсчетам, всегда будет большей, чем найденная по первым 10 отсчетам.

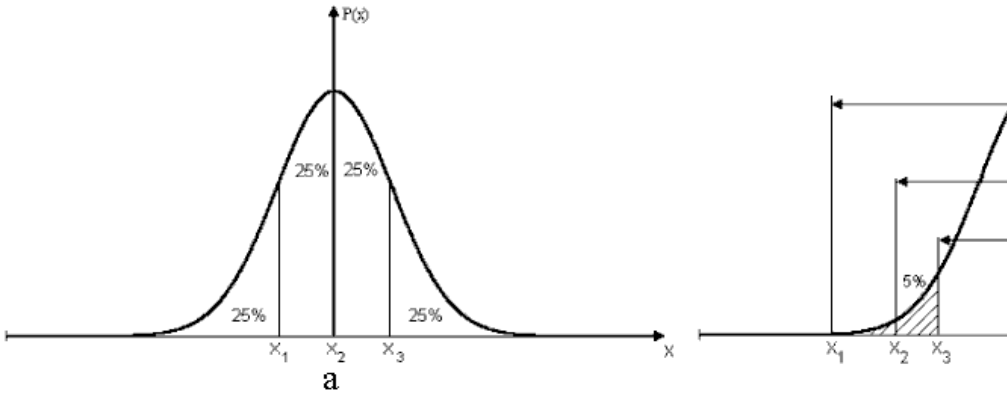
Главным недостатком такой оценки является бессмысленность арифметического суммирования таких «предельных» значений, так как получаемая сумма может превышать действительные погрешности в несколько раз.

4.6.2 Квантильные оценки случайной погрешности.

Площадь, заключенная под кривой плотности распределения

(рис.), согласно правилу нормирования, равна единице, т. е. отражает вероятность всех возможных событий. Эту площадь можно разделить на некоторые части вертикальными линиями. Абсциссы таких линий называют *квантилями.* Так, $x=x_1$ на рис., *а* есть 25%-ная квантиль, так как площадь под кривой $p(x)$ слева от нее составляет 25% всей площади, а справа — 75%. Между x_1 и x_3, т. е. 25%- и 75%-ной квантилями, которые принято называть *сгибами* (или квартилями) данного распределения, заключено 50% всех возможных значений погрешности, а остальные 50% лежат вне этого промежутка.

Медиана $(x=x_2$ на рис. *а)* —это 50%-ная квантиль, так как она дел площадь под кривой $p(x)$ на две равные части.



а

На рис. *б* $x = X_3$ есть 5%-ная квантиль, так как площадь под кривой $p(x)$ слева от нее составляет 5% всей площади. Соответственно значения x_1, x_2,, x_6, и x_7, на рис. б — это 1%-, 2,5%-, 97,5% и 99%-ная квантили. Их удобно обозначать соответственно как $x_{0,01}$, $x_{0,025}$,$x_{0,975}$ и $x_{0,99}$. Интервал значений x между $x_3=x_{0,05}$ и $x_4=x_{0,95}$ охватывает 90% всех возможных значений случайной величины и называется *интерквантильным промежутком с 90%-ной вероятностью.*

На основании такого подхода вводится понятие квантильных оценок погрешности, т. е. значений погрешности с заданной доверительной вероятностью $P_д$, как границ интервала

неопределенности $\pm\Delta_д=\pm P_д/2$, на протяжении которого встречается $P_д$ процентов всех значений погрешности, а $1-P_д$ процентов общего числа их значений остаются за границами этого интервала.

Таким образом, доверительное значение случайной погрешности есть ее максимальное значение с указанной доверительной вероятностью $P_д$, т. е. сообщение, что часть реализации погрешности с вероятностью $1-P_д=q$ может быть и больше указанного значения погрешности.

Так как квантили, ограничивающие доверительный интервал погрешности, могут быть выбраны различными, то при сообщении такой оценки должно одновременно обязательно указываться значение принятой доверительной вероятности $P_д$. Удобнее всего для этого обозначение доверительной погрешности снабжать индексом, численно равным принятой доверительной вероятности, т. е. писать, например $\Delta_{0,9}$ при $P_д=0,9$, $\Delta_{0,95}$ при $P_д=0,95$ и т.д.

Исторически сложилось так, что в разных областях знаний используют различные значения доверительной вероятности, равные 0,5; 0,8; 0,9; 0,95 и 0,99. Так, в высоко ответственной области расчета артиллерийской стрельбы общепринятой является так называемая *срединная ошибка,* т. е. погрешность с доверительной вероятностью $P_д=0,5$, когда 50% всех возможных отклонений меньше ее, а другие 50% — больше (см. рис. *а)*. Доверительная вероятность $P_д=0,8$ является общепринятой в теории и практике оценки надежности средств автоматики, электронной и измерительной техники.

Погрешность $\Delta_{0,9}$ обладает тем уникальным свойством, что для широкого класса наиболее употребляемых законов распределения вероятностей только она имеет однозначное соотношение со средним квадратическим отклонением в виде $\Delta_{0,9}=1,6\sigma$ вне зависимости от вида закона распределения. Поэтому ГОСТ 11.001—73 (сейчас не работает) при отсутствии данных о виде закона распределения для определения двусторонней доверительной вероятности предписывал использовать только $P_д=0,9$. При наличии у прибора, кроме чисто случайной составляющей погрешности, еще и систематической погрешности θ выход возможных значений погрешности за границы доверительного интервала $\pm(\theta+\Delta_{0,9})$ становится практически односторонним. Для

односторонней вероятности выхода за пределы интервала $\pm\Delta_\partial$ при отсутствии данных о виде закона распределения ГОСТ 11.001—73 предписывал использование доверительной вероятности $P_\partial=0,95$.

5. Статические измерения с многократными наблюдениями.

Очевидно, что результат каждого наблюдения отличается от истинного значения измеряемой величины из-за наличия случайной и систематической составляющих погрешности. Очевидно и то, что повторяя наблюдения мы получаем информацию о случайной погрешности. О систематической погрешности из самих наблюдений информацию извлечь нельзя. Чтобы оценить эту погрешность надо знать свойства используемых средств измерений, метод измерения и условия измерений. Вопросами определения систематических погрешностей и их суммированием со случайными мы займёмся отдельно. Сейчас обратим внимание на статистические измерения, при которых многократные наблюдения выполняются с целью уменьшения влияния случайных погрешностей. Таким образом, путём статистической обработки многократных отсчётов решаются три задачи:

- Оценивание случайной погрешности т.е. области неопределённости исходных экспериментальных данных;
- Нахождение более точного усреднённого результата измерений;
- Оценивание погрешности этого усреднённого результата, т.е. более узкой его области неопределённости.

Методы статистической обработки многократных отсчётов (при допущении о неизменности их закона распределения во всех точках модели исследуемого явления) оказываются сходными как в простейшем однофакторном, так и в сложных многофакторных

экспериментах и сводятся к определению числовых оценок параметров соответствующих законов распределения (координаты центра, оценок ширины и формы).

5.1 Достоверность определения доверительного значения погрешности по экспериментальным данным.

Достоинство доверительного значения погрешности состоит в том, что оно может быть достаточно просто оценено прямо по экспериментальным данным. Пусть проведена серия из n измерений. Из наблюдавшихся n случайных погрешностей составляют вариационный ряд, располагая их в порядке возрастания: $\Delta_{(1)} \le \Delta_{(2)} \le \Delta_{(3)} \le ..... \le \Delta_{(n)}$. Далее используется предположение, что каждый из членов вариационного ряда является оценкой соответствующих квантилей, которые делят весь интервал возможных вероятностей (от 0 до 1) на n+1 частей с равными значениями вероятности, иными словами, вероятности попадания значений погрешности в каждый из интервалов $(-\infty, \Delta_{(1)})$, $(\Delta_{(1)}, \Delta_{(2)})$, ... $(\Delta_{(n-1)}, \Delta_n)$ и $(\Delta_{(n)}, +\infty)$
предполагаются одинаковыми, а следовательно, равными *1/(n+1)*-Отсюда каждое из наблюдавшихся значений $\Delta_{(i)}$ может быть принято как оценка *[1/(n+1)]100%*-ной квантили.

Таким образом, практическое определение границ интервала неопределённости $\pm\Delta_\partial$ сводится к тому, что из всех полученных отсчетов отбрасываются наиболее удаленные от центра, а следовательно, самые ненадежные отсчеты. Если при переменном n отбрасывается постоянная относительная доля всех отсчетов, то определяемое по крайним членам оставшегося вариационного ряда значение Δ_∂, в отличие от Δ_m, с ростом длины n серии отсчетов не возрастает, а стабилизируется и оказывается тем более устойчивым, чем больше объем выборки n, не уступая по простоте своего определения «максимальному» значению Δ_m.

При этом следует иметь в виду, что по ограниченным экспериментальным данным мы получаем не точные доверительные

значения, а лишь их приближенные значения — *оценки*. Достоверность квантильных оценок резко повышается с понижением значений $P_д$, а при постоянном $P_д$—с ростом числа отсчетов n. Поэтому квантильные оценки с большими доверительными вероятностями могут быть найдены только при большом числе отсчетов. Действительно, так как вариационный ряд из n членов определяет границы $n + 1$ интервалов, вероятность попаданий в которые принимается нами одинаковой, то при отбрасывании лишь интервалов $(\infty, \Delta_{(1)})$ и $(\Delta_{(n)}, +\infty)$ оценка погрешности может быть определена с доверительной вероятностью, не большей, чем

$$P_д \le \frac{n-1}{n+1}$$

При небольших объемах выборки n фактическая доверительная вероятность может быть существенно меньшей, т. е. достоверность оценки $\Delta_д$, найденной таким путем, очень мала. Для определения оценки $\Delta_д$ с большей достоверностью с каждого из концов вариационного ряда должны быть отброшены не только пустые интервалы от $-\infty$ до $\Delta_{(1)}$ и от $\Delta_{(n)}$ до $+\infty$, но и какое-то число фактических отсчетов. Располагая рядом из n отсчетов и отбрасывая с каждого из концов ряда по $n_{отб}$ отсчетов, можно определить $\Delta_д$ с доверительной вероятностью, не большей, чем

$$P_д \le \frac{n-1-2n_{отб}}{n+1} \qquad (11)$$

Отсюда число отсчетов n, необходимое для определения по экспериментальным данным $\Delta_д$ с заданной вероятностью $P_д$, будет не меньшим, чем

$$P_д \le \frac{1+P_д+2n_{отб}}{1-P_д} \approx \frac{2(1+n_{отб})}{1-P_д} \qquad (12)$$

и для различных значений $P_д$ и $n_{отб}=1$ приведено ниже:

$P_д$	0,8	0,9	0,95	0,98	0,99	0,995

n	20	40	80	200	400	800

Таким образом, по экспериментальным данным легко определить значение Δ_∂ лишь с доверительной вероятностью $P_\partial \le 0{,}95\,(n \approx 80)$, а определение $\Delta_{0,99}$ или $\Delta_{0,997}$ практически трудноосуществимо (нужно>400—1333). При этом необходимо обратить внимание на то, что объем выборки *n,* рассчитанный по формуле (12), обеспечивает лишь выполнение неравенства (11), т. е., взяв, например, выборку объемов $n = 80$ и отбросив с каждой стороны по одному отсчету, получим, что доверительная вероятность не может быть большей, чем 0,95. При этом нет никаких оснований утверждать, что она равна 0,95,так же как утверждать, что она равна 0,8 или 0,3.

Тем не менее очень часто доверительные интервалы погрешности рассчитывают, вводя ничем не обоснованное предположение о том, что вид закона распределения погрешностей будто бы точно известен. В частности, используют прием, заключающийся в вычислении по небольшой выборке в 20—30 отсчетов оценки среднего квадратического отклонения, а затем указывают погрешность с доверительной вероятностью $P_\partial = 0{,}997$, равную $\Delta_{0,997} = 3\sigma$ на основании предположения о нормальности закона распределения.

Из приведенного выше анализа ясно, что такой прием является некорректным вне зависимости от того, допускается ли он сознательно или неосознанно. Следует иметь в виду, что реальные законы распределения погрешностей приборов весьма разнообразны и часто очень далеки от нормального. Для установления действительного хода кривой распределения на ее краях необходимо проведение испытаний, число которых должно быть тем больше, чем большим выбирается значение доверительной вероятности (см. формулу (12)). При малом числе отсчетов (20—30) какие-либо сведения о ходе кривой в области квантилей, соответствующих $P_\partial = 0{,}95—0{,}99$ (не говоря уже о $P_д == 0{,}997$), отсутствуют и утверждения о ходе кривой распределения в этой неисследованной области лишены каких-либо оснований.

Основным недостатком доверительного значения погрешности $\Delta_д$ при произвольно выбираемых $P_д$, как и «максимальной» погрешности $\Delta_м$, является невозможность их суммирования, так как доверительный интервал суммы не равен сумме доверительных интервалов слагаемых. Поэтому приведенные выше применительно к Δ_m, рассуждения остаются в силе и для $\Delta_д$.

5.2 Среднее квадратическое отклонение случайной величины. Закон сложения случайных погрешностей. Связь точности с числом наблюдений.

Среднее квадратическое отклонение σ случайной величины (сокращенно с. к. о.). Это положительное значение квадратного корня из ее дисперсии

$$\sigma = \sqrt{D} = \sqrt{\int_{-\infty}^{+\infty} (x - X_ц)^2 \, p(x) dx}$$

где D — дисперсия, т. е. второй центральный момент случайной величины, а p*(x)* — плотность распределения, $X_ц$ – координата центра распределения.. Для определения оценки дисперсии по экспериментальным данным пользуются соотношением

$$D = \sum_{i=1}^{n} \frac{(x_i - X_ц)^2}{n-1}$$

где x_i—значения отдельных отсчетов; *n*—объем выборки.

Отсюда оценка с. к. о. определяется как

$$\sigma = \sqrt{\sum_{i-1}^{n} \frac{(x_i - X_ц)^2}{n-1}}$$

Основным достоинством оценки разброса случайных величин средним квадратическим значением σ является возможность определения дисперсии суммы статистически независимых величин независимо от разнообразия законов распределения каждой из суммируемых величин и деформации законов распределения при

образовании композиций.

Таким образом для того, чтобы отдельные составляющие погрешности средств измерений можно было суммировать расчётным путём, они должны быть предварительно представлены своими средними квадратическими значениями σ, а не максимальными Δ_m или доверительными Δ_∂ значениями. При этом открывается возможность расчётным путём не только складывать любое число составляющих погрешности, что необходимо при анализе точности косвенных измерений или сложных измерительных устройств, но и достаточно точно вычитать погрешности, что необходимо при синтезе методов измерений или сложных устройств с заданной результирующей погрешностью. Действительно, если $\sigma_\Sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$, то $\sigma_2 = \sqrt{\sigma_\Sigma^2 - \sigma_1^2}$. Это правомерно для независимых случайных величин.

