

Л**МФТИ**

Е. Н. Аристова  
Н. А. Завьялова  
А. И. Лобанов

**Практические занятия  
по вычислительной  
математике  
в МФТИ**

**Часть I**

Министерство науки и высшего образования Российской Федерации  
Федеральное государственное автономное образовательное  
учреждение высшего образования  
«Московский физико-технический институт  
(национальный исследовательский университет)»

Е. Н. Аристова, Н. А. Завьялова, А. И. Лобанов

**ПРАКТИЧЕСКИЕ ЗАНЯТИЯ  
ПО ВЫЧИСЛИТЕЛЬНОЙ  
МАТЕМАТИКЕ В МФТИ**

**Часть 1**

Издание второе, исправленное и дополненное

Учебное пособие

МОСКВА  
МФТИ  
2021

УДК 519.6(075)

ББК 22.19я73

A81

**Рецензенты:**

Кафедра прикладной математики ИИТ федерального государственного бюджетного образовательного учреждения высшего образования

«МИРЭА – Российский технологический университет»

(зав. кафедрой – к.ф.-м.н. Р.И. Дзержинский)

Доктор физико-математических наук, доцент С.А. Ишанов

**Аристова, Елена Николаевна,  
Завьялова, Наталья Александровна,  
Лобанов, Алексей Иванович**

**A81** Практические занятия по вычислительной математике в МФТИ. Ч. 1. Изд. 2-е, испр. и дополн. : учеб. пособие / Е. Н. Аристова, Н. А. Завьялова, А. И. Лобанов. – Москва : МФТИ, 2021. – 242 с. – Библиогр.: с. 238-231.

ISBN 978-5-7417-0774-6 (Ч. I)

Учебное пособие включает материал по практическим занятиям по вычислительной математике, соответствующий материалу первого семестра третьего курса. Рассматриваются разделы, связанные с погрешностями вычислений, вопросами прикладной линейной алгебры, решения переопределенных систем, интерполяции функций, численного интегрирования, численного решения задач Коши для систем ОДУ.

Первое издание вышло в 2014 г. и завоевало популярность не только в МФТИ, но и в других вузах. Во втором издании ликвидированы замеченные опечатки, устранены некоторые возникшие дублирования в задачах. Расширен и дополнен справочный материал.

**УДК 519.6(075)**

**ББК 22.19я73**

*Печатается по решению Редакционно-издательского совета Московского физико-технического института (национального исследовательского университета)*

**ISBN 978-5-7417-0774-6 (Ч. I)** © Аристова Е. Н., Завьялова Н. А., Лобанов А. И., 2021

**ISBN 978-5-7417-0763-0** © Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт институт (национальный исследовательский университет)», 2021

# Оглавление

Предисловие .....	8
I. Погрешности вычислений .....	9
I.1. Введение.....	9
I.2. Погрешности вычислений. Теоретическая справка .....	9
I.3. Вычисление значения функции с помощью разложения в ряд Тейлора .....	12
I.4. Вычисление производной — задача численного дифференцирования .....	13
I.4.1. Двухточечные формулы численного дифференцирования .....	13
I.4.2. Формула второго порядка аппроксимации .....	15
I.4.3. Формула четвертого порядка аппроксимации .....	16
I.5. Стандарт представления числа с плавающей точкой.....	17
I.6. Задачи на доказательства.....	19
I.7. Примеры решения задач .....	21
I.8. Теоретические задачи .....	22
I.9. Практические задачи.....	29
I.10. Библиографический комментарий .....	31
II. Элементы прикладной линейной алгебры .....	32
II.1 Введение .....	32
II.1. Некоторые сведения о векторных пространствах .....	32
II.2.1. Согласованные и подчиненные нормы векторов и матриц.....	33
II.1.2. Другие нормы в $R^n$ . Теорема об эквивалентности норм .....	34
II.3. Обусловленность СЛАУ. Число обусловленности матрицы .....	35
II.4. Решение систем линейных алгебраических уравнений (СЛАУ). Прямые и итерационные методы .....	36
II.4.1. Прямые методы решения СЛАУ .....	36
II.3.2. Метод исключения Гаусса.....	36
II.4.3. LU-разложение .....	38
II.5. Итерационные методы решения СЛАУ .....	39

II.5.1. Метод простой итерации.....	39
II.5.2. Каноническая форма записи двухслойных итерационных методов	41
II.5.3. Методы Якоби, Зейделя, верхней релаксации .....	41
II.6. О спектральных задачах.....	45
II.7. Задачи на доказательство .....	47
II.8. Задачи с решениями .....	52
II.9. Теоретические задачи.....	57
II.10. Практические задачи .....	66
II.11. Библиографический комментарий .....	71
III. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ .....	72
III.1 Переопределенная система линейных алгебраических уравнений ...	72
III.2. Геометрический смысл метода наименьших квадратов .....	73
III.3. Задача неточной интерполяции функции.....	75
III.4. Теоретические задачи .....	77
III.5. Практические задачи.....	79
III.6. Библиографический комментарий .....	82
IV. ЧИСЛЕННОЕ РЕШЕНИЕ НЕЛИНЕЙНЫХ УРАВНЕНИЙ И СИСТЕМ .....	83
IV.1. Введение .....	83
IV.2. Метод деления отрезка пополам (дихотомии) .....	83
IV.3. Методы, основанные на интерполяции.....	84
IV.4. Метод простой итерации .....	86
IV.5. Метод Ньютона .....	87
IV.6. Метод простой итерации для систем нелинейных уравнений .....	89
IV.7. Метод Ньютона для систем нелинейных уравнений .....	90
IV.8. Критерии сходимости итераций .....	91
IV.9. Задачи на доказательство .....	92
IV.10. Задачи с решениями.....	94
IV.11. Теоретические задачи .....	100
IV.12. Практические задачи.....	106

IV.13. Библиографическая справка .....	108
<b>V. ЗАДАЧА ПОИСКА ЭКСТРЕМУМА ФУНКЦИИ .....</b>	<b>109</b>
V.1. Основные понятия .....	109
V.2. Метод перебора .....	111
V.3. Поиск минимума функции одного переменного.....	112
V.3.1. Метод деления отрезка пополам (метод дихотомии) .....	112
V.3.2. Метод золотого сечения .....	113
V.3.3. Метод парабол.....	114
V.3.4. Модифицированный метод Брэндта .....	115
V.4. Поиск минимума функции многих переменных .....	117
V.4.1. Методы спуска .....	117
V.4.1.1. Метод покоординатного спуска .....	117
V.4.2. Метод градиентного спуска .....	119
V.4.3. Метод наискорейшего спуска .....	120
V.4.4. Метод наискорейшего спуска для решения систем нелинейных уравнений .....	121
V.4.5. Динамический метод .....	123
V.5. Задачи с решениями.....	126
V.6. Теоретические задачи .....	128
V.7. Практические задачи .....	129
<b>VI. ТАБЛИЧНОЕ ЗАДАНИЕ И ИНТЕРПОЛИРОВАНИЕ ФУНКЦИЙ..</b>	<b>132</b>
VI.1. Задача интерполяции.....	132
VI.2. Алгебраическая интерполяция .....	132
VI.2.1. Непосредственное вычисление коэффициентов интерполяционного полинома .....	133
VI.2.2. Интерполяционный полином в форме Лагранжа. Интерполяционный полином в форме Ньютона.....	133
VI.2.3. Формула погрешности алгебраической интерполяции .....	134
VI.2.4. О сходимости интерполяционного процесса .....	135
VI.2.5. Обусловленность задачи интерполяции .....	136

VI.3. Тригонометрическая интерполяция .....	138
VI.3.1 Постановка задачи .....	138
VI.3.2. Обусловленность тригонометрической интерполяции .....	140
VI.4. Классическая кусочно-многочленная интерполяция.....	141
VI.4.1. Оценка неустойчивой погрешности при интерполяции .....	141
VI.4.2. Насыщаемость (гладкостью) кусочно-многочленной интерполяции .....	142
VI.4.3. Нелокальная гладкая кусочно-многочленная интерполяция .....	142
VI.5. Дробно-полиномиальные аппроксимации.....	145
VI.5.1. Рациональная интерполяция .....	145
VI.5.2. Аппроксимация Паде .....	147
VI.6. Задачи с решениями.....	149
VI.7. Задачи на доказательство .....	155
VI.8. Теоретические задачи .....	158
VI.9. Практические задачи.....	162
VI.10. Библиографический комментарий.....	169
<b>VII. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ .....</b>	<b>170</b>
VII.1. Квадратурные формулы Ньютона–Котеса (интерполяционного типа).....	170
VII.1.1. Оценка погрешности квадратурных формул.....	172
VII.1.2. Связь между формулами прямоугольников, трапеций и Симпсона .....	174
VII.2. Экстраполяция Ричардсона. Правило Рунге практического оценивания погрешности. Алгоритм Ромберга .....	175
VII.3. Квадратурные формулы Гаусса .....	176
VII.3.1. Квадратурные формулы Гаусса–Кристоффеля .....	178
VII.4. Приемы вычисления несобственных интегралов .....	178
VII.5. Вычисление интегралов от быстроосциллирующих функций .....	182
VII.6. Задачи на доказательство .....	183
VII.7. Задачи с решениями .....	184

VII.8. Теоретические задачи .....	190
VII.9. Практические задачи.....	194
VII.10. Библиографический комментарий.....	196
VIII. ЗАДАЧА КОШИ ДЛЯ СИСТЕМ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ .....	198
VIII.1. Аппроксимация, устойчивость, сходимость.....	198
VIII.2. Исследование устойчивости разностных схем для ОДУ .....	201
VIII.3. Явные методы Рунге–Кутты .....	205
VIII.4. Устойчивость явных методов Рунге–Кутты.....	207
VIII.5. Методы Адамса .....	209
VIII.6. Экстраполяция Ричардсона.....	210
VIII.7. Задачи на доказательство .....	211
VIII.8. Задачи с решениями.....	212
VIII.9. Теоретические задачи .....	216
VIII.10. Практические задачи .....	221
VIII.11. Устойчивость методов Рунге–Кутты на различных типах траекторий и практические задачи.....	223
VIII.12. Библиографический комментарий .....	224
Приложение. Некоторые системы ортогональных многочленов .....	225
Многочлены Чебышёва .....	225
Некоторые другие системы ортогональных многочленов .....	228
Ответы .....	231
Литература.....	238

# Предисловие

Дорогие читатели!

Вы держите в руках первую часть учебного пособия «Практические занятия по вычислительной математике в МФТИ». В пособии содержится материал, соответствующий первой части курса, обычно изучаемой в осен-нем семестре.

Практическая работа на семинарах – важная часть курса. Первый сборник задач для практических занятий был подготовлен на кафедре еще в 1974 году [1]. С тех пор менялось и совершенствовалось наполнение курсов, изданы книги, соответствующие курсам, читаемым в МФТИ [2–5]. Появилась необходимость в новых сборниках задач [6].

Кроме того, время изменилось. Если в 1974 году практические задачи с реализацией на компьютере занимали лишь малую часть нагрузки на каждого студента, то сейчас персональный компьютер легко доступен каждому. Поэтому возникла необходимость радикальной смены содержания практической части.

Пособие строится следующим образом. В начале каждой темы приводятся справочные материалы, необходимые для решения задач. Некоторые необходимые разделы, традиционно не освещаемые в лекциях, изложены достаточно подробно.

В каждой теме приводятся задачи с решениями. Задания для самостоятельной работы делятся на задачи на доказательства, теоретические задачи и практические задачи, предполагающие реализацию на компьютере.

При подготовке пособия использованы задачники [1, 7, 8]. Многие из приведенных задач являются авторскими. Составители данного пособия выражают благодарность авторам задач — это В. С. Рябенький, А.С. Холодов, В. Б. Пирогов, И. Б. Петров, В. И. Косарев, Т. К. Старожилова, М. В. Мещеряков, Л. А. Чудов, О. А. Пыркова и другие коллеги по кафедре вычислительной математики МФТИ. Особую благодарность авторы высказывают О. Н. Агахановой, А. А. Андреевой, В. Д. Иванову, Н. И. Караваевой за сделанные замечания и помочь в подготовке переиздания.

Первое издание пособия было встречено с большим интересом. Материал использовался на практических занятиях не только в МФТИ, но и в других институтах. При подготовке второго издания был учтен опыт использования задачника при проведении практических занятий. Исправлены некоторые опечатки. Добавлен справочный материал по ортогональным полиномам. Устранено некоторое дублирование задач.

Желаем всем читателям успехов в изучении курса!

# I. Погрешности вычислений

## I.1. Введение

При оценке достоверности результатов численного расчета ключевую роль играет анализ погрешностей, неизбежно возникающих при любом использовании компьютера. В первой главе Вы познакомитесь с основными источниками возникновения погрешности.

Весь классический математический анализ опирается на понятие действительного числа. При этом действительное число понимается как бесконечная, вообще говоря, непериодическая десятичная дробь. При работе на вычислительной системе бесконечные десятичные дроби заменяются их конечными приближениями. Ввиду того, что и мантисса числа при работе с числами в формате с плавающей точкой, и порядок ограниченны сверху, и снизу, мы имеем дело с конечным (хотя и очень большим) множеством чисел, которые могут быть представлены в машинной арифметике.

Возникает актуальная проблема соответствия результатов решения конечномерной задачи при решении ее на конечном подмножестве действительных чисел и точного решения задачи. Точное решение задачи, как правило, на ЭВМ невозможно в силу указанных причин.

Как будет показано ниже, результаты вычислений могут существенно меняться при изменении внутреннего машинного представления действительных чисел. Кроме того, дается краткая теоретическая справка по теории погрешностей.

Рассматривается также задача приближенного вычисления производных функций – задача численного дифференцирования. Рассматривается влияние погрешностей метода и погрешностей округления на результат вычислений.

## I.2. Погрешности вычислений. Теоретическая справка

Напомним некоторые понятия, связанные с погрешностями. Если  $a$  — точное значение некоторой величины,  $a^*$  — ее приближенное значение, то *абсолютной погрешностью* величины  $a^*$  обычно называют наименьшую величину  $\Delta(a^*)$ , про которую известно, что

$$|a^* - a| \leq \Delta(a^*).$$

*Относительной погрешностью* приближенного значения называют наименьшую величину  $\delta(a^*)$ , про которую известно, что

$$\left| \frac{(a^* - a)}{a^*} \right| \leq \delta(a^*).$$

В любой вычислительной задаче по некоторым входным данным требуется найти ответ на поставленный вопрос. Для вычисления значения функции  $y = f(x)$  при  $x = t$  входными данными задачи служат число  $x$  и закон  $f$ , по которому каждому значению аргумента  $x$  ставится в соответствие значение функции  $y = f(x)$ .

Если ответ можно дать с любой точностью, то погрешность отсутствует. Но обычно ответ удается найти лишь приближенно. Погрешность задачи вызывается тремя причинами.

Первая — неопределенность при задании входных данных, которая приводит к неопределенности в ответе. Ответ может быть указан лишь с погрешностью, которая называется *неустранимой*.

Проиллюстрируем понятие неустранимой погрешности на примере. Пусть функция  $f(x)$  известна приближенно, например, она отличается от  $\sin x$  не более чем на величину  $\varepsilon > 0$ :

$$\sin(x) - \varepsilon \leq f(x) \leq \sin(x) + \varepsilon. \quad (2.1)$$

Кроме того, пусть значение аргумента  $x = t$  получается приближенным измерением, в результате которого получаем  $x = t^*$ , причем известно, что  $t$  лежит в пределах

$$t^* - \delta \leq t \leq t^* + \delta, \quad (2.2)$$

где  $\delta > 0$  — число, характеризующее точность измерения (для определенности будем считать, что функция  $\sin t$  на отрезке (2.2) монотонно возрастает).

Величиной  $y = f(t)$  может оказаться координата любой точки отрезка  $y \in [a, b]$  (см. рис. 1), где  $a = \sin(t^* - \delta) - \varepsilon$ ,  $b = \sin(t^* + \delta) + \varepsilon$ . Понятно, что, приняв за приближенное значение величины  $y = f(x)$  значение в любой точке  $y^*$  отрезка  $[a, b]$ , можно гарантировать оценку погрешности:

$$|y - y^*| \leq |b - a|. \quad (2.3)$$

Эту гарантированную оценку погрешности нельзя существенно улучшить при имеющихся неполных входных данных. Самая малая погрешность получается, если принять за  $y^*$  середину отрезка  $[a, b]$ , положив

$$y^* = y_{\text{опт}}^* = \frac{|b - a|}{2}.$$

Тогда справедливая оценка

$$|y - y_{\text{опт}}^*| \leq \frac{|b-a|}{2}. \quad (2.4)$$

Таким образом,  $0.5|b-a|$  и есть та *неустранимая* (не уменьшаемая) *погрешность*, которую можно гарантировать при имеющихся неопределенных входных данных в случае самого удачного выбора приближенного решения  $y^*$ . Оптимальная оценка (2.4) ненамного лучше оценки (2.3). Поэтому не только о точке  $y^*$  опт, но и о любой точке  $y^* \in [a, b]$  условимся говорить, что она является приближенным решением задачи вычисления числа  $y(t)$ , найденным с неустранимой погрешностью, а вместо  $0.5|b-a|$  из (2.4) за величину неустранимой погрешности примем (условно) число  $|b-a|$ .

Вторая причина возникновения погрешности состоит в том, что при фиксированных входных данных ответ вычисляется с помощью приближенного метода. Возникает погрешность, связанная с выбором метода — *погрешность метода вычислений*. Проиллюстрируем это на следующем простом примере.

Положим  $y^* = \sin t^*$ . Точка  $y^*$  выбрана среди других точек отрезка  $[a, b]$  (см. выше по поводу неустранимой погрешности), так как она задается при помощи формулы, удобной для дальнейшего.

Воспользуемся разложением функции  $\sin t$  в ряд Тейлора в окрестности нуля (ряд Маклорена):

$$\sin t = t - \frac{t^3}{3!} + \frac{t^5}{5!} - \dots \quad (2.5)$$

Для вычисления приближенного значения функции  $y^*$  можно взять разное

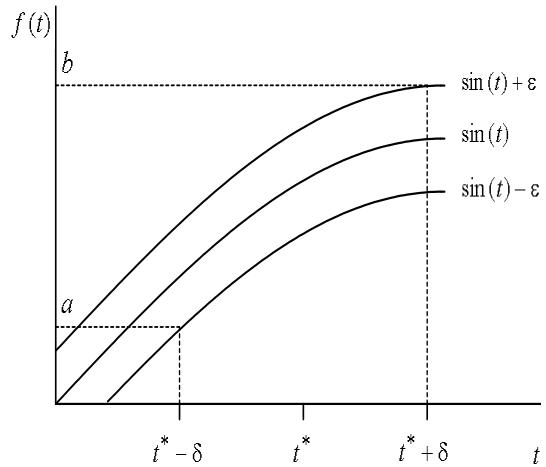


Рис. 1.1. К приближенному вычислению значения функции

количество членов разложения:

$$\begin{aligned}y^* &\approx y_1^* = t^*, \\y^* &\approx y_2^* = t^* - \frac{t^{*3}}{3!}, \\y^* &\approx y_n^* = \sum_{k=0}^{n-1} (-1)^k \frac{t^{*(2k+1)}}{(2k+1)!},\end{aligned}\tag{2.6}$$

что определяет метод вычисления.

Величина  $|y^* - y_n^*|$  — погрешность метода вычисления.

Фактически выбранный метод вычисления зависит от параметра  $n$  и позволяет добиться, чтобы погрешность метода была меньше любой наперед заданной величины за счет выбора этого параметра.

Очевидно, нет смысла стремиться, чтобы погрешность метода была существенно (во много раз) меньше неустранимой погрешности. Поэтому число  $n$  не стоит выбирать слишком большим. Однако если  $n$  слишком мало и погрешность метода существенно больше неустранимой погрешности, то выбранный способ не полностью использует информацию о решении, содержащуюся во входных данных. Часть этой информации теряется.

Наконец, сам выбранный приближенный метод реализуется неточно из-за ошибок округления при вычислениях на реальном компьютере. Так, при вычислении  $y_n^*$  по одной из формул (2.6) на реальном компьютере в результате ошибок округления мы получим значение  $\tilde{y}_n^*$ .

Величину  $|y_n^* - \tilde{y}_n^*|$  называют погрешностью округления. Она не должна быть существенно больше погрешности метода. В противном случае произойдет потеря точности метода за счет ошибок округления. Точность метода вычислений также целесообразно согласовывать с величиной ожидаемых ошибок округления.

Погрешность результата складывается, таким образом, из неустранимой погрешности метода и погрешности округления.

### I.3. Вычисление значения функции с помощью разложения в ряд Тейлора

Пусть требуется вычислить значения  $y = \sin t$ . Воспользуемся разложением функции  $\sin t$  в окрестности нуля в ряд Тейлора (2.5), радиус сходимости которого для данной функции равен бесконечности.

Для вычисления  $y$  можно воспользоваться одним из приближенных выражений (2.6). Выбирая для вычисления  $y$  одну из формул, мы тем самым

выбираем приближенный метод вычисления, точность которого определяется числом привлекаемых членов ряда  $n$ . Ряд Тейлора для функции  $\sin t$  является знакопеременным, сходится для любого значения  $t$ , а его частичная сумма отличается от точного значения функции не более, чем на величину первого отброшенного члена ряда. Выбирая  $n$  так, чтобы

$$\frac{t^{2n+1}}{(2n+1)!} \leq \varepsilon,$$

можно добиться любой наперед заданной точности  $\varepsilon$ .

Однако при вычислениях на реальном компьютере получить результат с требуемой точностью для  $t$  (которое существенно больше единицы) не удается из-за быстрого роста ошибок округления. Последние тем больше, чем больше  $t$ . Это связано с различным характером поведения величины членов ряда Тейлора при  $t > 1$  и  $t < 1$ . При  $t < 1$  члены ряда по абсолютной величине монотонно убывают в зависимости от  $n$ . При  $t > 1$  члены ряда по модулю сначала растут (тем сильнее, чем больше  $t$ ) и только потом, достигнув при некотором  $k = m$  максимума, начинают убывать и стремиться к нулю при  $n \rightarrow \infty$ . Для того чтобы обеспечить при вычислении, например,  $a_m$ -го (максимального по модулю) члена ряда абсолютную погрешность, не превосходящую  $\varepsilon$ , необходимо вычислить его с относительной погрешностью, не хуже чем

$$\delta(a_m) \leq \frac{\Delta a_m}{|a_m|} \leq \frac{\varepsilon}{|a_m|}.$$

Требуемая относительная точность тем выше, чем больше  $|a_m|$ , что можно обеспечить только увеличением длины мантиссы.

**Замечание.** В реальных расчетах методы вычисления значений функции через конечные суммы ряда Тейлора никогда не используются.

## I.4. Вычисление производной — задача численного дифференцирования

Пусть задана функция  $f(x)$ . Необходимо вычислить ее первую производную в некоторой точке  $x$ . Воспользуемся для этого формулами численного дифференцирования различного порядка аппроксимации.

### I.4.1. Двухточечные формулы численного дифференцирования

Простейшая формула численного дифференцирования получена на

основе определения производной функции, при этом приращение аргумента считается малым, но конечным:

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}. \quad (4.1)$$

Пусть известно, что  $|f''(\xi)| \leq M_2$ , тогда погрешность метода для этой формулы

$$|r_1| = \left| f'(x) - \frac{f(x+h) - f(x)}{h} \right| \leq \frac{M_2 h}{2}. \quad (4.2)$$

Говорят, что эта формула имеет первый порядок аппроксимации по  $h$ . Аккуратное определение порядка аппроксимации будет введено в главе VIII.

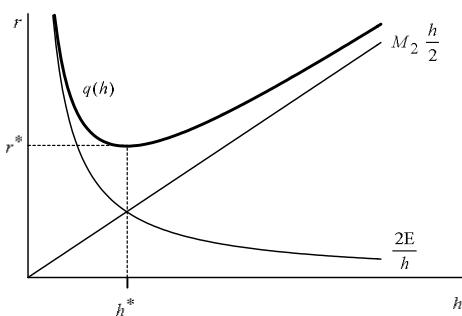


Рис. 1.2. К вычислению первой производной и определению оптимального шага численного дифференцирования

Пусть значения функции  $f(x)$  известны с погрешностью  $\varepsilon(x)$ ,  $|\varepsilon(x)| \leq E$ . Даже в случае отсутствия неустойчивой погрешности  $f$ , при вычислении значения функции на ЭВМ возникает погрешность за счет ошибок округления, и ее величина в этом случае зависит от представления чисел в машине. Обозна-

чим  $\varepsilon_{\text{маш}}$  — максимальное число, для которого в машинной арифметике справедливо равенство  $1 + \varepsilon_{\text{маш}} = 1$ . Ошибка, связанная с ошибкой округления значения  $f(x)$ , не превосходит величины  $E = M_0 \cdot \varepsilon_{\text{маш}}$ , где  $|f'(\xi)| \leq M_0$ . Тогда при вычислении производной по формуле (4.1) возникает погрешность  $r_2$ , причем

$$|r_2| \leq \frac{2E}{h}. \quad (4.3)$$

Для суммарной абсолютной погрешности  $r$  имеем оценку

$$|r| \leq |r_1| + |r_2| \leq q(h) = \frac{M_2 h}{2} + \frac{2E}{h}. \quad (4.4)$$

Для уменьшения погрешности метода необходимо, согласно оценке (4.2), уменьшить шаг  $h$ , но при этом растет второе слагаемое в (4.4).

На рис.1.2 представлен характер зависимости погрешности метода, погрешности вычисления функции и суммарной погрешности в зависимости от шага  $h$ . Минимум суммарной погрешности достигается в точке  $h^*$  экстремума функции  $q(h)$ :  $q'(h) = 0$ , причем в ней  $r_1 = r_2$ .

Тогда имеем для оптимального шага дифференцирования:

$$\frac{dq(h)}{dh} = 0, \quad h_{\text{опт}} = 2 \sqrt{\frac{E}{M_2}}. \quad (4.5)$$

При использовании формулы (4.1.1) нельзя рассчитывать на точность более высокую, чем

$$r_{\min} = 2\sqrt{EM_2}, \quad (4.6)$$

наличие предельной точности является следствием (4.4) при  $h = h_{\text{опт}}$ .

Следовательно, производную можно вычислить, в лучшем случае, с половиной верных знаков (если  $M_2$  и  $M_0 \approx 1$ ).

Рассмотрим теперь, как изменятся результаты при использовании формулы численного дифференцирования второго порядка аппроксимации.

#### I.4.2. Формула второго порядка аппроксимации

Формула второго порядка аппроксимации с центральной разностью может быть записана

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}. \quad (4.7)$$

Пусть известно, что  $|f'''(\xi)| < M_3$ ; тогда погрешность метода для этой формулы имеет второй порядок по  $h$ :

$$|r_1| = \left| f'(x) - \frac{f(x+h) - f(x-h)}{2h} \right| \leq \frac{M_3 h^2}{6}. \quad (4.8)$$

Пусть значения функции  $f(x)$  известны с погрешностью  $\varepsilon(x)$ ,  $|\varepsilon(x)| \leq E$ . Тогда при вычислении производной по формуле (4.7) возникает погрешность  $|r_2|$ , причем

$$|r_2| \leq \frac{E}{h}. \quad (4.9)$$

Для суммарной погрешности  $r$  имеем оценку:

$$|r| = |r_1| + |r_2| \leq q(h) = \frac{M_3 h^2}{6} + \frac{E}{h}. \quad (4.10)$$

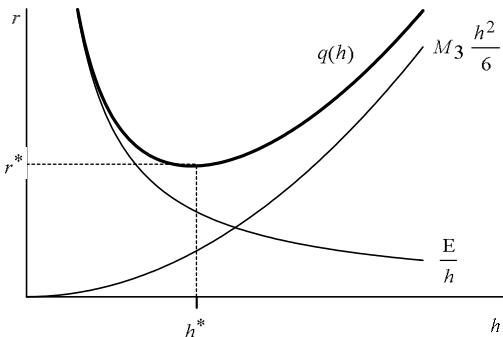


Рис. 1.3. К определению оптимального шага дифференцирования по формуле второго порядка

в точке  $h$  — экстремума функции  $q(h)$ :  $q'(h) = 0$ . Оптимальное значение шага численного дифференцирования есть

$$h_{\text{опт}} = \sqrt[3]{\frac{3E}{M_3}}. \quad (4.11)$$

Таким образом, при использовании формулы (4.7) нельзя рассчитывать на точность более высокую, чем

$$r_{\min} = \sqrt[3]{\frac{9E^2 M_3}{8}}. \quad (4.12)$$

#### I.4.3. Формула четвертого порядка аппроксимации

$$f' \approx \frac{f(x-2h) - 8f(x-h) + 8f(x+h) - f(x+2h)}{12h}. \quad (4.13)$$

Пусть известно, что  $|f''(\xi)| \leq M_5$ ; тогда погрешность метода для этой формулы имеет четвертый порядок по  $h$ :

$$|r_1| = \left| f' - \frac{f(x-2h) - 8f(x-h) + 8f(x+2h) - f(x+2h)}{12h} \right| \leq \frac{M_5 h^4}{30}. \quad (4.14)$$

Для уменьшения погрешности метода необходимо, согласно оценке (4.8), уменьшить шаг  $h$ , но при этом расчет второе слагаемое в (4.10). На рис. 3 представлен характер зависимости погрешности метода, погрешности вычисления функции и суммарной погрешности в зависимости от шага  $h$ . Минимум погрешности достигается

Пусть значения функции  $f(x)$  известны с погрешностью  $\varepsilon(x)$ ,  $|\varepsilon(x)| \leq E$ . Тогда при вычислении производной по формуле (4.13) возникает погрешность округления  $|r_2|$ , причем

$$|r_2| \leq \frac{3E}{2h}. \quad (4.15)$$

Для суммарной погрешности  $r$  получаем оценку

$$|r| \leq |r_1| + |r_2| \leq q(h) = \frac{M_5 h^4}{30} + \frac{3E}{2h}. \quad (4.16)$$

Для уменьшения погрешности метода необходимо, согласно (4.14), уменьшить шаг  $h$ , но при этом растет второе слагаемое в (4.16). На рис. 1.4 представлен характер зависимости погрешности метода, погрешности вычислений и суммарной погрешности в зависимости от  $h$ . Минимум погрешности достигается в точке  $h_{\text{опт}}$  экстремума функции  $q(h)$ :  $q'(h) = 0$ . Имеем для оптимального шага

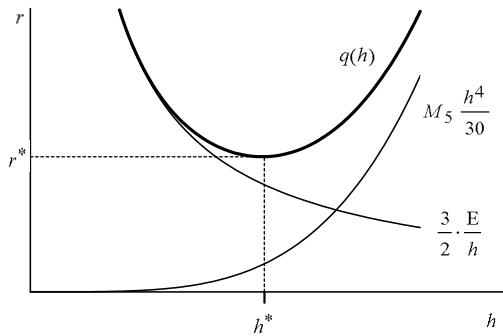


Рис. 1.4. Оптимальный шаг для формулы четвертого порядка

$$h_{\text{опт}} = \sqrt[5]{\frac{45E}{4M_5}}. \quad (4.17)$$

Таким образом, при использовании формулы (4.13) нельзя рассчитывать на точность более высокую, чем

$$r_{\min} = \frac{15}{8} \sqrt[5]{\frac{4E^4 M_5}{15}}.$$

## I.5. Стандарт представления числа с плавающей точкой

Стандарт представления чисел с плавающей точкой был разработан в 1985 году в Institute of Electrical and Electronics Engineers и носит название IEEE-арифметики. В настоящее время общепринятой считается новая вер-

сия стандарта 2019 года. В этом стандарте основной формой действительного числа является нормализованное представление одинарной и двойной точности, стандарт предполагает возможность представления субнормальных чисел для возможности корректного округления при математических операциях. Кроме того, в арифметике есть специальные величины бесконечность (Infinity) и так называемое «не число» (NaN) — от английского Not a number или Non-arithmetical number.