Из предыдущего следует, что $\sigma_\Sigma^2 = \sigma_1^2 + \sigma_2^2$. В случае сложения не двух, а большего числа дисперсий или с.к.о. независимых случайных величин закон сложения будет таким же. Следует обратить внимание на то, что как вы уже убедились, для нахождения суммарной погрешности следует складывать не сами погрешности, а их квадраты. В том случае. Если мы складываем вероятности, то закон сложения будет тем же. $P_\Sigma = \sqrt{P_1^2 + P_2^2}$.

Из закона сложения погрешностей следуют два очень важных вывода. Первый относится к роли каждой из погрешностей в общей погрешности результата. Он состоит в том, что значение отдельных погрешностей очень быстро падает по мере их уменьшения. Поясним сказанное примером: пусть *X* и Y - два слагаемых, определенных со средними квадратическими погрешностями σ_x и σ_y , причем известно, что σ_y. в два раза меньше, чем σ_x. Тогда погрешность суммы $Z=X+Y$ будет

$$\sigma_Z^2 = \sigma_x^2 + \sigma_y^2 = \sigma_x^2 + \left(\frac{\sigma_x}{2}\right)^2 = \frac{5}{4}\sigma_x^2$$

Откуда $\sigma_z \approx 1{,}1\sigma_x$. Следовательно, если одна из погрешностей в два раз меньше другой, то общая погрешность возросла за счет этой меньшей погрешности всего на 10%, что обычно играет очень малую роль. Это означает, что если мы хотим повысить точность измерений величины *Z*, то нам нужно в первую очередь стремиться уменьшить ту погрешность измерения, которая больше, т.е. погрешность измерения величины *X*. Если оставим точность измерения *X* неизменной, то, как бы мы ни повышали точность измерения слагаемого *Y*, нам не удастся уменьшить погрешность конечного результата измерений величины *Z* более чем на 10%.

Этот вывод всегда нужно иметь в виду, и для повышения точности измерений в первую очередь уменьшать погрешность, имеющую наибольшее значение. Конечно, если слагаемых много, а не два, как в нашем примере, то и малые погрешности могут внести заметный вклад в суммарную погрешность.

Если нужная нам величина *Z*; является разностью двух независимо измеряемых величин *X* и *Y*, то из выражения для суммы с.к.о. следует, что ее относительная погрешность

$$\frac{\Delta Z}{Z} = \frac{\sqrt{\Delta X^2 + \Delta Y^2}}{X - Y}$$

где *ΔX, ΔY, ΔZ* – погрешности измерений величин *X, Y, Z*.

Очевидно, что она будет тем больше, чем меньше $|X - Y|$, и относительная погрешность возрастает до бесконечности, если *X* стремиться к *Y*.

Это означает, что невозможно добиться хорошей точности определения какой-либо величины, строя измерения так, что она находится как небольшая разность результатов независимых измерений двух величин, существенно превышающих искомую. В противоположность этому относительная погрешность суммы

$$\frac{\Delta Z}{Z} = \frac{\sqrt{\Delta X^2 + \Delta Y^2}}{X + Y}$$

очевидно не зависит от соотношения величин *X* и Y.

Следующий вывод, вытекающий из закона сложения погрешностей, относится к определению погрешности среднего арифметического. Следует отметить, что среднее арифметическое из ряда измерений числом *n* отягощено меньшей погрешностью, чем результат каждого отдельного измерения. Запишем этот вывод в количественной форме. Пусть x_1, x_2, x_n результаты отдельных измерений, причем каждое из них характеризуется одной и той же дисперсией D . Образуем величину Y , равную

$$Y = \frac{1}{n}\sum_{1}^{n} x_i = \frac{1}{n}x_i + \frac{1}{n}x_2 + ... + \frac{1}{n}x_n$$

Дисперсии этой величины D_y определяются в соответствии с формулой сложения дисперсий

$$\sigma_{\Sigma}^2 = \sigma_1^2 + \sigma_2^2 \text{ как } D_y = \frac{D}{n}$$

Но y , по определению, это - среднее арифметическое из всех величин x_i и мы можем написать

$$\sigma_y = \sigma_{\bar{x}} = \frac{\sigma_i}{\sqrt{n}} \qquad (13)$$

Средняя квадратическая погрешность среднего арифметического равна средней квадратической погрешности отдельного результата измерений, деленной на корень квадратный из числа измерений. Это - фундаментальный закон возрастания точности при росте числа наблюдений. Мы его уже обсуждали в разделе 5.1. Из него следует, что, желая повысить точность измерений в 2 раза, мы должны сделать вместо одного - четыре измерения; чтобы повысить точность в 3 раза, нужно увеличить число измерений в 9 раз, и, наконец, увеличение числа наблюдений в 100 раз приведет к десятикратному увеличению точности измерений.

Разумеется, это рассуждение относится лишь к измерениям, при которых точность результата полностью определяется случайной погрешностью. В этих условиях, как уже указывалось, выбрав *n* достаточно большим, мы можем существенно уменьшить погрешность результата. Такой метод повышения точности широко

используется. Отметим, что повышение точности измерений целесообразно производить таким способом в том случае, если погрешность измерительного средства намного превышает цену деления шкалы отсчёта. В этом случае погрешность можно свести к значению цены деления. Очевидно, что получить точность выше цены деления не представляется возможным т.к. при отсчёте показаний округления производятся до целых делений шкалы. С помощью такого приёма легко снизить погрешность от вариации показаний.

При практической работе очень важно строго разграничивать применение средней квадратической погрешности отдельного измерения σ_i и средней квадратической погрешности среднего арифметического $\sigma_{\bar{x}}$

Последняя применяется всегда, когда нам нужно оценить погрешность того значения, которое мы получили в результате всех произведенных измерений.

В тех случаях, когда мы хотим характеризовать точность применяемого способа измерений, следует использовать погрешность σ_i , если n, достаточно велико.

5.3 .Статистические веса

Допустим, что одним и тем же методом с одинаковой степенью точности выполнено k серий измерений. В первой серии число измерений n_1, во второй – n_2, и т.д., в k-ой - n_k если каждое измерение характеризуется погрешностью σ, то погрешность среднего арифметического для серии с номером i будет в соответствии с формулой (13)

$$\sigma_i = \frac{\sigma}{\sqrt{n_i}}$$

Очевидно, что если в одной серии сделано в четыре раза

больше измерений, чем в другой, то погрешность результата одной серии будет соответственно в два раза меньше.

Если мы захотим для повышения точности результата усреднять его по средним значениям для обеих серий, то должны учитывать то обстоятельство, что один результат получен с вдвое меньшей погрешностью. С этой целью вводится понятие статистического веса или просто веса наблюдений. В приведенном примере за статистический вес P следует принять число, пропорциональное количеству наблюдений, выполненных в серии, в этом случае выражение для статистического веса серии измерений с номером i будет:

$$P_i = \frac{K}{\sigma_i^2}$$

здесь $K=k\sigma^2$, k - количество серий измерений

Если имеется ряд результатов измерений, вообще выполненных в разных условиях, причем для каждого результата известна средняя квадратическая погрешность σ_i , то и в этом случае можно для совместной обработки результатов приписать им соответствующие статистические веса P_i положив также

$$P_i = \frac{B}{\sigma_i^2}$$

Здесь B - произвольное число. Оно обычно выбирается таким, чтобы P_i. были по возможности небольшими целыми числами. Часто бывает, что σ_i, заранее неизвестны и отдельным измерениям приписываются веса на основании разного рода качественных соображений, связанных, например, с квалификацией наблюдателей, производивших

отдельные измерения, различием в точности измерительных инструментов, с которыми они производились, и т.п.

Введение статистических весов, определенных на глаз, разумеется нельзя считать строгим приемом, однако он дает возможность хоть как-то использовать всю совокупность наблюдений. Следует иметь в виду, что если веса отдельных наблюдений различаются в 10 и более раз (σ_i и σ_k различаются более чем в три раза), то обычно лучше просто отбросить из рассмотрения наблюдения с малыми весами, так как их учет может только испортить хорошие результаты.

Если нам известна совокупность ряда результатов x_i с соответствующими им статистическими весами P_i, то за наивероятнейшее значение измеряемой величины следует принять уже не среднее арифметическое, а взвешенное среднее, которое также обозначим $\overline{X}$

$$\overline{X} = \frac{\sum_1^n P_i x_i}{\sum_1^n P_i}$$

Разумеется, если $P_1=P_2=\ldots=P_n$, то выражение превращается в формулу для среднего арифметического.

Среднюю квадратическую погрешность для $\overline{X}$ можно получить из выражения

$$\sigma_{\overline{X}} = \sqrt{\frac{\sum_1^n P_i\left(\overline{X} - x_i\right)^2}{(n-1)\sum_1^n P_i}}$$

При выборе нужного числа измерений предполагаем, что систематическая погрешность метода достаточно мала.

5.4 Обнаружение промахов и грубых погрешностей

Промахи и грубые погрешности (разной литературе этот термин трактуется по разному) - это аномальные результаты измерений, которые не принадлежат рассматриваемой генеральной совокупности и нарушают однородность выборки. Промахи нужно обнаруживать и исключать. Признаками наличия промахов является резкое искажение симметричности функции плотности вероятностей, нарушение ее монотонности.

В математической статистике существует несколько критериев проверки возможности исключения промахов, например критерии Романовского, Ирвина, Граббса. Задача решается статистическими методами, основанными на том, что распределение, к которому относится рассматриваемая группа наблюдений, можно считать нормальным. При использовании критерия Романовского- первоначально по выборке вычисляют среднее арифметическое $\overline{X}$ и с. к. о. σ без учета члена ряда, предполагаемого как промах рассчитывают :

$$t = \frac{\overline{X} - x_1}{\sigma} \quad \text{или} \quad t = \frac{\overline{X} - x_n}{\sigma}$$

где x_1 и x_n – первый и последний члены ранжированного ряда.

Затем для заданного уровня значимости α и объема выборки *n* по табл. находят коэффициент t_α. Если $t > t_\alpha$, то с вероятностью P=1-α проверяемый член можно из выборки исключить. Аналогично для λ

$$\lambda = \frac{x_n - x_{n-1}}{\sigma}$$

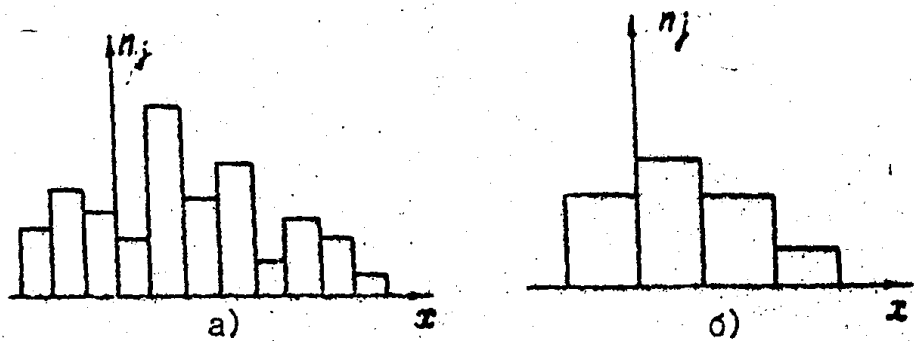
где x_n - член ряда, предполагаемый как промах, значение которого сравнивается с табличным. Указанные таблицы приводятся в литературе по мат статистике или в книгах: С.Г.Рабинович Погрешности измерений. Л.Энергия 1978; Е.П.Осадчий, В.И.Карпов Методы проведения эксперимента при проектировании измерительных элементов систем автоматики и телемеханики. Пензенский политехнический институт. Учебное пособие. Пенза 1988

В измерительной технике предлагаются различные правила обнаружения и исключения аномальных результатов измерения. Простейшим .является «правило 3σ»,при котором все $|x_i| \geq 3\sigma$ рассматриваются как промахи и исключаются.

Все рассмотренные рекомендации при установлении границ не учитывают вид закона распределения экспериментальных данных. Исходя из предположения о нормальности. .Однако результаты исследований последнего десятилетия показывают, что границы должны быть функциями той или иной характеристики вида закона распределения. Действительно, для такого ограниченного распределения как равномерное, весь диапазон которого заключен в границах 3σ, значение $x_i \geq 1{,}8\sigma$ уже является промахом; при нормальном же распределении это .значение лежит внутри границ промахов. В работах П.В.Новицкого предложен критерий $t\sigma$, где $t = 1 + \sqrt{2(\varepsilon - 1)}$; ε-эксцесс распределения (характеристика вида распределения). Для нормального закона ε=3 и t=3, т.е. приходим к «правилу3σ».

5.5 Способы, группирования данных. Методы установления вида закона распределения.

Графически интервальный вариационный ряд можно представать в виде гистограммы (рис.). На каждом интервале группирования как на основании строится прямоугольный столбец, высота которого равна либо числу результатов, попавших в *j*-тый интервал группирования *nj*, либо *nj/n* , либо *nj*/(n+Δx). Гистограмма является статистическим аналогом функции плотности распределения и позволяет судить о виде закона распределения. Соответствие вида гистограммы фактическому распределению зависит от числа интервалов группирования *m*. Если число интервалов *m* слишком велико, то гистограмма сильно изрезана, имеет провалы из-за малого числа точек, попавших в некоторые интервалы, ее вид будет значительно зависеть от случайностей экспериментальных данных (рис.а).



Рас.

Если число интервалов мало, то будут сглаживаться некоторые важные особенности распределения (рис.б). Очевидно, что должно существовать некоторое оптимальное число интервалов группирования, при котором гистограмма имеет плавный характер без разрывов и в то же время отражает особенности распределения. Рекомендации различных авторов по выбору числа интервалов группирования на основе анализа литературы приведены в табл.

Автор	Рекоменд. ф-ла
Старджес	$m=3,3\lg n+1$
Брукс и Каррузерс	$m=5\lg n$
Хайнхольд Гаеде	$m=\sqrt{n}$
Таушанов и Тонева	$m=4\lg n$

Приведенные рекомендации определяют число интервалов группирования в зависимости только от числа измерений n и не учитывают вид закона распределения. Однако для описания,

например, островершинного распределения нужны более мелкие интервалы, чем для описания равномерного. Следовательно, величина *m* должна зависеть от вида распределения. Один из путей учета вида распределения - выбор интервалов такой ширины, чтобы число попаданий значении результатов измерений в них было одинаковым. В этом случае автоматически в областях с высокой плотностью значений будут более узкие интервалы. Однако в подавляющем большинстве случаев группирование производится с интервалами одинаковой ширины. В этом случае величина *m* должна быть функцией не только *n*, но и какой-либо характеристики вида закона распределения. Целесообразно выполнять расчёт *m* по следующему выражению

$$m = \frac{1}{3}\sqrt[5]{\frac{n^2}{\Omega^8}}$$

где Ω - контрэксцесс, характеризующий вид распределения; *n* – число измерений. Значения контрэксцесса для различных видов закона распределения приводятся в табл.

Полную информацию об их значениях можно получить в книге Новицкого П.В. о которой мы уже говорили.

На основе анализа наиболее часто встречающихся распределений при исследовании метрологических характеристик средств измерения можно сказать, что

$$m_{\min} = 0{,}55n^{0{,}4} \qquad m_{\max} = 1{,}25n^{0{,}4}$$

(14)

Для точных расчётов значение *m* следует уточнить путем последовательных приближений.

Если из теории или из опыта известно, что распределение результатов измерения должно быть симметричным, то значение *m* необходимо округлить в большую сторону до ближайшего нечетного числа. Нечетное *m* обязательно при исследовании погрешностей, поскольку законы распределения погрешностей, как правило, симметричны.

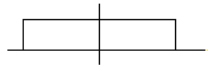
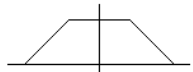
Таким образом, для ориентировочного расчёта значения m следует воспользоваться выражениями (14), после чего выполнить группирование данных. Для точного определения вида закона распределения следует уточнить значение контрэксцесса по таблице и повторить группирование.

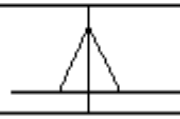
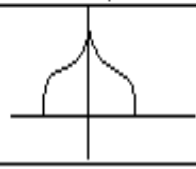
Ориентировочное суждение о виде закона распределения, как было указано, можно сделать по виду гистограммы. Гистограмму рекомендуется строить следующим образом. На ось измеряемой величины *x* наносят центр распределения (рис.) По формулам (14) приближенно оценивают число интервалов группирования. Затем определяют ширину интервала группирования; для этого диапазон значений результатов измерений (без промахов) делят на число интервалов группирования *m*. Первый интервал наносят на график так, чтобы центр распределения лежал в середине этого интервала. Так рекомендуется поступать потому, что в большинстве измерительных задач распределение результатов измерения является симметричным и одномодальным.

Симметрично располагают остальные интервалы группирования слева и справа относительно центрального интервала. Построение столбцов гистограммы производится в соответствие с предыдущими рекомендациями. По виду гистограммы в случае, показанном на рис.а закон распределения скорее всего нормальный, а в случае, показанном на рис. б, - равномерный. Однако такое определенное суждение можно высказать далеко не всегда. Часто возможны весьма широкие альтернативы.

Оценку вида закона распределения можно провести также по числовым характеристикам. Это более сложный способ, который ,однако, даёт более точные результаты.

Значения контрэксцесса для различных видов распределения.

№ п/п	Вид распределения	Ω
Трапецеидальное распределение.		
1	 (равномерное распределение)	0,745
2		0,728
3		0,704
4		0,677

5		0,645
Арксинусоидальное распределение		
6		0,816
7		0,752
8		0,667
Экспоненциальные распределения		
9	$P(x)=\frac{1}{48}e^{-4\sqrt{\lvert x\rvert}}$	0,0467
10	$P(x)=\frac{1}{12}e^{-3\sqrt{\lvert x\rvert}}$	0,0966
11	$P(x)=\frac{1}{4}e^{-\sqrt{\lvert x\rvert}}$	0,199
12	$P(x)=\frac{1}{2}e^{-\lvert x\rvert}$ (распределение Лапласа)	0,408
13	$P(x)=\frac{1}{\sqrt{\pi}}e^{-x^2}$ (Распределение Гаусса)	0,577
Распределение Стьюдента с числом степени свободы		
14	5	0,333
15	6	0,408
16	7	0,447
17	10	0,5
18	∞	0,577

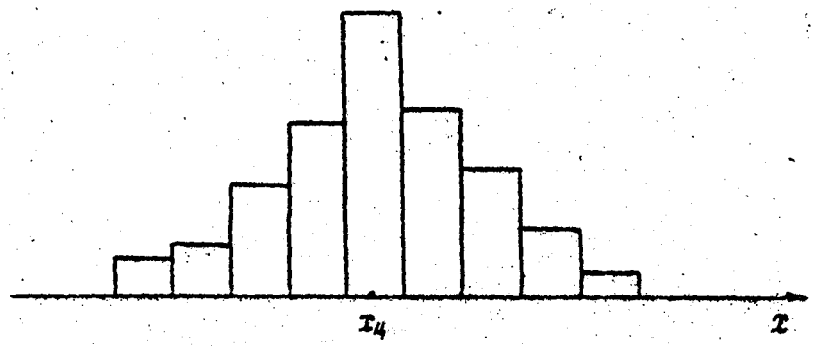


Рис а

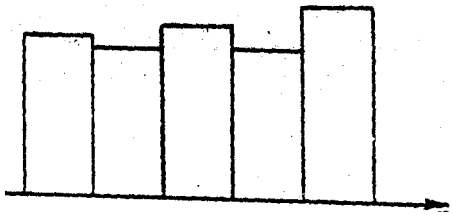


Рис б

5.6 Практические методы проверки нормальности распределения случайных погрешностей

Во многих случаях для расчётов погрешности закон распределения принимают нормальным, однако это утверждение требует соответствующей проверки. Обычно задача ставится так: имеется группа результатов наблюдений и высказывается гипотеза о том, что эти наблюдения можно считать реализациями случайной величины с формой функции распределения соответствующей

нормальной. Затем методами математической статистики эта гипотеза проверяется и либо принимается, либо отвергается.

При большом числе наблюдений *(n* > 50) лучшими критериями проверки данной гипотезы считают критерий согласия К. Пирсона (критерий χ^2) для группированных наблюдений и критерий Р. Мизеса — Н. В. Смирнова (критерий ω^2) для не группированных наблюдений.

Остановимся на критерии χ^2. Идея этого метода состоит в контроле отклонений гистограммы экспериментальных данных от гистограммы с таким же числом интервалов, построенной на основе нормального распределения. Сумма квадратов разностей частот по интервалам не должна превышать значений χ^2, для которых составлены таблицы в зависимости от уровня значимости критерия q и числа степеней свободы $k=m-3$, где m - число интервалов.

Вычисления ведутся по следующей схеме.

1. Вычисляют среднее арифметическое наблюдений и оценку среднего квадратического отклонений по формулам:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (15)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \qquad (16)$$

здесь x_i – значение наблюдений, n – их число.

2. Группируют наблюдения по интервалам. При числе наблюдений 40—100 обычно принимают. 5—9 интервалов. Для каждого интервала вычисляют середину x_{i0} и подсчитывают число наблюдений, попавшее в каждый интервал, φ_{in}

3. Вычисляют число наблюдений для каждого из интервалов, теоретически соответствующее нормальному распределению. Для этого сначала от реальных середин интервалов x_{i0} переходят к нормированным значениям:

$$z_i = \frac{x_{i0} - \bar{x}}{\sigma}$$

Затем для каждого значения z_i находят значение функции плотности вероятностей

$$f(z_1) = \frac{1}{2\pi} e^{-\frac{z_i^2}{2}}$$

Вычисление значения функции плотности вероятности нормированного нормального распределения ведётся с помощью таблицы, предлагаемой в справочниках по математике и другой специальной литературе (например в книге С..Рабинович Погрешности измерений. Энергия Л.1978.) Теперь можно вычислить ту часть φ_i, общего числа имеющихся наблюдений, которая теоретически должна была быть в каждом
из интервалов:

$$\varphi_i = n \frac{h}{\sigma} f(z_1)$$

здесь $h=X_{i0+1} - X_{i0}$ - длина интервала, принятая при построении гистограммы.

4. Если. в какой-либо интервал теоретически попадает меньше 5 наблюдений, то его в обеих гистограммах соединяют с соседним интервалом. Затем определяют число степеней свободы $k=m-3$. Здесь уже m – число интервалов после укрупнения.

5. Вычисляют показатель разности частот χ^2 :

$$\chi^2 = \sum_{i=1}^{m} \chi_i^2$$

где $\chi_i^2 = \frac{(\varphi_{in} - \varphi_i)^2}{\varphi_i}$

Выбирают уровень значимости критерия q. Уровень значимости должен быть достаточно малым, чтобы была мала вероятность отклонить правильную гипотезу. С другой стороны, слишком малое значение q также увеличивает вероятность принять ложную гипотезу. В связи с этим по уровню значимости q и числу

степеней свободы k по специальным таблицам находят границу критической области χ_q^2, отвечающую условию

$$P(\chi^2 > \chi_q^2) = q$$

Вероятность того, что. получаемое значение χ^2 превышает χ_q^2 равна q и мала. Поэтому, если оказывается, что $\chi^2 > \chi_q^2$, то гипотеза о нормальности отвергается. Если $\chi^2 < \chi_q^2$, то гипотеза о нормальности принимается.

Чем меньше q, тем при том же k больше значение χ_q^2, тем легче выполняется условие $\chi^2 < \chi_q^2$, и принимается проверяемая гипотеза. Но при этом увеличивается вероятность ошибки другого рода, о которой уже говорилось. Поэтому нецелесообразно выбирать q в интервале 0,02<q<0,01. При слишком большом *q*, как указывалось выше, возрастает вероятность ошибки и, кроме того, снижается чувствительность критерия.

По существу эти операции можно охарактеризовать как определение меры расхождения между эмпирическим и теоретическим распределением. Эта мера расхождения определяется величиной χ^2.

Следует отметить, что данный критерий позволяет проверять соответствие эмпирических данных любому теоретическому распределению, а не только нормальному. Однако этот критерий, как, впрочем, и другие критерии согласия, не позволяет установить вид распределения наблюдений, а лишь дает возможность проверить, допустимо ли отнести их к нормальному или иному, выбранному заранее распределению.

Критерий χ^2 рекомендуется применять при наличии большой выборки ($n \geq 100$), так как мера расхождения следует распределению χ^2 лишь при большом n . Число попаданий величины *x* в любой из интервалов группирования должно быть не менее 5..Для проверки гипотез о законе распределения применяются также критерий Колмогорова-Смирнова и критерий ω^2. Для проверки нормальности

распределения используют критерий Романовского. При этом меру расхождения рассчитывают так же, как и по критерию χ^2, а затем вычисляют значение

$$r = \frac{\left|\chi^2 - k\right|}{\sqrt{2k}}$$

где k - число степеней свободы такое же**,** как при применении критерия χ^2

Расхождение между теоретическим и эмпирическим распределениями считается несущественным, если r имеет абсолютное значение меньше трех.

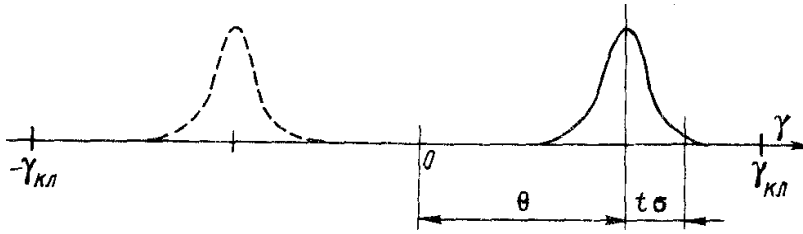
5.7 Систематические погрешности

5.7.1 Учёт систематических погрешностей при оценке результатов статистической обработки многократных отсчётов

Оценивая погрешность результата, полученного при статистической обработке многократных отсчетов, нельзя забывать о том, что при усреднении уменьшаются лишь случайные погрешности, в то время как систематическая погрешность, присутствовавшая во всех усредняемых отсчетах, остается без изменения.

Эту особенность систематических погрешностей следует иметь в виду как при ручной статистической обработке экспериментальных данных, так и, особенно, при организации усреднения многократных отсчетов в компьютерных программах. В этом случае случайная погрешность будет уменьшена в десятки или сотни раз и погрешность усредненного результата будет характеризоваться не этой ничтожной случайной погрешностью, а определяться не зависящим от числа усредняемых отсчетов значением систематической составляющей погрешности.

То или иное значение систематической составляющей погрешности, повторяющейся во всех отсчетах, а поэтому не усредняемой при статистической обработке, присутствует при любом измерении. В общем случае положение кривой плотности вероятности погрешности в границах предела нормированной классом прибора погрешности $\pm\gamma_{кл}$, может быть представлено графиком рис. где сплошной линией показано положение кривой плотности вероятности погрешности при положительном значении систематической составляющей θ, а штриховой кривой — при отрицательном. Таково вероятностное описание систематической и случайной погрешности.



Рассмотрим особенности суммирования систематической и случайной составляющих погрешности. В некоторых книгах по метрологии нередко утверждается, что складывать между собой случайные и систематические составляющие погрешности «нельзя с принципиальных позиций, так как систематические и случайные погрешности имеют разную природу». Однако это утверждение вряд ли бесспорно хотя бы потому, что разделение погрешности на систематическую и случайную составляющие мы вводим сами для облегчения анализа. Но после проведения такого анализа правомерна постановка и обратной задачи — задачи суммирования этих составляющих. Задача суммирования систематической и случайной составляющих погрешности тем более важна, что потребителя измерительной информации мало интересует структура погрешности, в первую очередь интересным является точность измерений.

Исходя из рис. наглядно виден механизм такого суммирования. Если доверительная граница с вероятностью *Рд* (равной, например, 0,9) для случайной составляющей определяется как $\gamma_{0,9}=t\sigma$, то с учетом систематической составляющей она будет выражаться как $|\theta|+t\sigma$ Но при $|\theta|>0{,}66\sigma$ выход погрешности за границы $\pm(|\theta|+t\sigma)$ будет происходить даже для распределения Лапласа практически только с одной стороны, т. е., например, при оценке случайной составляющей с *Рд*=0,9 доверительная вероятность выхода результата за границы $\pm(|\theta|+t\sigma)$ будет иметь значение *Рд* = 0,95.

Таким образом, механизм суммирования систематической и случайной составляющих резко отличается от механизма суммирования случайных погрешностей. Во-первых, систематическая погрешность может суммироваться только с доверительным значением погрешности, а отнюдь не со с. к. о., во-вторых, это суммирование происходит арифметически с модулем систематической погрешности (без учета ее знака) и, в-третьих, результирующая погрешность, указываемая как $\gamma_{\Sigma}=(|\theta|+t\sigma)$ при $|\theta|>0{,}66\sigma$, получается с уровнем значимости $q=\dfrac{1-P_{\partial}}{2}$ где *Рд* — доверительная вероятность, с которой была определена случайная составляющая погрешности.