В нормализованном виде число представляется в виде ненулевого старшего разряда, мантиссы после запятой и степени экспоненты. Например, число 176.243 в нормализованном виде будет представлено как  $1.76243 \cdot 10^2$ .

Основными типами представления чисел являются данные одинарной и двойной точности. Для представления данных одинарной точности отводится 32 бита, тогда один бит отводится под знак числа  $s$ , 23 бита отводится под мантиссу и 8 бит под показатель:

$s$	$f$	$e$
1	23	8

Старший ненулевой разряд нормализованного числа не хранится, поэтому по записи такого вида число в двоичном представлении восстанавливается как  $(-1)^s (1 + f \cdot 2^{-23}) 2^{(e - 127)}$ . Сдвиг экспоненты делается для того, чтобы не хранить еще и знак степени. С представлением чисел связаны три константы, важные в вычислительной математике:

OFL (Over Flow Limit) — порог переполнения, который есть максимальное представимое число, так что любое большее число полагается равным бесконечности. Для одинарной точности

$$\text{OFL} = (1,111\dots1_2) \cdot 2^{127} = (2 - 2^{-23}) \cdot 2^{127} \approx 10^{38}.$$

UFL (Under Flow Limit) — порог машинного нуля, — это нормализованное число, такое, что любое меньшее число полагается равным нулю:  $\text{UFL} = 1 \cdot 2^{-126} \approx 10^{-38}$ .

Машинное эпсилон определяется как максимальное число, которое в машинной арифметике обеспечивает справедливость равенства  $1 + \varepsilon_{\text{маш}} = 1$ . Для одинарной точности задания чисел

$$\varepsilon_{\text{маш}} = \frac{1}{2} \cdot 2^{-23} \approx 6 \cdot 10^{-8}.$$

Погрешность округления не затрагивает разрядную часть числа, поэтому погрешность округления числа с модулем  $|a|$  не превосходит  $|a| \varepsilon_{\text{маш}}$ .

Аналогично, для числа двойной точности имеем:

$s$	$f$	$e$
1	52	11

Число в двоичном представлении восстанавливается как

$$(-1)^s (1 + f \cdot 2^{-52}) 2^{(e - 1023)}.$$

Соответствующие константы

$$\text{OFL} = (1,111\dots1_2) \cdot 2^{1023} \approx 2^{1024} \approx 10^{308},$$

$$\text{UFL} = 1 \cdot 2^{-1022} \approx 10^{-308},$$

$$\varepsilon_{\text{маш}} = \frac{1}{2} \cdot 2^{-52} \approx 10^{-16}.$$

Присутствие в арифметике субнормальных чисел позволяет реализовать арифметику с правильным округлением при математических операциях. Тогда минимальное субнормальное число в арифметике одинарной точности равно  $2^{-23} \cdot 2^{-126} \approx 10^{-45}$ , а в арифметике двойной точности  $2^{-52} \cdot 2^{-1022} \approx 5 \cdot 10^{-324}$ .

В настоящее время стандарт IEEE-арифметики реализован на большинстве компьютеров.

## I.6. Задачи на доказательства

**I.6.1.** Показать, что предельная абсолютная погрешность<sup>1</sup> суммы или разности равна сумме предельных абсолютных погрешностей с точностью до членов второго порядка малости.

**I.6.2.** Показать, что предельная относительная погрешность<sup>2</sup> произведения или частного равна сумме предельных относительных погрешностей с точностью до членов второго порядка малости.

**I.6.3.** Пусть  $y^*$  — приближение к корню уравнения  $f(y) = 0$ . Вывести приближенное равенство

$$y - y^* \approx -\frac{f(y^*)}{f'(y^*)}.$$

**I.6.4.** Как известно, для вычисления функции  $\ln x$  можно использовать следующий ряд по  $x$ :

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots + (-1)^{k+1} \frac{x^k}{k} + \dots \quad (\text{a})$$

Можно представить  $1+x$  в виде  $1+x = 2^m \cdot z$ , где  $0.5 \leq z \leq 1$ , положив

$$y = \frac{1-z}{1+z},$$

---

<sup>1</sup> Под предельной абсолютной погрешностью понимается верхняя грань (супремум) абсолютной погрешности приближения числа.

<sup>2</sup> Под предельной относительной погрешностью понимается верхняя грань (супремум) относительной погрешности приближения числа.

для представления логарифма получаем ряд

$$\ln x = m \ln 2 - 2 \left( y + \frac{y^3}{3} + \dots + \frac{y^{2k-1}}{2k-1} + \dots \right). \quad (6)$$

В чем преимущества и недостатки использования ряда (б)? Как оценить погрешность метода при использовании каждого из этих разложений?

**I.6.5.** Какова относительная погрешность округления при представлении действительного числа в ЭВМ, если под хранение мантиссы отводится  $p$  бит?

**I.6.6.** Пусть функция  $f(x)$  задана таблично: заданы значения аргументов  $x_0 < x_1 < x_2 < \dots < x_N$  (расстояние между двумя соседними точками  $h$ ) и значения функции в них  $f_0, f_1, \dots, f_N$ .

Самостоятельно выведите формулу вычисления односторонней производной для приближенного вычисления  $f'(x)$  в точках  $x_0$  и  $x_N$  с точностью до  $O(h^2)$  и  $O(h^3)$ . Найдите оптимальные шаги численного дифференцирования. Сравните их с оценками для центральных разностей.

Указание. Для вывода формул используйте метод неопределенных коэффициентов, а именно, равенство

$$f'(x_0) \approx \frac{\alpha_0 f(x_0) + \alpha_1 f(x_1) + \alpha_2 f(x_2)}{h}.$$

Подберите  $\alpha_0$ ,  $\alpha_1$  и  $\alpha_2$  так, чтобы равенство выполнялось с точностью до  $O(h^2)$ . Сколько членов нужно взять, чтобы получить формулу третьего порядка аппроксимации?

**I.6.7.** Вторая и третья производные функции вычисляются по приближенным формулам:

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

и

$$f'''(x) = \frac{f(x+2h) - 2f(x+h) + 2f(x-h) - f(x-2h)}{2h^3}.$$

Найдите погрешность метода и неустранимую погрешность при вычислениях по этим формулам. Найдите оптимальные шаги численного дифференцирования и минимально возможную ошибку.

## I.7. Примеры решения задач

**I.7.1.** Найти абсолютную предельную погрешность, погрешность по производной, линейную погрешность для функции  $u = t^{10}$ , если заданы точка приближения  $t^* = 1$ , значение функции  $u^*$  в этой точке и погрешность  $\Delta t^* = 10^{-1}$ .

Решение. Обозначим

$$b = \sup_{|t-1| \leq 0.1} |u'_t(t)| = \sup_{|t-1| \leq 0.1} 10 \cdot t^9 = 10 \cdot (1.1)^9 \approx 23.58.$$

Абсолютная предельная погрешность может быть определена как  $D(u^*) = \sup_{|t-1| \leq 0.1} |t^{10} - 1| = (1.1)^{10} - 1 \approx 1.594$ .

Оценка погрешности  $u$  при вычислении значения функции по максимуму производной и линейная оценка соответственно будут:  $D_1(u^*) = b\Delta(t^*) = 2.358\dots; D_2(u^*) = |10 \cdot 1^9| \Delta(t^*) = 10 \cdot 0.1 = 1$ .

**I.7.2.** Дать линейную оценку погрешности при вычислении неявной функции  $\phi(u, t_1, t_2, \dots, t_n) = 0$ , если известны точка приближения  $\{t_1^*, \dots, t_n^*\}$ , значение функции в точке приближения  $u^*$  и погрешность в определении аргументов  $\Delta t_1^*, \dots, \Delta t_n^*$ .

Решение. Дифференцируя по  $t_j$ , получим

$$\frac{\partial \phi}{\partial u} \frac{\partial u}{\partial t_j} + \frac{\partial \phi}{\partial t_j} = 0,$$

откуда

$$\frac{\partial u}{\partial t_j} = - \left( \frac{\partial \phi}{\partial t_j} \right) \left( \frac{\partial \phi}{\partial u} \right)^{-1}.$$

При заданных  $\{t_1^*, \dots, t_n^*\}$  можно найти  $u^*$  как корень уравнения  $\phi(u, t_1, t_2, \dots, t_n) = 0$ , а затем — значения  $b_j(0) = - \left( \frac{\partial \phi}{\partial t_j} \right) \left( \frac{\partial \phi}{\partial u} \right)^{-1} \Big|_{(u^*, t_1^*, \dots, t_n^*)}$ , от-

куда можно получить линейную оценку погрешности функции  $D_2(u^*)$ .

**I.7.3.** Вычислить относительную погрешность в определении значения функции  $u = xy^2z^3$   $u = xy^2z^3$ , если  $x^* = 37.1$ ,  $y^* = 9.87$ ,  $z^* = 6.052$ ,  $\Delta x^* = 0.3$ ,  $\Delta y^* = 0.11$ ,  $\Delta z^* = 0.016$ .

Решение:

$$\delta = \frac{0.3}{37.1} \approx 0.81 \cdot 10^{-2}, \quad \delta_y = \frac{0.11}{9.87} \approx 1.12 \cdot 10^{-2}, \quad \delta_z = \frac{0.016}{6.052} \approx 0.26 \cdot 10^{-2},$$

$$\delta(u) = \delta(x^*) + 2\delta(y^*) + 3\delta(z^*) = 3.8 \cdot 10^{-2}.$$

**I.7.4.** Оценить погрешность в определении корней квадратного уравнения  $\varphi(u, t_1, t_2) = u^2 + t_1 u + t_2 = 0$ , если заданы приближения  $t_1^*, t_2^*, \Delta(t_1^*), \Delta(t_2^*)$ .

Решение:

Пусть  $u^*$  — решение уравнения

$$u^{*2} + t_1^* u^* + t_2^* = 0.$$

Воспользовавшись формулой из задачи 7.2

$$b_j(0) = - \left( \frac{d\varphi}{dt_j} \right) \left( \frac{d\varphi}{du} \right)^{-1} \Bigg|_{(u^*, t_1^*, \dots, t_n^*)},$$

получим

$$b_1(0) = \frac{du}{dt_1} \Bigg|_{(t_1^*, t_2^*)} = -\frac{u^*}{2u^* + t_1^*}, \quad b_2(0) = \frac{du}{dt_2} \Bigg|_{(t_1^*, t_2^*)} = -\frac{1}{2u^* + t_1^*}.$$

Следовательно, линейная оценка будет

$$D_2(u^*) = \frac{|u^*| \cdot \Delta(t_1^*) + \Delta(t_2^*)}{|2u^* + t_1^*|}.$$

## I.8. Теоретические задачи

**I.8.1.** Пусть  $y^*$  — корень кратности  $k$  уравнения  $y^2 + by + c = 0$  при заданных приближенных значениях коэффициентов  $b^*, c^*$  и их погрешностях  $\Delta_{b^*}$  и  $\Delta_{c^*}$ . Показать, что погрешность приближенного значения корня имеет порядок  $O(\rho^{1/k})$ , где  $\rho = \sqrt{\Delta_{b^*}^2 + \Delta_{c^*}^2}$ .

**I.8.2.** С каким числом знаков надо взять  $\lg 2$ , для того, чтобы вычислить корни уравнения  $x^2 - 2x + \lg 2 = 0$  с четырьмя верными знаками?

**I.8.3.** Найти абсолютную предельную погрешность числа  $a = 3.14$ , приближающего число  $\pi$ .

**I.8.4.** Найти абсолютную предельную погрешность, погрешность по производной и линейную оценку погрешности для функций

$$u = \sin t, u = 1/(t^2 - 5t + 6).$$

Заданы точка приближения  $t = t^*$  и погрешность  $\Delta t$ .

**I.8.5.** Найти погрешность по производной для функции  $u = \sqrt{t}$ , если заданы точка приближения  $t^* = 4$ , значение функции  $u^*$  в этой точке и погрешность  $\Delta t^* = 0.1$ .

**I.8.6.** Найти линейную оценку погрешности для функции  $u = t^5$ , если заданы точка приближения  $t^* = \sqrt{2}$ , значение функции  $u^*$  в этой точке и погрешность  $\Delta t^* = 0.1$ .

**I.8.7.** Каждое ребро куба, измеренное с точностью до 0.02 см, оказалось равным 8 см. Найти абсолютную и относительную погрешность при вычислении объема куба.

**I.8.8.** Стороны прямоугольника  $a \approx 5$  м и  $b \approx 6$  м. Какова допустимая предельная абсолютная погрешность при измерении этих сторон (одинаковая для обеих сторон), чтобы площадь  $S$  прямоугольника можно было определить с предельной абсолютной погрешностью  $\Delta(S) = 1\text{m}^2$ .

**I.8.9.** Найти абсолютную предельную погрешность для функции  $u = \sin t$ , если заданы точка приближения  $t^* = \pi/4$ , значение функции  $u^*$  в этой точке и погрешность  $\Delta t^* = 0.05$ .

**I.8.10.** Найти погрешность по производной для функции  $u = t^2$ , если заданы точка приближения  $t^* = 2$ , значение функции  $u^*$  в этой точке и погрешность  $\Delta t^* = 0.1$ .

**I.8.11.** Найти линейную оценку погрешности для функции  $u = \ln t$ , если заданы точка приближения  $t^* = 1$ , значение функции  $u^*$  в этой точке и погрешность  $\Delta t^* = 0.1$ .

**I.8.12.** Вычислить относительную погрешность в определении значения функции  $u = xy^2$ , если  $x^* = 9.87$ ,  $y^* = 37.1$ ,  $\Delta x^* = 0.11$ ,  $\Delta y^* = 0.1$ .

**I.8.13.** Вычислить относительную погрешность в определении значения функции  $u(x, y, z) = x^2y^2/z^4$ , если заданы  $x^* = 37.1$ ,  $y^* = 9.87$ ,  $z^* = 6.052$ ,  $\Delta(x^*) = 0.1$ ,  $\Delta(y^*) = 0.05$ ,  $\Delta(z^*) = 0.02$ .

**I.8.14.** Радиус круга равен 1 м. С какой точностью его надо измерить, чтобы погрешность площади круга была не больше 1 см<sup>2</sup>?

**I.8.15.** Величина  $y$  вычисляется по формуле  $y = f(x)$ , а величина  $x$  получается прямым измерением, которое осуществляется с погрешностью, не превосходящей некоторое заданное число  $\Delta_x$ .

Требуется найти наименьшее число  $\Delta_y$ , при котором для данного  $x^*$ , полученного в результате приближенного измерения величины  $x$ , справедлива оценка

$$|y^* - y| < \Delta_y; \quad y^* = f(x^*); \quad y = f(x).$$

Указать факторы, от которых зависит точность приближенной формулы  $\Delta_y = f'(x^*) \Delta_x$  для  $\Delta_y$ :

a)  $f(x) = \sin x$ ; б)  $f(x) = \ln x$  в)  $f(x) = 1 / (x^2 - 5x + 6)$ .

**I.8.16.** Пусть  $z = f(x, y)$ , причем величина  $x^*$  получается в результате приближенных измерений с неустранимой погрешностью  $\Delta_x = 10^{-3}$ .

С какой разумной точностью следует измерять  $y$  в предположении, что вклад от погрешности  $x$  и  $y$  в величину  $z$  примерно одинаков?

a)  $z = x + 10y$ ; б)  $z = xy + xy^2$ ; в)  $z = x/y$ .

Пусть при вычислении  $z$  нас интересует абсолютная погрешность в случае а) и относительная погрешность в б) и в).

**I.8.17.** Рассмотрим модель представления чисел в IEEE-арифметике следующего вида:

$S = \{\pm b_0, b_1 b_2 \cdot 2^{\pm a}\}$ , где числа  $a, b_1, b_2 \in \{0, 1\}$ , а число  $b_0 = 1$  всегда, кроме того случая, когда  $a = b_1 = b_2 = 0$ , в этом случае  $b_0 \in \{0, 1\}$ .

- а) Изобразить множество  $S$  на действительной оси. Сколько чисел в данной модели арифметики у Вас получилось?  
б) Чему равны машинные константы  $\varepsilon_{\text{маш}}$ , UFC, OFL в этой модели?

**I.8.18.** Рассмотрим модель представления чисел в IEEE-арифметике следующего вида:

$S = \{\pm b_0, b_1 b_2 b_3 \cdot 2^{\pm a}\}$ , где числа  $a, b_1, b_2, b_3 \in \{0, 1\}$ , а число  $b_0 = 1$  всегда, кроме того случая, когда  $a = b_1 = b_2 = b_3 = 0$ , в этом случае  $b_0 \in \{0, 1\}$ .

- а) Построить множество  $S$  на действительной оси. Сколько чисел в данной модели арифметики у Вас получилось?  
б) Чему равны машинные константы  $\varepsilon_{\text{маш}}$ , UFC, OFL в этой модели?

**I.8.19.** Пусть для вычисления функции  $u = f(t)$  используется частичная сумма ряда Маклорена  $u(t) \approx u(0) + \frac{u'(0)}{1!}t + \dots + \frac{u^{(n)}(0)}{n!}t^n$ , причем аргумент задан с погрешностью  $\Delta t = 10^{-3}$ .

Найти  $n$  такое, чтобы погрешность в определении функции  $u(t)$  по данной формуле не превышала  $\Delta t$ . Рассмотреть отрезки  $t \in [0, 1]$  и  $t \in [10, 11]$ . Предложить более совершенный алгоритм для вычисления функций  $u(t) = \sin t$ ,  $u(t) = e^t$  на отрезке  $t \in [10, 11]$ .

**I.8.20.** Пусть неустранимая погрешность при измерениях  $x$  не превосходит  $\Delta_x = 10^{-3}$ . Для вычисления заданной функции  $y = f(x)$  используется частичная сумма ряда Маклорена:

$$y \approx f(0) + \frac{f'(0)}{1!}x + \dots + \frac{f^{(n)}(0)}{n!}x^n.$$

а) Как выбрать  $n$ , чтобы погрешность аппроксимации функции  $f(x)$  отрезком ряда Маклорена не превосходила неустранимую погрешность? Рассмотреть функцию  $f(x) = \sin x$  на отрезке  $0 < x < 1$  и  $10 < x < 11$ .

Указание. Формула Стирлинга, приближающая факториал, дает хорошее приближение при  $n \geq 10$ ,  $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$ .

б) Каковы требования к относительным погрешностям округления слагаемых  $f^{(k)}(0) \cdot x^k / k!$ , чтобы абсолютная погрешность их вычисления не превосходила неустранимую погрешность при вычислении  $\sin x$ . Рассмотреть случаи  $0 \leq x \leq 1$  и  $10 \leq x \leq 11$ .

в) Не можете ли Вы предложить для вычисления  $\sin x$  на  $10 \leq x \leq 11$  более совершенную процедуру, чем задаваемую рядом Маклорена?

**I.8.21.** Оценить погрешность в определении корней уравнения  $ay^3 + d = 0$ , если величины  $a = 1$  и  $d = 8$  заданы с точностью  $\Delta(a) = 10^{-3}$  и  $\Delta(d) = 10^{-3}$ .

**I.8.22.** Оценить погрешность в определении вещественных корней уравнения  $ay^3 + cy + d = 0$ , если величины  $a = 1$ ,  $c = 2$  и  $d = 3$  заданы с точностью  $\Delta(a) = 10^{-3}$  и  $\Delta(c) = 10^{-3}$   $\Delta(d) = 10^{-3}$ .

**I.8.23.** Оценить погрешность в определении вещественных корней уравнения  $ay^3 + by^2 + d = 0$ , если величины  $a = 1$ ,  $b = 2$  и  $d = -3$  заданы с точностью  $\Delta(a) = 10^{-2}$  и  $\Delta(b) = 10^{-2}$   $\Delta(d) = 10^{-2}$ .

**I.8.24.** Оценить погрешность в определении положительного корня уравнения  $ay^3 + cy = 0$ , если величины  $a = 1$  и  $c = -4$  заданы с точностью  $\Delta(a) = 10^{-2}$  и  $\Delta(c) = 10^{-2}$ .

**I.8.25.** Пусть требуется вычислить производную функции  $f(x)$  в некоторой точке  $x$ . Причем известно, что  $|f''(x)| \leq 1$  при всех  $x$ .

Используется приближенная формула

$$f'(x) \approx \frac{f^*(x+h) - f^*(x)}{h},$$

где  $f^*(x)$  — приближенные значения функции  $f(x)$ , полученные в результате измерений с погрешностью, не превосходящей  $10^{-4}$ .

Какова наибольшая точность, с которой можно вычислить  $f'(x)$  по указанной формуле? Указать оптимальный выбор шага  $h$ .

**I.8.26.** Определить шаг  $\tau$ , при котором погрешность вычисления производной  $u'(t)$ , приближенно вычисляемой в соответствии с формулами

$$u'(t) \approx \frac{f(t+\tau) - f(t)}{\tau},$$

$$u'(t) \approx \frac{f(t+\tau) - f(t-\tau)}{2\tau},$$

не превосходит  $10^{-3}$ . Известно, что  $|u''(t)| \leq 1$ ,  $|u'''(t)| \leq 1$  для любых  $t$ .

**I.8.27.** (В.Б. Пирогов) В приведенной ниже таблице представлены значения функции  $f(x)$  с шагом  $h = 0.002$ , вычисленные на компьютере. Пусть известно, что  $\max |f^{(2)}(x)| \leq M_2 = 1$  и  $\max |f^{(3)}(x)| \leq M_3 = 1$ . Вычислить максимально точно значение первой производной функции  $f(x)$  в точке  $x = (i-1)*h$ . Дать оценку погрешности полученного результата при заданном  $i$

$x$	0	0.002	0.004	0.006	0.008	0.01	0.012	0.014	0.016	0.018
$f(x)$	1.000E01	1.000E01	1.0000E01	1.000E01	1.000E01	1.0000E01	0.9999	0.9999	0.9999	0.9998

а)  $i = 3$ , б)  $i = 5$ , в)  $i = 7$ .

**I.8.28.** Табличная функция  $\{f_n\}$  есть проекция на равномерную сетку с шагом  $h$  бесконечно дифференцируемой функции  $f(x)$ :  $f_n = f(x_0 + nh)$ . Используется приближенный метод вычисления первой производной:

$$f'(x_0) \approx \frac{-11f_0 + 18f_1 - 9f_2 + 2f_3}{6h}.$$

Каков порядок аппроксимации этой формулы? Указать оптимальный шаг численного дифференцирования и минимальную погрешность, с которой может быть найдено значение производной.

**I.8.29.** Табличная функция  $\{f_n\}$  есть проекция на равномерную стеку с шагом  $h$  бесконечно дифференцируемой функции  $f(x)$ :  $f_n = f(x_0 + nh)$ .

Используется приближенный метод вычисления первой производной:

$$f'(x_2) \approx \frac{f_0 - 6f_1 + 3f_2 + 2f_3}{6h}.$$

Каков порядок аппроксимации этой формулы? Указать оптимальный шаг численного дифференцирования и максимальную точность, с которой может быть найдено значение производной.

**I.8.30.** Табличная функция  $\{f_n\}$  есть проекция на равномерную стеку с шагом  $h$  функции  $f(x)$ . Известно, что  $|f'''(x)| \leq 1$ . Построить формулу для приближенного вычисления  $f'(x_0)$  со вторым порядком аппроксимации. Оценить погрешность метода. Найти оптимальный шаг численного дифференцирования.

**I.8.31.** Даны неравномерная сетка  $h_i = x_{i+1} - x_i$ ,  $h_{i-1} = x_i - x_{i-1}$ . Получить формулу для приближенного вычисления второй производной в точке  $x_i$ , оценить главный член погрешности аппроксимации. Найти оптимальный шаг численного дифференцирования при условии  $h_{i-1} = 2h_i$ . Какие требования необходимо наложить на гладкость функции?

**I.8.32.** Пусть функция  $f(x)$  задана в точках  $f(x + kh)$ ,  $k = 1, 2, 3$ .

Получить формулу для вычисления первой производной  $f'(x)$  функции в точке  $x$  с максимально высокой точностью. С какой максимальной точностью можно вычислить первую производную по этой формуле, если функция в точках задана с абсолютной погрешностью  $\varepsilon$ ? Считать, что необходимая для оценки производная не превышает по модулю 1.

**I.8.33.** Для вычисления первой производной функции  $f(x)$  в точке  $x + h$  используется формула  $(f(x+2h) - f(x-2h))/(4h)$ .

- Каков порядок аппроксимации этой формулы?
- Найти оптимальный шаг дифференцирования по этой формуле в произвольной точке для четырежды дифференцируемой функции.
- Оценить его численное значение для функции  $f(x) = \cos(x + \pi/4)$  в случае использования IEEE-арифметики одинарной и двойной точности.

**I.8.34.** Для вычисления первой производной функции  $f(x)$  в точке  $x - h$  используется формула  $(f(x+2h) - f(x-2h))/(4h)$ .

- Каков порядок аппроксимации этой формулы?
- Найти оптимальный шаг дифференцирования по этой формуле в произвольной точке для четырежды дифференцируемой функции.
- Оценить его численное значение для функции  $f(x) = \cos(x + 2\pi/3)$  в случае использования IEEE-арифметики одинарной и двойной точности.

**I.8.35.** Для вычисления второй производной функции  $f(x)$  в точке  $x + h$  используется формула  $(f(x+2h) - 2f(x) + f(x-2h)) / (4h^2)$ .

- Каков порядок аппроксимации этой формулы?
- Найти оптимальный шаг дифференцирования по этой формуле в произвольной точке для четырежды дифференцируемой функции.
- Оценить его численное значение для функции  $f(x) = \cos(x - \pi/6)$  в случае использования IEEE-арифметики одинарной и двойной точности.

**I.8.36.** Для вычисления второй производной функции  $f(x)$  в точке  $x - h$  используется формула  $(f(x+2h) - 2f(x) + f(x-2h)) / (4h^2)$ .

Выполнить пункты а), б) предыдущей задачи.

- Оценить его численное значение для функции  $f(x) = \cos(x - 7\pi/8)$  в случае использования IEEE-арифметики одинарной и двойной точности.

**I.8.37.** Пусть функция  $f(x)$  задана в точках  $f(x + kh)$ ,

- $k = 0, -1, -2$ . Получить формулу для вычисления второй производной  $f^{(2)}(x)$  функции в точке  $x$  с максимально высокой точностью.
- $k = -1, 0, +1$ . Получить формулу для вычисления первой производной  $f^{(1)}(x)$  функции в точке  $x$  с максимально высокой точностью.
- $k = 0, 1, 2, 3$ . Получить формулу для вычисления первой производной  $f^{(1)}(x)$  функции в точке  $x$  с максимально высокой точностью.
- $k = -1, 0, 1, 2$ . Получить формулу для вычисления первой производной  $f^{(1)}(x)$  функции в точке  $x$  с максимально высокой точностью.

С какой максимальной точностью можно вычислить требуемую производную по полученной формуле, если функция в точках задана с абсолютной погрешностью  $\varepsilon$ ? Считать, что необходимая для оценки производная не превышает по модулю 1.

**I.8.38.** Для функции, заданной таблично

$x$	1	2	3	5	7
$f$	0.5	0.25	0.25	0.2	0.1

вычислить значение первой производной с максимально возможной точностью. Воспользоваться методом неопределенных коэффициентов а) в точке  $x = 3$ , б) в точке  $x = 7$ .

**I.8.39.** Для функции, заданной таблично на отрезке, вычислить вторую производную со вторым порядком точности в точке  $x = 0$ , если известно, что на левой границе  $f^{(3)}(0) = 5$ ,

$x$	0	2	4
$f$	2	8	3

Пусть в двух произвольных точках функция задана с относительной погрешностью  $10^{-4}$ . Во всех остальных точках функция задана точно. Оценить ошибку округления при вычислении производной. Указать оптимальный шаг численного дифференцирования для формулы первого порядка точности в условиях задачи.

**I.8.40.** Для функции, заданной таблично на отрезке  $[1, 3]$ , вычислить ее первую производную с третьим порядком точности в точке  $x = 3$ , если известно, что на правой границе  $f^{(3)}(3) = 4$ ,

$x$	1	2	3
$f$	-3	-6	2

Пусть в двух произвольных точках функция задана с относительной погрешностью  $10^{-4}$ . Во всех остальных точках функция задана точно. Оценить ошибку округления при вычислении производной. Указать оптимальный шаг численного дифференцирования для формулы второго порядка в условиях задачи.

## I.9. Практические задачи

**I.9.1.** Написать программу для вычисления  $\exp(x)$ , пользуясь рядом Маклорена и конечностью разрядов машинной арифметики: ввести величину  $SUM = 1.$ , в цикле по  $I$  вычислять  $TERM = TERM * X / I$ , и если  $SUM + TERM$  равен  $SUM$ , то закончить вычисления и напечатать результат, а если не равен, то  $SUM = SUM + TERM$  и выполнять цикл далее. Вычислить и сравнить  $SUM$  и экспоненту от  $x$  для следующих аргументов:

$$x \in \{1, 5, 10, 15, 20, 25, -1, -5, -10, -15, -20, -25\}$$

при вычислениях с одинарной точностью. Объяснить результат. Предложить усовершенствованную процедуру для вычисления экспоненты отрицательного аргумента.

**I.9.2.** Написать программу для вычисления многочлена  $p(x) = \sum_{j=0}^N a_j x^j$ ,

пользуясь схемой Горнера:

$$p = a_N // \text{for } j = N - 1 \text{ to } 0 \text{ do } p = x \cdot p + a_j // \text{end for} // \text{write } x, p$$

для многочлена  $p(x) = (x - 2)^9$  на интервале  $[1.92, 2.08]$  с шагом  $10^{-4}$ . Объяснить полученный результат. Сравнить его с вычислением по формуле  $p(x) = (x - 2)^9$ . Почему алгоритм вычисления многочлена по схеме Горнера непригоден для численного определения нуля функций?

**I.9.3.** Вычислить постоянную Эйлера  $C = \lim_{n \rightarrow \infty} \left( \sum_{k=0}^n \frac{1}{k} - \ln n \right)$  с точностью  $10^{-12}$ .

Показать, что при вычислении частичной суммы гармонического ряда путем добавления очередного слагаемого в арифметике конечной разрядности эти суммы будут иметь предел. Оценить его для арифметики с двойной точностью.

Константа Эйлера вычисляется как предел разности двух больших чисел. В чем недостатки такого подхода? Преобразовать формулу для константы Эйлера, оценивая разность между суммами с  $N$  и  $N + 1$  слагаемыми.

**I.9.4.** а) Применить алгоритм Архимеда для нахождения числа  $\pi$  как предела последовательности периметров правильных  $2n$ -угольников, вписанных в окружность. Существует рекуррентная связь между периметрами двух последовательных многоугольников из этого класса:

$$p_{n+1} = 2^n \sqrt{2 \left( 1 - \sqrt{1 - \left( p_n / 2^n \right)^2} \right)}.$$

Вычислить значение  $p_n$  для значений  $n = 3, 4, \dots, 60$ . Попытайтесь объяснить результаты.

б) Формулу для вычисления  $p_n$  из предыдущего пункта задачи можно улучшить, чтобы устраниТЬ из нее вычитание. Запишем  $p_{n+1}$  в виде  $p_{n+1} = 2^n \sqrt{r_{n+1}}$ , где  $r_{n+1} = 2 \left( 1 - \sqrt{1 - \left( p_n / 2^n \right)^2} \right)$ ,  $r_3 = 2 / (2 + \sqrt{2})$ . Покажите, что  $r_{n+1} = r_n / (2 + \sqrt{4 - r_n})$ .

Полученную итерационную формулу используйте для вычисления  $p_n$  и  $r_n$  для значений  $n = 3, 4, \dots, 60$ . В конечном счете разность  $4 - r_n$  будет округляться до значения 4. Таким образом, последняя формула также подвержена влиянию ошибок округления при больших значениях  $n$ . Однако есть ли теперь основания для беспокойства?