Рассмотрим теперь распределение погрешности усредненного результата многократных отсчетов, оно также имеет вид кривой, показанной на последнем рис. При этом систематическая составляющая погрешности θ остается без изменения, а ширина разброса случайной составляющей погрешности $\dfrac{t\sigma_{xi}}{\sqrt{n}}$ уменьшается в $\sqrt{n}$ раз, где *n*—число усредненных отсчетов. Поэтому если *n* достаточно велико, то $t\sigma_{\bar{x}}<<\theta$ и результирующая погрешность усредненного результата определяется, по существу, только его систематической

погрешностью.

В этой связи ГОСТ 8.207—76 (сейчас работает) устанавливает, что если $\theta < 0{,}8\sigma_{\bar{x}}$, то следует пренебречь систематической составляющей погрешности и учитывать только случайную погрешность усредненного результата в виде $t\sigma_{\bar{x}}$. Если же $\theta > 0{,}8\sigma_{\bar{x}}$, то, наоборот, следует пренебречь случайной составляющей и усредненный результат характеризовать лишь его систематической погрешностью θ.

На основе изложенного сформулируем следующие выводы:

1. Возможность повышения точности путем усреднения ограничена, так как наличие неисключенной систематической погрешности делает практически бессмысленным использование статистического усреднения.

2. При оценке погрешности результата статистического усреднения крайне важен всесторонний анализ и учет неисключенных систематических погрешностей, которые не уменьшаются при статистическом усреднении, о чем часто забывают, увлекшись изящностью методов статистической обработки.

3. Практически реализовать все возможности статистического повышения точности можно лишь тогда, когда одновременно со статистическим усреднением случайных погрешностей производится достаточно полное исключение систематических погрешностей.

5.7.2 Методы оценки центра распределения и их сравнительная эффективность

Очевидно, что при выполнении многократных измерений мы получаем вариационный ряд, представляющий собой численное описание распределения значений результатов измерения. Очевидно и то, что для оценки систематической и случайной составляющих

погрешностей необходимо оценить где лежит центр распределения. Центр распределения можно оценить различными методами, которые мы сейчас и рассмотрим.

Центр размаха определяется с помощью формулы

$$x_p = \frac{x_1 + x_2}{2}$$

где: x_1 *и* x_2 - минимальное и максимальное значения членов вариационного ряда.

Очевидно, что с помощью среднего арифметического значения $\bar{x}$ также можно определить центр распределения. Для простой совокупности n результатов измерений x_i запишем:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Для данных, представленных в виде безынтервального вариационного ряда:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j n_j$$

где n_j - число одинаковых значений величины x_j .

Для интервального вариационного ряда

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{m} \bar{x}_j n_j$$

где $\bar{x}_j$ середина j-того интервала; n_j- число значений, попавших в j- тый интервал, m – число интервалов группирования.

Медиана x_m также может характеризовать центр распределения. Для простой совокупности результатов измерения медиана - это центральный член ранжированного ряда при нечетном числе членов и полусумма двух центральных членов при четном их числе. Если данные сгруппированы в интервальный вариационный ряд, то:

$$x_m = x_j + \frac{\Delta x}{n_{j+1}}\left(\frac{n}{2} - \sum_{k=1}^{k=j} n_k\right)$$

где x_j - нижняя граница интервала, в. котором лежит медиана, т.е. интервала, накопленная частота, в .котором, переходит через n/2; Δx –ширина интервала; n_{j+1} – число значений, попавших в медианный интервал;n_k число значений, попавших в k - ый интервал; j - номер интервала, предшествующего, медианному.

Пример1 Пусть результаты измерений записаны в порядке их получения:

89, 90. 99, 90, 85, 91, 96, 91, 84, 91.

Ранжируем данные:

84, 85, 89, 90, 90, 91, 91, 91, 96, 99.

Находим центр размаха:

$$x_p = \frac{84+99}{2} = 91{,}5$$

Среднее арифметическое: $\bar{x}$ = 1/10 (84+...+99) = 90,6.

Медиана: x_m = (90 + 91)/2 = 90,5.

Пример 2. При исследовании погрешностей манометра получен интервальный вариационный ряд, представленный в табл. В этой же таблице подсчитана накопленная частота и даны значения середин интервалов

Таблица

№ . интервала	Интервалы	Частота nj	Накопленная частота $\sum_{k=1}^{k=j} n_k$	$\bar{x}\bar{j}$
I	8,6- 9,4	2	*2*	9,0

2	9,4-10,2	6	8	9,8
3	10,2-11,0	15	23	10,6
4	11,0-11,8	23	46	11,4
5	11,8-12,6	25	71	12,2
6	12,6-13,4	17	88	13,0
7	13,4-14,2	7	95	13,8
8	14,2-15,9	5	100	14,6

Общее число результатов измерений n=100; ширина интервала Δx=О,8. Находим оценки центра:

$$x_p = \frac{9{,}0 + 14{,}6}{2} = 11{,}8$$

Среднее арифметическое

$$\bar{x} = \frac{1}{100}(9{,}0 \cdot 2 + 9{,}8 \cdot 8 + \ldots + 13{,}8 \cdot 7 + 14{,}6 \cdot 5) = 12{,}13$$

Медиана

$$x_m = 11{,}8 + \frac{0{,}8}{25}(50 - 46) = 11{,}93$$

Из табл. видно, что медианный интервал пятый, в нем накопленная частота переходит через $n/2$= 50.

Недостатком оценок X_p и $\bar{x}$ является то, что они чувствительны к промахам. Особенно чувствительна к промахам оценка Xp. -Не чувствительной к промахам является медиана и другие квантильные оценки.

В связи с тем, что квантиль - это величина, отсекающая заданную долю членов ранжированного ряда, а медиана делит ряд

пополам, т.е. отсекает 50% членов ряда, то медиану можно считать 50%-ной квантилью. Если отбросить по *25%* членов в начале и конце ряда, то границами оставшейся совокупности данных будут 25%-ная и 75%-ная квантили, которые обозначим X_{25} и X_{75} соответственно. Оценка центра, использующая эти величины, определится из формулы:

$$X_c = \frac{X_{25} + X_{75}}{2}$$

эта оценка называется центром серединного размаха. Её недостатком является то, что точность ее получения прямо определяется погрешностями измерений и не увеличивается с увеличением числа измерений. Оценкой, которая сочетает защищенность квантильных оценок от промахов и возможность уточнения с ростом числа измерений среднего арифметического, является среднее арифметическое из 50% измерений в центре ряда, расположенных между X_{25} и X_{75}:

$$\bar{x}_{0,5} = \frac{2}{n} \sum_{i=n/4}^{i=n3/4} x_i$$

Таким образом, существует целый ряд оценок центра распределения. Оказывается, что в зависимости от вида распределения значений результатов измерения эффективной может оказаться та или иная оценка. Проиллюстрируем это на ряде простых примеров.

При симметричных двухмодальных распределениях результатов измерения (рис.А) экспериментальные точки неустойчивы на границах диапазона рассеивания, поэтому оценка *Xp* будет обладать большой дисперсией. Медиана располагается в области малой плотности результатов и поэтому оценивается весьма неточно. Среднее арифметическое также оценивается с большой дисперсией, поскольку числа точек, попавших в две модальные области, могут существенно отличаться особенно при малых, выборках. Реально для таких двухмодальных распределений экспериментальные данные группируются около

мод, которые в свою очередь соответствуют 25%-ной и 75%-ной квантилям распределения.. Поскольку X_{25} и X_{75} располагаются в областях с наибольшей плотностью значении результатов, то они определяются наиболее точно. Следовательно, оценка X_c в данном случае является наиболее эффективной.

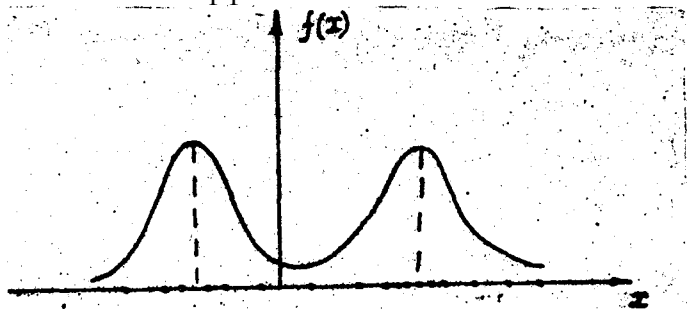
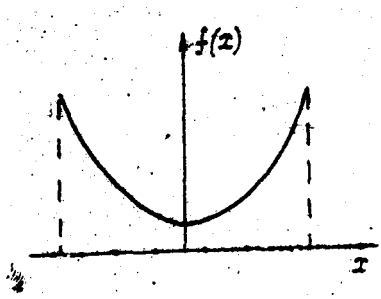
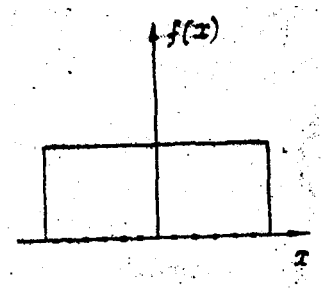


Рис. А



Б



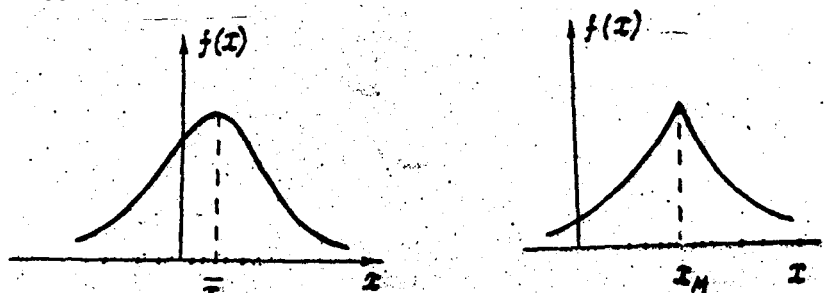
В

При эволюции распределения в сторону таких ограниченных распределений как арксинусоидальное (рис.Б) и равномерное (рис. В) более эффективной является оценка *Хр*. Для арксинусоидального распределения это видно из рис.Б, показывающего наибольшую плотность расположения данных на границах распределения. Здесь границы распределения определяются с наименьшей дисперсией.

Результаты исследований эффективности применения различных оценок показывают, что для нормального и треугольного законов распределения более эффективна оценка центра средним арифметическим, а для равномерного и

арксинусоидального законов значительно более эффективен центр размаха. Интересно отметить, что эффективность оценки X_p растет с ростом числа измерений. Очевидно, что опенки X_m и Xc в данном случае являются менее эффективными.

Для законов распределения, слизких к нормальному (рис.а), эффективной является оценка центра средним арифметическим.



а) **б)**

При дальнейшей эволюции закона распределения в сторону островершинных экспоненциальных законов (рис. 6) эффективной оценкой центра становится медиана. Действительно, значения результатов на "хвостах" распределения являются неустойчивыми, оценки X_p и X_c имеют большую дисперсию. Эта неустойчивость, особенно при малых выборках, существенно влияет и на величину дисперсии. Наиболее плотно результаты измерений группируются вокруг медианы, вследствие чего она оценивается с наименьшей дисперсией.

Таким образом, если имеется какая-то информация о характере распределения значений результатов измерения, то для определения центра при двухмодальных распределениях следует пользоваться оценкой X_c, при ограниченных распределениях - оценкой, Xp , при распределениях близких к нормальному - оценкой $\bar{x}$, при островершинных -оценкой $X_м$.

Если априорных данных о характере распределения нет, то вычисляют все четыре оценки X_c ,X_p, $\bar{x}$, $X_м$.и добавляют к ним оценку $X_{0,5}$, сочетающую защищенность от промахов с возможностью уточнения путем увеличения n, Рекомендуется расположить эти пять оценок в порядке возрастания их значений и выбрать медиану из полученного ряда, т.е. взять третью оценку,

которую и целесообразно использовать в качестве центра группирования во всех последующих расчетах. Достоинство такой оценки состоит в том, что она полностью защищена от наличия промахов.

5.8 Интервальные оценки погрешностей

5.8.1 Доверительные интервалы

В разделе 5.1 мы говорили об определении доверительного значения погрешности непосредственно по экспериментальным данным. Существует и иной способ определения доверительного значения и доверительного интервала. Отметим, что построение доверительного интервала для истинного значения измеряемой величины необходимо для того, чтобы определить на сколько это значение будет изменяться при последующих измерениях.

Доверительным интервалом называется интервал, который с заданной вероятностью, называемой доверительной накрывает истинное значение измеряемой величины. В общем случае доверительные интервалы можно строить на основе неравенства Чебышева. При этом не требуется знать вида распределения наблюдений, но требуется знать с.к.о. - σ. Однако получаемые с помощью неравенства Чебышева интервалы оказываются слишком широкими для практики, и они не получили применения.

Обычно доверительные интервалы строят, основываясь на распределении Стьюдента, о котором мы говорили в разделе 4.5.5. Существует и более простая формула, для этого распределения, которое назовём t.

$$t = \frac{\overline{x} - x_{ист}}{\sigma(\overline{x})}$$

Здесь $x_{ист}$ – истинное значение измеряемой величины;

$\sigma(\bar{x}) = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}}$ - оценка среднего квадратического отклонения от среднего арифметического.

Доверительный интервал $\left[\bar{x} - t_q\sigma(\bar{x}), \quad \bar{x} + t_q\sigma(\bar{x})\right]$ соответствует доверительной вероятности

$$P_д\left\{\left|\bar{x} - x_{ист}\right| \le t_q\sigma(\bar{x})\right\} = \alpha \qquad (17)$$

Здесь t_q – q-процентная квантиль распределения Стьюдента, значения которой находятся по табл.по числу степеней свободы $k=n-1$ и уровню квантили $q=1-\alpha$. Указанные таблицы приведены в справочной и специальной литературе. Например, в книге С.Г.Рабиновича о кот. уже говорилось.

Таким образом, способ, приведённый в разделе 5.6 позволяет проверить гипотезу о нормальности распределения, а выражение (17) оценить доверительный интервал.

5.8.2 Толерантные интервалы.

Толерантным интервалом называется интервал, который с заданной вероятностью P содержит не менее чем заданную часть β всей совокупности случайной величины. Таким образом, толерантный интервал – интервал для случайной величины, и этим он в принципе отличается от доверительного интервала, который строится для того, чтобы накрыть неслучайную величину. Толерантный интервал следует использовать при нормальном законе распределения погрешности.

Границы толерантного интервала

$$l_1 = M(x) - K\sigma \qquad l_2 = M(x) + K\sigma$$

где $M(x)$ – мат. ожидание случайной величины, K – толерантный коэффициент, σ - с.к.о.

Для наиболее употребляемых уровней P и β составлены таблицы, которые приводятся в специальной литературе. Так зависимость этого коэффициента от числа наблюдений n при $\beta=0,95$ и

P=0,9 приведена в таблице

n	16	20	24	30	40
K	2,6	2,5	2,4	2,4	2,3

Как правило, толерантный интервал используется для оценки случайной погрешности нестандартизованных средств измерений.