**I.9.5.** Одним из классических методов решения кубического уравнения является формула Кардано. Кубическое уравнение  $x^3 + a x^2 + b x + c = 0$  заменой  $x = y - a / 3$  преобразуется к виду  $y^3 + p y + q = 0$ , в котором коэффициенты  $p = b - a^3 / 3$ ,  $q = c - ab/3 + 2(a/3)^3$ . Один вещественный корень исходного уравнения определяется следующим образом:  $s = \sqrt{\left( p / 3 \right)^3 + \left( q / 2 \right)^2}$ ,  $y_1 = \sqrt[3]{s - q / 2} + \sqrt[3]{-s - q / 2}$ , и тогда вещественный корень исходного уравнения есть величина  $x_1 = y_1 - a / 3$ . Два других корня можно найти, разделив исходное уравнение на  $x - x_1$  и решив получившееся квадратное уравнение.

а) Воспользуйтесь методом Кардано для вещественного корня уравнения  $x^3 + 3x^2 + \alpha^2 x + 3\alpha^2 = 0$  при различных значениях  $\alpha$ . Исследуйте потерю точности из-за округления при  $\alpha$  порядка величины, обратной к  $\epsilon_{\text{маш}}$ .

б) Можно ли модифицировать формулу Кардано, чтобы избежать потерю точности при больших значениях параметра  $\alpha$ ?

в) Для решения нелинейного уравнения можно использовать итерационный метод Ньютона (см. раздел IV, в котором изложены условия и скорость сходимости метода). Пусть задано начальное приближение  $x^{(0)}$  к решению уравнения  $f(x) = 0$ . Следующее приближение к решению находится

по формуле  $x^{s+1} = x^s - \frac{f(x^s)}{f'(x^s)}$ . Примените метод Ньютона к уравнению с

теми же значениями параметра  $\alpha$ . Исследуйте эффект ошибок округления и выбора начального значения.

**I.9.6\***. Написать и полностью оттестировать программу, вычисляющую евклидову норму вектора по заданным компонентам. Наиболее очевидная и неудовлетворительная (почему?) реализация на псевдокоде выглядит так:

```
G = 0  
FOR I = 1 TO N  
    G = G + XI2  
ENDFOR  
G = SQRT(G)
```

Алгоритм вычисления евклидовой нормы вектора должен обладать совокупностью следующих желательных свойств:

1) Результат должен вычисляться с высокой точностью, т.е. почти все разряды результата должны быть верными, если  $\|x\|$  не находится (почти) за пределами области нормализованных чисел с плавающей точкой.

2) Алгоритм должен быть в большинстве случаев почти же быстр, что и приведенная выше программа.

3) Алгоритм должен работать на любой разумной машине, включая, возможно, и те, арифметика которых отлична от IEEE-арифметики. Это означает, что работа алгоритма не может приводить к останову, если  $\|x\|$  не (почти) превосходит наибольшего числа с плавающей точкой.

Вероятно, вы не сможете одинаково успешно удовлетворить всем выдвинутым требованиям. Здесь есть пространство маневра: полнее удовлетворить одним требованиям за счет ослабления каких-то других.

## I.10. Библиографический комментарий

Изложение элементарной теории погрешностей в данном пособии следует книгам [2, 5, 9]. Представление машинных чисел подробно описано в действующем стандарте [10].

## II. Элементы прикладной линейной алгебры

### II.1 Введение

Одна из самых важных и хорошо разработанных областей вычислительной математики – вычислительная линейная алгебра. В нее входят такие традиционные разделы, как методы решения систем линейных алгебраических уравнений (СЛАУ), методы поиска собственных чисел матриц и собственных векторов, задачи повышения эффективности матричных операций (например, быстрое перемножение матриц чрезвычайно большого размера), алгоритмы работы с матрицами специального вида (например, с ленточными матрицами, разреженными матрицами).

Первый раздел коснется основных методов и идей прикладной линейной алгебры. Рассматриваются простейшие прямые и итерационные методы решения СЛАУ. К численному решению систем линейных алгебраических уравнений сводятся многие задачи математической физики. Математические модели, представляющие собой СЛАУ большой размерности, встречаются в математической экономике, биологии и т. п.

К другим методам линейной алгебры мы вернемся при рассмотрении методов решения сеточных уравнений, возникающих при аппроксимации разностными методами дифференциальных уравнений в частных производных эллиптического типа.

По прикладной линейной алгебре существует обширная литература, а программы, реализующие наиболее популярные алгоритмы вычислительной линейной алгебры, являются неотъемлемой частью прикладного программного обеспечения, в частности, современных математических пакетов.

### II.1. Некоторые сведения о векторных пространствах

**Норма.** Будем ставить в соответствие каждому элементу  $n$ -мерного векторного пространства  $\mathbf{A}$  неотрицательное число  $m(\mathbf{A})$ , называемое нормой. Оно должно удовлетворять следующим трем свойствам (аксиомам нормы).

$$1. \ m(\mathbf{A}) = 0 \Leftrightarrow \mathbf{A} = 0 .$$

Если первая аксиома не выполняется и ноль может соответствовать и ненулевому элементу, то это число — полунорма.

2. Для любого скалярного множителя  $\alpha$  выполнено  $m(\alpha\mathbf{A}) = |\alpha|m(\mathbf{A})$ .

3.  $m(\mathbf{A} + \mathbf{B}) \leq m(\mathbf{A}) + m(\mathbf{B})$  (неравенство треугольника).

Ниже мы увидим, что норму в векторном пространстве можно задать неединственным способом.

### П.2.1. Согласованные и подчиненные нормы векторов и матриц

В векторном  $n$ -мерном линейном нормированном пространстве введем следующие нормы вектора:

$$\text{кубическая: } \|\mathbf{u}\|_1 = \max_{1 \leq i \leq n} |u_i|, \quad (2.1a)$$

$$\text{октаэдрическая: } \|\mathbf{u}\|_2 = \sqrt{\sum_{i=1}^n |u_i|^2}, \quad (2.1b)$$

евклидова (в комплексном случае — эрмитова)

$$\|\mathbf{u}\|_3 = \sqrt{\sum_{i=1}^n |u_i|^2} = \sqrt{(\mathbf{u}, \mathbf{u})}. \quad (2.1b)$$

Такое обозначение соответствует традициям научных школ, сформировавшихся в МФТИ. Такие обозначения приняты ниже во всех задачах.

Рассмотрим квадратную матрицу  $\mathbf{A}$  и связанное с ней линейное преобразование  $\mathbf{v} = \mathbf{A}\mathbf{u}$  где  $\mathbf{v}, \mathbf{u} \in \mathbb{R}^N$  ( $\mathbb{R}^N$  —  $N$ -мерное линейное нормированное пространство). Норма матрицы определяется как действительное неотрицательное число, характеризующее это преобразование:

$$\|\mathbf{A}\| = \sup_{\|\mathbf{u}\| \neq 0} \frac{\|\mathbf{A}\mathbf{u}\|}{\|\mathbf{u}\|}. \quad (2.2)$$

Введенную таким образом норму матрицы называют *подчиненной* соответствующей норме вектора (или *операторной*). Конкретный вид нормы матрицы в этом случае зависит от выбранной нормы вектора. Укажем некоторые свойства нормы матрицы:

$$\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|,$$

$$\|\lambda\mathbf{A}\| = |\lambda| \|\mathbf{A}\|,$$

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|,$$

$$\|\mathbf{A}\| = 0 \text{ тогда и только тогда, когда } \mathbf{A} = \mathbf{0}.$$

Говорят, что норма матрицы  $\mathbf{A}$  согласована с нормой вектора  $\mathbf{u}$ , если выполнено условие  $\|\mathbf{Au}\| \leq \|\mathbf{A}\| \|\mathbf{u}\|$ .

Нетрудно видеть, что подчиненная норма согласована с соответствующей метрикой векторного пространства. В самом деле

$$\|\mathbf{A}\| = \sup_{\|\mathbf{u}\| \neq 0} \frac{\|\mathbf{Au}\|}{\|\mathbf{u}\|} \geq \frac{\|\mathbf{Au}\|}{\|\mathbf{u}\|}, \text{ откуда } \|\mathbf{Au}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{u}\|.$$

Подчиненные введенным выше нормам векторов нормы матриц будут определяться следующим образом:

$$\|\mathbf{A}\|_1 = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|, \quad (2.3a)$$

$$\|\mathbf{A}\|_2 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|, \quad (2.3b)$$

$$\|\mathbf{A}\|_3 = \sqrt{\max_{1 \leq i \leq n} \lambda^i (\mathbf{A}^* \mathbf{A})}. \quad (2.3c)$$

В качестве примеров норм матриц, не подчиненных никаким векторным, можно привести норму Фробениуса и так называемую max-норму:

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2}, \quad (2.4a)$$

$$\|\mathbf{A}\|_{MAX} = \max_{i,j} |a_{ij}|. \quad (2.4b)$$

### П.1.2. Другие нормы в $\mathbf{R}^n$ . Теорема об эквивалентности норм

Рассмотрим следующее выражение:  $x(\mathbf{u}) = \sqrt[m]{\sum_{i=1}^n |u_i|^m}$ . Нетрудно убедиться, что при любом натуральном  $m$  для величины  $x(\mathbf{u})$  выполнены все аксиомы нормы. Выше слушаю (2.1a) соответствует предел при  $m \rightarrow \infty$ , норме (2.1б) соответствует  $m = 1$  и норме (2.1в) —  $m = 2$ . Часто такие нормы обозначают  $\|x\|_m$  в соответствии со значением параметра, при котором сумма вычисляется. Если не оговорено иное, такая нотация в задачах не применяется. Существуют и другие нормы в линейном векторном пространстве.

Для конечномерных пространств справедливо следующее утверждение. Каковы бы ни были две нормы  $\|\cdot\|_a$  и  $\|\cdot\|_b$ , то существуют положительные числа  $\gamma_1, \gamma_2$  такие, что для всех элементов рассматриваемого пространства выполняется  $\gamma_1\|\mathbf{u}\|_a \leq \|\mathbf{u}\|_b \leq \gamma_2\|\mathbf{u}\|_a$ . Числа  $\gamma_1$  и  $\gamma_2$  называются константами эквивалентности.

Это утверждение называется теоремой об эквивалентности норм в конечномерных пространствах.

В силу теоремы об эквивалентности норм все утверждения теорем верны для любых норм, поэтому ниже выбор нормы не конкретизируется. Естественно, в задачах требуется выбирать конкретную норму так, чтобы решение получалось самым легким способом.

### II.3. Обусловленность СЛАУ. Число обусловленности матрицы

Понятия согласованных норм матриц и векторов позволяют оценить погрешности, возникающие при численном решении СЛАУ. Пусть и матрица, и правая часть системы уравнений заданы с некоторой погрешностью, тогда наряду с системой

$$\mathbf{A}\mathbf{u} = \mathbf{f} \quad (3.1)$$

рассматривается возмущенная система

$$(\mathbf{A} + \Delta\mathbf{A})(\mathbf{u} + \Delta\mathbf{u}) = \mathbf{f} + \Delta\mathbf{f}$$

**Теорема 1.** Пусть правая часть и невырожденная матрица СЛАУ (3.1) вида  $\mathbf{A}\mathbf{u} = \mathbf{f}$ ,  $\mathbf{u} \in R^n$ ,  $\mathbf{f} \in R^n$ , получили приращения  $\Delta\mathbf{f}$  и  $\Delta\mathbf{A}$  соответственно. Пусть существует обратная матрица  $\mathbf{A}^{-1}$  и выполнены условия  $\|\mathbf{A}\| \neq 0$ ,  $\mu\|\Delta\mathbf{A}\|/\|\mathbf{A}\| < 1$ , где  $\mu = \|\mathbf{A}\|\cdot\|\mathbf{A}^{-1}\|$ . В этом случае оценка относительной погрешности решения  $\|\Delta\mathbf{u}\|/\|\mathbf{u}\|$  удовлетворяет неравенству

$$\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leq \frac{\mu}{1 - \mu} \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|} + \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}.$$

При  $\Delta\mathbf{A} = 0$  получаем оценку при наличии погрешности только правых частей:

$$\frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \leq \mu \frac{\|\Delta\mathbf{f}\|}{\|\mathbf{f}\|}. \quad (3.2)$$

Это важное соотношение показывает, насколько возрастают относительные ошибки решения СЛАУ в случае наличия относительных погрешностей

задания правых частей и элементов матриц. Величина  $\mu = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$  называется **числом обусловленности** матрицы  $\mathbf{A}$ . Она играет существенную роль во всех задачах прикладной линейной алгебры.

Почти очевидно, что всегда  $\mu \geq 1$ . Действительно:

$$1 = \|\mathbf{E}\| = \|\mathbf{A}^{-1}\mathbf{A}\| \leq \|\mathbf{A}^{-1}\| \|\mathbf{A}\| = \mu.$$

При известной (фиксированной) правой части оценку (3.2) можно улучшить (см. II.8.3).

## П.4. Решение систем линейных алгебраических уравнений (СЛАУ). Прямые и итерационные методы

Рассмотрим СЛАУ  $\mathbf{A}\mathbf{u} = \mathbf{f}$ , где  $\mathbf{A}$  — невырожденная ( $\det \mathbf{A} \neq 0$ ) квадратная матрица размером  $n \times n$ ,  $\mathbf{u} = \{u_1, \dots, u_n\}^T$  — вектор-столбец решения,  $\mathbf{f} = \{f_1, \dots, f_n\}^T$  — вектор-столбец правой части.

Так как матрица системы невырожденная,  $\Delta = \det \mathbf{A} \neq 0$ , то решение системы (2.1) существует и единственno.

*Прямые методы* позволяют в предположении отсутствия ошибок округления (при проведении расчетов на идеальном, т. е. бесконечноразрядном компьютере) получить точное решение задачи за конечное число арифметических действий. *Итерационные методы*, или методы последовательных приближений, позволяют вычислить последовательность  $\{\mathbf{u}^k\}$ , сходящуюся к решению задач при  $k \rightarrow \infty$  (на практике, разумеется, ограничиваются конечным  $k$ , в зависимости от требуемой точности).

### П.4.1. Прямые методы решения СЛАУ

К прямым методам решения СЛАУ относятся правило Крамера, метод исключения Гаусса, поиск решения с помощью обратной матрицы и метод сопряженных градиентов. Правило Крамера неэкономично для систем размерности выше трех. Метод сопряженных градиентов, который является прямым методом решения СЛАУ, для систем большой размерности может использоваться как итерационный, т.е. вычисления прекращают, не завершая полный цикл вычислений. Неприятным свойством метода сопряженных градиентов является его возможная неустойчивость. Наиболее употребительным прямым методом решения СЛАУ является метод Гаусса.

### П.3.2. Метод исключения Гаусса

Прямой ход метода Гаусса состоит в следующем.

Положим, что  $a_{11} \neq 0$  и исключим  $u_1$  из всех уравнений, начиная со второго, для чего ко второму уравнению прибавим первое, умноженное на  $-a_{21}/a_{11} = \eta_{21}$ , к третьему прибавим первое, умноженное на  $-a_{31}/a_{11} = \eta_{31}$ , и т.д. После этих преобразований получим эквивалентную систему, коэффициенты и правые части которой определяются следующим образом:

$$a_{ij}^1 = a_{ij} - \eta_{i1}a_{1j}; f_i^1 = f_i - \eta_{i1}f_1; i, j = 2, \dots, n.$$

Без ограничения общности считаем, что  $a_{22}^1 \neq 0$ . В противном случае меняем местами второе уравнение «новой» системы и первое уравнение, в котором элемент во втором столбце отличен от нуля.

Аналогично исключаем  $u_2$  из последних  $(n-2)$  уравнений системы. В результате преобразований получим новую эквивалентную систему уравнений в которой  $a_{ij}^2 = a_{ij}^1 - \eta_{i2}a_{2j}^1$ ;  $f_i^2 = f_i^1 - \eta_{i2}f_2^1$ ;  $i, j = 3, \dots, n$ . Продолжая алгоритм, т.е. исключая  $u_i$  ( $i = k+1, \dots, n$ ), приходим на  $n-1$  шаге к системе с треугольной матрицей.

Обратный ход метода Гаусса позволяет определить решение системы линейных уравнений. Из последнего уравнения системы находим  $u_n$ ; подставляем это значение в предпоследнее уравнение, получим  $u_{n-1}$ . Поступая так и далее, последовательно находим  $u_{n-2}, u_{n-3}, \dots, u_1$ .

Вычисления компонент вектора решения проводятся по формулам

$$u_n = f_n^{(n-1)} / a_{nn}^{(n-1)},$$

...

$$u_k = \frac{1}{a_{kk}^{(k-1)}} \left( f_k^{(k-1)} - a_{k,k+1}^{(k-1)}u_{k+1} - \dots - a_{kn}^{(k-1)}u_n \right), \quad k = n-1, n-2, \dots, 1,$$

...

$$u_1 = \frac{1}{a_{11}} \left( f_1 - a_{12}u_2 - \dots - a_{1n}u_n \right).$$

Этот алгоритм прост и легко реализуем при условии, что  $a_{11} \neq 0$ ,  $a_{22} \neq 0, \dots$ . Количество арифметических действий прямого хода:  $\approx n^3/3$  умножений и  $n^3/3$  сложений, обратного  $\approx n^2$ .

В реальных вычислениях используются методы с выбором главного (или *вседущего*) элемента. Выбор главного элемента *по столбцам* реализуется следующим образом: перед исключением  $u_1$  отыскивается  $\max_i |a_{1i}|$ . Пусть максимум достигается при  $i = k$ . В этом случае меняются местами первое и  $k$ -е уравнения и реализуется процедура исключения. Затем отыскивается  $\max_i |a_{i2}^{(1)}|$ , процедура

поиска главного элемента в столбцах повторяется. Так же реализуется выбор главного элемента *по строкам*: перед исключением  $u_i$  отыскивается  $\max_j |a_{kj}|$ .

Если максимум достигается при  $i = k$ , то у  $u_1$  и  $u_k$  меняются номера, то есть максимальный элемент из коэффициентов первого уравнения окажется на месте  $a_{11}$ , и т.д. Наиболее устойчивым является метод Гаусса с выбором главного элемента по всей матрице.

При использовании поиска главного элемента по строке необходимо запоминать совершенные перестановки, так как после завершения процедуры решения потребуется перестановка компонент векторного решения.

Во многих методах важным является условие *диагонального преобразования*

$$|a_{ii}| \geq \sum_{j=1, j \neq i}^n |a_{ij}|$$

для  $i = 1, \dots, n$ . При выполнении этого условия проблемы с устойчивостью не возникают. Если для всех строк матрицы выполняются строгие неравенства, то говорят о *строгом диагональном преобладании*.

### П.4.3. LU-разложение

Среди прямых методов численного решения СЛАУ широко используется также LU-разложение матрицы **A**, эквивалентное методу Гаусса, и метод Холецкого (или метод квадратного корня).

Если матрица **A** представима в виде произведений матриц **LU**, то СЛАУ может быть представлена в виде **(LU)u = f**.

Перепишем исходную систему, вводя вспомогательный вектор **v**, в следующем виде:

$$\mathbf{L}\mathbf{v} = \mathbf{f}, \quad \mathbf{U}\mathbf{u} = \mathbf{v}.$$

Решение СЛАУ свелось к последовательному решению двух систем с треугольными матрицами. Первый этап решения системы **Lv = f**:

$$v_1 = f_1,$$

$$l_{21}v_1 + v_2 = f_2,$$

...

$$l_{n1}v_1 + l_{n2}v_2 + \dots + l_{n,n-1}v_{n-1} + v_n = f_n,$$

откуда можно вычислить все  $v_k$  последовательно по формулам  
 $v_k = f_k - \sum_{j=1}^{k-1} l_{kj} v_j ; k = 2, \dots, n$ . Далее, рассмотрим систему  $\mathbf{U}\mathbf{u} = \mathbf{v}$  или

$$d_{11}u_1 + d_{21}u_2 + \dots + d_{n1}u_n = v_1,$$

$$d_{22}u_2 + \dots + d_{n2}u_n = v_2,$$

$$\dots \\ d_{nn}u_n = v_n,$$

решение которой находятся в обратном порядке, т.е. при  $k = n-1, \dots, 1$  по очевидным формулам  $u_k = d_{kk}^{-1}(v_k - \sum_{j=k+1}^n d_{kj}u_j)$ .

Условия существования такого разложения даются следующей теоремой.

**Теорема 2.** *Если все главные миноры квадратной матрицы  $\mathbf{A}$  отличны от нуля, то существуют единственные нижняя и верхняя треугольные матрицы  $\mathbf{L} = \{l_{ij}\}$  и  $\mathbf{U} = \{d_{ij}\}$  такие, что  $\mathbf{A} = \mathbf{LU}$ . При этом все диагональные коэффициенты матрицы  $\mathbf{L}$  фиксированы и равны единице.*

Для симметричных положительно определенных матриц существует модификация LU-разложения матрицы  $\mathbf{A}$ , называемая методом Холецкого, или методом квадратного корня. В этом методе матрица системы представляется в виде  $\mathbf{A} = \mathbf{LL}^T$ , где

$$\mathbf{L} = \begin{pmatrix} l_{11} & 0 & \dots & 0 \\ l_{12} & l_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ l_{1n} & l_{2n} & \dots & l_{nn} \end{pmatrix}, \mathbf{L}^T = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1n} \\ 0 & l_{22} & \dots & l_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & l_{nn} \end{pmatrix}.^3 \quad (4.1)$$

## II.5. Итерационные методы решения СЛАУ

### II.5.1. Метод простой итерации

Рассмотрим систему линейных алгебраических уравнений  $\mathbf{Au} = \mathbf{f}$ .

Проведем несколько равносильных преобразований. Умножим обе части системы на один и тот же скалярный множитель  $\tau$ , затем прибавим к правой и левой частям системы вектор  $\mathbf{u}$ . Систему уравнений можно теперь записать в виде, удобном для итераций

$$\mathbf{u} = \mathbf{Ru} + \mathbf{F},$$

---

<sup>3</sup> Обратите внимание, что стандартная индексация элементов матрицы лишь у  $\mathbf{L}^T$ .

где  $\mathbf{R} = \mathbf{E} - \tau\mathbf{A}$ ,  $\mathbf{F} = \tau\mathbf{f}$ .  $\mathbf{R}$  называется *матрицей перехода*.

Построим последовательность приближений к решению системы. Выберем произвольный вектор  $\mathbf{u}^{(0)}$  — начальное приближение к решению. Часто его просто полагают нулевым вектором. Скорее всего, начальное приближение не удовлетворяет исходной системе. При подстановке его в исходное уравнение возникает невязка  $\mathbf{r}^{(0)} = \mathbf{f} - \mathbf{Au}^{(0)}$ . Вычислив невязку, можно уточнить приближение к решению

$$\mathbf{u}^{(1)} = \mathbf{u}^{(0)} + \tau\mathbf{r}^{(0)}.$$

По первому приближению снова вычисляется невязка, процесс продолжается. В ходе итерации получаем  $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \tau\mathbf{r}^{(k)}$ ,  $\mathbf{r}^{(k)} = \mathbf{f} - \mathbf{Au}^{(k)}$ . Эквивалентная формулировка метода, называемого методом простых итераций, заключается в следующем. Решение системы  $\mathbf{Au} = \mathbf{f}$  находится как предел последовательности  $\{\mathbf{u}^{(0)}, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \dots\}$  приближений, члены которой связаны рекуррентным соотношением

$$\mathbf{u}^{(k+1)} = \mathbf{Ru}^{(k)} + \mathbf{F},$$

$\mathbf{u}^{(0)} = 0$  (или любому произвольному вектору). Если предел такой последовательности существует, то говорят о **сходимости** итерационного процесса к решению СЛАУ.

Существуют другие формы записи метода итераций, например

$$\mathbf{u}^{(k+1)} = (\mathbf{E} - \tau\mathbf{A})\mathbf{u}^{(k)} + \tau\mathbf{f}.$$

**Теорема 3 (достаточное условие сходимости метода простой итерации).** Итерационный процесс  $\mathbf{u}^{(k+1)} = \mathbf{Ru}^{(k)} + \mathbf{F}$  сходится к решению  $\mathbf{U}$  СЛАУ  $\mathbf{Au} = \mathbf{f}$  со скоростью геометрической прогрессии при выполнении условия:  $\|\mathbf{R}\| \leq q < 1$ .

**Доказательство.** Пусть  $\mathbf{U}$  — точное решение системы. Вычитая из точного равенства  $\mathbf{AU} = \mathbf{f}$  равенство  $\mathbf{u}^{(k+1)} = \mathbf{Ru}^{(k)} + \mathbf{F}$ , получим  $\mathbf{u}^{(k)} - \mathbf{U} = \mathbf{R}(\mathbf{u}^{(k-1)} - \mathbf{U})$ . Обозначив погрешность  $\boldsymbol{\varepsilon}^{(k)} = \mathbf{u}^{(k)} - \mathbf{U}$ , получим для эволюции погрешности уравнение  $\boldsymbol{\varepsilon}^{(k)} = \mathbf{R}\boldsymbol{\varepsilon}^{(k-1)}$ .

Справедлива цепочка неравенств:

$$\|\mathbf{u}^{(k)} - \mathbf{U}\| = \|\boldsymbol{\varepsilon}^{(k)}\| \leq \|\mathbf{R}\| \cdot \|\boldsymbol{\varepsilon}^{(k-1)}\| \leq q \|\boldsymbol{\varepsilon}^{(k-1)}\| \leq \dots \leq q^k \|\boldsymbol{\varepsilon}^{(0)}\| = q^k \|\mathbf{u}^{(0)} - \mathbf{U}\|,$$

где  $0 < q \leq \|\mathbf{R}\|$ . Отсюда следует, что при  $q < 1$   $\lim_{k \rightarrow \infty} \mathbf{u}^{(k)} = \mathbf{U}$ . ■

Из неравенства  $\|\boldsymbol{\varepsilon}^{(k)}\| \leq q^k \|\boldsymbol{\varepsilon}^{(0)}\|$  можно получить оценку количества итераций, необходимых для достижения точности  $\varepsilon$ , т.е. для выполнения условия  $\|\mathbf{u}^{(k)} - \mathbf{U}\| = \|\boldsymbol{\varepsilon}^{(k)}\| \leq \varepsilon$ . Эта оценка имеет вид  $k \geq \ln(\varepsilon / \|\boldsymbol{\varepsilon}_0\|) / \ln q$ .

У этой оценки есть недостаток — она использует неизвестную величину — погрешность нулевой итерации  $\|\boldsymbol{\varepsilon}^{(0)}\|$ . Как заканчивать итерационный процесс с гарантированной точностью, анализируя только итерационную поправку, описано в главе IV на примере окончания итерационных процессов для решения нелинейных уравнений.

*Теорема 4 (критерий сходимости метода простой итерации).* Пусть СЛАУ имеет единственное решение. Тогда для сходимости итерационного процесса  $\mathbf{u}^{(k+1)} = \mathbf{R}\mathbf{u}^{(k)} + \mathbf{F}$  необходимо и достаточно, чтобы все собственные значения матрицы  $\mathbf{R}$  по абсолютной величине были меньше единицы.

### II.5.2. Каноническая форма записи двухслойных итерационных методов

**Канонической формой записи двухслойного итерационного метода** называется следующая:

$$\mathbf{B}_{k+1} \frac{\mathbf{u}^{(k+1)} - \mathbf{u}^{(k)}}{\tau_{k+1}} + \mathbf{A}\mathbf{u}^{(k)} = \mathbf{f}.$$

При  $\mathbf{B}_k = \mathbf{E}$ ,  $\tau_k = \tau$  последняя формула соответствует однопараметрическому итерационному процессу — рассмотренному выше *методу простых итераций*. При  $\mathbf{B}_k = \mathbf{E}$ ,  $\tau_k = \{\tau_k, k = 1, \dots, n\}$  —  $n$ -шаговому явному итерационному процессу, при  $\mathbf{B}_k = \mathbf{B}'$ ,  $\tau_k = 1$  — методу простой итерации без итерационного параметра. В случае, когда  $\mathbf{B} \neq \mathbf{E}$ , итерационный метод называется *неявным* — для вычисления следующего приближения к решению придется решать (как правило, более простую, чем исходную) систему линейных уравнений.

### II.5.3. Методы Якоби, Зейделя, верхней релаксации

Рассмотрим СЛАУ:

$$a_{11}u_1 + a_{12}u_2 + \dots + a_{1n}u_n = f_1,$$

$$a_{21}u_1 + a_{22}u_2 + \dots + a_{2n}u_n = f_2,$$

...

$$a_{n1}u_1 + a_{n2}u_2 + \dots + a_{nn}u_n = f_n.$$

В *методе Якоби* из  $i$ -го уравнения выражается  $i$ -компонент решения,

при этом все остальные компоненты берутся с предыдущей итерации:

$$u_i^{(s+1)} = -\frac{1}{a_{ii}} \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij} u_j^{(s)} + \frac{f_i}{a_{ii}}. \quad (5.1)$$

В реальных вычислениях все уже вновь вычисленные компоненты решения на новой итерации можно использовать при вычислении новой итерации. Это не приводит к дополнительному расходу ресурсов. Такой метод называется *методом Гаусса–Зейделя*:

$$x_i^{(s+1)} = -\frac{1}{a_{ii}} \left( \sum_{j=1}^{i-1} a_{ij} x_j^{(s+1)} + \sum_{j=i+1}^n a_{ij} x_j^{(s)} - f_i \right). \quad (5.2)$$

Формулы (5.1) – (5.2) являются основными расчетными формулами этих методов.

Для доказательства сходимости этих методов удобно представить их с использованием разложения матрицы  $\mathbf{A}$  на сумму трех матриц

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U},$$

где  $\mathbf{L}$  и  $\mathbf{U}$  — нижняя и верхняя треугольные матрицы с нулевыми элементами на главной диагонали,  $\mathbf{D}$  — диагональная матрица. Рассматриваемая СЛАУ может быть переписана в следующем эквивалентном виде:

$$\mathbf{L}\mathbf{u} + \mathbf{D}\mathbf{u} + \mathbf{U}\mathbf{u} = \mathbf{f}.$$

Построим два итерационных процесса:

$$\mathbf{L}\mathbf{u}^{(k)} + \mathbf{D}\mathbf{u}^{(k+1)} + \mathbf{U}\mathbf{u}^{(k)} = \mathbf{f}$$

и

$$\mathbf{L}\mathbf{u}^{(k+1)} + \mathbf{D}\mathbf{u}^{(k+1)} + \mathbf{U}\mathbf{u}^{(k)} = \mathbf{f},$$

или, соответственно,

$$\mathbf{u}^{(k+1)} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{u}^{(k)} + \mathbf{D}^{-1}\mathbf{f}$$

и

$$\mathbf{u}^{(k+1)} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}\mathbf{u}^{(k)} + (\mathbf{L} + \mathbf{D})^{-1}\mathbf{f}.$$

Очевидно, что эти формулы описывают итерационные методы вида  $\mathbf{u}^{(k+1)} = \mathbf{R}\mathbf{u}^{(k)} + \mathbf{F}$ , если положить в первом случае  $\mathbf{R} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ ,  $\mathbf{F} = \mathbf{D}^{-1}\mathbf{f}$  или  $\mathbf{R} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}$ ,  $\mathbf{F} = (\mathbf{L} + \mathbf{D})^{-1}\mathbf{f}$  во втором. Эти итерационные методы соответствуют методам Якоби и Зейделя соответственно.

**Теорема 5 (достаточное условие сходимости метода Якоби).** Итерационный метод Якоби сходится к решению соответствующей СЛАУ, если выполнено условие диагонального преобладания

$$|a_{ii}| > \sum_{j=1, j \neq i}^n |a_{ij}|, \quad i = 1, \dots, n.$$

**Теорема 6 (критерий сходимости итерационного метода Якоби).** Для сходимости итерационного метода Якоби необходимо и достаточно, чтобы все корни уравнения

$$\begin{vmatrix} \lambda a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & \lambda a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & \lambda a_{nn} \end{vmatrix} = 0$$

по модулю не превосходили единицы.