Следует отметить, что толерантный интервал часто интерпретируют как допусковый интервал, а толерантные границы как границы поля допуска. Однако в такой трактовке есть существенная неточность. Допуск или границы поля допуска устанавливают, как правило, до изготовления изделия и делают это таким образом, что в случае, если интересующий нас параметр изделия выходит за пределы поля допуска, то изделие признаётся негодным. Иными словами, границы поля допуска – жёсткие границы, не связанные ни с какими вероятностными соотношениями. Толерантный же интервал определяют на основе исследований уже изготовленных изделий и вычисляют его границы так, чтобы с заданной вероятностью в этот интервал попадали параметры заданной доли всего возможного числа изделий. Таким образом, границы толерантного интервала, также как и границы доверительного интервала , - случайные величины и этим они отличаются от допусковых границ, или допусков, которые являются неслучайными.

5.9 Обработка результатов прямых измерений с многократными наблюдениями

При статистической обработке группы результатов наблюдений следует выполнить следующие операции:

1. Исключить известные систематические погрешности из результатов наблюдений.
2. Вычислить среднее арифметическое исправленных результатов наблюдений, принимаемое за результат измерения.
3. Вычислить оценку среднего квадратического отклонения

результата измерения.

4. Проверить гипотезу о том, что результаты наблюдений принадлежат нормальному распределению.
5. Вычислить доверительные границы случайной погрешности (случайной составляющей погрешности) результата измерения.
6. Вычислить границы неисключённой систематической погрешности (неисключённых остатков систематической погрешности) результата измерения.
7. Вычислить доверительные границы погрешности результата измерения.

5.9.1 Результат измерения, оценка его среднего квадратического отклонения и доверительных границ случайной погрешности

Вычисления по первым двум пунктам не вызывают затруднений. Вычисление оценки с.к.о. рекомендуется производить по формуле

$$\sigma(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n(n-1)}}$$

здесь x_i – i-тый результат наблюдения; $\bar{x}$ - результат измерения т.е. среднее арифметическое исправленных результатов наблюдений; n – число результатов наблюдений.

Проверка гипотезы о том, что результаты наблюдений принадлежат нормальному распределению рекомендуется производить после исключения выбросов. При числе результатов наблюдений $n>50$ предпочтительной является проверка по критериям χ^2, Пирсона или ω^2 Мизеса-Смирнова при этом уровень значимости q

(см. раздел 5.6) рекомендуется выбирать в интервале от 10% до 2%. При числе результатов измерений 50>n>15 для проверки принадлежности их к нормальному распределению предпочтительным является составной критерий, алгоритм которого приведён в ГОСТ. Суть этого критерия мало отличается от критериев, рассмотренных в разделе 5.6. При числе результатов наблюдений $n \le 15$ принадлежность их к нормальному распределению не проверяют. При этом нахождение доверительных границ случайной погрешности возможно только в случае, если заранее известно, что результаты наблюдений принадлежат нормальному распределению.

Доверительные границы случайной погрешности в соответствии с этим ГОСТ устанавливают для результатов наблюдений, принадлежащих нормальному распределению. Если это условие не выполняется, то ГОСТ рекомендует в методике выполнения измерений приводить отдельный алгоритм расчёта. Для определения доверительных границ случайной погрешности стандарт рекомендует выбирать доверительную вероятность P=0,95. Однако ГОСТ не запрещает также использование P=0,99, например, при обработке результатов измерений, имеющих значение для здоровья людей. Доверительные границы Δ_∂ (без учёта знака) случайной погрешности результата измерений следует вычислять по формуле

$$\Delta_\partial = t\sigma(\bar{x}) \qquad (18)$$

где t – коэффициент Стьюдента, который в зависимости от доверительной вероятности P_∂ (см. раздел 5.1) и числа результатов наблюдений n находят по таблице, приведённой в ГОСТ.

5.9.2 Доверительные границы неисключённой систематической погрешности

В качестве границ составляющих неисключённой систематической погрешности принимают, например, пределы допускаемых основных и дополнительных погрешностей средств измерений, если случайные составляющие погрешности пренебрежимо малы.

При суммировании составляющих неисключённой систематической погрешности результата измерения неисключённые систематические погрешности средств измерений каждого типа и погрешности поправок рассматривают как случайные величины. При отсутствии данных о виде распределения случайных величин их распределения принимают за равномерные.

Границы неисключённой систематической погрешности θ результата измерения вычисляют путём построения композиции неисключённых систематических погрешностей средств измерений, метода и погрешностей, вызванных другими источниками. При равномерном распределении неисключённых систематических погрешностей эти границы (без учёта знака) вычисляются по формуле:

$$\theta = K\sqrt{\sum_{i=1}^{m}\theta_i^2} \qquad (19)$$

Здесь θ_i – граница i – той неисключённой систематической погрешности; K – коэффициент, определяемый принятой доверительной вероятностью.

Коэффициент K принимают равным 1,1 при $P_д$=0,95. При $P_д$=0,99 коэффициент K принимают равным 1,4, если число суммируемых неисключённых систематических погрешностей более 4 (m>4). Если же $m \le 4$, то коэффициент определяют по графику, приведённому в стандарте.

5.9.3 Граница погрешности и форма записи результата измерений

В том случае, если $\frac{\theta}{\sigma(\bar{x})} < 0{,}8$, то неисключёнными систематическими погрешностями, по сравнению со случайными пренебрегают и принимают, что граница погрешности результата

$\Delta=\Delta_\partial$. Если $\frac{\theta}{\sigma(\bar{x})} > 0{,}8$, то случайной погрешностью по сравнению с систематической пренебрегают и принимают, что граница погрешности результата $\Delta=\theta$.

В том случае, если неравенства не выполняются, границу погрешности результата измерения рекомендуется находить путём построения композиции распределения случайных и неисключённых систематических погрешностей, рассматриваемых как случайные величины в соответствии с формулой (19). Если доверительные границы случайных погрешностей найдены по формуле (18), то допускается границы погрешности результата измерения Δ (без учёта знака) вычислить по формуле

$$\Delta=kS_\Sigma$$

где $k=\frac{\Delta_\partial+\theta}{\sigma(\bar{x})+\sqrt{\sum_{i=1}^{m}\frac{\theta_i^2}{3}}}$ – коэффициент, зависящий от соотношения случайной и неисключённой систематической погрешностей;

$S_\Sigma=\sqrt{\sum_{i=1}^{m}\frac{\theta_1^2}{3}+\sigma^2(\bar{x})}$ - оценка суммарного среднего квадратического отклонения результата измерения; θ_1 – составляющая систематической погрешности по числовому значению наиболее отличающаяся от других.

Оформление результатов измерений производится следующим образом. При симметричной доверительной погрешности $\bar{x}\pm\Delta$. Причём числовое значение результата измерения должно оканчиваться цифрой того же разряда, что и значение погрешности Δ. При отсутствии данных о виде функции распределений составляющих погрешности результата и необходимости дальнейшей обработки результатов или анализа погрешностей, результаты измерений представляют в форме $\bar{x}$; $\sigma(\bar{x})$; n; θ. В случае, если границы

неисключённой систематической погрешности вычислены в соответствии с формулой (19) , следует дополнительно указывать доверительную вероятность Р. Следует отметить то, что оценки $\sigma(\bar{x})$ и θ могут быть выражены в абсолютной и относительной формах.

6. Косвенные измерения

6.1 Предварительные замечания и классификация

В соответствии с РМГ 29-99 Косвенные измерения — это определение искомого значения физической величины на основании результатов прямых измерений других физических величин, функционально связанных с искомой величиной. Таким образом, при косвенных измерениях мы имеем функцию и аргумент.

Значения аргументов чаще всего находят в результате прямых измерений, но иногда — в результате совместных, совокупных или косвенных, в свою очередь, измерений.

Искомое значение физической величины A следует находить на основании результатов измерений аргументов $a_1 \ldots . a_m$, связанных с искомой величиной уравнением

$$A=f(a_1,\ldots . a_m) \qquad (20)$$

Вид функции f должен быть известен из теоретических предпосылок или установлен экспериментально с погрешностью, которой можно пренебречь.

Случаи неявной зависимости между A и α нетипичны. По виду функциональной, зависимости (20) будем различать косвенные измерения с линейной зависимостью между измеряемой величиной и измеряемыми аргументами, косвенные измерения с нелинейной зависимостью между этими величинами и косвенные измерения с зависимостью между величинами смешанного типа. При линейной зависимости .уравнение (20) имеет вид

$$A = \sum_{i=1}^{m} b_i \tilde{a}_i \qquad (21)$$

где b_i — постоянный коэффициент аргумента ; $\tilde{a}_i$ - результат измерения a_i –того аргумента, m— число слагаемых.

Косвенные измерения при линейной зависимости между величинами будем называть линейными косвенными измерениями, а при нелинейной зависимости—нелинейными косвенными измерениями.

В общем случае уравнение (20) при нелинейных косвенных измерениях можно представить как произведение некоторых функций

$$A = \prod_{i=1}^{m} f_i(a_i)$$

В частном случае $A=f(a_i)$. В случае зависимости между величинами смешанного типа уравнение (20) принимает вид

$$A = \prod_{i=1}^{m} f_i(a_i) + ... + \prod_{l=1}^{r} f_l(a_l)$$

где l – знак, соответствующий величинам линейного типа

Если известны методы обработки результатов наблюдений для нелинейных и линейных косвенных измерений, то аналогичная задача для случая с зависимостью смешанного типа элементарно сводится к двум предыдущим. Поэтому специально этот вид косвенных измерений можно не .рассматривать.

Косвенные измерения, так же как и прямые, делят на статические и динамические. Статические косвенные измерения могут быть весьма различными в зависимости

от свойств измеряемых аргументов. Если измеряемые аргументы можно считать неизменными во времени, то неизменна и косвенно измеряемая величина, т. е. имеем обычную статическую ситуацию.

Однако измеряемая величина может быть неизменной и тогда, когда аргументы изменяются. Например, измеряем сопротивление резистора методом амперметра и вольтметра, и по условиям измерения напряжение источника изменяется во времени. Хотя измеряемые аргументы изменяются, измеряемая величина остается неизменной. Такие косвенные измерения назовем квазистатическими.

Для получения правильного результата в рассматриваемом случае аргументы необходимо измерять такими приборами, чтобы за время установления показаний приборов изменения аргументов можно было считать незначительными.

Косвенные измерения в принципе возможны и тогда, когда изменяются во времени и измеряемые аргументы, и сама косвенно измеряемая величина.

Косвенные измерения, при которых средства измерений или часть их находятся в динамическом режиме, в соответствии с общим определением динамических, измерений надо считать динамическими.

Специфическим приемом выполнения косвенных измерений является одновременное измерение аргументов. Последнее позволяет подставить одновременно полученные значения аргументов в соотношение, связывающее с ними измеряемую величину, и получить таким образом, мгновенное значение измеряемой величины, отвечающее моменту времени измерения аргументов. Совокупность таких значений ничем не отличается от совокупности мгновенных значений величины, полученной при прямых измерениях.

Получаемые рассмотренным путем совокупности

мгновенных значений естественно обрабатывать так же, как и совокупности данных, получаемые при прямых измерениях.

Приведение косвенных измерений к прямым целесообразно не только при динамических, но — в случае сложной нелинейной зависимости измеряемой величины от измеряемых аргументов— и при статических косвенных измерениях. Необходимым условием осуществления этого приема является согласованное, например одновременное, измерение аргументов.

Способ выполнения косвенных измерений, при котором получают группу значений измеряемой величины и, обрабатывая ее как группу наблюдений при прямых измерениях, находят результат косвенного измерения, будем называть методом приведения, а погрешность, возникающую при этом методической.

6.2 Определение результатов измерения и оценивание погрешностей при косвенных измерениях

6.2.1 Общие положения

Определим результаты измерений и оценку их погрешностей при условии, что аргументы, от которых зависит измеряемая величина, являются постоянными физическими величинами; известные систематические погрешности результатов измерений аргументов исключены, а неисключённые систематические погрешности распределены равномерно внутри заданных границ $\pm\theta$.

При оценивании доверительных границ погрешностей результата косвенного измерения будем принимать вероятность, равную 0,95 или 0,99.

6.2.2 Косвенные измерения при линейной зависимости

Искомое значение A связано с m измеряемыми аргументами уравнением

$$A=b_1a_1+b_2a_2+\ldots+b_ma_m$$

Здесь b постоянные коэффициенты при аргументах a.

Корреляция между погрешностями измерений аргументов отсутствует. Результат косвенного измерения вычисляют по формуле (21).

Оценку с.к.о. результата косвенного измерения определяют по формуле

$$\sigma(A)=\sqrt{\sum_{i=1}^{m}b_i^2\sigma^2(a_i)}$$

здесь $\sigma(a_i)$ – оценка с.к.о. результата измерения a_i аргумента

Доверительные границы случайной погрешности результата косвенного измерения, при условии, что результаты измерений аргументов распределены по нормальному закону без учёта знака вычисляется по формуле

$$\Delta_\partial(p)=t\sigma(A) \qquad (22)$$

где t – коэффициент Стьюдента, соответствующий доверительной вероятности $P=1-q$, (q – процент квантили распределения) и числу степеней свободы, вычисляемому по формуле

$$N=\frac{\left(\sum_{i=1}^{m}b_i^2\sigma^2(a_i)\right)^2-2\left(\sum_{i=1}^{m}\frac{b_i^4\sigma^4(a_i)}{n_i+1}\right)}{\sum_{i=1}^{m}\frac{b_i^4\sigma^4(a_i)}{(n_i+1)}}$$

n_i – число измерений при определении a_i – того аргумента.

Границы неисключённой систематической погрешности вычисляются следующим образом.

Если неисключённые систематические погрешности

результатов измерений аргументов заданы границами θ, то доверительные границы неисключённой систематической погрешности, без учёта знака, при вероятности *P* вычисляются из формулы:

$$\theta(p) = k\sqrt{\sum_{i=1}^{m} b_i^2 \theta_i^2}$$

k-поправочный коэффициент. В том случае, если число суммируемых составляющих m>4 при *p*=0,95 коэффициент *k*=1,1 , при *p*=0,99, коэффициент *k*=1,45. В других случаях этот коэффициент находится из кривых, приведённых в РД.

В том случае, если границы неисключённых систематических погрешностей результатов измерений аргументов заданы доверительными границами, соответствующими вероятностям Р, то границы неисключённой систематической погрешности результата косвенного измерения вычисляют по формуле

$$\theta(p) = k\sqrt{\sum_{i=1}^{m} b_i^2 \frac{\theta_i^2}{k_i^2}}$$

k_i^2 соответствует значению *k* для i-той систематической погрешности. Для *P*=0,95 *k*=1,1.