**Теорема 7 (критерий сходимости итерационного метода Зейделя).** Для сходимости итерационного метода метода Зейделя необходимо и достаточно, чтобы все корни уравнения

$$\begin{vmatrix} \lambda a_{11} & a_{12} & \dots & a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda a_{n1} & \lambda a_{n2} & \dots & \lambda a_{nn} \end{vmatrix} = 0$$

по модулю не превосходили единицы.

**Теорема 8 (достаточное условие сходимости метода Зейделя).** Пусть **А** — вещественная, симметричная, положительно определенная матрица. В этом случае итерационный метод Зейделя сходится.

Развитием метода Зейделя является *метод последовательной релаксации*. В предположении, что метод Зейделя меняет каждый компонент вектора решения в правильном направлении, введем параметр, который позволяет «пройти» по этому пути несколько дальше:

$$\mathbf{u}_i^{(k+1)} = \mathbf{u}_i^{(k)} + \omega (\mathbf{z}_i^{(k+1)} - \mathbf{u}_i^{(k)}), \quad (5.3)$$

где  $\mathbf{z}_i^{(k)}$  —  $i$ -я компонента решения, полученная методом Зейделя. (5.3) можно переписать в эквивалентном виде:

$$\mathbf{u}_i^{(k+1)} = \mathbf{u}_i^{(k)} (1 - \omega) + \omega \mathbf{z}_i^{(k+1)}. \quad (5.4)$$

Из (5.4) немедленно следует координатное представление метода последовательной верхней релаксации:

$$u_i^{(s+1)} = (1-\omega)u_i^{(s)} - \frac{\omega}{a_{ii}} \left( \sum_{j=1}^{i-1} a_{ij}u_j^{(s+1)} + \sum_{j=i+1}^n a_{ij}u_j^{(s)} - f_i \right), \quad i = 1, 2, \dots, n.$$

В этом методе введен *параметр релаксации*  $\omega$ .

Как можно видеть, новое приближение во всех трех методах находится последовательно для каждого компонента решения. Метод Якоби допускает любую перестановку действий, следовательно, возможна его параллельная реализация. При некоторых модификациях возможна и параллельная реализация метода последовательной релаксации.

Метод релаксации может быть представлен в матричной форме:

$$(\omega \mathbf{L}\mathbf{u}^{(k+1)} + \mathbf{D}\mathbf{u}^{(k+1)}) + (\omega - 1)\mathbf{D}\mathbf{u}^{(k)} + \omega \mathbf{U}\mathbf{u}^{(k)} = \omega \mathbf{f}.$$

Выбором  $\omega$  можно существенно изменять скорость сходимости итерационного метода.

Выразим  $\mathbf{u}^{(k+1)}$ :

$$\mathbf{u}^{(k+1)} = -(\mathbf{D} + \omega \mathbf{L})^{-1} [(\omega - 1)\mathbf{D} + \omega \mathbf{L}] \mathbf{u}^{(k)} + \omega (\mathbf{D} + \omega \mathbf{L})^{-1} \mathbf{f}.$$

В общем случае задача вычисления  $\omega_{\text{опт}}$  (оптимального итерационного параметра) не решена, однако известно, что  $0 < \omega_{\text{опт}} < 2$ . Для положительно определенных матриц  $1 < \omega_{\text{опт}} < 2$ . В этом случае итерационный метод называется методом последовательной верхней релаксации или SOR — Successive over relaxation. Иногда встречается термин «сверхрелаксация» при  $1 < \omega_{\text{опт}} < 2$ . При  $0 < \omega < 1$  имеем метод нижней релаксации.

Для очень важного частного случая, когда 1) матрица  $\mathbf{A}$  симметрична и положительно определена, и 2) существует перестановка переменных  $\mathbf{P}$ , такая что матрица  $\mathbf{A}$  после перестановки имеет вид  $\mathbf{T}$ :

$$\mathbf{T} = \mathbf{P}\mathbf{A}\mathbf{P}^T = \begin{pmatrix} \mathbf{D}_1 & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{D}_2 \end{pmatrix},$$

где  $\mathbf{D}_1, \mathbf{D}_2$  — диагональные матрицы, оптимальное значение релаксационного параметра можно определить через спектральный радиус матрицы перехода метода Якоби  $\rho(\mathbf{R}_J)$ ,  $\mathbf{R}_J = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ :

$$\omega_{\text{опт}} = \frac{2}{1 + \sqrt{1 - \rho^2(\mathbf{R}_J)}}.$$

Условие существование перестановки  $\mathbf{P}$  означает, что вектор переменных распадается на два класса, так что каждое уравнение системы содержит одну переменную одного класса и, возможно, все переменные другого класса.

## II.6. О спектральных задачах

Спектральные задачи — вычислительно наиболее трудоемкие задачи в прикладной линейной алгебре. Различают полную и частичную проблемы собственных значений. В первом случае необходимо отыскать ВСЕ собственные числа матрицы, во втором — лишь максимальное по абсолютной величине собственное число. Различают также самосопряженную спектральную задачу и задачу для произвольной матрицы. Очевидно, самосопряженная проблема решается проще — спектр самосопряженной матрицы всегда действительный.

Рассмотрим два алгоритма для самосопряженных матриц. Первый — *степенной* алгоритм, для вычисления наибольшего по абсолютной величине собственного числа. Выбираем произвольный ненулевой вектор  $\mathbf{u}^{(0)}$  и строим последовательность векторов:

$$\mathbf{u}^{(k+1)} = \mathbf{A}\mathbf{u}^{(k)}.$$

Легко показать, что выражение

$$\lambda \approx \frac{(\mathbf{A}\mathbf{u}^{(k)}, \mathbf{u}^{(k)})}{(\mathbf{u}^{(k)}, \mathbf{u}^{(k)})} = \frac{(\mathbf{u}^{(k+1)}, \mathbf{u}^{(k)})}{(\mathbf{u}^{(k)}, \mathbf{u}^{(k)})}$$

приближает максимальное по абсолютной величине собственное значение с точностью  $O(\lambda_{N-1}/\lambda_N)^k$ . Здесь  $\lambda_{N-1}/\lambda_N$  — обратное отношение самого большого по модулю собственного числа матрицы к следующему по абсолютной величине.

Для решения полной самосопряженной проблемы собственных значений применяется **метод вращений**.

Определение собственных значений самосопряженной матрицы  $\mathbf{A}$  эквивалентно отысканию такой ортогональной матрицы  $\mathbf{T}$ , что

$$\mathbf{\Lambda} = \mathbf{T}^T \mathbf{A} \mathbf{T},$$

матрица  $\mathbf{\Lambda}$  — диагональная.

Среди всех ортогональных преобразований данное минимизирует сумму квадратов внедиагональных элементов исходной матрицы. Построим итерационный метод, минимизирующий эту сумму на каждой итерации. Пусть каждое преобразование подобия на каждой итерации содер-

жит лишь одну матрицу вращения  $\hat{\mathbf{A}} = \mathbf{T}'_{ij} \mathbf{A} \mathbf{T}_{ij}$ , где матрица  $\mathbf{T}_{ij}$  есть матрица поворота в плоскости  $(u_i, u_j)$  на угол  $\alpha$ . Эта матрица отличается от матрицы  $\mathbf{A}$  только двумя строками и двумя столбцами (с номерами  $i$  и  $j$ ). Так как норма Фробениуса матрицы (2.4а) не изменяется при ортогональных преобразованиях, то легко получить соотношение между суммами квадратов внедиагональных элементов старой и новой матриц:

$$\sum_{k \neq l} \hat{a}_{kl}^2 = \sum_{k \neq l} a_{kl}^2 - 2a_{ij}^2 + \frac{1}{2}((a_{jj} - a_{ii})\sin 2\alpha + 2a_{ij} \cos 2\alpha)^2.$$

Очевидны условия минимизации суммы в левой части последнего равенства: справа следует вычитать как можно больше и прибавлять как можно меньше. Следует на текущей итерации выбирать индексы так, чтобы выполнялось условие  $|a_{ij}| = \max_{k \neq l} |a_{kl}|$ . Тогда угол поворота выбирается из условия

$$0 = ((a_{jj} - a_{ii})\sin 2\alpha + 2a_{ij} \cos 2\alpha)^2.$$

Этот угол удовлетворяет уравнению  $\operatorname{tg} 2\alpha = \frac{2a_{ij}}{a_{ii} - a_{jj}}$ ,  $|\alpha| \leq \frac{\pi}{4}$ .

Независимо от наличия кратных собственных значений метод вращений обладает квадратичной сходимостью. Это означает, что для нормы внедиагональных элементов матрицы  $\mathbf{A}$ :  $\operatorname{off}(\mathbf{A}) = \sqrt{\sum_{1 \leq j < k \leq n} a_{jk}^2}$ , за число шагов

$N = n(n-1)/2$ , достаточное, чтобы выбрать каждый наддиагональный элемент по одному разу, эта норма уменьшается квадратично:  $\operatorname{off}(\mathbf{A}_{i+N}) = O(\operatorname{off}^2(\mathbf{A}_i))$ .

Выбор максимального по модулю внедиагонального элемента — затратная операция, поэтому часто реализуется метод вращений с барьерами. Его идея состоит в следующем. При переборе внедиагональных значений вращение производится тогда, когда значение элемента по абсолютной величине превосходит некоторую величину (барьер). Если все элементы меньше барьера, его значение уменьшается, например, на порядок, и снова начинается циклический перебор внедиагональных элементов. Подробнее о методе вращений смотри в [11, 15].

Метод обратной итерации применяется для поиска собственного числа, наиболее близкого к данному. Суть его заключается в следующем. Рассмотрим равенство  $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$ . Зафиксируем параметр  $a$ . Тогда  $(\mathbf{A} - a\mathbf{E})\mathbf{u} = (\lambda - a)\mathbf{u}$ . Верно будет и равенство  $(\mathbf{A} - a\mathbf{E})^{-1}\mathbf{u} = \mathbf{u}/(\lambda - a)$ . Но

если мы интересуемся собственным числом, наиболее близким к  $a$ , то среди собственных чисел матрицы  $(\mathbf{A} - a\mathbf{E})^{-1}$  именно  $1/(\lambda - a)$  будет наибольшим по абсолютной величине. Для его вычисления можно использовать степенной метод, но так как нам фактически необходимы степени обратной матрицы, то в степенной метод вносится модификация, теперь  $\mathbf{u}_k = (\mathbf{A} - a\mathbf{E})\mathbf{u}_{k+1}$ . Вот она, обратная итерация — для поиска следующего приближения в степенном методе надо решать СЛАУ. Причем, чем ближе находится искомый корень характеристического уравнения к выбранному параметру  $a$ , тем эта СЛАУ ближе к вырожденной со всеми вытекающими отсюда трудностями.

## II.7. Задачи на доказательство

II.7.1. Является ли выражение

$$\min(|x_1| + 5|x_2|, 5|x_1| + |x_2|)$$

нормой вектора  $\mathbf{x}$  в  $\mathbb{R}^2$ ?

II.7.2. Нормы  $\|\cdot\|_1$  и  $\|\cdot\|_2$  называются эквивалентными, если для всех  $\mathbf{x} \in \mathbb{R}^n$  справедливы неравенства с постоянными  $\gamma_1, \gamma_2$ , не зависящими от выбора вектора  $\mathbf{x}$ :

$$\gamma_1\|\mathbf{x}\|_1 \leq \|\mathbf{x}\|_2 \leq \gamma_2\|\mathbf{x}\|_1.$$

Найти константы эквивалентности, связывающие три основные нормы векторов.

II.7.3. Доказать, что если  $\mathbf{C}$  — симметричная положительно определенная матрица, то  $\sqrt{(\mathbf{C}\mathbf{x}, \mathbf{x})}$  можно принять за норму вектора  $\mathbf{x}$ . Найти константы эквивалентности, связывающие эту норму с евклидовой нормой вектора.

II.7.4. Доказать утверждения (2.3а), (2.3б), (2.3в).

II.7.5. Показать, что модуль любого собственного значения матрицы не больше любой ее нормы.

II.7.6. Показать, что для подчиненных норм матриц справедливо неравенство

$$\|\mathbf{A}\|_3^2 \leq \|\mathbf{A}\|_1 \|\mathbf{A}\|_2.$$

II.7.7. Доказать, что для вектора  $\mathbf{x} = (x_1, x_2)$  и  $h > 0$  выражение  $\|\mathbf{x}\|_h = \max(|x_1|, |x_2 - x_1|/h)$  является нормой. Найти матричную норму, подчиненную этой векторной норме.

**П.7.8.** Нормой Фробениуса матрицы называется  $N(\mathbf{A}) = \sqrt{\sum_{i,j=1}^n a_{ij}^2}$ . Показать, что эта норма согласована сама с собой, т.е.  $N(\mathbf{AB}) \leq N(\mathbf{A}) N(\mathbf{B})$ , и найти константы эквивалентности, связывающие эту норму матрицы с нормами матриц, подчиненными трем основным нормам векторов. Показать, что норма Фробениуса не является операторной нормой.

**П.7.9.** Мах нормой матрицы называется  $\eta(\mathbf{A}) = \max_{i,j} |a_{ij}|$ . Показать, что эта норма не является согласованной сама с собой, а выражение  $M(\mathbf{A}) = n \cdot \eta(\mathbf{A})$  вводит самосогласованную норму матриц. Для  $M(\mathbf{A})$  найти константы эквивалентности с нормами матрицы, подчиненными трем основным нормам векторов. Показать, что норма  $M(\mathbf{A})$  не является операторной нормой.

**П.7.10.** Пусть числа  $d_k > 0$ ,  $k = 1, \dots, n$ . Доказать, что  $\max(d_k |x_k|)$  есть норма вектора  $\mathbf{x}$ . Найти норму матрицы, подчиненную этой векторной норме.

**П.7.11.** Пусть числа  $d_k > 0$ ,  $k = 1, \dots, n$ . Доказать, что  $\sum_{k=1}^n d_k |x_k|$  есть норма вектора  $\mathbf{x}$ . Найти норму матрицы, подчиненную этой векторной норме.

**П.7.12.** Пусть числа  $d_k > 0$ ,  $k = 1, \dots, n$ . Доказать, что  $\sqrt{\sum_{k=1}^n d_k x_k^2}$  есть норма вектора  $\mathbf{x}$ . Найти норму матрицы, подчиненную этой векторной норме.

**П.7.13.** Доказать, что  $\max_{1 \leq i \leq n} \left| \sum_{k=1}^i x_k \right|$  есть норма вектора  $\mathbf{x}$ . Найти норму матрицы, подчиненную этой векторной норме.

**П.7.14.** Проверить, что  $\|\mathbf{x}\|_p = \left( \sum_{i=1}^N |x_i|^p \right)^{1/p}$ ,  $p \geq 1$  является нормой в пространстве  $\mathbf{C}^N$  векторов с комплексными координатами. Показать, что при  $\mathbf{x} \in \mathbf{C}^N$  справедливо неравенство  $\|x\|_p \leq c (\|\operatorname{Re} \mathbf{x}\|_p + \|\operatorname{Im} \mathbf{x}\|_p)$ ,  $c = \text{const}$ . Найти такую постоянную  $c_0$ , что  $c_0 (\|\operatorname{Re} \mathbf{x}\|_2 + \|\operatorname{Im} \mathbf{x}\|_2) \leq \|\mathbf{x}\|_2$  для всех  $\mathbf{x} \in \mathbf{C}^N$ .

**П.7.15.** Пусть  $\|\cdot\|$  — некоторая норма в  $\mathbf{R}^N$ . Доказать, что равенство  $\|\mathbf{x}\|_* = \max_{\mathbf{y} \neq 0} ((\mathbf{x}, \mathbf{y}) / \|\mathbf{y}\|)$  также задает норму в  $\mathbf{R}^N$ , называемую двойственной к  $\|\cdot\|$ .

**П.7.16.** Пусть  $\mathbf{B}$  — невырожденная матрица,  $\|\cdot\|$  — некоторая норма в пространстве векторов размерности  $N$ . Доказать, что  $\|\mathbf{x}\|^* = \|\mathbf{Bx}\|$  также является нормой в пространстве векторов. Какая норма в пространстве матриц порождается нормой  $\|\mathbf{x}\|^*$  в пространстве векторов?

**П.7.17.** Показать, что если  $\mathbf{A}$  — невырожденная матрица, то  $\left\| \mathbf{A}^{-1} \right\|^{-1} = \inf_{\mathbf{x} \neq 0} \left( \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \right)$ .

**П.7.18.** Доказать неравенство  $\|\mathbf{A}\|_3 \leq \|\mathbf{A}\|^{1/2} \|\mathbf{A}^T\|^{1/2}$  для любой нормы  $\mathbf{A}$ , подчиненной какой-либо векторной норме.

**П.7.19.** Доказать, что если  $\mathbf{A} = \mathbf{A}^T$ , то  $\|\mathbf{A}\|_3 = \max_{\mathbf{y} \neq 0} \left( (\mathbf{Ay}, \mathbf{y}) / \|\mathbf{y}\|_3^2 \right)$ .

**П.7.20.** Пусть  $\|\cdot\|$  — норма в пространстве матриц, подчиненная некоторой норме векторов. Доказать согласованность этой матричной нормы, т.е. справедливость неравенства  $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \|\mathbf{B}\|$ .

**П.7.21.** Пусть  $\mathbf{A} = \mathbf{A}^T > 0$  и  $\|\mathbf{x}\|_{\mathbf{A}} = (\mathbf{Ax}, \mathbf{x})^{1/2}$ . Доказать, что для произвольного многочлена  $p_m(t)$  степени  $m \geq 0$  верно равенство  $\|p_m(\mathbf{A})\|_{\mathbf{A}} = \|p_m(\mathbf{A})\|_3$ .

**П.7.22.** Привести пример положительно определенной матрицы, спектр которой не является вещественным.

**П.7.23.** Пусть  $\mathbf{A} = \mathbf{A}^T > 0$  и  $F(\mathbf{x}) = 0.5 \cdot (\mathbf{Ax}, \mathbf{x}) - (\mathbf{b}, \mathbf{x})$  — квадратичная функция. Доказать, что:

- 1)  $F(\mathbf{x}) = 1/2 \left\| \mathbf{x} - \mathbf{x}^* \right\|_{\mathbf{A}}^2 - 1/2 \left\| \mathbf{x}^* \right\|^2$ , где  $\mathbf{x}^*$  — точное решение системы  $\mathbf{Ax} = \mathbf{b}$ ;
- 2) равенство  $F(\mathbf{x}^*) = \min_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{x})$  выполнено тогда и только тогда, когда  $\mathbf{x}^*$  — точное решение системы  $\mathbf{Ax} = \mathbf{b}$ ;
- 3) для градиента функции  $F(\mathbf{x})$  справедлива формула  $\nabla F(\mathbf{x}) = \mathbf{Ax} - \mathbf{b}$ .

**П.7.24.** Доказать, что max норма матрицы (2.4б) и норма Фробениуса (2.4а) не подчинены никаким векторным нормам.

**П.7.25.** Можно ли утверждать, что если определитель матрицы мал, то матрица плохо обусловлена?

**П.7.26.** Доказать, что  $\mu(\mathbf{A} \cdot \mathbf{B}) \leq \mu(\mathbf{A}) \cdot \mu(\mathbf{B})$  для любой из норм матриц, согласованных с нормами векторов, и любых квадратных матриц ( $\mathbf{A}, \mathbf{B}$  — квадратные матрицы).

Численно показать справедливость этого неравенства для матриц:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}; \quad \mathbf{B} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

**П.7.27.** Показать, что если  $\mathbf{A}$  — нормальная матрица ( $\mathbf{A}^T \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{A}^T$ ), то  $\|\mathbf{A}\|_3 = R(\mathbf{A})$ , где  $R(\mathbf{A})$  — спектральный радиус матрицы. Вычислить спектральный радиус и число обусловленности матрицы  $\mathbf{A} = \begin{pmatrix} 1 & 10 \\ 100 & 1001 \end{pmatrix}$ .

**П.7.28.** Доказать, что при условии наличия диагонального преобладания у матрицы системы метод Зейделя сходится, причем быстрее метода Якоби.

**П.7.29.** Используя покомпонентную запись метода верхней релаксации

$$\mathbf{u}_i^{(k+1)} = \mathbf{u}_i^{(k)} + \omega (\mathbf{z}_i^{(k+1)} - \mathbf{u}_i^{(k)}),$$

где  $\mathbf{z}_i^{(k)}$  —  $i$ -я компонента решения, полученная методом Зейделя, получить матричное представление метода верхней релаксации:

$$(\omega \mathbf{L}\mathbf{u}^{(k+1)} + \mathbf{D}\mathbf{u}^{(k+1)}) + (\omega - 1)\mathbf{D}\mathbf{u}^{(k)} + \omega \mathbf{U}\mathbf{u}^{(k)} = \omega \mathbf{f}.$$

**П.7.30.** Для СЛАУ при заданном  $\mathbf{f} = (f_1, f_2)$  найти наименьшее число  $v(\mathbf{f})$ , при котором независимо от  $\Delta\mathbf{f}$  выполнено:  $\|\delta\mathbf{u}\| \leq v(\mathbf{f})\|\delta\mathbf{f}\|$ . Для этой СЛАУ найти тот вектор  $\mathbf{f}$ , которому соответствует наименьшее число  $v(\mathbf{f})$ , а также само значение  $v(\mathbf{f})$  для трех норм векторов:  $\|\cdot\|_1$ ,  $\|\cdot\|_2$ ,  $\|\cdot\|_3$ .

$$a) \begin{cases} u_1 + \sqrt{3}u_2 = f_1 \\ -\sqrt{3}u_1 + u_2 = f_2 \end{cases}, \quad b) \begin{cases} u_1 + u_2 = f_1 \\ u_1 - u_2 = f_2 \end{cases}, \quad b) \begin{cases} u_1 + 0.99u_2 = f_1 \\ 0.99u_1 + u_2 = f_2 \end{cases}.$$

**П.7.31.** Для системы линейных алгебраических уравнений  $\mathbf{Ax} = \mathbf{f}$ ,  $\mathbf{A} = \mathbf{A}^* > 0$  найти такую правую часть  $\mathbf{f}$ , чтобы при фиксированной относительной ошибке задания правой части  $\delta = \|\delta\mathbf{f}\| / \|\mathbf{f}\|$  относительная погрешность решения была бы минимальна и выполнялась оценка  $\|\delta\mathbf{x}\| / \|\mathbf{x}\| \leq v \|\delta\mathbf{f}\| / \|\mathbf{f}\|$ . Найти значение  $v$ . Используется евклидова норма векторов.

**П.7.32.** Параметр  $v(\mathbf{f})$  определяется как  $v(\mathbf{f}) = \|\mathbf{f}\| \|\mathbf{A}^{-1}\| / \|\mathbf{u}\|$  в оценке П.7.31. Показать, что при использовании евклидовой нормы векторов его минимальное и максимальное значения соответствуют векторам правых частей СЛАУ  $\mathbf{Au} = \mathbf{f}$ , коллинеарным собственным векторам  $\mathbf{u}_k$  матрицы системы  $\mathbf{A}$ , соответствующим максимальному и минимальному собственным значениям. Найти эти собственные значения  $\lambda_{\max}$  и  $\lambda_{\min}$ , а также их собственные

векторы  $\mathbf{u}_{\max}$  и  $\mathbf{u}_{\min}$  для СЛАУ

$$\text{а) } \begin{cases} u_1 + u_2 = f_1, \\ u_1 - u_2 = f_2, \end{cases}, \quad \text{б) } \begin{cases} u_1 + 0.99u_2 = f_1, \\ 0.99u_1 + u_2 = f_2. \end{cases}$$

**П.7.33.** Докажите, что при любом начальном векторе  $(x^{(0)}, y^{(0)}, z^{(0)})^T$  последовательности векторов  $(x_1^{(k)}, y_1^{(k)}, z_1^{(k)})^T$  и  $(x_2^{(k)}, y_2^{(k)}, z_2^{(k)})^T$ , определяемые равенствами

$$x_1^{(k+1)} = 0.1x_1^{(k)} + 0.2y_1^{(k)} - 3,$$

$$y_1^{(k+1)} = 0.2x_1^{(k)} - 0.1y_1^{(k)} + 0.1z_1^{(k)} + 2,$$

$$z_1^{(k+1)} = -0.3x_1^{(k)} + 0.2z_1^{(k)} - 1$$

и

$$x_2^{(k+1)} = (2y_2^{(k)} - 30)/9,$$

$$y_2^{(k+1)} = (2x_2^{(k)} + z_2^{(k)} + 20)/11,$$

$$z_2^{(k+1)} = -(3x_2^{(k)} + 10)/8,$$

сходятся, причем к одному и тому же предельному вектору  $(x^*, y^*, z^*)^T$ . Запишите линейную систему стандартного вида, решением которой является этот предельный вектор. За сколько итераций по данным формулам можно получить предельный вектор с точностью до  $\varepsilon = 10^{-6}$  (в кубической норме), если за начальное приближение принять нулевой вектор?

**П.7.34.** Вывести расчетные формулы для элементов матриц при использовании разложения Холецкого (4.1).

**П.7.35.** Найти число обусловленности матрицы  $\mathbf{A}$  в матричной норме, подчиненной евклидовой норме вектора, выразив его через число обусловленности матрицы  $\mathbf{B}$ , если  $\mathbf{A} = \mathbf{B} \cdot \mathbf{B} > 0$ .

**П.7.36.** Пусть  $\mathbf{A} = \mathbf{A}^T > 0$ ,  $\lambda_A \in [m, M]$  и  $\mathbf{A} \neq \beta \mathbf{E}$ . Доказать, что  $\mu(\mathbf{A} + \alpha \mathbf{E})$  монотонно убывает по  $\alpha > 0$ .

**П.7.37.** Найти область сходимости метода Якоби и метода Зейделя для систем с матрицами вида:

$$\text{а) } \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}, \text{ б) } \begin{pmatrix} \alpha & 0 & \beta \\ 0 & \alpha & 0 \\ \beta & 0 & \alpha \end{pmatrix}, \text{ в) } \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix}, \text{ г) } \begin{pmatrix} \alpha & \alpha & 0 \\ \alpha & \beta & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

**П.7.38.** Показать, что число обусловленности СЛАУ с симметричной матрицей  $\mathbf{A}\mathbf{u} = \mathbf{f}$  равно  $\mu = \frac{\max_k |\lambda_A^k|}{\min_k |\lambda_A^k|}$ .

**П.7.39.** Доказать, что для систем линейных уравнений второго порядка ( $n = 2$ ) методы Якоби и Гаусса–Зейделя сходятся и расходятся одновременно.

**П.7.40.** Показать, что существует система уравнений третьего порядка, для которой метод Якоби сходится, а метод Гаусса–Зейделя расходится.

**П.7.41.** Показать, что существует система уравнений третьего порядка, для которой метод Гаусса–Зейделя сходится, а метод Якоби расходится.

## П.8. Задачи с решениями

**П.8.1.** Даны система линейных уравнений

$$\begin{pmatrix} 18 & -6 & -7 \\ -6 & 6 & 0 \\ -7 & 0 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -13 \\ 6 \\ 6 \end{pmatrix}.$$

Вычислить число обусловленности матрицы  $\mathbf{A}$  в трех нормах.

**Решение.** Для этого необходимо вычислить обратную матрицу и собственные значения матрицы (очевидно, что матрица самосопряженная!)  
Обратная матрица

$$\begin{pmatrix} 18 & -6 & -7 \\ -6 & 6 & 0 \\ -7 & 0 & 6 \end{pmatrix}^{-1} = \frac{1}{138} \begin{pmatrix} 36 & 36 & 42 \\ 36 & 59 & 42 \\ 42 & 42 & 72 \end{pmatrix}.$$

Число обусловленности в первой и второй нормах:

$$(18 + |-6| + |-7|) \cdot (42 + 42 + 72) = 31 \cdot 156 / 138 \approx 35.$$

Собственные значения определяются уравнением

$$(18 - \lambda)(6 - \lambda)^2 - 49(6 - \lambda) - 36(6 - \lambda) = 0.$$

Его решения:  $\lambda = 1$ ,  $\lambda = 6$ ,  $\lambda = 23$ . Число обусловленности в евклидовой норме равно 23, и оно минимально из чисел обусловленности в трех нормах.

### II.8.2. Для системы линейных алгебраических уравнений $\mathbf{Ax} = \mathbf{f}$

$$\mathbf{A} = \begin{pmatrix} -2 & 1 \\ 0 & 2 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$$

построить сходящийся вариант метода простых итераций. Оценить оптимальное значение итерационного параметра. Оценить число итераций, необходимое для достижения точности  $10^{-3}$  по невязке, если начальное приближение к решению  $\mathbf{x}^0 = (0, 0)^T$ .

**Решение.** Матрица системы — несамосопряженная, не положительная. Можно сделать ее самосопряженной положительной матрицей, умножив обе части системы на сопряженную.

Тогда система перейдет в равносильную линейную систему  $\mathbf{Bx} = \mathbf{g}$ ,

$$\mathbf{B} = \begin{pmatrix} 4 & -2 \\ -2 & 5 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} -6 \\ 5 \end{pmatrix}.$$

Собственные значения матрицы  $\mathbf{B}$  есть  $(9 \pm \sqrt{17})/2$ . Итерационный метод есть  $\mathbf{x}^{n+1} = (\mathbf{E} - \tau\mathbf{B})\mathbf{x}^n + \tau\mathbf{f}$ . Оптимальное значение параметра  $2/9$ . За одну итерацию невязка убывает в  $\sqrt{17}/9$  раз. В силу того, что начальное приближение нулевое, норма начальной невязки равна норме правой части,  $\|\mathbf{R}\|_0 = \sqrt{6^2 + 5^2} = \sqrt{61}$ . Для числа итераций в третьей норме имеем уравнение  $(\sqrt{17}/9)^N \cdot \sqrt{61} = 10^{-3}$ . Проводя вычисления, получим  $N = 13$ .

### II.8.3. Получить оценку относительной погрешности решения СЛАУ при точном задании матрицы системы и фиксированной относительной погрешности правой части в зависимости от правой части системы.

**Решение.** Обычно для СЛАУ правая часть известна, но для погрешности правой части мы имеем только некоторые оценки. Обусловленность конкретной системы (в отличие от обусловленности матрицы), вообще говоря, должна зависеть от  $\mathbf{f}$  и от  $\Delta\mathbf{f}$ . Тогда число обусловленности системы удовлетворяет неравенству

$$v(\mathbf{f}, \Delta\mathbf{f}) \geq \frac{\delta u}{\delta f} = \frac{\|\Delta\mathbf{u}\|}{\|\mathbf{u}\|} \cdot \frac{\|\mathbf{f}\|}{\|\Delta\mathbf{f}\|}.$$

Его можно определить как точную верхнюю грань отношения  $\frac{\delta u}{\delta f}$  по  $\Delta f$ , что соответствует наихудшей ситуации. Тогда

$$v(f) = \sup_{\Delta f} \left( \frac{\|\Delta u\|}{\|u\|} \frac{\|f\|}{\|\Delta f\|} \right) = \sup_{\Delta f} \left( \frac{\|f\|}{\|u\|} \frac{\|\Delta u\|}{\|\Delta f\|} \right) = \frac{\|f\|}{\|u\|} \sup_{\Delta f} \frac{\|\mathbf{A}^{-1} \Delta f\|}{\|\Delta f\|} = \frac{\|f\|}{\|u\|} \|\mathbf{A}^{-1}\|.$$

Оценим точную верхнюю грань и точную нижнюю грань этого выражения:

$$\sup_f \left( \|\mathbf{A}^{-1}\| \cdot \frac{\|f\|}{\|u\|} \right) = \|\mathbf{A}^{-1}\| \cdot \sup_f \frac{\|\mathbf{A}u\|}{\|u\|} = \|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\| = \mu(\mathbf{A}),$$

с другой стороны,

$$\inf_f v(f) = \inf_f \left( \|\mathbf{A}^{-1}\| \cdot \frac{\|f\|}{\|u\|} \right) = \|\mathbf{A}^{-1}\| \left( \sup_f \frac{\|u\|}{\|f\|} \right)^{-1} = \|\mathbf{A}^{-1}\| \left( \sup_f \frac{\|\mathbf{A}^{-1} f\|}{\|f\|} \right)^{-1} = 1.$$

Параметр  $v(f)$ , характеризующий обусловленность системы, зависит от правых частей. Более тонкая его оценка есть  $v(f) = \|\mathbf{A}^{-1}\| \frac{\|f\|}{\|u\|}$ , причем

$1 \leq v(f) \leq \mu$ . Так как такую оценку провести не всегда возможно, то чаще используется точная верхняя грань  $\|\mathbf{A}^{-1}\| \|\mathbf{A}\|$ .

**П.8.4.** При решении СЛАУ методом Гаусса из-за погрешностей округления возникла ненулевая невязка. Предложить процедуру уточнения численного решения СЛАУ, если невязка известна.

Решение. Полученное решение можно улучшить следующим образом. Пусть  $\mathbf{r}^1 = \mathbf{f} - \mathbf{Au}^1$  есть невязка, допущенная при решении рассматриваемой системы ( $\mathbf{u}^1$  — полученное численное решение) за счет ошибки округлений. Очевидно, что погрешность  $\boldsymbol{\varepsilon}^1 = \mathbf{u} - \mathbf{u}^1$  удовлетворяет СЛАУ  $\mathbf{A}\boldsymbol{\varepsilon}^1 = \mathbf{r}^1$ , так как  $\mathbf{A}\boldsymbol{\varepsilon}^1 = \mathbf{Au} - \mathbf{Au}^1 = \mathbf{f} - \mathbf{Au}^1$ . Решив последнюю систему, получаем  $\boldsymbol{\varepsilon}^1$ , после чего уточняем решение

$$\mathbf{u}^2 = \mathbf{u}^1 + \boldsymbol{\varepsilon}^1.$$

Если после такого уточнения невязка велика, то эту процедуру можно продолжить.

**П.8.5.** Записать формулы метода Гаусса в виде последовательности умножения исходной матрицы на соответствующие матрицы элементарных преобразований.

**Решение.** Рассмотрим метод Гаусса с позиции операций с матрицами. Пусть  $\mathbf{A}_1$  — матрица системы после исключения первого неизвестного:

$$\mathbf{A}_1 = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^1 & a_{23}^1 & \dots & a_{2n}^1 \\ 0 & a_{32}^1 & a_{33}^1 & \dots & a_{3n}^1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & a_{n2}^1 & a_{n3}^1 & \dots & a_{nn}^1 \end{pmatrix}, \quad \mathbf{f}_1 = \{f_1, f_2^1, \dots, f_n^1\}^T.$$

Введем новую матрицу

$$\mathbf{N}_1 = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -\eta_{21} & 1 & 0 & \dots & 0 \\ -\eta_{31} & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -\eta_{n1} & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Очевидно  $\mathbf{A}_1 = \mathbf{N}_1 \mathbf{A}$ ,  $\mathbf{f}_1 = \mathbf{N}_1 \mathbf{f}$ . Аналогично, после второго шага система приводится к виду  $\mathbf{A}_2 \mathbf{u} = \mathbf{f}_2$ , где  $\mathbf{A}_2 = \mathbf{N}_2 \mathbf{A}_1$ ,  $\mathbf{f}_2 = \mathbf{N}_2 \mathbf{f}_1$ ,

$$\mathbf{A}_1 = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^1 & a_{23}^1 & \dots & a_{2n}^1 \\ 0 & 0 & a_{33}^2 & \dots & a_{3n}^2 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & a_{n3}^2 & \dots & a_{nn}^2 \end{pmatrix},$$

$$\mathbf{N}_2 = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & -\eta_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & -\eta_{n2} & 0 & \dots & 1 \end{pmatrix}, \quad \mathbf{f}_2 = \{f_1, f_2^1, f_3^2, \dots, f_n^2\}^T.$$

После  $n-1$  шага получим  $\mathbf{A}_{n-1} \mathbf{u} = \mathbf{f}_{n-1}$ ,  $\mathbf{A}_{n-1} = \mathbf{N}_{n-1} \cdot \mathbf{A}_{n-2}$ ,  $\mathbf{f}_{n-1} = \mathbf{N}_{n-1} \mathbf{f}_{n-2}$ ,  $\mathbf{f}_{n-1} = \{f_1, f_2^1, f_3^2, \dots, f_n^{n-1}\}^T$ .

$$\mathbf{A}_{n-1} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22}^1 & a_{23}^1 & \dots & a_{2n}^1 \\ 0 & 0 & a_{33}^2 & \dots & a_{3n}^2 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & a_{nn}^{(n-1)} \end{pmatrix},$$

$$\mathbf{N}_{n-1} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & -\eta_{n,n-1} & 1 \end{pmatrix}.$$

В итоге получаются матрица и вектор  $\mathbf{A}_{n-1} = \mathbf{N}_{n-1} \dots \mathbf{N}_2 \mathbf{N}_1 \mathbf{A}$ ,  $\mathbf{f}_{(n-1)} = \mathbf{N}_{n-1} \dots \mathbf{N}_2 \mathbf{N}_1 \mathbf{f}$ , откуда  $\mathbf{A} = \mathbf{N}_1^{-1} \mathbf{N}_2^{-1} \dots \mathbf{N}_{n-1}^{-1} \cdot \mathbf{A}_{n-1}$ . При этом

$$\mathbf{N}_1^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \eta_{21} & 1 & 0 & \dots & 0 \\ \eta_{31} & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \eta_{n1} & 0 & 0 & \dots & 1 \end{pmatrix},$$

$$\mathbf{N}_2^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & \eta_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \eta_{n2} & 0 & \dots & 1 \end{pmatrix},$$

$$\mathbf{N}_{n-1}^{-1} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & \eta_{n,n-1} & 1 \end{pmatrix}.$$

После введения обозначений  $\mathbf{U} = \mathbf{A}_{n-1}$ ,  $\mathbf{L} = \mathbf{N}_1^{-1} \mathbf{N}_2^{-1} \dots \mathbf{N}_{n-1}^{-1}$ , где

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \eta_{21} & 1 & 0 & \dots & 0 \\ \eta_{31} & \eta_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \eta_{n1} & \eta_{n2} & \eta_{n3} & \dots & 1 \end{pmatrix},$$

получим  $\mathbf{A} = \mathbf{LU}$ .

## II.9. Теоретические задачи

**II.9.1.** Даны системы линейных алгебраических уравнений  $\mathbf{Ax} = \mathbf{b}$ :

a)  $\mathbf{A} = \begin{pmatrix} 101 & 110 \\ 110 & 122 \end{pmatrix}; \mathbf{b} = \begin{pmatrix} 312 \\ 342 \end{pmatrix};$  б)  $\mathbf{A} = \begin{pmatrix} 101 & -110 \\ -110 & 122 \end{pmatrix}; \mathbf{b} = \begin{pmatrix} 92 \\ -98 \end{pmatrix};$

в)  $\mathbf{A} = \begin{pmatrix} 82 & 90 \\ 90 & 101 \end{pmatrix}; \mathbf{b} = \begin{pmatrix} 254 \\ 281 \end{pmatrix};$  г)  $\mathbf{A} = \begin{pmatrix} 101 & -90 \\ -90 & 82 \end{pmatrix}; \mathbf{b} = \begin{pmatrix} 112 \\ -98 \end{pmatrix}.$

Найти  $\tau_{\text{опт}}$ , при котором метод простой итерации  $\mathbf{x}^{k+1} = (\mathbf{E} - \tau \mathbf{A})\mathbf{x}^k + \tau \mathbf{b}$  будет сходиться быстрее всего. Оценить скорость сходимости.

Пусть  $\tau$  принадлежит интервалу  $0 < \tau < \tau_{\text{опт}}$ . Получить оценку скорости сходимости в этом случае. Можно ли так задать вектор начального приближения  $\mathbf{x}^0$ , чтобы скорость сходимости в этом случае была бы выше, чем при оптимальном  $\tau_{\text{опт}}$ ? Если это возможно, то указать такое  $\mathbf{x}^0$ .

**II.9.2.** Даны системы линейных алгебраических уравнений  $\mathbf{Ax} = \mathbf{b}$ .

Оценить максимально точно относительную погрешность  $\|\Delta \mathbf{x}\| / \|\mathbf{x}\|$  в заданной норме. Найти вектор ошибки  $\Delta \mathbf{b}$ , на котором эта оценка достигается. При каком  $\Delta \mathbf{b}$  относительная ошибка  $\|\Delta \mathbf{x}\| / \|\mathbf{x}\|$  будет минимальной? Найти ее.

а)  $\mathbf{A} = \begin{pmatrix} 101 & 110 \\ 110 & 122 \end{pmatrix}; \mathbf{b} = \begin{pmatrix} 312 \\ 342 \end{pmatrix}; \frac{\|\Delta \mathbf{b}\|_1}{\|\mathbf{b}\|_1} = 0.01, \quad \|\mathbf{x}\|_1 = \max_i |x_i|;$

б)  $\mathbf{A} = \begin{pmatrix} 101 & -110 \\ -110 & 122 \end{pmatrix}; \mathbf{b} = \begin{pmatrix} 92 \\ -98 \end{pmatrix}; \frac{\|\Delta \mathbf{b}\|_2}{\|\mathbf{b}\|_2} = 0.01, \quad \|\mathbf{x}\|_2 = \sum_i |x_i|;$

в)  $\mathbf{A} = \begin{pmatrix} 101 & -90 \\ -90 & 82 \end{pmatrix}; \mathbf{b} = \begin{pmatrix} 112 \\ -98 \end{pmatrix}; \frac{\|\Delta \mathbf{b}\|_2}{\|\mathbf{b}\|_2} = 0.01, \quad \|\mathbf{x}\|_2 = \sum_i |x_i|;$

$$\text{г) } \mathbf{A} = \begin{pmatrix} 65 & 72 \\ 72 & 82 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} 137 \\ 154 \end{pmatrix}; \quad \frac{\|\Delta\mathbf{b}\|_3}{\|\mathbf{b}\|_3} = 0.01, \quad \|\mathbf{x}\|_3 = \sqrt{(\mathbf{x}, \mathbf{x})};$$

$$\text{д) } \mathbf{A} = \begin{pmatrix} 50 & 70 \\ 70 & 101 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} 120 \\ 171 \end{pmatrix}; \quad \frac{\|\Delta\mathbf{b}\|_3}{\|\mathbf{b}\|_3} = 0.01, \quad \|\mathbf{x}\|_3 = \sqrt{(\mathbf{x}, \mathbf{x})}.$$

**П.9.3.** Данна система линейных алгебраических уравнений  $\mathbf{Ax} = \mathbf{b}$ :

$$\text{а) } \mathbf{A} = \begin{pmatrix} 65 & 80 \\ 80 & 101 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} 145 \\ 181 \end{pmatrix}; \quad \text{б) } \mathbf{A} = \begin{pmatrix} 65 & 72 \\ 72 & 82 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} 137 \\ 154 \end{pmatrix};$$

$$\text{в) } \mathbf{A} = \begin{pmatrix} 50 & 70 \\ 70 & 101 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} 120 \\ 171 \end{pmatrix}.$$

Найти  $\tau_{\text{опт}}$ , при котором метод простой итерации  $\mathbf{x}^{k+1} = (\mathbf{E} - \tau\mathbf{A})\mathbf{x}^k + \tau\mathbf{b}$  будет сходиться быстрее всего. Оценить скорость сходимости.

Пусть  $\tau$  принадлежит интервалу  $\tau_{\text{опт}} < \tau < \tau_{\text{max}}$ . Получить оценку скорости сходимости в этом случае. Можно ли так задать вектор начального приближения  $\mathbf{x}^0$ , чтобы скорость сходимости в этом случае была бы выше, чем при оптимальном  $\tau_{\text{опт}}$ ? Если это возможно, то указать такое  $\mathbf{x}^0$ .

**П.9.4.** Найти области сходимости методов простой итерации и Зейделя для систем  $\mathbf{x} = \mathbf{Rx} + \mathbf{g}$ , где  $\mathbf{R} = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$ .

**П.9.5.** Для системы уравнений выписать итерационные формулы вычисления решения, используя диагональное преобладание. Какой метод при этом получается? Сколько итераций достаточно, чтобы уменьшить погрешность исходного приближения в тысячу раз?

$$10x + y - z = 1,$$

$$x - 20y + 3z = 2,$$

$$2x + 3y - 10z = -1.$$

**П.9.6.** Предположим, что некоторая система размерности  $n \times n$  вида  $\mathbf{x} = \mathbf{Rx} + \mathbf{b}$  с матрицей, имеющей норму  $\|\mathbf{R}\| \approx 0.5$ , решается методом простых итераций с уровнем абсолютных погрешностей округления арифметических операций порядка  $10^{-6}$ . Допустим, что при этом первая итерационная поправка имеет норму  $\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \approx 1$ . Каким числом следует ограничить количество итераций, чтобы вычислительная погрешность не стала существенно превышать погрешность метода?

### II.9.7. Для системы уравнений

$$10^{-3}u_1 + u_2 = f_1,$$

$$u_1 - u_2 = f_2$$

ответить на следующие вопросы.

а) Каково число обусловленности  $\mu$  системы, если в качестве нормы произвольного вектора  $\mathbf{u}$  используется  $\|\mathbf{u}\| = \max\{|u_1|, |u_2|\}$ ?

б) Какова допустимая относительная погрешность при задании  $\mathbf{f} = (f_1, f_2)^T$ , при которой относительная погрешность решения не превосходит  $10^{-2}$ ?

в) Пусть  $f_1 = 2, f_2 = 1$ . С каким числом знаков надо ввести вычисления по методу Гаусса без выбора главного элемента, чтобы  $\{u_1, u_2\}$  имели хотя бы по одному верному десятичному знаку?

Тот же вопрос для метода Гаусса с выбором главного элемента.

### II.9.8. Используя метод Гаусса, найти численно решения двух СЛАУ:

$$\begin{cases} u + 3v &= 4, \\ u + 3.00001v &= 4.00001 \end{cases} \text{ и } \begin{cases} u + 3v &= 4, \\ u + 2.9999v &= 4.00001 \end{cases}$$

и объяснить результат. Как ответ будет зависеть от длины мантиссы, отведенной под хранение числа?

### II.9.9. Для СЛАУ

$$\left\{ \begin{array}{lcl} 10u_1 + u_2 & = 1, \\ u_1 + 10u_2 + u_3 & = 2, \\ u_2 + 10u_3 + u_4 & = 3, \\ & \dots & \dots \\ u_{98} + 10u_{99} + u_{100} & = 99, \\ u_1 + u_2 + \dots + u_{99} + u_{100} & = a, \end{array} \right.$$

где  $a$  – параметр, описать алгоритм метода Гаусса без выбора главного элемента при  $a = 100$ . Предложить алгоритм экономичного решения данной системы уравнений, если нужно получить решение системы для набора значений параметра  $a$ .

II.9.10. Найти область значений итерационного параметра  $\tau$ , при которых итерационный процесс  $\mathbf{x}^{k+1} = (\mathbf{E} - \tau\mathbf{A})\mathbf{x}^k + \tau\mathbf{f}$  сходится, если  $\operatorname{Re}\{\lambda(\mathbf{A})\} \geq \delta > 0$ .

**П.9.11.** Запишите итерационный процесс Якоби нахождения решения системы

$$\begin{cases} 5x_1 + 2x_2 - x_3 + x_4 = 9, \\ x_1 - 4x_2 + 2x_4 = 10, \\ 2x_1 + 3x_2 - 9x_3 - x_4 = -10, \\ 3x_1 + x_3 - 6x_4 = -5. \end{cases}$$

Каким должен быть критерий окончания процесса итерирования, чтобы максимальная из абсолютных погрешностей компонент приближенного решения не превышала заданного малого  $\varepsilon$ ?

**П.9.12.** Предложить способ решения системы методом простых итераций с оптимальным параметром. Найти оптимальный параметр и количество итераций, необходимое для достижения точности  $10^{-4}$ :

a)  $\begin{cases} 3x + z = 6, \\ x - y + z = 2, \\ 3y - z = 3. \end{cases}$  б)  $\begin{cases} 2x - 0.5y = 2, \\ 0.5x - 3y + 0.5z = -12, \\ 4z - 0.5y = 6. \end{cases}$

В качестве начального приближения берется  $\mathbf{x}^0 = (0, 0, 0)^T$ .

**П.9.13.** Система уравнений  $\mathbf{Ax} = \mathbf{f}$ , где

$$\mathbf{A} = \begin{pmatrix} 0.5 & -0.5 & 0.5 \\ 1 & 2 & -1 \\ -0.5 & 0.5 & 3.5 \end{pmatrix}, \mathbf{f} = (0, 3, 2)^T,$$

решается с помощью метода  $\mathbf{x}^{n+1} = (\mathbf{E} - 2/5\mathbf{A})\mathbf{x}^n + 2/5 \mathbf{f}$ , начальное приближение нулевое. Оценить число итераций, необходимое для уменьшения первоначальной невязки в  $10^4$  раз.

Выписать расчетные формулы метода Якоби. Исследовать его на сходимость.

**П.9.14.** Система уравнений  $\mathbf{Ax} = \mathbf{f}$ , где

$$\mathbf{A} = \begin{pmatrix} 3 & -1 & 1 \\ 2 & 6 & -2 \\ -1 & 1 & 9 \end{pmatrix}, \mathbf{f} = (4, -4, 2)^T,$$

решается с помощью метода простых итераций, начальное приближение нулевое. При каких значениях итерационного параметра метод будет сходиться

а) при вычислениях с бесконечным числом бит в мантиссе?

- б) при длине мантиссы 52 бита?  
 в) при каком значении итерационного параметра сходимость будет самая быстрая?  
 г) Найти значение оптимального параметра для первого шага решения системы методом наискорейшего спуска.

**П.9.15.** Проверить выполнение необходимых условий сходимости методов Якоби и Зейделя, примененных к системе

$$\begin{aligned}x_1 + x_2 &= 2, \\x_1 + 2x_2 + x_3 &= 4, \\x_2 + 2x_3 &= 3.\end{aligned}$$

**П.9.16.** Пусть методом Якоби решение системы

$$b_i x_{i-1} + c_i x_i + a_i x_{i+1} = f_i, \quad i = 1, 2, \dots, n; \quad b_1 = a_n = 0$$

с нужной точностью достигается за  $k$  шагов. Существуют ли такие  $k$  и  $n$ , для которых применение метода Якоби в этой ситуации эффективнее метода прогонки по числу арифметических операций?

**П.9.17.** Для линейной системы

$$10x_1 + 2x_2 + 3x_3 + 4x_4 = 1,$$

$$2x_1 + 5x_2 + x_3 = -6,$$

$$3x_1 + x_2 + 10x_3 - x_4 = -7,$$

$$4x_1 + x_3 + 10x_4 = -6$$

выписать формулы метода Зейделя в компонентах, исследовать его сходимость. Выписать также расчетные формулы метода последовательной верхней релаксации.

**П.9.18.** А) Проанализируйте сходимость степенного метода в случае, когда  $\lambda_1$  — кратное вещественное наибольшее по модулю собственное число  $n$ -мерной матрицы простой структуры.

Как можно найти все соответствующие ему собственные векторы в зависимости от показателя кратности?

Б) Что можно сказать о поведении последовательности приближений вычисления максимального по модулю собственного значения степенным методом, если  $\lambda_1 = -\lambda_2$ ,  $|\lambda_2| > |\lambda_3| \geq \dots \geq |\lambda_n|$ ,  $\lambda_i \in R$  ?

В) Рассмотрите и объясните поведение степенного метода в случае, когда данная матрица  $A$  — диагональная.

**П.9.19.** Исследовать на сходимость метод Якоби для решения системы уравнений с матрицей  $\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 1 & 6 \end{pmatrix}$ .

**П.9.20.** При каких значениях параметра  $\tau$  метод

$$\mathbf{x}^{k+1} = (\mathbf{E} - \tau \mathbf{A}) \mathbf{x}^k + \tau \mathbf{b}$$

сходится с произвольно взятого начального приближения для системы линейных уравнений  $\mathbf{Ax} = \mathbf{b}$  с матрицей  $\mathbf{A} = \begin{pmatrix} 3 & 8 \\ 2 & 9 \end{pmatrix}$ ?

**П.9.21.** (В.Б. Пирогов). Данна система линейных уравнений  $\mathbf{Ax} = \mathbf{b}$ .

1. Для заданной относительной погрешности правой части найти границы для относительной погрешности  $\|\Delta \mathbf{x}\|/\|\mathbf{x}\|$  решения заданной системы в той же норме, в которой задана погрешность правой части.

2. Исследовать на сходимость и оценить скорость сходимости метода простой итерации  $\mathbf{x}^{(k+1)} = (\mathbf{E} - \tau \mathbf{A}) \mathbf{x}^{(k)} + \tau \mathbf{b}$  при заданном параметре  $\tau$ .

3. Найти  $\tau_{\text{опт}}$  и дать оценку скорости сходимости при этом  $\tau_{\text{опт}}$ .

4. Задано начальное приближение  $\mathbf{x}^{(0)}$ ; найти для него скорость сходимости при заданном  $\tau$ , а также новое значение оптимального параметра  $\tau'_{\text{опт}}$  (почему оно возникает?) и оценку скорости сходимости при этом  $\tau'_{\text{опт}}$ .

5. Выписать формулы для итерационного процесса Якоби и доказать его сходимость.

6. Выписать формулы для итерационного процесса Зейделя и доказать его сходимость.

a)  $\mathbf{A} = \begin{pmatrix} 18 & -6 & -7 \\ -6 & 6 & 0 \\ -7 & 0 & 6 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} -13 \\ 6 \\ 6 \end{pmatrix},$

$$\frac{\|\Delta \mathbf{b}\|_1}{\|\mathbf{b}\|_1} = 0.01, \quad \|\mathbf{x}\|_1 = \max_k |x_k|, \quad \tau = 0.02, \quad \mathbf{x}^{(0)} = (17, 2, -12)^T;$$

б)  $\mathbf{A} = \begin{pmatrix} 4 & -3 & 0 \\ -3 & 16 & 6 \\ 0 & 6 & 4 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 4 \\ 3 \\ 4 \end{pmatrix},$

$$\frac{\|\Delta \mathbf{b}\|_2}{\|\mathbf{b}\|_2} = 0.01, \quad \|\mathbf{x}\|_2 = \sum_k |x_k|, \quad \tau = 0.01, \quad \mathbf{x}^{(0)} = (2, 5, 4)^T;$$

в)  $\mathbf{A} = \begin{pmatrix} 13 & -4 & 4 \\ -4 & 21 & 8 \\ 4 & 8 & 7 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 9 \\ 17 \\ 12 \end{pmatrix},$

$$\|\Delta \mathbf{b}\|_2 / \|\mathbf{b}\|_2 = 0.01 (\|\mathbf{x}\|_2 = \sum_i |x_i|), \quad \tau = 0.03.$$

$$\frac{\|\Delta \mathbf{b}\|_2}{\|\mathbf{b}\|_2} = 0.01, \quad \|\mathbf{x}\|_2 = \sum_k |x_k|, \quad \tau = 0.03, \quad \mathbf{x}^{(0)} = (2, 6, 3)^T.$$

**П.9.22.** (В.Б. Пирогов). Данна система линейных уравнений  $\mathbf{Ax} = \mathbf{b}$ . Для этой СЛАУ исследовать на сходимость и найти оптимальное значение параметра для метода простой итерации  $\mathbf{x}^{(k+1)} = (\mathbf{E} - \tau \mathbf{A})\mathbf{x}^{(k)} + \tau \mathbf{b}$ , проводимого от заданного начального приближения  $\mathbf{x}^{(0)}$ , при вычислениях

- 1) на идеальном, т.е. бесконечноразрядном, компьютере;
- 2) с одинарной точностью.

а)  $\mathbf{A} = \begin{pmatrix} 18 & -6 & -7 \\ -6 & 6 & 0 \\ -7 & 0 & 6 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} -13 \\ 6 \\ 6 \end{pmatrix}, \quad \mathbf{x}^{(0)} = \begin{pmatrix} 6 \\ 14 \\ 1 \end{pmatrix};$

б)  $\mathbf{A} = \begin{pmatrix} 4 & -3 & 0 \\ -3 & 16 & 6 \\ 0 & 6 & 4 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 4 \\ 3 \\ 4 \end{pmatrix}, \quad \mathbf{x}^{(0)} = \begin{pmatrix} 4 \\ 1 \\ 0 \end{pmatrix};$

в)  $\mathbf{A} = \begin{pmatrix} 13 & -4 & 4 \\ -4 & 21 & 8 \\ 4 & 8 & 7 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 9 \\ 17 \\ 12 \end{pmatrix}, \quad \mathbf{x}^{(0)} = \begin{pmatrix} 4 \\ 2 \\ -1 \end{pmatrix}.$

**П.9.23.** При каком векторе  $\mathbf{b}$  и произвольной погрешности  $\Delta \mathbf{b}$ , допущенной при его задании, достигается  $\mu$  — максимальная величина в оценке относительной погрешности решения  $\frac{\|\Delta \mathbf{x}\|_3}{\|\mathbf{x}\|_3} \leq \mu \frac{\|\Delta \mathbf{b}\|_3}{\|\mathbf{b}\|_3}$  для системы линейных уравнений  $\mathbf{Ax} = \mathbf{b}$ , где  $\|\mathbf{x}\|_3 = \sqrt{(\mathbf{x}, \mathbf{x})}$ .

Чему в этом случае равно  $\mu$ ?

$$\mathbf{A} = \begin{pmatrix} -6 & 8 \\ -8 & 6 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

**П.9.24.** Найти все возможные значения параметра  $a$ , при которых, с учетом верно выбранного  $\tau$ , метод простых итераций решения системы уравнений

$$\begin{pmatrix} a & a-1 \\ 2 & a+2 \end{pmatrix} \mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{x}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

имеет наибольшую возможную скорость сходимости к точному решению.

**П.9.25.** Для системы линейных алгебраических уравнений  $\mathbf{Ax} = \mathbf{b}$

$$\mathbf{A} = \begin{pmatrix} 5 & -1 & 2 \\ 1 & 4 & -1 \\ 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 8 \\ -4 \\ 4 \end{pmatrix}$$

построить итерационный метод Зейделя. Найти первое и второе приближения по методу Зейделя, задав в качестве начального приближения нулевой вектор. Доказать сходимость метода.

**П.9.26.** Система уравнений  $\mathbf{Au} = \mathbf{f}$  решается с помощью метода простой итерации. Выбрано нулевое начальное приближение. Найдите, при каких значениях итерационного параметра метод будет сходиться, а также его оптимальное значение

- а) при вычислениях с бесконечным числом бит в мантиссе?
- б) при длине мантиссы 50 бит?

Рассмотреть следующие варианты задания матрицы системы и вектора правой части:

$$1) \quad \mathbf{A} = \begin{pmatrix} 10 & -3 & -2 \\ -3 & 5 & -3 \\ -2 & -3 & 10 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 4 \\ 7 \\ -8 \end{pmatrix}, \quad 2) \quad \mathbf{A} = \begin{pmatrix} 10 & -3 & -2 \\ -3 & 5 & -3 \\ -2 & -3 & 10 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 22 \\ -11 \\ 0 \end{pmatrix},$$

$$3) \quad \mathbf{A} = \begin{pmatrix} 10 & -3 & -2 \\ -3 & 5 & -3 \\ -2 & -3 & 10 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 11 \\ -11 \\ 11 \end{pmatrix}, \quad 4) \quad \mathbf{A} = \begin{pmatrix} 10 & -3 & -2 \\ -3 & 5 & -3 \\ -2 & -3 & 10 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 5 \\ -1 \\ 5 \end{pmatrix},$$

$$5) \quad \mathbf{A} = \begin{pmatrix} 10 & 9 & -6 \\ 9 & 10 & -6 \\ -6 & -6 & 13 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 14 \\ 0 \\ 14 \end{pmatrix}.$$

**П.9.27.** Исследовать на сходимость метод Зейделя для решения системы уравнений с матрицей  $\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 4 & 5 \end{pmatrix}$ .

**П.9.28.** Для следующих СЛАУ определить те, для которых можно найти оптимальный параметр метода верхней релаксации и найти его. Для этих систем выписать формулы метода верхней релаксации в правильной последовательности вычисления компонент и сделать по три итерации от нулевого начального вектора с параметром релаксации, близким к оптимальному. Сравнить результат с тремя итерациями метода Зейделя.

$$a) \begin{cases} 18x - 6y - 7z = 5, \\ -6x + 6y = 0, \\ -7x + 6z = -1. \end{cases} b) \begin{cases} -9x + 7y + 5z = 4, \\ 7x + 8y + 9z = -7, \\ 5x + 9y + 8z = 19. \end{cases} v) \begin{cases} 4x - 3y = -7, \\ -3x + 16y + 6z = 13, \\ 6y + 4z = 2. \end{cases}$$

**П.9.29.** Пусть вещественная матрица  $\mathbf{A}$  системы линейных уравнений порядка  $m$   $\mathbf{Ax} = \mathbf{f}$ ,  $\mathbf{x} = \{x_i\}$ ,  $i = 1, \dots, m$  симметрична, и ее наименьшее и наибольшее собственные числа  $\lambda_{\min}$  и  $\lambda_{\max}$  положительны. Введена норма  $\|\mathbf{y}\| = (y_1^2 + y_2^2 + \dots + y_m^2)^{1/2}$ .

a) Подобрать параметр  $\tau$  так, чтобы в методе последовательных приближений

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \tau (\mathbf{Ax}^n - \mathbf{f}), n = 0, 1, 2, \dots$$

$\mathbf{x}^0$  — задан, норма погрешности  $\boldsymbol{\varepsilon}^n = \mathbf{x}^n - \mathbf{x}^*$ , где  $\mathbf{x}^*$  — вектор-решение, убывала наиболее быстро.

б) Подобрать пару итерационных параметров  $\tau_1, \tau_2$  так, чтобы в методе последовательных приближений

$$\mathbf{z} = \mathbf{x}^{(n)} - \tau_1 (\mathbf{Ax}^{(n)} - \mathbf{f}), \quad \mathbf{x}^{(n+1)} = \mathbf{z} - \tau_2 (\mathbf{Az} - \mathbf{f}), \quad n = 0, 1, \dots,$$

вектор  $\mathbf{x}^{(0)}$  задан, норма погрешности  $\boldsymbol{\varepsilon}^n$  убывала возможно быстрее.

в) Пусть  $\lambda_{\min} = 1, \lambda_{\max} = 10$ . Во сколько раз больше арифметических операций потребуется для уменьшения первоначальной погрешности в заданное число раз при использовании первого итерационного алгоритма по сравнению со вторым?

**П.9.30.** Степенной метод поиска максимального по абсолютной величине собственного значения матрицы применен к несамосопряженной матрице, все собственные числа которой действительны. Получится ли при этом адекватное приближение собственного числа? Какова точность данного приближения?