Погрешность результата косвенного измерения оценивают на основе композиции распределений случайных и неисключённых систематических погрешностей. Оценку производят аналогично той, которая произведена в п. 5.9.3.

6.2.3 Косвенные измерения при нелинейной зависимости

Для косвенных измерений при нелинейных зависимостях и некоррелированных погрешностях измерений аргументов используют метод линеаризации. Этот метод предполагает разложение нелинейной функции в ряд Тейлора:

$$A = f(a_1,...a_m) = f(\tilde{a}_1,...,\tilde{a}_m) - \sum_{i=1}^{m} \frac{\partial f}{\partial a_i} \Delta a_i + R$$

где $f(a_1,...a_m)$ – нелинейная функциональная зависимость измеряемой величины A от измеряемых аргументов a_i ; $\partial f/\partial a_i$ – первая производная от функции f по a_i –тому аргументу, вычисленная в точке $\tilde{a}_1,...,\tilde{a}_m$; Δa_i – отклонение отдельного результата измерения a_i – того аргумента от его среднего арифметического; R – остаточный член.

Остаточным членом $R = \frac{1}{2} \sum_{i=1}^{m} \frac{\partial^2 f}{\partial a_i^2} (\Delta a_i^2)$ пренебрегают, если

$$R < 0{,}8 \sqrt{\sum_{i=1}^{m} \left(\frac{\partial f}{\partial a} \right)^2 \sigma^2(a_i)}$$

$\sigma(a_i)$ – оценка с.к.о. случайных погрешностей результата измерения a_i – того аргумента.

Результат измерения вычисляют по формуле

$$A = f(\tilde{a}_1,...,\tilde{a}_m)$$

Оценку с.к.о. случайной погрешности результата косвенного измерения вычисляют по формуле

$$\sigma(A) = \sqrt{\sum_{i=1}^{m} \left(\frac{\partial f}{\partial a_i} \right)^2 \sigma^2(\tilde{a}_i)}$$

Доверительные границы случайной погрешности результата косвенного измерения при условии, что распределение погрешностей не противоречит нормальному распределению вычисляют в соответствии с формулой (22), подставляя вместо коэффициентов b_i первые производные $\partial f/\partial a_i$ соответственно.

Границы неисключённой систематической погрешности результата косвенного измерения вычисляют в соответствии с

рекомендациями, сделанными в разделе 6.2.2, подставляя вместо коэффициентов b_i производные $\partial f/\partial a_i$. Погрешность результата косвенного измерения оценивают на основе композиции распределений случайных и неисключённых систематических погрешностей. Оценку производят аналогично той, которая произведена в п. 5.9.3.

6.2.4 Метод приведения

При неизвестных распределениях погрешностей измерений аргументов и при наличии корреляции между ними для определения результата косвенного измерения и его погрешности используют метод приведения, который предполагает наличие ряда отдельных значений измеряемых аргументов, полученных в результате многократных измерений.

Метод основан на приведении ряда отдельных значений косвенно измеряемой величины к ряду прямых измерений. Получаемые сочетания отдельных результатов измерения аргументов подставляют в выражение (20) и вычисляют отдельные значения измеряемой величины A_j. Результат косвенного измерения вычисляют по формуле

$$A = \sum_{j=1}^{n} \frac{A_j}{n}$$

Здесь n – число отдельных значений измеряемой величины. A_j –j-тое отдельное значение измеряемой величины, полученное в результате подстановки j-того сочетания согласованных результатов измерений.

Оценку среднего квадратического отклонения случайных погрешностей результата косвенного измерения вычисляют по формуле

$$\sigma(A) = \sqrt{\sum_{j=1}^{n} \frac{(A_j - A)^2}{n(n-1)}}$$

Доверительные границы случайной погрешности для результата косвенного измерения вычисляют (без учёта знака) в соответствии с ГОСТ 8.207-76. Границы неисключённой систематической погрешности результата косвенных измерений при линейной зависимости вычисляют в соответствии с рекомендациями раздела 6.2.2., при нелинейной зависимости в соответствии с рекомендациями раздела 6.2.3.

7. Динамические погрешности

7.1 Методы оценки динамических погрешностей

Оценивание динамических погрешностей обладает рядом особенностей, и на основных из них необходимо остановиться.

Прежде всего, нужно заметить, что хотя динамические погрешности в конкретных ситуациях учитываются с давних пор, общая теория оценивания динамических погрешностей измерений, как и вообще теория динамических измерений, в настоящее время находится еще в стадии формирования.

В соответствии с РМГ динамическая погрешность средства измерений определяется как возникающая при измерении изменяющейся в процессе измерения физической величины. Часто динамическую погрешность интерпретируют как разность погрешности прибора в динамическом режиме и статической погрешности. Последнюю можно трактовать как следствие отклонения действительного коэффициента преобразования средства измерений от его номинального значения. Динамическими погрешностями обладают как первичные чувствительные преобразователи, так и промежуточные преобразователи измерительной информации. При проектировании измерительного средства необходимо учитывать особенности этих преобразователей в отдельности, а при исследовании динамических погрешностей прибора рассматривать измерительный тракт в целом.

Типичным случаем измерения, для которого существенна

динамическая погрешность, является измерение с регистрацией сигнала, изменяющегося во времени. В этом случае, в соответствии с общим определением абсолютной погрешности, для динамической погрешности можно написать

$$\Delta_{дин}(t) = \frac{y(t)}{K} - x(t) \qquad (23)$$

где $x(t)$, $y(t)$ — сигналы на входе и соответственно на выходе средства измерений, K — коэффициент преобразования.

Связь между сигналами на входе и выходе средства измерений можно представить операторным уравнением

$$y=Bx \qquad (24)$$

где B — оператор средства измерений.

Оператор в общем виде выражает всю совокупность динамических свойств средств измерений. Последние зависят от того, по отношению к какому воздействию они рассматриваются. Так, динамические свойства по отношению к изменяющейся влияющей величине или к помехе, действующей не на входе средства измерений, могут быть другими, чём по отношению к входному сигналу. В уравнении (24) оператор B рассматривается по отношению к входному сигналу.

При проектировании средств измерений обычно добиваются независимости коэффициента преобразования от уровня входных воздействий. Тогда средства измерений можно описать линейной моделью, причем, как правило, удается рассматривать линейные модели с сосредоточенными параметрами.

Таким образом, задача определения динамических погрешностей сводится к восстановлению формы входного сигнала, при условии знания выходного.

Следует отметить, что выходной сигнал средства измерений в конечном счетё всегда имеет убывающий по интенсивности с ростом частоты спектральный состав. Амплитудно-частотная характеристика средства измерений (которое, естественно, является устойчивой системой) на высоких частотах также приближается к оси частот.

Таким образом, по двум функциям с убывающими спектрами требуется найти третью (входной сигнал), которая их однозначно связывает. В области низких и средних (для данных функций) частот, где интенсивности спектров высоки, удается достаточно достоверно определить искомый сигнал, причем неизбежные погрешности исходных данных и процедуры вычислений действуют «регулярным» образом, т. е. искажают решение, не лишая его физического смысла. В области высоких частот интенсивности спектров падают настолько, что их влияние на решение оказывается соизмеримым с влиянием погрешностей исходных данных. Влияние этих погрешностей может быть так велико, что истинное решение оказывается совершенно подавленным. Методы решения таких некорректно поставленных задач (методы регуляризации) активно разрабатываются в математике (А. Н. Тихоновым и его последователями), математической физике, геофизике, теории автоматического управления. Существо методов регуляризации состоит в фильтрации искажений на основе априорной информации об истинном решении. При этом основным является вопрос об установлении оптимальной степени фильтрации с тем, чтобы отфильтровать помехи, не исказив истинного решения. Различные методы регуляризации требуют различного объема и формы априорной информации.

По приведенным соображениям восстановление формы (с сохранением параметров) входного сигнала для оценивания динамических погрешностей применяется редко и на практике данная задача решается иначе.

В тех случаях, когда регистрация однородных по форме сигналов выполняется неоднократно, создают специальный тип средств измерений (или выбирают такой из имеющихся) и затем оценивают и нормируют предел возникающей динамической погрешности. Оценку погрешности можно найти экспериментальным путем; если имеется менее инерционное средство измерений, или расчетным путем. Применение специализированных – нестандартизованных средств измерений существенно упрощает задачу.

При применении универсальных средств измерений для

оценивания погрешности составляют суждение о форме входного сигнала и ее возможных изменениях. После этого, зафиксировав параметры входного сигнала, т. е. выбрав конкретную его модель, пользуясь выражением (24), находят соответствующий сигнал на выходе средства измерений. Далее на основе уравнения (23) получают выражение (или график) для динамической погрешности, которая характеризует погрешности регистрации выбранного входного сигнала, Для нескольких входных сигналов (двух-трех) находят таким образом, несколько функций динамических погрешностей.

Но оперировать с погрешностью как функцией времени неудобно. Поэтому обычно динамическую погрешность регистрации стремятся охарактеризовать параметром, который для всей функции принимает какое-то одно значение. Чаще всего для этого берут максимальную по модулю погрешность или ее среднее квадратическое отклонение.

Из полученных таким образом (двух-трех) значений динамической погрешности затем обычно в качестве общей характеристики берут худшую, т. е. наибольшую.

Нужно заметить, что приведенная схема вычислений применительно к различным задачам измерений может видоизменяться. Так, часто не имеет значения сдвиг во времени выходного сигнала относительно входного. В этом случае, возможно искусственно располагать сигналы таким образом, чтобы норма погрешности стала минимальной.

Однако рассматриваемая нами задача в целом состоит в оценивании динамической погрешности измерения. В частном случае это может быть измерение, которое как промежуточный этап содержит в себе регистрацию. Следовательно, оценка измеряемой величины составляется по записи некоторого сигнала путем ее соответствующей обработки. Иными словами, известен функционал, преобразующий сигнал в оценку измеряемой величины. Для решения задачи нужно иметь оценки возможных форм входного сигнала. Для этих сигналов, зная оператор средства измерений, находят соответствующие выходные сигналы. Затем, обработав эти сигналы, определяют значения измеряемой величины по входному сигналу $A_{вх}$

и по выходному $A_{вых}$. Их разность дает динамическую погрешность измерения

$$\Delta_{дин}=A_{вых} - A_{вх}$$

Последнюю следует представить в форме относительной погрешности

$$\Delta=\Delta_{дин}/A_{вх}$$

Следует отметить, что динамическая погрешность возникает также при измерении параметров периодических и нестационарных процессов с помощью показывающих приборов. При некоторой типичной форме входного сигнала приборы градуируют. Динамическая погрешность возникает, если форма, входного сигнала при измерении оказывается отличной от той, при которой прибор градуировали. Очевидно, что, основываясь только на показаниях прибора, оценить эту .погрешность невозможно. Для решения задачи необходимо иметь оценку формы входного сигнала, имевшего место при измерении. Тогда, зная форму входного сигнала при градуировке прибора и оператор прибора, по приведенной выше схеме можно оценить динамическую погрешность измерения. Расчеты целесообразно выполнять при таких параметрах входного сигнала, которые соответствуют фактическому показанию прибора.

Рассмотренный путь решения задачи требует много информации и расчетов. Поэтому практика выработала другой подход к ее решению.

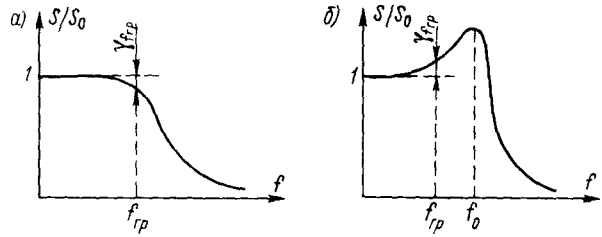
Входной сигнал можно представить некоторой моделью, характеризуемой рядом параметров. Один из них — информативный, измеряемый, остальные неинформативные. Средства измерений проектируют так, чтобы сделать их нечувствительными ко всем неинформативным параметрам входного сигнала. Полностью, однако, этого достичь не удается, и в общем случае влияние неинформативных параметров только ослабляется. Далее для всех неинформативных параметров можно определить такие границы, что при изменении неинформативных параметров внутри этих границ суммарная погрешность средств измерений будет изменяться незначительно. Это позволяет установить нормальные области

значений неинформативных параметров. Если какой-либо неинформативный параметр выходит за границы области его нормальных значений, то возникающую погрешность рассматривают как дополнительную. Влияние каждого неинформативного параметра нормируют по отдельности как и воздействия влияющих величин. Таким образом, вместо оценивания динамической погрешности в целом приходят к оцениванию ряда погрешностей, что упрощает задачу.

Нормирование влияния неинформативных параметров и оценивание возникающих из-за них погрешностей выполняется на основе тех же положений, которые развиты для учета дополнительных погрешностей, вызванных воздействием внешних влияющих величин.

7.2 Простейшая оценка динамических погрешностей при использовании аналоговых средств регистрации

Как правило, регистрацию быстропротекающих процессов производят с помощью компьютера или иного устройства, включающего микроконтроллер, однако во многих случаях в этих целях используют универсальные аналоговые регистраторы. Эти приборы способны зарегистрировать процессы, протекающие с частотами от долей Гц до 50 кГц, в зависимости от их типа. Широко используются аналоговые средства регистрации: самопишущие приборы с чернильной записью, светолучевые и электронные осциллографы с фото приставками.



а *б*

Динамические погрешности таких приборов, а также используемых в комплекте с ними датчиков и усилителей принято нормировать указанием их амплитудно-частотной характеристики, представляющей собой график зависимости от частоты f отношения их чувствительности S при частоте f к чувствительности S при f=O.

В большинстве случаев эти характеристики имеют вид кривых, показанных на рис. При апериодической частотной характеристике (рис. *а)* (показывающие приборы, усилители, датчики температуры и т. п.) чувствительность S прибора или датчика монотонно понижается с ростом частоты f воспринимаемого процесса. Частотная погрешность γ_f есть разность между ординатами частотной характеристики и постоянным уровнем $S/S_0=1$, показанным на рис. *а* штриховой прямой. Она всегда отрицательна и увеличивается с ростом частоты. Ее численное значение может быть найдено из этого графика для любой частоты и использовано для оценки точности регистрации или введения поправки в результат измерения.

Частотная характеристика, изображенная на рис. *б,* характерна для колебательных систем с малым успокоением (гальванометров, светолучевых осциллографов, датчиков манометров, акселерометров и т. д.). Она имеет резонансный пик вблизи собственной частоты колебательной системы и положительную частотную погрешность γ_f. Для приборов и датчиков с такими частотными характеристиками нормируется рабочий диапазон частот, простирающийся от f=0 до такой частоты $f_{гр}$, где γ_f достигает некоторого граничного значения $\gamma_{fгр}$ Так как граничное значение частотной погрешности достигается только в конце рабочего диапазона частот, то внутри его частотные погрешности оказываются много меньше этого значения.