## II.10. Практические задачи

II.10.1. Используя степенной метод, оценить спектральный радиус матрицы  $\mathbf{A}$  с погрешностью  $\epsilon = 0.1$ :

$$\text{а) } \mathbf{A} = \begin{pmatrix} -7 & 4 & 5 \\ 4 & -6 & -9 \\ 5 & -9 & -8 \end{pmatrix}, \text{ б) } \mathbf{A} = \begin{pmatrix} -9 & 7 & 5 \\ 7 & 8 & 9 \\ 5 & 9 & 8 \end{pmatrix}, \text{ в) } \mathbf{A} = \begin{pmatrix} 5 & 5 & 3 \\ 5 & -4 & 1 \\ 3 & 1 & 2 \end{pmatrix},$$

$$\text{г) } \mathbf{A} = \begin{pmatrix} 8 & 2 & -1 \\ 2 & -5 & -8 \\ -1 & -8 & -5 \end{pmatrix}, \text{ д) } \mathbf{A} = \begin{pmatrix} 0 & -7 & 7 \\ -7 & -9 & -5 \\ 7 & -5 & -1 \end{pmatrix}.$$

II.10.2. Сделайте по пять итераций методов Якоби и Зейделя для системы

$$10x_1 + x_2 - 2x_3 = 10,$$

$$x_1 - 5x_2 + x_3 = 10,$$

$$3x_1 - x_2 + 2x_3 = -5.$$

Сколько верных знаков можно гарантировать в приближенных решениях, полученных тем и другим способом?

II.10.3. Найдите степенным методом грубые приближения к собственным числам матрицы и уточните их обратными итерациями со сдвигом

$$\text{а) } \mathbf{A} = \begin{pmatrix} 5 & 2 & -3 \\ 4 & 5 & -4 \\ 6 & 4 & -4 \end{pmatrix}, \text{ б) } \mathbf{A} = \begin{pmatrix} 4 & 2 & -1 \\ 2 & 4 & 1 \\ -1 & 1 & 3 \end{pmatrix}.$$

II.10.4. Для системы линейных алгебраических уравнений  $\mathbf{Ax} = \mathbf{f}$

- 1) вычислить число обусловленности системы в нормах, подчиненных кубической, октаэдрической и евклидовой норме вектора.
- 2) привести формулы и выполнить три итерации методов Якоби, Зейделя и верхней релаксации, выбрав итерационный параметр, близкий к оптимальному. За начальное приближение взять вектор  $\mathbf{x} = (0,0)^T$ .
- 3) провести три шага вычислений для определения максимального по модулю собственного значения матрицы системы степенным методом, взяв в качестве начального приближения вектор  $\mathbf{x} = (1,0)^T$ .

$$\text{а) } \mathbf{A} = \begin{pmatrix} 6 & 1 \\ 1 & 6 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 4 \\ -11 \end{pmatrix}, \text{ б) } \mathbf{A} = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

$$\text{в) } \mathbf{A} = \begin{pmatrix} -3 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} -1 \\ 5 \end{pmatrix}, \quad \text{г) } \mathbf{A} = \begin{pmatrix} 10 & -6 \\ -6 & 10 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 4 \\ 4 \end{pmatrix}.$$

**П.10.5.** Решить методами Гаусса и Зейделя, найти  $\lambda_{\min}$ ,  $\lambda_{\max}$ , определить число обусловленности матрицы  $\mu = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ . Сделать печать невязок обоих методов. Указать критерий останова итераций метода Зейделя.

**a)**  $n = 99$ ,  $a_i = c_i = 1$ ,  $b_i = 10$ ,  $p_i = 1$ ,  $f_i = i$ .

$$\left\{ \begin{array}{ll} b_1 x_1 + c_1 x_2 & = f_1 \\ a_2 x_1 + b_2 x_2 + c_2 x_3 & = f_2 \\ a_3 x_2 + b_3 x_3 + c_3 x_4 & = f_3 \\ & \dots \quad \dots \\ a_n x_{n-1} + b_n x_n + c_n x_{n+1} & = f_n \\ p_1 x_1 + p_2 x_2 + \dots + p_{n-1} x_{n-1} + p_n x_n + p_{n+1} x_{n+1} & = f_{n+1} \end{array} \right.,$$

**б)** система п. а) при

$$n = 9, b_1 = 1, c_1 = 0, f_1 = 1, a_i = c_i = 1,$$

$$b_i = -2, p_i = 2, f_i = 2 / i^2, i = 2, 3, \dots, n, f_{n+1} = -n / 3, p_1 = p_{n+1} = 1,$$

**в)** система п. а) при

$$n = 19, b_1 = 1, c_1 = 0, f_1 = 1,$$

$$a_i = c_i = 1, b_i = -2, p_i = 2, f_i = 2 / i^2, i = 2, 3, \dots, n,$$

$$f_{n+1} = -n / 3, p_1 = p_{n+1} = 1,$$

**г)**  $x_1 + x_2 + \dots + x_{98} + x_{99} + x_{100} = 100$ ,

$$x_1 + 10x_2 + x_3 = 99,$$

$$x_2 + 10x_3 + x_4 = 98,$$

$\dots$

$$x_{98} + 10x_{99} + x_{100} = 2,$$

$$x_{99} + x_{100} = 1,$$

д)  $a = 10$ ,

$$\left\{ \begin{array}{rcl} ax_1 + x_2 + x_3 + x_4 + x_5 & = & 1 \\ x_1 + ax_2 + x_3 + x_4 + x_5 + x_6 & = & 2 \\ x_1 + x_2 + ax_3 + x_4 + x_5 + x_6 + x_7 & = & 3 \\ x_1 + x_2 + x_3 + ax_4 + x_5 + x_6 + x_7 + x_8 & = & 4 \\ x_1 + x_2 + x_3 + x_4 + ax_5 + x_6 + x_7 + x_8 + x_9 & = & 5 \\ x_2 + x_3 + x_4 + x_5 + ax_6 + x_7 + x_8 + x_9 + x_{10} & = & 6 \\ \dots & & = \dots \\ x_{k-4} + x_{k-3} + x_{k-2} + x_{k-1} + ax_k + x_{k+1} + x_{k+2} + x_{k+3} + x_{k+4} & = & k \\ \dots & & = \dots \\ x_{93} + x_{94} + x_{95} + x_{96} + ax_{97} + x_{98} + x_{99} + x_{100} & = & 97 \\ x_{94} + x_{95} + x_{96} + x_{97} + ax_{98} + x_{99} + x_{100} & = & 98 \\ x_{95} + x_{96} + x_{97} + x_{98} + ax_{99} + x_{100} & = & 99 \\ x_{96} + x_{97} + x_{98} + x_{99} + ax_{100} & = & 100 \end{array} \right.$$

е)  $a = 10$ ,

$b_{ij} = 1 / i, f_i = i$ ,

$$\left\{ \begin{array}{rcl} ax_1 + b_{1,2}x_2 & = & f_1 \\ b_{2,1}x_1 + ax_2 + b_{2,3}x_3 & = & f_2 \\ b_{3,1}x_1 + b_{3,2}x_2 + ax_3 + b_{3,4}x_4 & = & f_3 \\ \dots & & \dots \\ b_{99,1}x_1 + b_{99,2}x_2 + \dots + b_{99,98}x_{98} + ax_{99} + b_{99,100}x_{100} & = & f_{99} \\ b_{100,1}x_1 + b_{100,2}x_2 + \dots + b_{100,98}x_{98} + b_{100,99}x_{99} + ax_{100} & = & f_{100} \end{array} \right.$$

ж) система п. д) при  $a = 20$ .

з)  $a = 10, b_{ij} = 1 / i, f_i = i$ ,

$$\left\{ \begin{array}{l} ax_1 + b_{1,2}x_2 = f_1 \\ b_{2,1}x_1 + ax_2 + b_{2,3}x_3 = f_2 \\ b_{3,1}x_1 + b_{3,2}x_2 + ax_3 + b_{3,4}x_4 = f_3 \\ \dots \quad \dots \\ b_{99,1}x_1 + b_{99,2}x_2 + \dots + b_{99,98}x_{98} + ax_{99} + b_{99,100}x_{100} = f_{99} \\ b_{100,1}x_1 + b_{100,2}x_2 + \dots + b_{100,98}x_{98} + b_{100,99}x_{99} + ax_{100} = f_{100} \end{array} \right.$$

**и)** система п. 3) при  $a = 100$ ,  $b_{ij} = i / j$ ,  $f_i = i$ ,

**к)**  $n = 10$ ,  $a_{ii} = 1$ ,  $a_{ij} = 1/(i+j)$  ( $i \neq j$ ),  $f_i = 1/i$ ,

$$\left\{ \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = f_1 \\ \dots \quad \dots \quad \dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = f_n \end{array} \right.$$

**л)** система п. к) при  $n = 20$ ,  $a_{ii} = 10$ ,  $a_{ij} = 1/(i+j)$  ( $i \neq j$ ),  $f_i = 1/i$ ,

**м)**  $n = 20$ ,  $c_i = 10$ ,  $f_i = i$ ,  $i = 1, 2, \dots, n$ ,

$$b_{i+1} = d_i = 1, \quad i = 1, 2, \dots, n-1,$$

$$a_{i+2} = e_i = 0.1, \quad i = 1, 2, \dots, n-2,$$

$$\left\{ \begin{array}{l} c_1x_1 + d_1x_2 + e_1x_3 = f_1 \\ b_2x_1 + c_2x_2 + d_2x_3 + e_2x_4 = f_2 \\ a_3x_1 + b_3x_2 + c_3x_3 + d_3x_4 + e_3x_5 = f_3 \\ a_4x_2 + b_4x_3 + c_4x_4 + d_4x_5 + e_4x_6 = f_4 \\ \dots \quad \dots \quad \dots \\ a_mx_{m-2} + b_mx_{m-1} + c_mx_m + d_mx_{m+1} + e_mx_{m+2} = f_m \\ \dots \quad \dots \quad \dots \\ a_{n-1}x_{n-3} + b_{n-1}x_{n-2} + c_{n-1}x_{n-1} + d_{n-1}x_n = f_{n-1} \\ a_nx_{n-2} + b_nx_{n-1} + c_nx_n = f_n \end{array} \right.$$

**н)** система п. м) при

$$n = 20, \quad c_i = 10, \quad f_i = i, \quad i = 1, 2, \dots, n$$

$$b_{i+1} = d_i = 1, i = 1, 2, \dots, n-1$$

$$a_{i+2} = e_i = 0, i = 1, 2, \dots, n-2,$$

**o)**  $n = 100, a = b = 10,$

$$\left\{ \begin{array}{lll} ax_1 + x_2 + x_3 / b & = & 1 \\ x_1 + ax_2 + x_3 + x_4 / b & = & 2 \\ x_2 + ax_3 + x_4 + x_5 / b & = & 3 \\ \dots & \dots & \dots \\ x_{m-1} + ax_m + x_{m+1} + x_{m+2} / b & = & m \\ \dots & \dots & \dots \\ x_{n-2} + ax_{n-1} + x_n & = & n-1 \\ x_{n-1} + ax_n & = & n \end{array} \right. ,$$

**п)**  $n = 99, a_i = c_i = 1, b_i = 10 + i, p_1 = p_{n+1} = 1,$

$$p_i = 2, f_i = i / n, i = 1, 2, \dots, n+1,$$

$$\left\{ \begin{array}{lll} b_1 x_1 + c_1 x_2 & = & f_1 \\ a_2 x_1 + b_2 x_2 + c_2 x_3 & = & f_2 \\ a_3 x_2 + b_3 x_3 + c_3 x_4 & = & f_3 \\ \dots & \dots & \dots \\ a_n x_{n-1} + b_n x_n + c_n x_{n+1} & = & f_n \\ p_1 x_1 + p_2 x_2 + \dots + p_{n-1} x_{n-1} + p_n x_n + p_{n+1} x_{n+1} & = & f_{n+1} \end{array} \right. ,$$

**п)** система п. п) при

$$n = 49, a_i = c_i = 1, b = 5, p_i = 1, f_i = i,$$

**с)** система п. д) при  $a = 30.$

**т)** система п. з) при  $a = 10, b_{ij} = j / (i+j), f_i = i + 2,$

**у)**  $n = 12, a_{ij} = 1, a_{ij} = 1 / (i^2 + j) (i \neq j), f_i = 1 / i,$

$$\left\{ \begin{array}{lll} a_{11} x_1 + a_{12} x_2 + \dots + a_{1n} x_n & = & f_1 \\ \dots & \dots & \dots \\ a_{n1} x_1 + a_{n2} x_2 + \dots + a_{nn} x_n & = & f_n \end{array} \right. .$$

## **П.11. Библиографический комментарий**

Изложение элементарной теории численных методов линейной алгебры можно найти в [2, 4, 5]. Для лучшего понимания некоторых эффектов при решении СЛАУ полезно проделать соответствующую лабораторную работу из [9]. Вопросам вычислительной линейной алгебры посвящены книги [11–15], в частности, в них разбираются вопросы специфики машинных вычислений. Многие методы прикладной линейной алгебры развиты в научной школе, сформировавшейся в Новосибирске [17, 18]. В частности, разработаны высокоточные методы решения спектральных задач. В данный раздел авторы включили как оригинальные, так и хорошо известные «старые» задачи. Часть задач взята из книг [1, 2, 5, 7, 8, 19–21].

### III. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Развитие метода наименьших квадратов связано с именами К.Ф. Гаусса и А.-М. Лежандра. Первый предложил метод решения насущной практической задачи о введении минимальных изменений в геодезические данные, второй плодотворно занимался задачами интерполяции.

#### III.1 Переопределенная система линейных алгебраических уравнений

Каноническая запись переопределенной системы линейных алгебраических уравнений имеет следующий вид:

$$\begin{aligned} a_{11}b_1 + \dots + a_{1s}b_s &= f_1, \\ &\dots \\ a_{n1}b_1 + \dots + a_{ns}b_s &= f_n, \quad n > s. \end{aligned} \tag{1.1}$$

Введем пространства  $\mathbb{R}^s$  и  $\mathbb{R}^n$ , состоящие из элементов вида  $\mathbf{b} = (b_1, \dots, b_s)^T$ ,  $\mathbf{f} = (f_1, \dots, f_n)^T$  и имеющие размерности  $s$  и  $n > s$  соответственно. Обозначим через  $\mathbf{A}$  прямоугольную матрицу системы (1.1):

$$\mathbf{A} = \begin{vmatrix} a_{11} & \dots & a_{1s} \\ a_{21} & \dots & a_{2s} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{ns} \end{vmatrix}. \tag{1.2}$$

Тогда систему (1.1) можно записать в виде

$$\mathbf{Ab} = \mathbf{f}, \quad \mathbf{b} \in \mathbb{R}^s, \quad \mathbf{f} \in \mathbb{R}^n. \tag{1.3}$$

Введем в  $\mathbb{R}^n$  «основное» скалярное произведение, положив

$$(\mathbf{f}, \mathbf{g})^{(n)} = \sum_{k=1}^n f_k g_k. \tag{1.4}$$

Скалярное произведение в  $\mathbb{R}^n$  можно ввести множеством других способов. Именно, произвольной симметричной и положительно определенной матрице  $\mathbf{B} = \mathbf{B}^* > 0$ , т. е.  $(\mathbf{B}\mathbf{f}, \mathbf{f}) > 0$ , для любого вектора  $\mathbf{f} \neq 0$ , соответствует

скалярное умножение

$$(\mathbf{f}, \mathbf{g})_{\mathbf{B}} = (\mathbf{B}\mathbf{f}, \mathbf{g}); \quad \mathbf{f}, \mathbf{g} \in \mathbb{R}^n. \quad (1.5)$$

Известно, что любое скалярное произведение в пространстве  $\mathbb{R}^n$  можно записать формулой (1.5), подобрав соответствующий самосопряженный оператор  $\mathbf{B} = \mathbf{B}^* > 0$ .

Система (1.1), как правило, не имеет классического решения, т. е. не существует такого набора чисел  $b_1, \dots, b_s$ , который обращает каждое из  $n$  уравнений (1.1) в тождество.

**Определение.** *Фиксируем  $\mathbf{B} = \mathbf{B}^* > 0$ ,  $\mathbf{B}: \mathbb{R}^n \rightarrow \mathbb{R}^n$ . Введем функцию от  $\mathbf{b} \in \mathbb{R}^s$ , положив*

$$\Phi(\mathbf{b}) = (\mathbf{Ab} - \mathbf{f}, \mathbf{Ab} - \mathbf{f})_{\mathbf{B}}. \quad (1.6)$$

*Примем за обобщенное решение системы (1.1) вектор  $\mathbf{b} \in \mathbb{R}^s$ , придающий наименьшее значение квадратичной форме (1.6).*

**Замечание.** Выбор  $\mathbf{B} = \mathbf{B}^* > 0$  зависит от исследователя. Матрица  $\mathbf{B}$  имеет смысл «весовой» матрицы и выбирается из тех или иных соображений о том, какую цену придать невязке системы (1.1) при заданном  $(b_1, b_2, \dots, b_s)$ .

**Теорема 1.** *Пусть столбцы матрицы  $\mathbf{A}$  линейно независимы, т.е. ранг матрицы  $\mathbf{A}$  равен  $s$ . Тогда существует одно и только одно обобщенное решение  $\mathbf{b}$  системы (1.1). Обобщенное решение системы (1.1) является классическим решением системы уравнений*

$$\mathbf{A}^* \mathbf{B} \mathbf{A} \mathbf{b} = \mathbf{A}^* \mathbf{B} \mathbf{f}, \quad (1.7)$$

*которая содержит  $s$  скалярных уравнений относительно  $s$  неизвестных  $b_1, b_2, \dots, b_s$ .*

В дальнейшем будем иногда использовать обозначение

$$\mathbf{C} = \mathbf{A}^* \mathbf{B} \mathbf{A}.$$

### III.2. Геометрический смысл метода наименьших квадратов

Переопределенную систему  $\mathbf{Ab} = \mathbf{f}$ , где  $\mathbf{A} = \|a_{ij}\|$ ,  $1 \leq i \leq n$ ,  $1 \leq j \leq s$ ,  $n > s$ , можно записать в виде

$$b_1 \mathbf{V}_1 + b_2 \mathbf{V}_2 + \dots + b_s \mathbf{V}_s = \mathbf{f},$$

где  $\mathbf{V}_i \in \mathbb{R}^n$  —  $i$ -й столбец матрицы  $\mathbf{A}$ ,  $\mathbf{f} = (f_1, f_2, \dots, f_n)^T \in \mathbb{R}^n$ , а вектор  $\mathbf{b} = (b_1, b_2, \dots, b_s)^T \in \mathbb{R}^s$ .

Требуется найти коэффициенты  $b_1, b_2, \dots, b_s$  линейной комбинации  $b_1\mathbf{V}_1 + b_2\mathbf{V}_2 + \dots + b_s\mathbf{V}_s$  так, чтобы эта линейная комбинация наименее отличалась от  $\mathbf{f}$ :

$$\left\| \mathbf{f} - \sum b_k \mathbf{V}_k \right\|_{\mathbf{B}} \rightarrow \min.$$

Обозначим через  $R^s(\mathbf{V}) \subset R^n$  подпространство размерности  $s$  пространства  $R^n$ , состоящее из всевозможных линейных комбинаций векторов  $\mathbf{V}_1, \dots, \mathbf{V}_s$ .

Пусть  $b_1, b_2, \dots, b_s$  — обобщенное решение переопределенной системы. Тогда линейная комбинация  $\sum b_k \mathbf{V}_k$  — ортогональная в смысле скалярного умножения  $(\cdot, \cdot)_{\mathbf{B}}$  проекция вектора  $\mathbf{f} \in R^n$  на подпространство  $R^s(\mathbf{V})$ , так как любой вектор из  $R^s(\mathbf{V})$  имеет вид:

$$\mathbf{A}\delta = \delta_1 \mathbf{V}_1 + \dots + \delta_s \mathbf{V}_s \in R^s(\mathbf{V}), \quad \delta \in R^s.$$

Наименее уклоняется от  $\mathbf{f}$  элемент  $\sum b_k \mathbf{V}_k$  подпространства  $R^s(\mathbf{V})$ , имеющий вид  $\mathbf{Ab}_B$ , где  $\mathbf{b}_B$  — решение системы (1.7) методом наименьших квадратов (МНК).

В силу  $(\mathbf{Ab}_B - \mathbf{f}, \mathbf{A}\delta)_B = (\mathbf{B}(\mathbf{Ab}_B - \mathbf{f}), \mathbf{A}\delta)^{(n)} = 0$  элемент  $\mathbf{f} - \mathbf{Ab}_B$  ортогонален любому элементу  $\mathbf{A}\delta \in R^s(\mathbf{V})$ .

Если в пространстве  $R^s$  вместо базиса  $\mathbf{V}_1, \dots, \mathbf{V}_s$  выбрать какой-либо другой базис  $\mathbf{V}'_1, \dots, \mathbf{V}'_s$ , то система (1.7) заменится системой

$$\mathbf{C}'\mathbf{b}' = \mathbf{f} \tag{2.1}$$

с матрицей  $\mathbf{C}' = \|c'_{ij}\|$ , где  $c'_{ij} = (\mathbf{V}'_i, \mathbf{V}'_j)_B$ ,  $i, j = 1, 2, \dots, s$ , и правой частью  $i$ -я компонента которой  $\mathbf{f}'_i = (\mathbf{f}, \mathbf{V}'_i)_B$ .

Вместо решения  $\mathbf{b}_B$  системы (1.7) получим новое решение  $\mathbf{b}'_B$  системы (2.1), но проекция  $\mathbf{f}$  на  $R^s$  останется прежней.

Если нас интересует проекция заданного вектора  $\mathbf{f}$  на заданное подпространство  $R^s \in R^n$ , то естественно стремиться к выбору базиса  $\mathbf{V}_1, \mathbf{V}_2, \dots, \mathbf{V}_s$  этого подпространства, по возможности мало отличающегося от ортонормированного. Искомая проекция от выбора базиса в  $R^s$  не зависит, а система МНК (1.7) в случае такого базиса будет иметь хорошо обусловленную матрицу.

### III.3. Задача неточной интерполяции функции

Эта задача возникает, когда есть необходимость найти функциональную связь между переменной и значениями функции в некоторых выбранных точках, задаваемых таблицей:

$x$	$x_0$	$x_1$	$\dots$	$x_n$
$y$	$y_0$	$y_1$	$\dots$	$y_n$

Эта задача может ставиться как задача интерполяции, т.е. как задача нахождения функции из заданного класса, проходящая через все точки  $(x_i, y_i)$ . Однако, если значения  $y_i$  известны неточно, полученная в результате интерполяции функция может иметь большую ошибку между узлами интерполяции.

Зачастую мы хотим, чтобы функция  $y = f(x)$  передавала зависимость «в среднем». Обычно в этом случае вид зависимости  $y \approx \varphi(x)$  выбирается из каких-то внешних соображений. Для семейства  $m$ -параметрических функций  $\varphi(x, a_1, a_2, \dots, a_m)$  параметры подбираются так, чтобы сумма квадратов отклонений приближенно вычисленной функции от табличных значений была минимальна:

$$\Phi(x_i, a_1, a_2, \dots, a_m) = \sum_{i=0}^n (\varphi(x_i, a_1, a_2, \dots, a_m) - y_i)^2 \rightarrow \min. \quad (3.1)$$

Такая функция будет наилучшей аппроксимацией  $f(x)$  в смысле метода наименьших квадратов. Естественным требованием здесь также является  $n > m$ . Функционал  $\Phi$  называется целевым функционалом.

Необходимым условием экстремума является обращение в нуль производных функционала  $\Phi(x_i, a_1, a_2, \dots, a_m)$  по параметрам:

$$\begin{aligned} \frac{\partial \Phi}{\partial a_1} &\equiv 2 \sum_{i=0}^n (\varphi(x_i, a_1, a_2, \dots, a_m) - y_i) \frac{\partial \varphi}{\partial a_1} = 0, \\ &\dots \\ \frac{\partial \Phi}{\partial a_n} &\equiv 2 \sum_{i=0}^n (\varphi(x_i, a_1, a_2, \dots, a_m) - y_i) \frac{\partial \varphi}{\partial a_n} = 0. \end{aligned} \quad (3.2)$$

Если функция  $\varphi(x, a_1, a_2, \dots, a_m)$  представляет собой линейную функцию своих параметров  $a_1, a_2, \dots, a_m$ , то система (3.2) будет линейной. В общем случае эта система может быть нелинейной, что может повлечь за собой трудности в ее решении.

В практике для такой интерполяции в смысле МНК часто используются следующие двухпараметрические и трехпараметрические семейства функций:

$$\begin{aligned} y &= ax + b, \quad y = a + b \ln x, \quad y = a + b \lg x, \\ y &= a x^b, \quad y = a e^{bx}, \quad y = a 10^{bx}, \\ y &= a + b/x, \quad y = 1/(a + bx), \quad y = x/(a + bx), \\ y &= ax^2 + bx + c, \quad y = a x^b + c, \quad y = a e^{bx} + c. \end{aligned}$$

При изучении периодических явлений применяют тригонометрические функции.

Для самого распространенного случая поиска линейного приближения  $y = ax + b$  имеем линейную систему уравнений для поиска коэффициентов:

$$\sum_{i=0}^n (ax_i + b - y_i) \cdot x_i = 0, \quad \sum_{i=0}^n (ax_i + b - y_i) \cdot 1 = 0.$$

Ее решение находится тривиально:

$$a = \frac{(n+1)\sum x_k y_k - \sum x_k \sum y_k}{(n+1)\sum x_k^2 - (\sum x_k^2)^2}, \quad b = \frac{\sum y_k \sum x_k^2 - \sum x_k y_k \sum x_k}{(n+1)\sum x_k^2 - (\sum x_k^2)^2}.$$

Если зависимость приближающей функции от параметров является линейной комбинацией базисных функций  $\varphi(x, a_1, a_2, \dots, a_m) = \sum_{i=1}^m a_i \cdot \varphi_i(x)$ , то применение метода наименьших квадратов приводит к СЛАУ

$$(\varphi_1, \varphi_1)a_1 + (\varphi_1, \varphi_2)a_2 + \dots + (\varphi_1, \varphi_n)a_n = (\varphi_1, y),$$

...

$$(\varphi_n, \varphi_1)a_1 + (\varphi_n, \varphi_2)a_2 + \dots + (\varphi_n, \varphi_n)a_n = (\varphi_n, y).$$

Здесь под скалярным произведением сеточных функций понимается величина

$$(\varphi_l, \varphi_k) = \sum_{i=0}^n \varphi_l(x_i) \varphi_k(x_i). \quad (3.3)$$

Это система с матрицей Грама — симметричной, положительно определенной, следовательно, решение такой системы существует и единственno.

Если в силу каких-либо причин мы хотим по-разному учитывать влияние погрешности задания данных в различных точках интервала (например, чтобы уменьшить влияние краев отрезка), в определение целевого функционала (3.1) можно ввести веса

$$\Phi = \sum_{i=0}^n \rho_k r_k^2 \rightarrow \min, \quad r_k = \varphi(x_i, a_1, a_2, \dots, a_m) - y_i.$$

Введение весов точек  $\rho_k$  аналогично использованию весовой матрицы  $\mathbf{B}$  при решении переопределенных СЛАУ.

При приближении непрерывной функции другой из выбранного класса в смысле МНК скалярное произведение (3.3) должно быть заменено на интеграл

$$(\varphi_l, \varphi_k) = \int_{x_0}^{x_n} \varphi_l(x) \varphi_k(x) dx. \quad (3.4)$$

### III.4. Теоретические задачи

**III.4.1.** Доказать, что  $\forall \varphi, \forall L \in N : \sum_{k=0}^{N-1} \exp\left(i\left(\frac{2\pi L k}{N} + \varphi\right)\right) = 0$ . Получить из этой формулы следствия для действительной и мнимой частей ( $L \neq 0, N \geq 2$ ).

**III.4.2.** Пользуясь результатом предыдущей задачи, доказать, что если набор узлов  $x_k, k = 0, 1, \dots, n$  определяется нулями многочлена Чебышева  $n+1$  порядка:  $T_{n+1}(x_k) = 0, \quad k = 0, 1, \dots, n$ , то  $\forall l, m \leq n$ :

$\sum_{k=0}^n T_l(x_k) T_m(x_k) = \frac{n+1}{2} \delta_{lm}$ , т.е. многочлены Чебышева ортогональны на нулях более старших многочленов.

**III.4.3.** Предложить алгоритм проведения на плоскости окружности через четыре и более точек методом наименьших квадратов.

**III.4.4.** Напряженность магнитного поля  $H$  и магнитная индукция  $B$  связаны соотношением  $B = H / (a + bH)$ . По результатам следующих экспериментальных измерений определить  $a$  и  $b$ :

$H$	8	10	15	20	30	40	60	80
$B$	13.0	14.0	15.4	16.3	17.2	17.8	18.5	18.8

**Указание.** Непосредственное применение МНК к задаче нахождения коэффициентов приводит к нелинейной системе. Линейную систему уравнений для нахождения коэффициентов можно получить, проделав тождественные преобразования в функционале, приведя его к виду, пригодному для применения метода итерированного веса.

**III.4.5.** Измерения трех углов плоского треугольника привели к значениям:  $A_1 = 54^\circ 5'$ ,  $A_2 = 50^\circ 1'$ ,  $A_3 = 76^\circ 6'$ . Сумма углов треугольника  $A_1 + A_2 + A_3 = 180^\circ 12'$  дает невязку в  $12'$ , происходящую от погрешностей наблюдений. Ликвидировать невязку, следуя предписанию наименьших квадратов.

**III.4.6.** Периодическая с периодом  $2\pi$  функция  $y = f(x)$  задана в узлах  $x_k = k \cdot 2\pi/N$ ,  $k = 0, 1, \dots, N-1$ ;  $y_k = f(x_k)$ . Число  $N$  — нечетное. Построить тригонометрический многочлен

$$P_m(x) = C_{-m} e^{-imx} + C_{-m+1} e^{-i(m-1)x} + \dots + C_m e^{imx},$$

интерполирующий функцию  $y = f(x)$  в смысле метода наименьших квадратов в случае  $N = 5$ ;  $m = 0, 1, 2$ .

**III.4.7.** Функцию  $y = \sqrt{1 + \sin^2(x-1)}$  решено приближенно заменить тригонометрическим полиномом

$$P_2(x) = a + a_1 \sin x + b_1 \cos x + a_2 \sin 2x + b_2 \cos 2x,$$

который наименее в смысле метода наименьших квадратов удаляется от таблицы значений этой функции, вычисленной в некоторых десяти заданных точках  $x_0, x_1, \dots, x_9$ .

а) Опишите алгоритм для отыскания коэффициентов  $a, a_1, a_2, b_1, b_2$ .

б) Какие (существенные!) упрощения можно сделать в случае, если  $x_k = k \cdot 2\pi/10$ ,  $k = 0, 1, \dots, 9$ .

**III.4.8.** Пусть замеры функции  $y = f(x)$  осуществлены в точках  $x_k = \cos \frac{\pi(1+2k)}{2(n+1)}$ ,  $k = 0, 1, \dots, n$ , являющихся нулями многочлена Чебышёва  $T_{n+1}(x)$ , и собраны в таблицу

$x$	$x_0$	$x_1$	$\dots$	$x_{n-1}$	$x_n$
$y$	$y_0$	$y_1$	$\dots$	$y_{n-1}$	$y_n$

Среди многочленов степени не выше заданного  $k$ ,  $0 \leq k \leq n$ , указать тот многочлен  $P_k(x)$ , который наилучшим (в смысле метода наименьших квадратов) образом приближает заданную функцию.

Указание. Искать  $P_k(x)$  в виде  $P_k(x) = \sum_{j=0}^k C_j T_j(x)$  и воспользоваться тем,

что многочлены Чебышева  $T_k(x)$ ,  $k = 0, 1, \dots, n + 1$  образуют ортогональную систему векторов на множестве точек  $x_0, x_1, \dots, x_n$  (см. Приложение).

Провести вычисления в случае  $n = 3$ .

$x$	$x_0$	$x_1$	$x_2$	$x_3$
$y$	2	1	3	0

### III.5. Практические задачи

**III.5.1.** Для таблично заданной функции путем решения нормальной системы МНК найти приближающий многочлен первой степени:

a)

$i$	0	1	2	3	4	5
$x_i$	-1.0	0.0	1.0	2.0	3.0	4.0
$y_i$	-0.5	0.0	0.5	0.86603	1.0	0.86603

б)

$i$	0	1	2	3	4	5
$x_i$	-1.0	0.0	1.0	2.0	3.0	4.0
$y_i$	0.86603	1.0	0.86603	0.50	0.0	-0.50

**III.5.2.** Задана таблица приближенных значений функции  $y(x) = a + bx^3$ . Найти  $y(1)$  с помощью метода наименьших квадратов.

$x$	-2	-1	0	1	2
$y$	-8	-1	15	1	8

**III.5.3.** Задана таблица приближенных значений функции  $y(x) = a + bx + cx^2$ . С помощью МНК определить коэффициенты  $a, b, c$ .

$x$	-3	-2	-1	0	1	2	3
$y$	-10	-5	-2	0	2	5	10

**III.5.4.** Построить обобщенный многочлен  $P_4(x) = \sum_{n=-2}^2 C_n e^{inx}$  наилучшего среднеквадратичного приближения для заданной таблицы значений функции

$x$	0	$\pi/4$	$\pi/2$	$3\pi/4$	$\pi$	$5\pi/4$	$3\pi/2$	$7\pi/4$
$y$	1	0	5	2	-1	2	5	0

**III.5.5.** Для заданной таблицы значений функции, где в качестве узлов  $\{x_k\}$  выбраны нули многочлена Чебышева  $T_6(x)$ , определить элемент

$y = a_3x^3 + a_2x^2 + a_1x + a_0$  ( $a_i, i = 0, 1, 2, 3$  – постоянные) наилучшего среднеквадратичного приближения.

$x$	$x_0$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$y$	0	-1	0	0	1	0

**III.5.6.** Найти обобщенное в смысле наименьших квадратов решение определенной системы уравнений:

$$x + y = 3.0; \quad 2x - y = 0.2; \quad x + 3y = 7.0; \quad 3x + y = 5.0.$$

**III.5.7.** Составить программу для вычисления многочлена из задачи III.4.6 в заданной точке  $x = x^* \in [-1, 1]$  при произвольном  $n$  и произвольном наборе чисел  $y_0, y_1, \dots, y_n$ .

**III.5.8.** Известно, что некоторая величина  $J$  зависит от времени  $t$  следующим образом:  $J = ae^{-pt}$ . Ее измерения, проведенные с одинаковой точностью, дали следующую таблицу зависимости  $J$  от  $t$ :

$t$	0	1	2	3
$J$	2.010	1.210	0.740	0.450

Найти значения параметров  $a$  и  $p$  для этой зависимости.

**III.5.9.** Сопротивление проволоки  $R$  линейно зависит от температуры  $t$ :  $R = a_0 + a_1t$ . По результатам следующих экспериментов определить  $a_0$  и  $a_1$ .

$t$	19.1	25.0	30.1	36.0	40.0	45.1	50.0
$R$	76.3	77.8	79.75	80.80	82.35	83.90	85.10

**III.5.10.** Выполнить квадратичное и линейное приближения по методу наименьших квадратов для таких исходных экспериментальных данных:

а)

$x$	0.2	0.3	0.7	0.8	1.2	1.4	1.8
$y$	2.229	2.180	1.972	1.887	1.696	1.590	1.332

Определить  $y$  для  $x = 0.578; x = 0.882; x = 1.356$ ;

б)

$x$	1.0	2.0	2.5	3.0	4.0	4.5	5.0	6.0
$y$	1.88	0.96	-0.13	-2.08	-6.72	-10.67	-14.13	-22.80

Определить  $y$  для  $x = 1.326; x = 3.712; x = 4.698$ ;

в)

$x$	0.0	0.5	1.0	2.0	2.2	2.8	3.0
$y$	2.354	2.307	2.915	5.457	6.300	8.893	10.062

Определить  $y$  для  $x = 0.87; 2.54; 2.17; 2.91$ ;

г)

$x$	-0.5	-0.3	-0.1	0.2	0.6	0.8	1.0
$y$	3.241	2.563	2.138	1.914	2.514	3.149	3.985

Вычислить значения функции в точках  $x = 0, 0.378, 0.521, -0.435$ .

**III.5.11.** Постройте наилучшую среднеквадратическую линейную аппроксимацию для функции

- а)  $f(x) = x^{1/2}$  при  $x \in [0, 1]$ ; б)  $f(x) = 1/x$  при  $x \in [1, 2]$ ;
- в)  $f(x) = \ln(1+x)$  при  $x \in [0, 1]$ ; г)  $f(x) = \sin x$  при  $x \in [0, \pi]$ ;
- д)  $f(x) = x^2$  при  $x \in [0, 1]$ ; е)  $f(x) = e^x$  при  $x \in [0, 1]$ ;
- ж)  $f(x) = \sin x$  при  $x \in [0, \pi/2]$ .

**III.5.12.** Согласно переписи население США менялось следующим образом:

1910 – 92 228 496 человек,  
 1920 – 106 021 537,  
 1930 – 123 202 624,  
 1940 – 132 164 569,  
 1950 – 151 325 798,  
 1960 – 179 323 175,  
 1970 – 203 211 926,  
 1980 – 226 545 805,  
 1990 – 248 709 873,  
 2000 – 281 421 906.

Используя многочлены степеней  $N = 2, 3, 4, 5$  построить аппроксимацию этих данных в смысле МНК. Аппроксимацию можно строить на основе базисных многочленов следующих видов:

$$\text{а) } f(x) \approx \sum_{n=0}^N c_n x^n, \quad \text{б) } f(x) \approx \sum_{n=0}^N c_n (x - 1910)^n,$$

$$\text{в) } f(x) \approx \sum_{n=0}^N c_n (x - 1955)^n, \quad \text{г) } f(x) \approx \sum_{n=0}^N c_n ((x - 1955)/45)^n.$$

Какое представление базисных многочленов является наилучшим?

Используйте построенные приближения для предсказания численности населения США в 2010 году и сравните с точным результатом в 308 745 538 человек.

**III.5.13.** В следующей таблице представлены уровни смертности (число смертей на сто тысяч человек) для возрастов 20–45 лет в Англии начала столетия:

20	21	22	23	24	25	26	27	28
431	409	429	422	530	505	459	499	526
29	30	31	32	33	34	35	36	37
563	587	595	647	669	746	760	778	828
38	39	40	41	42	43	44	45	
846	836	916	956	1014	1076	1134	1024	

а) Построить прямую метода наименьших квадратов для аппроксимации этих данных и исходные данные. Хорошо ли прямая аппроксимирует данные?

б) График исходных данных позволяет предположить, что для возрастных интервалов [20, 28], [28, 39] и [39, 45] данные можно приблизить различными прямыми. Методом наименьших квадратов постройте приближения на трех этих отрезках отдельно.

в) Приближение, построенное в пункте б), не является непрерывным в точках 28 и 39. Один из способов обеспечить это свойство — выбрать базисные функции, обладающие этим свойством. Поскольку для задания трех

прямых линий нужно шесть коэффициентов, а непрерывность в точках 28 и 39 накладывает две связи на коэффициенты, то потребуется  $6 - 2 = 4$  базисные функции, в качестве которых мы используем четыре функции, изображенные на Рис. 5.1.

Все эти функции определены и непрерывны на отрезке [20, 45], и этим же свойством обладает и любая их линейная комбинация. Постройте разложение в смысле метода наименьших квадратов по этим базисным функциям. Какая из трех

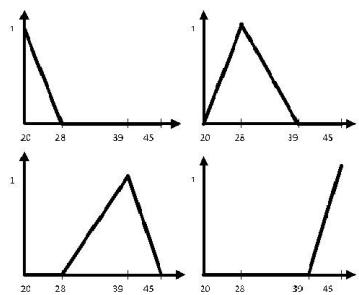


Рис. 5.1. Базисные функции МНК для задачи III.5.13

аппроксимаций: а), б) или в) дает наилучшее приближение в смысле минимизации функционала невязок?

### III.6. Библиографический комментарий

Изложение элементарной теории погрешностей в данном пособии следует книгам [2, 5, 9]. Отметим, что метод наименьших квадратов и среднеквадратичные приближения функций являются основой современных алгоритмов машинного обучения<sup>4</sup>. О современных реализациях МНК с использованием QR-разложения матриц см., например, [12].

<sup>4</sup> См., например, В.В. Вьюгин. Математические основы машинного обучения и прогнозирования. Москва : Изд-во МЦНМО, 2018 – 384 с. и курс лекций [http://www.machinelearning.ru/wiki/index.php?Title=Машинное\\_обучение\\_\(курс\\_лекций\\_K\\_V\\_Воронцова\)](http://www.machinelearning.ru/wiki/index.php?Title=Машинное_обучение_(курс_лекций_K_V_Воронцова)) (режим доступа 22 января 2020 года).

## **IV. ЧИСЛЕННОЕ РЕШЕНИЕ НЕЛИНЕЙНЫХ УРАВНЕНИЙ И СИСТЕМ**

### **IV.1. Введение**

Многие задачи математической физики приводят к необходимости решать нелинейные уравнения или системы нелинейных уравнений, в том числе очень большой размерности. В случае нахождения решения нелинейного уравнения одного переменного методы решения делятся на двухточечные, т.е. использующие информацию о локализации корня (об отрезке, на концах которого функция имеет различные знаки), и одноточечные, в такой информации не нуждающиеся. На случай системы нелинейных уравнений могут быть обобщены только одноточечные методы, т.к. в многомерном случае не существует понятия локализация корня.

Нелинейные уравнения и системы уравнений решаются с применением итерационных методов. *Итерационные методы* (их называют также *методами последовательных приближений*) состоят в том, что решение  $x^*$  находится как предел последовательных приближений  $x^n$  при числе итераций  $n$ , стремящемся к бесконечности. Критерии окончания итерационных процессов различны для методов разного порядка сходимости. Этот вопрос будет рассмотрен для каждого метода в отдельном разделе.

При решении нелинейных уравнений возникают две задачи: указание областей, в каждой из которых находится единственное решение (задача локализации корней), и задача отыскания какого-либо корня с заданной точностью (задача уточнения корней). Для локализации корней не существует общих приемов. Можно использовать построение графиков функции, отыскание участков ее монотонности, участков, на которых функция меняет знак, теорема Руше (ТФКП) и другие частные приемы.

### **IV.2. Метод деления отрезка пополам (дихотомии)**

Метод деления отрезка пополам не требует от функции  $F(x)$ , нуль которой мы ищем, ничего, кроме непрерывности. Это метод двухточечный.

Пусть на концах отрезка  $x \in [a, b]$  функция  $F(x)$  имеет разные знаки, т.е. выполнено

$$F(a) F(b) < 0.$$

Возьмем среднюю точку отрезка  $c = (a + b) / 2$ . Если  $F(a) F(c) < 0$ , то переходим к отрезку  $[a_1, b_1] = [a, c]$ , в противном случае к отрезку

$[a_1, b_1] = [c, b]$ . Итерации продолжаются до тех пор, пока не будет достигнута заданная точность, т.е. длина очередного отрезка локализации  $[a_n, b_n]$  не станет меньше желаемой точности:

$$|a_n - b_n| < \varepsilon.$$

Метод обладает линейной скоростью сходимости:

$$|a_n - b_n| = 2^{-n} |a_0 - b_0|.$$

Он исключительно прост в реализации, всегда сходится к решению, однако методом дихотомии невозможно найти корни четной кратности. Это справедливо для всех методов, использующих локализацию корней.

### IV.3. Методы, основанные на интерполяции

**1. Метод секущих** основан на линейной интерполяции функции, нуль которой мы ищем. Метод также является двухточечным.

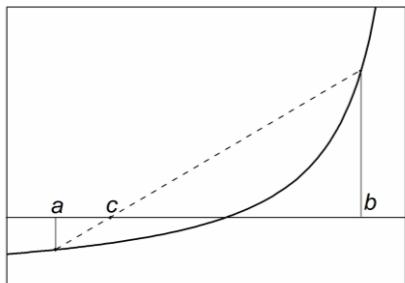
Метод секущих можно рассматривать как приближенную реализацию метода Ньютона (см. ниже) с разностным вычислением первой производной, если аналитическое вычисление производной в силу различных причин невозможно.

Этот метод существует в двух вариантах — с проверкой знаков и без проверки знаков.

В методе *с проверкой знаков* предполагаем, что задан отрезок локализации корня  $x \in [a, b]$ . Через крайние точки проводится прямая

$$y = y_n + \frac{F(b) - F(a)}{b - a} (x - a),$$

Рис. 3.1. Метод секущих с проверкой знаков



пересечение которой с осью будет новым приближением к корню (рис. 3.1):

$$c = \frac{aF(b) - bF(a)}{F(b) - F(a)}.$$

В качестве следующего отрезка локализации выбирается либо отрезок  $[a, c]$ , либо отрезок  $[c, b]$ , на концах которого функция имеет разные знаки.

*Метод секущих без проверки знаков.* Метод не требует обязательной локализации корня на отрезке.

Через две точки  $(x^{n-1}, F(x^{n-1}))$  и  $(x^n, F(x^n))$  проводится прямая. Абсцисса точки пересечения данной прямой с осью  $x$  и является новым приближением  $x^{n+1}$  к решению нелинейного уравнения (рис. 3.2).

В случае реализации метода секущих без проверки знаков возможен вариант, при котором очередная точка приближения лежит вне области определения функции  $F(x^n)$ .

Если корень найден с погрешностью  $\varepsilon$ , то линейная интерполяция устраниет члены порядка  $O(\varepsilon)$ , новая погрешность должна быть порядка  $O(\varepsilon^2)$ , т.е. сходимость должна быть квадратичной. На практике даже около простого корня скорость сходимости хуже и имеет порядок приблизительно 1.5. Кроме того, этот метод крайне ненадежен: он может сходиться медленнее метода деления отрезка пополам или не сходиться вовсе.

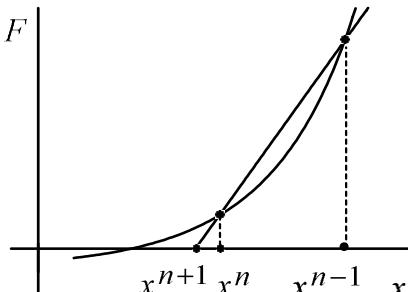


Рис. 3.2 Метод секущих без проверки знаков

**2. Метод парабол** базируется на квадратичной интерполяции функции. Этот метод является трехточечным, т.е. для построения очередного приближения к нулю функции нам необходимо знать три предыдущие точки приближений:  $x^n$ ,  $x^{n-1}$ ,  $x^{n-2}$ . По трем точкам проводится парабола, из двух точек пересечения которой с осью  $x$  выбирается та, которая ближе к последнему приближению.

**3. Адаптированный метод Брендта** является двухшаговым: из исходной пары точек, локализующих корень, за два шага строится еще две точки, также локализующие корень. Целью метода является построение двух точек, лежащих ближе к корню и по разные стороны от него. Система проверок гарантирует, что скорость сходимости метода будет не хуже, чем у метода деления отрезка пополам.

1 шаг . Строится линейная интерполяция по точкам  $a$  и  $b$ , как в методе секущих с проверкой знаков:

$$c = \frac{aF(b) - bF(a)}{F(b) - F(a)}.$$

2 шаг. Пусть выпуклость функции такова, что  $F(a)F(c) > 0$ . Тогда строится линейная интерполяция по точкам  $a$  и  $c$ :

$$d = \frac{cF(a) - aF(c)}{F(a) - F(c)}.$$

Далее делаются проверки полученных точек, обеспечивающие скорость сходимости не ниже деления отрезка пополам:

1) Если  $d > (b + c)/2$ , то полагают  $d = (b + c)/2$ .

2) Проверяют, что  $d > c + \varepsilon$ , где  $\varepsilon$  – заданная точность, в противном случае полагают  $d = c + \varepsilon$ .

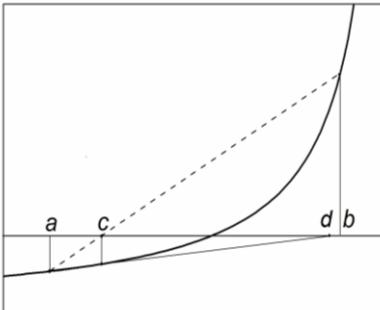


Рис. 3.3. Метод Брэндта

При заданной выпуклости функции корень лежит правее  $c$ , поэтому брать точку левее  $c + \varepsilon$  бессмысленно.

В идеале имеем  $F(c)F(d) < 0$ , т.е. корень локализован на отрезке  $[c, d]$ . Если из-за сдвига точки  $d$  получилось  $F(c)F(d) > 0$ , тогда корень локализован на  $[d, b]$  и сходимость не хуже линейной. В этом случае функция не может быть представлена линейной функцией с небольшой квадратичной поправкой. Если

$F(a)F(c) < 0$ , то  $F(b)F(c) > 0$ , то делают все то же самое с заменой  $a$  на  $b$ :

$$d = \frac{cF(a) - aF(c)}{F(a) - F(c)},$$

$d < (b + c)/2$ , то  $d = (b + c)/2$ ,       $F(c)F(d) < 0$ , следующий отрезок  $[c, d]$ ,  
 $d > c - \varepsilon$ , то  $d = c - \varepsilon$ ,       $F(c)F(d) > 0$ , следующий отрезок  $[a, d]$ .

#### IV.4. Метод простой итерации

Пусть известно, что интересующий нас корень  $x^*$  уравнения  $F(x) = 0$  лежит в интервале  $Y = \{x | a < x < b\}$ . Приведем уравнение  $F(x) = 0$  к равносильному уравнению

$$x = f(x). \quad (4.1)$$

Для поиска решения  $x^*$ , принадлежащего интервалу  $Y$ , зададим  $x^0$ , а затем вычислим последующие приближения  $x^n$  по формуле

$$x^{n+1} = f(x^n), \quad n = 0, 1, 2, \dots \quad (4.2)$$

По построению метода очевидно, что метод является одноточечным и не требует отрезка локализации корня. Методы вида (4.2) называются *методом простой итерации* (МПИ).

**Теорема 1.** *Если функция  $f(x)$  удовлетворяет условию Липшица с константой  $q < 1$*

$$|f(x) - f(y)| < q |x - y|,$$

*то метод простой итерации (4.2) сходится и справедлива оценка*

$$|x^{k+1} - x^*| < q^k |x^0 - x^*|,$$

*или*

$$|x^{k+1} - x^*| < q^k / (1 - q) |x^1 - x^0|.$$

Способов приведения к виду (4.1) существует множество. Можно положить, например,

$$f(x) = x + \tau F(x), \quad (4.3)$$

где  $\tau = \text{const}$ . Такой вариант метода простых итераций иногда также называют *методом релаксации*.

По Теореме 1 МПИ сходится при  $|1 + \tau F'(x)| < 1$ , т.е. при

$$-2 < \tau F'(x) < 0.$$

Если в некоторой окрестности корня  $F'(x) < 0$  и имеет место оценка  $0 < m < |F'(x)| < M$ , то метод релаксации сходится при  $\tau < 2/M$ . Наиболее быстрая сходимость будет достигнута при выборе

$$\tau = 2/(M + m). \quad (4.4)$$

## IV.5. Метод Ньютона

Метод Ньютона является одноточечным, т.е. для построения следующего приближения нам нужно знать только одно значение приближенного решения. Пусть приближение  $x^n$  к корню  $x^*$  уравнения  $F(x) = 0$  уже найдено. Воспользуемся приближенной формулой

$$F(x) \approx F(x^n) + F'_x \cdot (x - x^n), \quad (5.1)$$

точность которой возрастает при приближении  $x^n$  к  $x^*$ . Вместо исходного уравнения  $F(x) = 0$  воспользуемся линейным уравнением

$$F(x^n) + F'_x(x^n) \cdot (x - x^n) = 0.$$

Решение этого уравнения примем за приближение  $x^{n+1}$ :

$$x^{n+1} = x^n - [F'_x(x^n)]^{-1} \cdot F(x^n), \quad n = 0, 1, 2, \dots \quad (5.2)$$

Метод линеаризации Ньютона допускает простую геометрическую интерпретацию (рис. 5.1). График функции  $F(x)$  заменяется касательной к нему в точке  $(x^n, F(x^n))$ . За приближение  $x^{n+1}$  принимается точка пересечения полученной прямой с осью абсцисс.

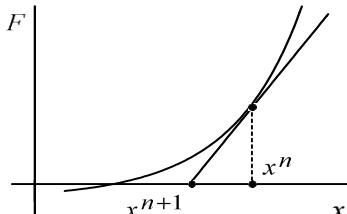


Рис. 5.1. Метод Ньютона

Формулу (5.2) можно интерпретировать как метод простой итерации с функцией  $f(x) = x - [F'_x]^{-1} F(x)$ . В точке простого корня  $x^*$  уравнения  $F(x) = 0$  выполняется равенство

$$f'_x = F(x^*) F''(x^*)/F'^2(x^*) = 0,$$

поэтому неравенство  $|f'_x| < q$  верно для любого положительного фиксированного значения  $q$  в достаточно малой окрестности корня.

Следовательно, асимптотически последовательность погрешностей  $\varepsilon^n = |x^* - x^n|$  метода Ньютона убывает быстрее последовательности членов геометрической прогрессии.

Справедлива теорема о квадратичной сходимости метода Ньютона.

**Теорема 2.** Пусть функция  $F(x)$  определена на интервале

$$U_r : a - r < x < a + r, \quad r > 0,$$

и удовлетворяет следующим условиям:

- 1)  $F(x)$  дважды непрерывно дифференцируема на этом интервале;
- 2) для всех точек интервала  $F'(x) \neq 0$  и существуют конечные значения

$$M_1 = \sup_{U_r} |[F'(x)]^{-1}|, \quad M_2 = \sup_{U_r} |F''(x)|, \quad M_2 > 0;$$

3) уравнение  $F(x) = 0$  имеет корень  $x^*$ :

$$a - r \leq x^* - \frac{2}{M} < x^* < x^* + \frac{2}{M} \leq a + r,$$

где  $M = M_1 M_2$ . Тогда для любого значения  $x^0$  из интервала

$$x^* - \frac{2}{M} \leq x^0 \leq x^* + \frac{2}{M},$$

итерационный процесс сходится к  $x^*$ , причем

$$|x^n - x^*| \leq \left(\frac{M}{2}\right)^{2^n-1} |x^0 - x^*|^{2^n}.$$

На практике более привлекательна такая формулировка условий сходимости метода Ньютона, для которой не нужна никакая информация о решении уравнения. Примером формулировки может служить следующая теорема.

**Теорема 3.** Пусть функция  $F(x)$  определена и дважды непрерывно дифференцируема на интервале  $|x - x^0| < r$  ( $r > 0$ ). Пусть также  $F(x^0) \neq 0$ ,  $F'(x^0) \neq 0$ , существует конечное значение  $M = \sup/[F'(x^0)]^{-1}F''(x) / > 0$  и

$$2M \cdot \left| \frac{F(x^0)}{F'(x^0)} \right| < 1, \quad 2 \cdot \left| \frac{F(x^0)}{F'(x^0)} \right| < r.$$

Тогда итерации процесса Ньютона сходятся к некоторому решению уравнения  $x^*$ . При этом для погрешности справедлива оценка

$$|x^n - x^*| \leq \frac{1}{2^n M} \cdot \left( 2 \left| \frac{F(x^0)}{F'(x^0)} \right| M \right)^{2^n}.$$

По аналогии с методом Ньютона могут строиться методы *высших порядков сходимости*. Пусть имеется итерационный процесс

$$x^{n+1} = f(x^n).$$

Решение уравнения есть неподвижная точка такого отображения  $x^* = f(x^*)$ . Вычтем второе равенство из первого:

$$x^{n+1} - x^* = f(x^n) - f(x^*)$$

и разложим правую часть в ряд Тейлора в окрестности точки  $x^*$ :

$$x^{n+1} - x^* = f'(x^*)(x^n - x^*) + \frac{f''(x^*)}{2}(x^n - x^*)^2 + \frac{f'''(x^*)}{6}(x^n - x^*)^3 + \dots$$

Если в точке  $x^*$  решения нелинейного уравнения  $F(x) = 0$  обращается в нуль не только коэффициент при  $(x^n - x^*)$ , но и при  $(x^n - x^*)^2$ , то метод имеет третий порядок сходимости. Однако методы высоких порядков сходимости используются довольно редко в силу повышенных требований к гладкости функции и жестких условий на начальное приближение, обеспечивающих сходимость метода.

## IV.6. Метод простой итерации для систем нелинейных уравнений

МПИ может быть обобщен для решения систем уравнений

$$\mathbf{F}(\mathbf{u}) = 0. \tag{6.1}$$

Пусть эту систему можно представить в эквивалентном виде

$$\mathbf{u} = \mathbf{f}(\mathbf{u}), \quad (6.2)$$

тогда следующие приближения к решению можно строить как последовательность итераций

$$\mathbf{u}^{n+1} = \mathbf{f}(\mathbf{u}^n).$$

**Определение.** Отображение  $\mathbf{v} = \mathbf{f}(\mathbf{u})$  называется *сжимающим* в замкнутой выпуклой области  $\Omega \subseteq \mathbb{R}^N$ , если существует  $q$ ,  $0 < q < 1$ , такое что

$$\rho(\mathbf{f}(\mathbf{u}^1), \mathbf{f}(\mathbf{u}^2)) \leq q \cdot \rho(\mathbf{u}^1, \mathbf{u}^2)$$

для любых двух точек  $\mathbf{u}^1, \mathbf{u}^2 \in \Omega$ .

**Теорема 4.** Если отображение  $\mathbf{v} = \mathbf{f}(\mathbf{u})$  является сжимающим в замкнутой выпуклой области  $\Omega$ , то уравнение (6.2) имеет единственное решение  $\mathbf{u}^*$  и справедлива оценка  $\rho(\mathbf{u}^{n+1}, \mathbf{u}^*) \leq \frac{q^n a}{1-q}$ , где  $a = \rho(\mathbf{u}^1, \mathbf{u}^0)$ .

**Теорема 5 (достаточное условие сходимости МПИ).** Пусть область  $\Omega \subseteq \mathbb{R}^N$  выпуклая, а компоненты  $f_i(\mathbf{u})$  вектор-функции  $\mathbf{f}(\mathbf{u}) = (f_1, \dots, f_n(\mathbf{u}))^\top$  имеют равномерно непрерывные производные первого порядка. Пусть норма матрицы Якоби

$$\tilde{\mathbf{J}} = \frac{d\mathbf{f}}{d\mathbf{u}} = \begin{pmatrix} \frac{\partial f_1}{\partial u_1} & \dots & \frac{\partial f_1}{\partial u_n} \\ \dots & \dots & \dots \\ \frac{\partial f_n}{\partial u_1} & \dots & \frac{\partial f_n}{\partial u_n} \end{pmatrix} \quad (6.3)$$

не превосходит некоторого числа  $q$ ,  $0 < q < 1$ , т.е.  $\|\tilde{\mathbf{J}}\| \leq q \quad \forall \mathbf{u} \in \Omega$ , тогда отображение  $\mathbf{v} = \mathbf{f}(\mathbf{u})$  является сжимающим в  $\Omega$ .

## IV.7. Метод Ньютона для систем нелинейных уравнений

Метод Ньютона для систем нелинейных уравнений (6.1) является обобщением метода Ньютона для одного нелинейного уравнения. Линеаризуем систему уравнений (6.1) в окрестности предыдущего приближения

$$\mathbf{F}(\mathbf{u}^{k+1}) \approx \mathbf{F}(\mathbf{u}^k) + \mathbf{J} \cdot (\mathbf{u}^{k+1} - \mathbf{u}^k) = 0.$$

Тогда может быть построено следующее приближение к корню:

$$\mathbf{u}^{k+1} \approx \mathbf{u}^k - \mathbf{J}^{-1} \cdot \mathbf{F}(\mathbf{u}^k).$$

$\mathbf{J}$  — матрица Якоби исходной системы (не путать с (6.3)!) вычисляется как

$$\mathbf{J} = \frac{d\mathbf{F}}{d\mathbf{u}} = \begin{pmatrix} \frac{\partial F_1}{\partial u_1} & \dots & \frac{\partial F_1}{\partial u_n} \\ \dots & \dots & \dots \\ \frac{\partial F_n}{\partial u_1} & \dots & \frac{\partial F_n}{\partial u_n} \end{pmatrix}.$$

Достаточное условие сходимости метода Ньютона имеет сложный вид, и проверить его на практике почти никогда не удается. Можем заметить лишь, что в достаточно малой окрестности решения скорость сходимости метода квадратичная.

## IV.8. Критерии сходимости итераций

В методах, использующих локализацию корня, останов итераций при заданной точности  $\varepsilon > 0$  вычисления положения корня осуществить просто: итерационный процесс останавливается, когда длина очередного отрезка локализации

$$|a_n - b_n| < \varepsilon.$$

Для одноточечных линейно сходящихся методов (МПИ) итерационный процесс следует прекратить при выполнении оценки

$$\|\mathbf{u}^* - x^n\| < \varepsilon. \quad (8.1)$$

Так как точное решение  $x^*$  неизвестно, то это условие на практике в явном виде проверить невозможно. Здесь поступают следующим образом. Введем величину  $\|x^n - x^{n-1}\|$ , которая называется *итерационной поправкой*. Иногда требуют просто, чтобы итерационная поправка не превосходила заданного значения точности:

$$\|x^n - x^{n-1}\| < \varepsilon. \quad (8.2)$$

В реальных больших задачах скорость сходимости линейно сходящегося итерационного процесса может быть очень медленной, поэтому критерий (8.2) не является удовлетворительным. Для обеспечения критерия сходимости (8.1) по итерационной поправке используем следующий прием:

$$\begin{aligned} \|x^n - x^*\| &= \|x^n - x^\infty\| = \|x^n - x^{n+1} + x^{n+1} - x^{n+2} + x^{n+2} - x^{n+3} + \dots - x^\infty\| \leq \\ &\leq \|x^n - x^{n+1}\| + \|x^{n+1} - x^{n+2}\| + \|x^{n+2} - x^{n+3}\| + \dots \leq \\ &\leq \|x^n - x^{n+1}\| \cdot (1 + q + q^2 + \dots) = \|x^n - x^{n+1}\| \frac{1}{1-q} \leq \varepsilon. \end{aligned} \quad (8.3)$$

Таким образом, мы видим, что для достижения заданной точности критерий (8.1) в терминах итерационной поправки нужно переформулировать:

$$\|x^n - x^{n-1}\| \leq \varepsilon(1-q). \quad (8.4)$$

При  $q$ , близких к единице, это может существенно увеличить количество итераций для достижения заданной точности по сравнению с (8.2).

Величина  $q$  в оценке (8.4) должна уточняться на итерациях:

$$q = \|x^n - x^{n-1}\| / \|x^{n-1} - x^{n-2}\|. \quad (8.5)$$

В случае достаточной малости величины  $q$  в (8.4) можно пренебречь множителем  $(1-q)$  и перейти к оценке (8.2).

Для метода Ньютона нахождения простого корня нелинейного уравнения выполнение условия (8.2) на итерационную погрешность является достаточным в силу квадратичной сходимости метода. Для кратных корней необходимо переходить к оценке (8.4).

Можно использовать также критерий сходимости по самой функции. Считается, что итерационный процесс сошелся, если выполнено условие  $\|F(x^n)\| < \varepsilon$ . Для кратных корней такой критерий является неудовлетворительным.

## IV.9. Задачи на доказательство

**IV.9.1.** Доказать формулу (4.4).

**IV.9.2.** Доказать, что метод простой итерации  $x^{(n+1)} = \phi(x^{(n)})$  сходится для любого начального приближения  $x_0 \in R$ :

- a)  $\phi(x) = \cos(x)$ ;
- б)  $\phi(x) = a \sin^2 x + b \cos^2 x + c$ , где  $|a - b| < 1$ ;
- в)  $\phi(x) = a \exp(-bx^2) + c$ , где  $b \geq 0$ ,  $2a^2b < e$ .