Располагая частотной характеристикой прибора или датчика, можно найти частотную погрешность для любого значения частоты регистрируемого процесса внутри рабочего диапазона частот. Так, например, при частотной характеристике, приведенной на рис. б, частотная погрешность может быть рассчитана по формуле

$$\gamma_f \approx (1-2\xi^2)(f / f_0)^2$$

где ξ - степень успокоения колебательной системы; f_0 - ее собственная частота.

При отсутствии успокоения ($\xi \approx 0$), что характерно для датчиков, не имеющих специальных средств успокоения, частотная погрешность

$$\gamma_f \approx (f / f_0)^2 \qquad (25)$$

Так же легко может быть вычислена частотная погрешность и для апериодических преобразователей невысоких порядков. Так, например, термопара или термометр сопротивления могут быть представлены апериодическим звеном первого порядка (с одной постоянной времени). Для них частотная погрешность может быть приближенно выражена как

$$\gamma_f \approx -\frac{1}{2}(f / f_c)^2 \qquad (26)$$

где $f_c = \frac{1}{2\pi\tau}$ частота среза частотной характеристики, а τ — постоянная времени.

Практическое использование формул (25) и (26) рассмотрим на двух конкретных примерах.

1. Пусть для регистрации пульсирующего давления используется мембранный датчик (тензометрический, пьезоэлектрический, емкостный или индуктивный) с собственной частотой f_0=5 кГц. Какие процессы и с какой погрешностью могут быть им измерены? Полагая, что степень успокоения датчика $\xi \approx 0$, и используя соотношение (25), получаем, что при частоте измеряемого процесса f=50 Гц его частотная погрешность γ_f= 0,01%, но при f=100 Гц уже γ_f =0,04%, при f=500 Гц γ_f=1%, а при f=1000 Гц γ_f =4%, т. е. рабочий диапазон частот датчика оказывается уже исчерпанным.

2. Пусть периодические колебания температуры измеряются с помощью термопары или термометра сопротивления средней инерционности с постоянной времени τ=1 мин =60 с. Спрашивается,

каков рабочий диапазон частот такого датчика? Для этого преобразуем формулу (26), заменив *f* на 1/T, где *T* — период измеряемого процесса; тогда получим

$$\gamma_f \approx -\frac{1}{2}\frac{f^2}{f_c^2} = -\frac{1}{2}\frac{(2\pi\tau)^2}{T^2} = -\frac{2\pi^2\tau^2}{T^2}$$

Подставляя в это выражение разные значения периода *T* измеряемых колебаний, получим частотную погрешность γ_f =0,14% при периоде колебаний T=2 ч, γ_f =0,5% —при *T = 60* мин, γ_f=2% — при T=30 мин, γ_f =5% — при T=20 мин, т. е. рабочий диапазон частот можно считать исчерпанным.

Соотношения (25) и (26) показывают, что частотная погрешность возрастает пропорционально квадрату частоты, что, приводит к очень неблагоприятным соотношениям при регистрации несинусоидальных процессов.

8. Организация и планирование измерительных процедур

При нормировании метрологических характеристик приборов, разработке технических условий на них, программ и методик метрологической аттестации и поверки приборов, а также методик выполнения измерений необходимо учитывать некоторые специфические особенности, которые влияют на организационно-технические аспекты, отражаемые в этих документах. Рассмотрим эти особенности подробнее.

8.1 Изменение погрешности средств измерения во время их эксплуатации.

Как бы тщательно ни был изготовлен и отрегулирован прибор к моменту выпуска его на приборостроительном заводе, с течением времени в элементах схемы и механизме неизбежно протекают разнообразные процессы старения и погрешность его возрастает. Поэтому нормирование гарантированных в паспорте СИ пределов допускаемой погрешности производится заводом-изготовителем с

1,25—2,5-кратным запасом на старение. Такое превышение пределов допускаемой погрешности над фактическим значением погрешности СИ в момент их выпуска с производства или из ремонта является по существу единственным практическим способом обеспечения долговременной метрологической стабильности средств измерений.

Это обстоятельство должно быть четко известно потребителю средств измерений, так как его приходится принимать во внимание при решении многих вопросов организации процессов измерений, поддержания СИ в работоспособном состоянии, оценки допускаемых при измерении погрешностей и т. д.

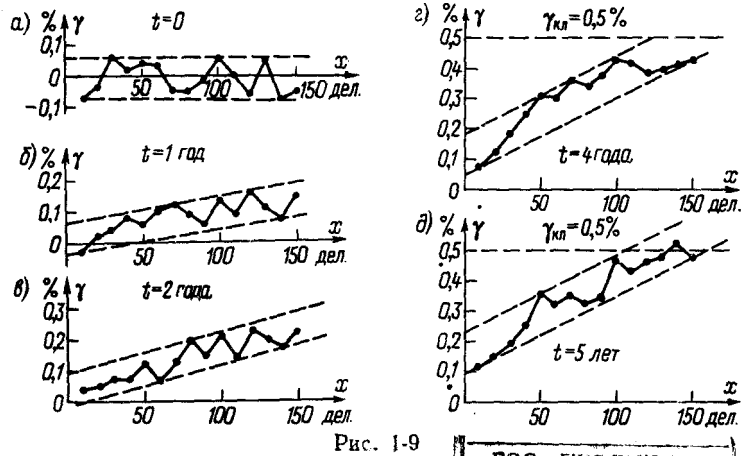


Рис. 1-9

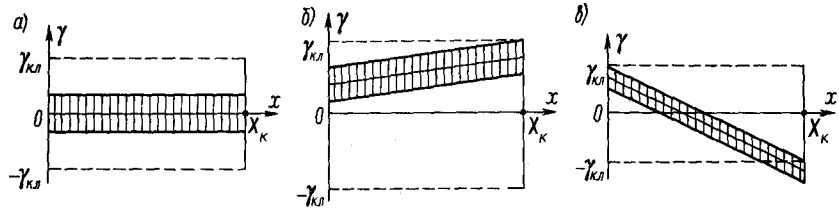
Характер возрастания погрешности СИ во времени показан на рис. где приведены результаты поверок на всех цифровых отметках шкалы прибора типа М105 класса точности 0,5 за первые пять лет его эксплуатации. У нового, только что изготовленного прибора (рис. *а* при t=0) полоса его погрешностей располагается симметрично относительно нуля в границах ±0,09%. Систематическая погрешность отсутствует, так как она устранена благодаря только что проведенной

на заводе регулировке или градуировке шкалы прибора, а случайная погрешность составляет одну пятую часть от нормированного предела $\gamma_{кл}$ = 0,5%.

Изменение погрешности с возрастом t прибора, наблюдаемое при последующих ежегодных поверках, происходит в виде прогрессирующего смещения и поворота полосы погрешностей, т. е. в виде непрерывного возрастания систематической составляющей погрешности прибора, в то время как размер случайной составляющей определяется шириной полосы погрешностей и остается практически неизменным.

Из кривых рис. видно, как постепенно с возрастом t прибора расходуется обеспеченный при изготовлении запас погрешности на старение. Так, если при t=0 он составлял 0,4% (из нормированного значения $\gamma_{кл}$ = 0,5%, то в возрасте прибора t=2 года максимальная погрешность прибора на 120-м делении шкалы достигла 0,23% и запас стал лишь двукратным. При t = 4 года запас на 100-м делении составлял лишь 0,07%, т. е. всего 1/7 от нормированного $\gamma_{кл}$ = 0,5%, а при t=5 лет запас был уже полностью израсходован и погрешность прибора на 140-м делении превысила допускаемую.

Аналогичный характер имеет и процесс накопления прогрессирующей погрешности с возрастом цифровых приборов и измерительных каналов измерительных информационных систем (ИИС) или измерительно-вычислительных комплексов (ИВК). Как правило, ИИС и ИВК выполняются достаточно высококачественно, т. е. при изготовлении полосе погрешностей канала стремятся придать вид, показанный на рис., *а.* Однако накопление прогрессирующей погрешности приводит, как и у других СИ, к смещению и повороту их полосы погрешностей, т. е. к постепенному расходованию запаса погрешности, созданного при изготовлении. Так, поверка одной из ИИС типа К200 показала, что через несколько лет после выпуска полоса погрешностей имела вид, представленный на рис. *б,* т. е. система находилась на пороге выхода из нормированного допуска. А поверка одной из больших ИИС через 5 лет после ее выпуска дала для каналов картину, представленную на рис. *в.*



Таким образом, характер проявления прогрессирующей погрешности с возрастом СИ является единым для всех СИ и пользователь средств измерений не может его игнорировать.

8.2 Метрологическая аттестация нестандартизованных средств измерения

8.2.1 Условия проведения эксперимента и его организация

Очевидно, что кроме значительного числа приборов, занесённых в государственный реестр средств измерений и определяемых как стандартизованные, промышленность выпускает и много нестандартизованных средств измерений (НСИ). Как правило, это приборы и измерительные комплексы всех подвижных объектов, а также приборы, и комплексы, созданные для решения каких либо частных задач. Следует отметить, что как стандартизованные, так и НСИ проходят метрологическую аттестацию и периодическую поверку, их метрологические характеристики отражаются в технических условиях на изделия. Следует отметить, что подход к аттестации стандартизованных приборов, которые выпускаются крупносерийно и нестандартизованных, выпускаемых единично, принципиально различный. Эта разница в подходе обусловлена тем, что законы распределения случайных погрешностей и значения систематических погрешностей стандартизованных приборов известны заранее т.к. они исследованы на значительных партиях этих средств измерений, чего нельзя сказать о нестандартизованных изделиях измерительной техники. Для метрологической аттестации средств измерений разрабатываются специальные программы и методики, с помощью которых определяются и рассчитываются

погрешности средств измерений, для оценки методических погрешностей разрабатывают методики выполнения измерений. Методы расчёта, содержащиеся в этих документах регламентируются ГОСТами лишь отчасти. В целом документ разрабатывается предприятием – изготовителем приборов и утверждаются Главным метрологом этого предприятия. Оценка случайных погрешностей стандартизованных приборов производится на основе тех методов, которые мы изучали (для этого, как правило, используются известные законы распределения погрешностей и коэффициент Стьюдента). Подход к аттестации нестандартизованных приборов иной, рассмотрим его подробнее.

При метрологической аттестации НСИ в качестве образцового средства измерений допускается использовать средство, предел погрешности которого составляет не более 1/3 предела допускаемой погрешности исследуемого НСИ, установленного в техническом задании на разработку. В случае выполнения этого требования влиянием погрешности образцового средства измерений на получаемую оценку погрешности и другие метрологические характеристики аттестуемого НСИ можно пренебречь. Условия проведения экспериментальных исследований при определении значений метрологических характеристик НСИ должны по возможности наиболее близко приближаться к реальным условиям применения НСИ.

При проведении эксперимента всегда решается задача выбора числа и распределения исследуемых точек по диапазону измерений. Рекомендуется исследуемыми точками выбирать точки, соответствующие отметкам шкалы прибора, а для преобразователей – точки, равноотстоящие в диапазоне преобразований. Число точек должно быть не менее пяти и они должны соответствовать ориентировочно 10, 25, 50, 75 и 100% диапазона измерений.

Очевидно, что число наблюдений в конкретной точке диапазона измерений определяет достоверность получаемых оценок метрологических характеристик и зависит от вклада случайной составляющей погрешности исследуемой метрологической

характеристики. При метрологической аттестации не всегда удаётся уверенно оценить соотношение случайной и систематической составляющих оцениваемой погрешности, поэтому выбор числа наблюдений в одной точке приходится производить из предположения о преимущественном влиянии случайной составляющей погрешности и некоторой допустимой погрешности в оценке исследуемой метрологической характеристики. В этом случае искомое число наблюдений n следует определять по формуле

$$n \geq \frac{\beta^2}{2q^2}$$

здесь q – принимаемая допустимая относительная погрешность (в долях) определения среднего квадратического отклонения случайной составляющей погрешности (например 0,1; 0,2; 0,25; и т.д.), β - коэффициент выбираемый из таблицы

Значение доверительной вероятности P	0,80	0,90	
Значение β	1,29	1,64	

Для большинства случаев метрологической аттестации НСИ рекомендуется принимать $0,1 \leq q \leq 0,3$ и доверительную вероятность P равной 0,9 или 0,95. Для случая, когда q=0,3 и P=0,9 получаем, что число наблюдений в одной исследуемой точке диапазона измерений должно быть не менее 15. Число наблюдений может быть уменьшено, если до проведения аттестации известны некоторые данные о составляющих погрешности НСИ, например: преимущественное влияние систематической погрешности, или дисперсия случайной составляющей постоянна в диапазоне измерений.

8.2.2 Определение значений метрологических

характеристик

Одна из задач проведения метрологической аттестации это определение рабочего диапазона средства измерения. В соответствии с РМГ диапазон измерений это область значений величины, в пределах которой нормированы допускаемые пределы погрешности средства измерений. Очевидно, что при аттестации прибора его диапазон стремятся сделать как можно шире, однако нижняя граница диапазона будет ограничена погрешностью, поэтому границу диапазона измерений необходимо определять значением, превышающим удвоенное значение предела нормируемой погрешности, выраженного в единицах измеряемой величины. Исключение составляет случай, когда нормируется постоянным значением относительная погрешность НСИ и указанная выше граница не имеет ограничений, связанных с пределом погрешности. Таким образом, границы диапазона измерений при метрологической аттестации определяются на основе требований технического задания на разработку НСИ и с учётом фактических значений погрешностей, полученных в процессе метрологической аттестации.

Оценка вариации в точке x_k диапазона измерения должна производиться по формуле

$$b_k = \left|\overline{\Delta}_M - \overline{\Delta}_Б\right| \qquad (27)$$

здесь $\vec{\Delta}_M = \dfrac{\sum_{i=1}^{n} \Delta_{мi}}{n}$ среднее значение погрешности в точке x_k диапазона измерения при медленных изменениях входного сигнала со стороны меньших значений до значения x_k; $\vec{\Delta}_Б = \dfrac{\sum_{i=1}^{n} \Delta_{бi}}{n}$ среднее значение погрешности в точке x_k диапазона измерения при медленных изменениях входного сигнала со стороны больших значений до значения x_k; $\Delta_{мi}(\Delta_{бi})$ – i-тая реализация погрешности Δ_i – при

изменениях входного сигнала со стороны меньших (больших) значений до значения x_k ; n – число реализаций погрешности в точке x_k при возрастании или убывании значений входного сигнала.

Вариацию можно считать малой и не нормировать, если выполняется неравенство $b \le 0{,}2 \max|\Delta_i|$,

где $\max|\Delta_i|$ - максимальное по модулю значение погрешности Δ_i, реализовавшееся в ходе эксперимента и принятое к обработке (не выброс или промах).

Оценка неисключённой систематической погрешности в К-той точке диапазона измерений выполняется по формуле

$$\theta_k = \frac{\overline{\Delta}_M + \overline{\Delta}_Б}{2} \qquad (28)$$

Здесь $\overline{\Delta}_M$ и $\overline{\Delta}_Б$, величины из формулы (27)

Граница случайной составляющей абсолютной погрешности измерений в К-той точке диапазона измерений определяется по результатам многократных наблюдений по формуле

$$\Delta_{абсК} = K_T \sigma_к(\Delta)$$

Здесь K_T – толерантный коэффициент (в разделе 5.8.2 мы обозначали его как K)

$\sigma_к(\Delta) = \sqrt{\dfrac{\sum_{i=1}^{n/2}(\Delta_{мi} - \overline{\Delta}_M)^2 + \sum_{i=1}^{n/2}(\Delta_{бi} - \overline{\Delta}_Б)^2}{2\left(\frac{n}{2} - 1\right)}}$ - оценка среднего квадратического отклонения случайной составляющей погрешности измерений в К-той точке диапазона измерений; n – число измерений.

Значения толерантного коэффициента приводятся в разделе 5.8.2.

Границу суммарной погрешности измерений определяют в соответствие с рекомендациями раздела 5.9.3.

Чувствительность измерительного преобразователя γ определяется в исследуемых точках диапазона по формуле: $\gamma = \frac{Y}{X}$, где X – значение измеряемой величины на входе преобразователя, Y – значение минимального информативного параметра выходного сигнала. Чувствительность измерительного преобразователя в диапазоне измерений может меняться, что является причиной нелинейности градуировочной характеристики. В случаях, когда нелинейная градуировочная характеристика заменяется на линейную необходимо определить среднее значение чувствительности в диапазоне измерений:

$$\gamma_{ср} = \frac{\gamma_{\max} - \gamma_{\min}}{2}$$

Здесь γ_{max} и γ_{min} – соответственно максимальное и минимальное значения чувствительности, полученные в исследованных точках диапазона измерений.

Относительная нелинейность градуировочной характеристики определяется из выражения: $\delta_{\gamma} = \frac{\gamma_{\max} - \gamma_{\min}}{\gamma_{\max} + \gamma_{\min}} 100\%$

В случаях необходимости учёта возможного влияния погрешности образцовых средств измерений на погрешность аттестуемого НСИ рекомендуется руководствоваться следующим:

-если систематическая составляющая погрешности образцового средства измерений является существенной, то искомый предел погрешности НСИ с учётом влияния погрешности образцового средства измерения определяется по формуле

$$\Delta_{пр} = \pm(\theta_{k\max} + \Delta_1)$$

где θ_{kmax} – максимальное значение систематической погрешности, вычисленное по формуле (28), Δ_1 – предел погрешности НСИ без учёта влияния образцового средства измерения.

- если соотношение между составляющими погрешности образцового средства измерений не определено, то искомый предел погрешности НСИ определим по формуле:

$$\Delta_{пр} = \pm\sqrt{\Delta^2{}_{обр} + \Delta^2{}_1}$$

Таким образом, при расчётах необходимо первоначально исключить из результатов расчёта погрешности НСИ погрешность образцового средства. Полученное, таким образом значение $\Delta_{пр}$ следует принимать в качестве предела соответствующей погрешности и записывать его в свидетельстве о метрологической аттестации.

8.3 Разработка методик выполнения измерений.

Как мы уже обсуждали ранее, метрологическая аттестация приборов использующих прямые и косвенные измерения производится по-разному. Очевидно, что инструментальная погрешность прибора определяется экспериментально, при этом, последовательность операций, при выполнении экспериментальной работы и расчётов погрешности определяется документами, которые называются программа и методика метрологической аттестации. При этом мы получаем информацию об инструментальной погрешности, которая записывается в формуляр прибора. В случае косвенных измерений необходимо определить методическую погрешность и суммарную погрешность измерения. Для определения методической погрешности разрабатывается методика выполнения измерений (МВИ), учитывающая функцию, связывающую измеряемую прибором величину с физической величиной, измеряемой первичным преобразователем. По результатам метрологической аттестации выпускается аттестат на МВИ. Общие требования к стандартизации и аттестации методик выполнения измерений определяются ГОСТ 8.010-72. Аттестацию МВИ и оформление аттестатов производят: метрологические организации Госстандарта или органы ведомственных метрологических служб. Метрологические организации Госстандарта проводят аттестацию МВИ особо точных и ответственных измерений по представлению министерств и ведомств. Аттестацию остальных МВИ проводят органы ведомственных метрологических служб.

В аттестате методики выполнения измерений указывают:

- назначение и область применения методики;
- Типы и номера экземпляров средств измерений, используемых для проведения измерений;
- Технические характеристики вспомогательных устройств, необходимых для выполнения измерений ;
- Метод измерений;
- Алгоритм операций подготовки и выполнения измерений;
- Численные значения показателей точности измерений;
- Межповерочные интервалы для средств измерений и номенклатура нормативных документов, согласно которым должна проводиться их поверка;
- Требования к квалификации операторов;
- Требования техники безопасности.

Следует отметить, что аттестация проводится по программе, утверждаемой руководителем организации, проводящей аттестацию. Об аттестации МВИ составляется технический отчёт, утверждаемый руководителем организации, проводившей аттестацию. В связи с тем, что во многих случаях функциональная связь при косвенных измерениях имеет математическое выражение, то исследования погрешностей этой функциональной связи может носить также аналитический характер. Указанные исследования включаются в отчёт об аттестации МВИ.

Литература

- Анфилатов В. С., Емельянов А. А., Кукушкин А. А. Системный анализ в управлении. — М. Финансы и статистика, 2002. — 368 с.
- ↑ Экономика и менеджмент — Высокие статистические технологии
- ↑ Статистические методы — Высокие статистические технологии
- ↑ Перегудов Ф. И., Тарасевич Ф. П. Введение в системный анализ. — М.: Высшая школа, 1989. — 367 с.
- • Бахрушин В.Є. Методи аналізу даних. — Запоріжжя, КПУ, 2011
- ↑ Ильясов Ф. Н. Шкалы и специфика социологического измерения // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 1. С. 3-16.
- • (1993) «Nominal, ordinal, interval, and ratio typologies are misleading». *The American Statistician* (American Statistical Association) **47**: 65–72. DOI:10.2307/2684788.
- ↑ Scaling : a sourcebook for behavioral scientists. — AldineTransaction, [2007]. — ISBN 9780202361758.
- ↑ *Bela O. Baker, Curtis D. Hardyck, Lewis F. Petrinovich* Weak Measurements vs. Strong Statistics: An Empirical Critique of S. S. Stevens' Proscriptions nn Statistics (англ.) // Educational and Psychological Measurement. — 1966-07-01. — Vol. 26, iss. 2. — P. 291–309. — ISSN 0013-1644. — DOI:10.1177/001316446602600204.
- ↑ *Edgar F. Borgatta, George W. Bohrnstedt* Level of Measurement: Once Over Again (англ.) // Sociological Methods & Research. — 1980-11-01. — Vol. 9, iss. 2. — P. 147–160. — ISSN 0049-1241. — DOI:10.1177/004912418000900202.
- ↑ *Louis Guttman* What is Not What in Statistics // Journal of the Royal Statistical Society. Series D (The Statistician). — 1977. — Т. 26, вып. 2. — С. 81–107. — DOI:10.2307/2987957.
- ↑ (December 1953) «On the Statistical Treatment of Football Numbers». *American Psychologist* **8**: 750–751. DOI:10.1037/h0063675.
- ↑ *Mosteller Frederick.* Data analysis and regression : a second course in statistics. — Reading, Mass: Addison-Wesley Pub. Co, 1977. — ISBN 978-0201048544.
- ↑ Wolman, Abel G (2006). «Measurement and meaningfulness in conservation science». *Conservation biology*.

- ↑ What is the difference between categorical, ordinal and interval variables?. *Institute for Digital Research and Education*. University of California, Los Angeles.

- ↑ Суппес П.[en], Зиннес Д. Основы теории измерений // Психологические измерения. М.: 1967. С. 9-110.

- Справочник по электроизмерительным приборам / Под ред. К. К. Илюнина — Л.: Энергоатомиздат, 1983

Нормативно-техническая документация

- РМГ 29-2013 ГСИ. Метрология, Основные термины и определения
- ГОСТ 5365-83 Приборы электроизмерительные. Циферблаты и шкалы. Общие технические требования
- ГОСТ 25741-83 Циферблаты и шкалы манометрических термометров. Технические требования и маркировка

1. Арутюнов П. А. Теория и применение алгоритмических измерений. - М.: Энергоатомиздат, 1990.
2. Болынев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1983.
3. Бурдун Г.Д. Марков Б.Н. Основы метрологии. - М.: Изд. стандартов, 1985.
4. Воинов В.Г., Никулин М.С. Несмещенные оценки и их применение. - М.: Наука, 1989.
5. Драхсел Р. Основы электроизмерительной техники. - М: Энергоатомиздат, 1982.
6. Иванов В.И., Машкович В.П., Цептер Э.М. Международная система единиц (СИ) в атомной науке и технике: Справочное руководство. -М.: Энергоиздат, 1981.
7. Камке Д., Кремер К. Физические основы измерений. - М.: Мир, 1980.
8. Каргаполов М.И., Мерзляков Ю.И. Основы теории групп. - М.: Наука, 1972.
9. Кендалл М.Дж., Стюарт А. Статистические выводы и связи. - М: Наука, 1973.
10. Корн Г., Корн Т. Справочник по математике. - М.: Наука, 1973.
11. Крамер Г. Математические методы статистики. - М.: Мир, 1975.
12. Кунце Х.И. Методы физических измерений. - М.: Мир, 1989.

13. МИ 199-79: Методика установления вида математической модели распределения погрешностей. -М: Изд. стандартов, 1980.
14. Митропольский А.К. Техника статистических вычислений. М.: Наука, 1971.
15. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. - М:. Энергоатомиздат, 1985.
16. Новоселов О.Н., Фомин А.Ф., Основы теории и расчета информационно-измерительных систем. - М.: Машиностроение, 1980.
17. Пфанцагль И. Теория измерений. - М.: Мир, 1976.
18. Романов В.Н. Планирование эксперимента: Учебное пособие. - СПб.: СЗПИ, 1992.
19. Романов В.Н., Соболев В.С., Цветков Э.И. Интеллектуальные средства измерений. - М.: РИЦ «Татьянин день», 1994.
20. Справочник по специальным функциям/ Под ред. М. Абрамовича, И. Стиган. - М.: Наука, 1979.
21. Стахов А.П. Введение в алгоритмическую теорию измерений. - М.: Сов. радио, 1977.
22. Супес П., Зинес Д. Основы теории измерений // Психологические измерения. - М.: Мир, 1967.
23. Феферман С. Числовые системы. - М.: Наука, 1971.
24. Хованов Н.В. Математические основы теории шкал измерения качества. - Л.: Изд. ЛГУ, 1982.
25. Хофман Д. Измерительно-вычислительные системы обеспечения качества. - М.: Энергоатомиздат, 1991.
26. Хьюбер П. Робастность в статистике. - М. Мир, 1984.
27. Цветков Э. И. Основы теории статистических измерений. - Л.: Энергия, 1979.

[1] Броневич А.Г., Каркищенко А.Н. Описание нечетких мер в рамках вероятностного подхода // Нечеткие системы и мягкие вычисления, том 2, № 7, 2007, стр. 7 30.
[2] Bronevich A.G. An investigation of ideals in the set of fuzzy measures // Fuzzy Sets and Systems, v. 152, 2005, pp. 271 288.
[3] Cozmari F. A brief introduction to the theory of sets of probability measures, http: // www.cs.crnu.edu/ fgcozman/qBayes.htnil. 1999.
[4] Walley P. Statistical reasoning with imprecise probabilities. Chapman and Hall, London, 1991.
[5] Кузнецов В.П. Интервальные статистические модели. - М.: Радио и связь, 1991.

[6] Klir G.J. Uncertainty and Information: Foundations of Generalized Information Theory. Wiley-Interscience, Hoboken, NJ, 2006.
[7] Bronevich A.G., Klir G.J. Measures of uncertainty for imprecise probabilities: An axiomatic approach // International Journal of Approximate Reasoning, vol. 51, 2010, pp. 365 390.
[8] Dempster A.P. Upper and lower probabilities induced by multivalued mapping // Ann. Math. Statist., 1967, v.38, pp.325 339.
[9] Shafer G. A mathematical theory of evidence. Princeton University Press, Princeton, N.J., 1976.
[10] Bronevich A.G., Lepskiy A.E. Geometrical fuzzy measures in image processing and pattern recognition. Proc. of the 10th IFSA World Congress, 2003, Istanbul, Turkey, pp. 151 154.
[11] Черников С.Н. Линейные неравенства. - М.: Наука, 1968.
[12] Walley P. Coherent lower (and upper) probabilities. Technical Report 22, Department of Statistics, University of Warwick. U.K.. 1981.
[13] Bronevich A.G. On the closure of families of fuzz}- measures under eventwise aggregations // Fuzzy sets and systems, v. 153, 2005. pp. 45 70-
[14] Jaffray J.Y. Bayesian updating and belief functions. IEEE Tr. Systems Man and Cybernetics, 1992, v. 22, pp. 1144 1152.
[15| Fagin R., Halpern J.Y. A new approach to updating beliefs. Uncertainty in Art. Int. 1991, v. 6, 347 374. .
[16] Smets P. Belief functions and the transferable belief model.
http://www.sipta.org/documentation/belief/belief.pdf, 1999.
[17) Дюбуа Д., Прад А. Теория возможностей. Приложения к представлению знаний в информатике. М.: Радио и связь, 1990.
[18] Denneberg D. Non-additive measure and integral, basic concepts and their role for applications. In M. Grabisch, T. Murofishi, M. Sugeno (eds.) Fuzzy measures arid integrals Theory and applications. Studies on fuzzyness and soft computing, Physica-Verlag, Heidelberg, 2000.
[19] De Cooman G. Precision-imprecision equivalence in abroad class of imprecise hierarchical uncertainty models // Journal of Statistical Planning and Inference, v. 105, 2002, pp 175 198.
[20] Dermeberg D. Conditioning (updating) non-additive measures // Annals of Operations Research, v. 52, 1994. pp. 21 42.
[21] Denneberg D. Totally monotone core and products of monotone measures // International Journal of Approximate Reasoning, v. 24, 2000, pp. 273 281.