**IV.9.3.** Использовать формулу метода Ньютона для решения уравнения  $x^2 = a$ ,  $a > 0$ . Доказать, что если за начальное приближение принять произвольное  $x_0 > 0$ , то все последующие приближения  $x_k$ , больше, чем  $\sqrt{a}$ .

**IV.9.4.** Известно, что уравнение  $F(x) = 0$  имеет решение на отрезке  $\alpha \leq x \leq \beta$ . На этом отрезке  $F'(x) > 0$ ,  $F''(x) > 0$ . Показать, что задачу можно решить численно методом Ньютона, положив  $x_0 = \beta$ . Построить пример, в котором при выборе в качестве начального приближения числа  $x_0 = \alpha$  сходимость отсутствует.

**IV.9.5.** Известно, что уравнение  $F(x) = 0$  имеет решение на отрезке  $\alpha \leq x \leq \beta$ . Как следует выбирать начальные приближения  $x_0$ , чтобы была гарантирована сходимость метода Ньютона в следующих случаях:

- на отрезке  $\alpha \leq x \leq \beta$  всюду  $F'(x) > 0, F''(x) < 0,$
- на отрезке  $\alpha \leq x \leq \beta$  всюду  $F'(x) < 0, F''(x) > 0,$
- на отрезке  $\alpha \leq x \leq \beta$  всюду  $F'(x) < 0, F''(x) < 0.$

**IV.9.6.** Доказать, что уравнение  $x + 0.5 \sin x + a = 0$  имеет единственный корень при любом  $a$ . Найти его значение с точностью  $\varepsilon = 10^{-3}$  для  $a = \pm 1, \pm 2, \pm 3$ .

**IV.9.7.** Какие условия теоремы о сходимости простых итераций для уравнения  $x = \varphi(x)$ ,  $\varphi(x) = x/2 - 4/5$ ,  $x_0 = 0$  на отрезке  $[-1, 1]$  не выполнены? Будет ли это уравнение иметь решение на этом отрезке?

**IV.9.8.** Пусть уравнение  $f(x) = 0$ , где  $f(x)$  — дважды дифференцируемая функция, имеет на отрезке  $[a, b]$  корень  $z$  кратности  $p$ .

а) Показать, что при этих условиях метод Ньютона сходится линейно со скоростью геометрической прогрессии со знаменателем  $(p-1)/p$ .

б) Построить модификацию метода Ньютона, имеющую квадратичную скорость сходимости.

Указание. Исследовать указанную модификацию среди методов вида  $x^{(n+1)} = x^{(n)} - \alpha (f'(x^{(n)}))^{-1} f(x^{(n)})$  и найти значение параметра  $\alpha$ , обеспечивающее квадратичную сходимость.

**IV.9.9.** Для нахождения простого нуля  $z$  функции  $f(x) \in C^4$  используется итерационный процесс

$$x^{(n+1)} = 0.5(u^{(n+1)} + v^{(n+1)}),$$

где

$$u^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}, \quad v^{(n+1)} = x^{(n)} - \frac{g(x^{(n)})}{g'(x^{(n)})}, \quad g(x) = \frac{f(x)}{f'(x)}.$$

Доказать, что если метод сходится, то скорость сходимости кубическая.

**IV.9.10.** Пусть уравнение  $f(x) = 0$  имеет единственный корень на отрезке  $[a, b]$ . Для его численного решения используются метод деления отрезка пополам и другой похожий метод, в котором отрезок делится на три равные части. Пусть каждое вычисление функции выполняется за время  $O(1)$ , а сравнение знаков выполняется мгновенно. Доказать, что метод деления пополам сходится к корню быстрее.

**IV.9.11.** Пусть  $f(x)$  дважды непрерывно дифференцируемая функция, и  $f(\xi) = 0$ . Пусть  $\exists X, X > \xi : \forall x \in [\xi, X]$  выполнено  $f'(x) > 0, f''(x) > 0$ . Доказать, что если  $x^{(0)} \in [\xi, X]$ , то последовательность, определяемая методом Ньютона  $x^{(s+1)} = x^{(s)} - \frac{f(x^{(s)})}{f'(x^{(s)})}$  сходится к  $\xi$ , причем в окрестности корня сходимость квадратичная.

Указание. Доказать, что для нового приближения выполнены два неравенства:  $x^{(s+1)} > \xi$  и  $x^{(s+1)} < x^{(s)}$ , после чего легко доказать, что последовательность сходится. Замечание. Интервал не обязан быть конечным.

**IV.9.12.** Какие еще комбинации знаков первой и второй производной (см. IV.9.11) гарантируют сходимость метода Ньютона для интервала правее корня?

**IV.9.13.** Какие комбинации знаков первой и второй производной гарантируют сходимость для интервала левее корня?

**IV.9.14.** Пусть  $f(x)$  дважды непрерывно дифференцируемая функция, и  $f(\xi) = 0, \xi > 0$ . Пусть выполнено  $f'(x) > 0, f''(x) > 0 \quad \forall x \in (0, \infty)$ . Доказать, что если  $x^{(0)} \in (0, \infty)$ , то итерационный процесс метода Ньютона сходится к  $\xi$ , причем в окрестности корня сходимость квадратичная.

Указание. Доказать, что если начальное приближение лежит левее корня, то первая итерация метода Ньютона дает значение правее корня и можно воспользоваться результатом задачи IV.9.11.

## IV.10. Задачи с решениями

**IV.10.1.** Методом простой итерации найти ширину функции  $y = x e^{-x}, x > 0$  на полувысоте с точностью  $\varepsilon = 10^{-2}$ .

Решение. Найдем максимальное значение функции:

$y' = e^{-x} - x e^{-x}$  обращается в ноль при  $x = 1$ , при этом максимальное значение функции  $y_{\max} = e^{-1}$  (рис. 10.1).

Необходимо найти два значения аргумента, при которых функция принимает значения, равные половине от максимального:

$$x e^{-x} = 1/2e. \quad (10.1)$$

Уравнение (10.1) может быть очевидным образом приведено к форме метода простой итерации двумя способами:

1 способ.  $x^{k+1} = 0.5\exp(x^k - 1)$ , в этом случае правая часть метода простой итерации  $f(x) = 0.5\exp(x^k - 1)$ . Для сходимости метода необходимо выполнение условия  $|f'(x)| < 1$ :

$f'(x) = 0.5\exp(x^k - 1)$ , и эта величина меньше единицы при выполнении условия  $x < 1 + \ln 2 \approx 1.7$ . Этот итерационный процесс пригоден для поиска левого корня уравнения (10.1).

2 способ.  $x^{k+1} = \ln(2ex^k)$ ,  $f'(x) = 1/x$ , производная  $|f'(x)| < 1$  при  $x > 1$ . Этот итерационный процесс пригоден для поиска правого корня (10.1).

Каждый из корней должен быть найден с точностью  $\varepsilon/2$ . С учетом критерия сходимости (8.4) имеем для обоих итерационных процессов, стартующих от точки максимума функции:

Левый корень

$$\begin{aligned} x^0 &= 1. \\ x^1 &= 0.5 \\ x^2 &= 0.30326 \\ x^3 &= 0.24910 \\ x^4 &= 0.23597 \\ x^5 &= 0.23289 \\ x^6 &= 0.23217 \end{aligned}$$

Правый корень

$$\begin{aligned} x^0 &= 1. \\ x^1 &= 1.69314 \\ x^2 &= 2.21973 \\ x^3 &= 2.49053 \\ x^4 &= 2.60564 \\ x^5 &= 2.65083 \\ x^6 &= 2.66802 \\ x^7 &= 2.67448 \\ x^8 &= 2.67690 \\ x^9 &= 2.67780 \\ x^{10} &= 2.67814 \end{aligned}$$

Тогда искомая величина  $\Delta_{1/2} = x_2 - x_1 = 2.678 - 0.232 = 2.446$ .

**IV.10.2.** Оценить число итераций метода Ньютона для нахождения положительного корня уравнения  $\sin x + x^2 - 1 = 0$  с точностью  $\varepsilon = 10^{-4}$ .

**Решение.** Для начала локализуем корень. Графически можно показать, что функции  $\sin x$  и  $1 - x^2$  пересекаются на отрезке  $[0, 1]$ . Для более точного определения отрезка локализации для оценки производных заменим в уравнении  $\sin x \approx x$ :

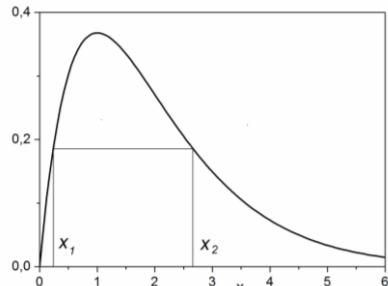


Рис. 10.1. К задаче IV.10.1

$x^2 + x - 1 = 0$ , положительный корень  $x = 0.5(-1 + \sqrt{5}) \approx 0.618$ . В качестве отрезка локализации возьмем интервал  $[\pi/6, \pi/4]$  а в качестве начального приближения величину  $x^0 \approx 0.7$ .

Тогда

$$M_1 = \sup_{\pi/6 \leq x \leq \pi/4} |(F'(x))^{-1}| = \left( \inf_{\pi/6 \leq x \leq \pi/4} |F'(x)| \right)^{-1} = \\ = \left( \inf_{\pi/6 \leq x \leq \pi/4} |2x + \cos x| \right)^{-1} \leq \frac{1}{\pi/3 + \cos \pi/4} \approx 0.4063.$$

$$M_2 = \sup_{\pi/6 \leq x \leq \pi/4} |F''(x)| = \sup_{\pi/6 \leq x \leq \pi/4} |2 - \sin x| = 2 - \sin \pi/4 \approx 1.2929$$

Показатель квадратичного убывания  $q = 0.5$   $M_1 M_2 = 0.2626$ .

По теореме 2 справедлива оценка  $|x^n - x^*| \leq (q)^{2^n-1} |x^0 - x^*|^{2^n} \leq \varepsilon$ . Подставляя грубую оценку  $|x^0 - x^*| \approx 0.5$ , будем иметь для необходимого числа итераций

$$(0.5 \cdot 0.2626)^{2^n-1} \leq 2\varepsilon, \text{ или } (0.1313)^{2^n-1} \leq 2 \cdot 10^{-4}, n \geq 3.$$

Начиная от выбранного начального приближения  $x^0 \approx 0.7$ , получим следующие приближения метода Ньютона:  $x^1 \approx 0.63801$ ,  $x^2 \approx 0.63673$ ,  $x^3 \approx 0.63673$ .

**IV.10.3.** Доказать, что если метод

$$x^{n+1} = x^n - \frac{F(x^n)}{F'_x(x^n)} - \frac{F''(x^n)F^2(x^n)}{2(F'_x(x^n))^3}$$

сходится, то порядок сходимости — третий.

**Решение.** Используем представление этого метода в качестве метода простой итерации  $x^{n+1} = f(x^n)$  с функцией

$$f(x) = x - \frac{F(x)}{F'_x(x)} - \frac{F''(x)F^2(x)}{2(F'_x(x))^3}.$$

Вычислим производные этой функции в ряд Тейлора в корне нелинейного уравнения  $F(x^*) = 0$ :

$$f'(x) = 1 - \frac{F'^2 - FF''}{F'^2} - \frac{F'''F^2 F' + 2F''F'^2 F - 3F''^2 F'^2 F^2}{2F'^4},$$

или

$$\begin{aligned} f'(x) &= \frac{FF''}{F'^2} - \frac{F'''F^2 F' + 2F''F'^2 F - 3F''^2 F^2}{2F'^4} = \\ &= -\frac{F'''F^2}{2F'^3} + \frac{3F''^2 F^2}{2F'^4} = F^2 \left( -\frac{F'''}{2F'^3} + \frac{3F''^2}{2F'^4} \right). \end{aligned}$$

Множитель  $F^2$  перед скобкой обеспечивает второй порядок нуля производной  $f'(x)$  в корне, поэтому сходимость не ниже квадратичной.

Вторую производную правой части МПИ в корне можно не вычислять — она обращается в нуль при  $x = x^*$  в силу множителя  $F^2$  перед скобкой. Таким образом, сходимость не хуже кубической. Можно показать, что третья производная в корне в нуль не обращается, т.е. если сходимость имеет место, то с кубической скоростью.

**IV.10.4.** В середине XIX века Ферхольст для описания динамики популяционной системы предложил измерять ежегодно численность особей  $u_k$ , где  $k$  — номер года. Относительная численность  $u_{k+1}$  полагалась пропорциональной численности в  $k$  год, однако она начинает убывать, когда животных становится много ( $u_k$  сравнимо с 1>):

$$u_{k+1} = \lambda u_k (1 - u_k), \quad u_0 = a$$

Пусть численность популяции нормирована на максимальную возможную численность. Так как  $u_k \in [0, 1]$ , то  $\lambda \in [0, 4]$ . Функция в правой части называется **логистическим отображением**.

Исследовать свойства логистического отображения при значениях параметра  $0 < \lambda \leq 1 + \sqrt{6}$ . Представить графически последовательность итераций логистического отображения при разных значениях параметра.

**Решение.** Рассмотрим подробнее свойства отображения

$$u_{k+1} = \lambda u_k (1 - u_k), \quad u_0 = a.$$

Заметим, что  $f(0) = f(1) = 0$  и  $\max f(u) = f(0.5) = \lambda/4$ , то при  $0 < \lambda < 4$  интервал  $X = [0, 1]$  отображается в себя,  $u \in X$ .

Рассмотрим вначале случай  $0 < \lambda < 1$ . На  $X = [0, 1]$  существует только одна предельная (или неподвижная) точка  $x = 0$ . Любая последовательность  $\{f^k(u_0)\}_{k=0}^\infty$  сходится к предельной точке рассматриваемого отображения  $x = 0$ .

Если рассматривается популяционная модель, то это означает, что рассматриваемая популяция не может выжить.

Из теоремы о сжимающем отображении следует, что последовательность  $\{u_k\}_{n=0}^\infty$  сходится к своей пре-

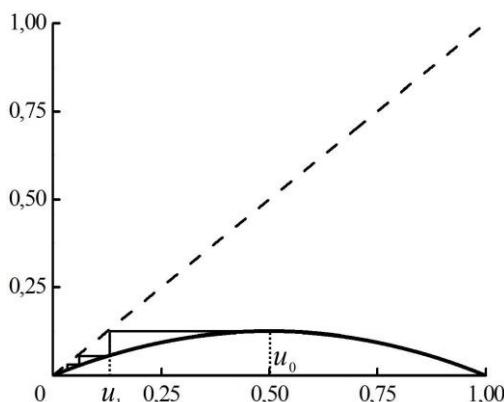


Рис. 10.2. Монотонная сходимость к нулю

дельной точке, если  $|f'_u| \leq 1$ . В этом случае точка называется притягивающей. При  $0 < \lambda < 1$  точка  $x = 0$  — притягивающая.

Графическое изображение траектории (лесенка Ламерей) представлено на рис. 10.2.

Теперь рассмотрим случай  $1 < \lambda < 3$ . При выполнении условия  $|f'_u| > 1$  точка называется отталкивающей. В случае, когда  $\lambda > 1$ , неподвижная точка  $u = 0$  становится отталкивающей, поскольку  $|f'(0)| > 1$ , а на отрезке  $[0, 1]$  появляется другая неподвижная точка  $u_1 = 1 - \lambda^{-1}$ .

Рассмотрим открытое множество  $X = (0, 1)$ . В окрестности неподвижной точки производная для логистического отображения  $|f'(u_1)| = |2 - \lambda| < 1$ . Тогда точка  $u_1$  при  $1 < \lambda \leq 3$  является притягивающей.

Отметим, что при  $1 < \lambda \leq 2$   $f'(u_1) > 0$  и

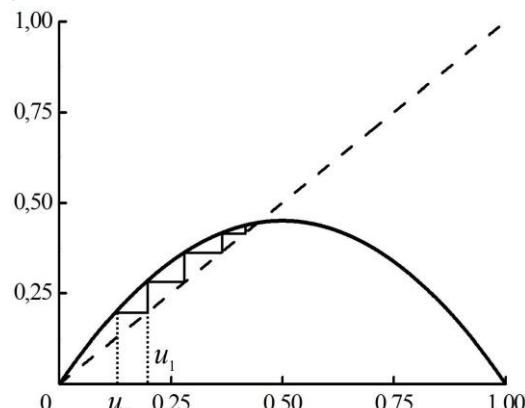


Рис. 10.3. Монотонная сходимость к ненулевому корню

траектория  $\{f^k(u_0)\}_{k=1}^\infty$  стремится монотонно к  $u_1$  (рис. 10.3); при  $2 < \lambda \leq 3$   $f'(u_1) < 0$  и траектория приближается к  $u_1$  немонотонно, поочередно принимая значения то меньше, то больше этого значения.

При  $\lambda = 3$  точка  $u_1$  остается притягивающей, но значение производной в этой точке является предельным:  $|f'(u_1)| = 1$ .

При значениях параметра логистического отображения  $\lambda = 1$  и  $\lambda = 3$  неподвижная точка этого отображения теряет устойчивость и появляется либо другая устойчивая неподвижная точка, как это произошло в первом случае, либо притягивающий цикл. Качественное изменение поведения решения (траектории отображения) при изменении параметра называется *бифуркацией*.

Пусть теперь  $3 < \lambda \leq 1 + \sqrt{6}$ . Рассмотрим теперь корни  $u_3, u_4$  уравнения  $f^2(u) = u$ , или  $\lambda^2 u^2 - \lambda(\lambda+1)u + (\lambda+1) = 0$ .

Если  $u_1$  — предельная точка отображения  $f(u) = u$ , то она является также и предельной точкой отображения  $f^2(u) = u$ . Действительно,  $f^2(u_1) = f(f(u_1)) = f(u_1) = u_1$  где  $u_1$  — любая предельная точка рассматриваемого отображения, отличная от корней уравнения  $f^2(u) = u$ . Тогда, зная два корня уравнения  $f^2(u) = u$  — точки  $u_3, u_4$  легко находятся как корни квадратного уравнения. Они есть

$$u_{3,4} = \frac{(\lambda+1) \pm \sqrt{2\lambda - 3\lambda^2 - 3}}{2\lambda}.$$

Эти корни связаны равенствами

$$f(u_3) = u_4, f(u_4) = u_3.$$

В данном случае говорят, что отображение имеет *цикл периода 2*, который будем обозначать  $P_2$ . Его наличие, например, в популяционной модели говорит об изменении численности особей с периодом в 2 единицы времени. Траектория для случая такого цикла изображена на рис. 10.4. Можно считать, что неподвижная (предельная) точка отображения есть цикл периода 1.

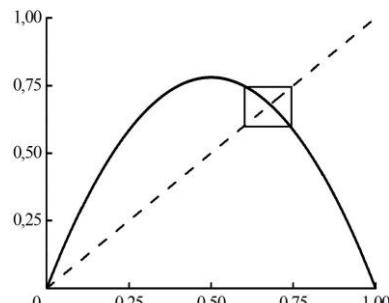


Рис. 10.4. Цикл периода 2

## IV.11. Теоретические задачи

**IV.11.1.** Определить область изменения параметров  $a, b$  и  $c$ , при которых метод простой итерации  $x^{(n+1)} = \phi(x^{(n)})$  сходится для любого начального приближения  $x_0 \in R$ :

- а)  $\phi(x) = a \sin x + b \cos x + c;$
- б)  $\phi(x) = a \cos(bx) + c;$
- в)  $\phi(x) = a \sin^2 x + b \cos^2 x + c;$
- г)  $\phi(x) = a \operatorname{arctg}(bx) + c;$
- д)  $\phi(x) = a \exp(-b^2 x^2) + c.$

**IV.11.2.** Уравнение  $F(x) = 0.001x^3 + x^2 - x + 0.24 = 0$  на отрезке  $0 \leq x \leq 1$  имеет два близких друг к другу корня. Предложить экономичный способ последовательных приближений, позволяющий найти точку, лежащую между этими корнями. Вычислить значение этих корней с заданной точностью  $\varepsilon = 10^{-6}$ .

Указание. Сначала решить уравнение  $F'(x) = 0$ .

**IV.11.3.** Построить метод Ньютона для вычисления числа  $1/a$  так, чтобы расчетные формулы не содержали операций деления. Определить область сходимости метода при  $a > 0$ .

**IV.11.4.** Требуется найти оба корня уравнения  $x = \ln(x + 2)$ .

а) Показать, что для отыскания положительного корня можно воспользоваться итерационным процессом  $x^{(n+1)} = \ln(x^{(n)} + 2)$ ,  $x^{(0)}$  — произвольное.

б) Можно ли указать  $x^{(0)}$ , не совпадающее с отрицательным корнем данного уравнения, таким образом, чтобы итерационный процесс  $x^{(n+1)} = \ln(x^{(n)} + 2)$ ,  $x^{(0)}$  задано, сходился к отрицательному корню?

в) Указать способ вычисления отрицательного корня.

**IV.11.5.** Пусть уравнение  $f(x) = 0$  имеет на отрезке  $[a, b]$  корень  $z$  неизвестной кратности  $p > 1$ , причем  $f'(x)$  — дважды дифференцируемая функция.

а) Построить модификацию метода Ньютона с квадратичной скоростью сходимости.

б) Предложить способ численной оценки кратности корня.

Указание. Исследовать кратность корня  $z$  функции  $g(x) \equiv \frac{f(x)}{f'(x)}$ .

**IV.11.6.** Пусть уравнение  $f(x) = 0$  имеет на отрезке  $[a, b]$  корень  $z$  неизвестной кратности  $p > 1$ , причем  $f'(x)$  — дважды дифференцируемая функция.

а) Построить модификацию метода Ньютона с квадратичной скоростью сходимости. Указание. Исследовать кратность корня  $z$  функции  $g(x) = f(x)/f'(x)$ .

б) Предложить способ численной оценки кратности корня.

**IV.11.7.** Определить порядок сходимости метода решения нелинейного уравнения  $f(x) = 0$

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})} - \frac{f(x^{(n)}) - f(x^{(n)})/f'(x^{(n)})}{f'(x^{(n)})}.$$

**IV.11.8.** Построить итерационный метод решения уравнения  $\operatorname{tg} x + e^x = 0$ . Для поиска корня, локализованного на третьем периоде тангенса  $x \in (5\pi/2, 7\pi/2)$ , применяются методы Ньютона и простых итераций. Выписать расчетные формулы, оценить число итераций, необходимых для достижения точности  $\varepsilon = 0.01$  для метода простой итерации.

**IV.11.9.** Для нелинейной системы уравнений

$$\begin{aligned} xy - x^2 &= 1.03, \\ -2x^3 + y^2 &= 1.98 \end{aligned}$$

известны приближенные значения корней  $x^0 = 1$ ,  $y^0 = 2$ .

Показать, что для уточнения корней можно воспользоваться итерационной схемой

$$x_{k+1} = \sqrt[3]{(y_k^2 - 1.98)/2}, \quad y_{k+1} = x_k + 1.03/x_k.$$

Оценить количество итераций, достаточное для уменьшения первоначальной погрешности не менее чем в  $10^4$ .

**IV.11.10.** Будем решать методом Ньютона уравнение  $F(x) = f(x) - g(x) = 0$ , где  $f(x)$  и  $g(x)$  — заданные функции. Выбираем  $x^0$ . Показать, что приближение  $x_1$  имеет геометрический смысл абсциссы точки пересечения касательных к графикам  $y = f(x)$  и  $y = g(x)$ , проведенным при  $x = x^0$ .

**IV.11.11.** Пронумеруем корни  $x(n)$ ,  $n = 0, 1, \dots$  нелинейного уравнения  $e^{-x} = \cos x$  в порядке возрастания. Показать, что итерации

$$x^{k+1} = x^k - F(x^k)/F'(x^k), \quad F(x) = e^{-x} - \cos x$$

сходятся к корню  $x(n)$ , если за  $x^0(n)$  принять число  $x^0(n) = \pi n/2$ .

Указание. Воспользоваться результатами предыдущей задачи.

**IV.11.12.** Показать, что положительное решение уравнения  $x = 0.5 \cos x$  можно приближенно вычислить, пользуясь итерационной формулой

$x^{n+1} = 0.5 \cos x^n$ ,  $x^0 \geq 0$  — произвольно. Положим  $x^0 = 0$ . Найти такое  $n$ , чтобы погрешность приближения  $x_n$  не превосходила  $10^{-6}$ .

**IV.11.13.** (В.С. Рябенький) Показать, что для решения методом Ньютона следующих уравнений за  $x^0$  можно принять любое  $x^0 > 0$  (см. IV.9.14):

a)  $e^x = 1/x$ ,

б)  $\ln x - 1/x = 0$ .

**IV.11.14.** Для нахождения положительного корня нелинейного уравнения предложено несколько вариантов МПИ. Исследовать эти методы и сделать выводы о целесообразности использования каждого из них.

$$\text{а) } x \ln(x+2) + x^2 - 1 = 0, \quad \begin{cases} 1. x_{n+1} = (1 - x_n^2) / \ln(x_n + 2) \\ 2. x_{n+1} = \exp(1/x_n - x_n) - 2 \\ 3. x_{n+1} = \sqrt{1 - x_n \ln(x_n + 2)} \\ 4. x_{n+1} = 1/x_n - \ln(x_n + 2) \end{cases},$$

$$\text{б) } x + \ln x = 0, \quad \begin{cases} 1. x_{n+1} = -\ln x_n \\ 2. x_{n+1} = \exp(-x_n) \\ 3. x_{n+1} = (x_n + \exp(-x_n)) / 2 \\ 4. x_{n+1} = (3x_n + 5\exp(-x_n)) / 8 \end{cases},$$

$$\text{в) } x^6 - 5x - 2 = 0, \quad \begin{cases} 1. x_{n+1} = (x_n^6 - 2) / 5 \\ 2. x_{n+1} = 5/x_n^4 + 2/x_n^5 \\ 3. x_{n+1} = \sqrt[6]{5x_n + 2} \\ 4. x_{n+1} = \sqrt{5/x_n^3 + 2/x_n^4} \end{cases},$$

$$\text{г) } \ln(x+1) - 2x^2 + 1 = 0, \quad \begin{cases} 1. x_{n+1} = \sqrt{(\ln(x_n + 1) + 1) / 2} \\ 2. x_{n+1} = \exp(2x_n^2 - 1) - 1 \\ 3. x_{n+1} = (\ln(x_n + 1) + 1) / (2x_n) \\ 4. x_{n+1} = x_n + \ln(x_n + 1) - 2x_n^2 + 1 \end{cases},$$

д)  $\sin x - x^2 + 1 = 0$ , 
$$\begin{cases} 1. x_{n+1} = \arcsin(x_n^2 - 1) \\ 2. x_{n+1} = \sqrt{\sin x_n + 1} \\ 3. x_{n+1} = (\sin x_n + 1) / x_n \\ 4. x_{n+1} = x_n + 0.1 \cdot (\sin x_n - x_n^2 + 1) \end{cases}$$

е)  $xe^x + x^2 - 1 = 0$ , 
$$\begin{cases} 1. x_{n+1} = \ln(1/x_n - x_n) \\ 2. x_{n+1} = \sqrt{1 - x_n e^{x_n}} \\ 3. x_{n+1} = (1 - x_n^2) e^{-x_n} \\ 4. x_{n+1} = 1/x_n - e^{x_n} \end{cases}$$

**IV.11.15.** Для нелинейной системы уравнений

$$2x^2 - xy - 5x + 1 = 0,$$

$$x + 3 \lg x - y^2 = 0$$

известны приближенные значения корней  $x^0 = 3.5$ ,  $y^0 = 2.2$ . Показать, что для уточнения корней можно воспользоваться итерационной схемой:

$$x_{k+1} = \sqrt{((x_k(y_k + 5) - 1)/2)}, \quad y_{k+1} = \sqrt{(x_k + 3 \lg x_k)}.$$

Оценить количество итераций, достаточное для уменьшения первоначальной погрешности не менее чем в  $10^4$ .

**IV.11.16.** Для нелинейной системы уравнений

$$x^3 + y^3 - 6x + 3 = 0,$$

$$x^3 + y^3 - 6y + 2 = 0$$

известны приближенные значения корней  $x^0 = y^0 = 0.5$  показать, что для уточнения корней можно воспользоваться итерационной схемой:

$$x_{k+1} = (x_k^3 + y_k^3)/6 + 1/2,$$

$$y_{k+1} = (x_k^3 + y_k^3)/6 + 1/3.$$

Оценить количество итераций, достаточное для уменьшения первоначальной погрешности не менее чем в  $10^4$ .

**IV.11.17.** Дано нелинейное уравнение

а)  $16x^5 + 24x^3 - 2x^2 - 11 = 0$ ,

б)  $24x^5 + 16x^3 - 3x^2 - 10 = 0$ ,

- в)  $8x^5 + 8x^3 - x^2 - 9 = 0$ ,  
 г)  $8x^5 + 16x^3 - x^2 - 10 = 0$ ,  
 д)  $8x^5 + 24x^3 - x^2 - 11 = 0$ ,  
 е)  $16x^5 + 8x^3 - 2x^2 - 9 = 0$ ,  
 ж)  $24x^5 + 8x^3 - 3x^2 - 9 = 0$ .

Отделить корни. Предложить сходящийся метод простой итерации для уточнения корня. Оценить, хотя бы грубо, скорость сходимости. Построить итерационный процесс метода Ньютона.

**IV.11.18.** Даны система нелинейных уравнений

$$x^2 + y^2 - 25 = 0, \quad (x+1)^3 - y + 1 = 0.$$

Приближенно определить области, где лежат решения системы. Выписать расчетные формулы для поиска решений методом простой итерации, проверить условия сходимости.

**IV.11.19.** Даны система нелинейных уравнений

$$\text{а)} \begin{cases} x^2 + y^2 - 9 = 0, \\ (y-1)^3 - x + 1 = 0. \end{cases} \quad \text{б)} \begin{cases} x^2 + y^2 - 25 = 0, \\ (x+1)^3 - y + 1 = 0. \end{cases}$$

Приближенно определить области, где лежат решения системы. Выписать расчетные формулы для поиска решений методом Ньютона, проверить условия сходимости.

**IV.11.20.** Предложить сходящийся вариант метода простой итерации для поиска **всех** корней нелинейного уравнения  $\operatorname{tg} x = x$ .

**IV.11.21.** Выписать формулу метода простых итераций для поиска корня нелинейного уравнения. Начальное приближение к корню определить графически. Оценить априорно число итераций, необходимое для достижения точности  $\varepsilon = 0.0001$ .

- а)  $x^3 - x + 1 = 0$ ;  
 б)  $\sin x - x^2 + 1 = 0$ ;  
 в)  $\operatorname{tg} x - 5x^2 + 1 = 0$ ,  $x \in [-1, 1]$ ;  
 г)  $x^3 - x + 1 = 0$ .

**IV.11.22.** Выписать формулу метода Ньютона для поиска корня нелинейного уравнения. Начальное приближение к корню определить графически. Оценить априорно число итераций, необходимое для достижения точности  $\varepsilon = 0.00001$ .

- а)  $\ln(x+2) - x^2 = 0$ ;  
 б)  $e^x - 2x - 2 = 0$ ;  
 в)  $\ln(x+2) - x^2 = 0$ .  
 г)  $\exp x - 2x - 2 = 0$ .

**IV.11.23.** Выписать формулы метода простых итераций для поиска решения системы нелинейных уравнений, лежащего в первой четверти:

- a)  $x^2 + x + 2y^2 - 3 = 0, \quad 2x^2 + y - xy - 7 = 0;$
- б)  $x^2 - x + y^2 - 1 = 0, \quad y - \operatorname{tg} x = 0;$
- в)  $2x - \cos y = 0, \quad 2y - e^x = 0;$
- г)  $e^{xy} + x - 4 = 0, \quad x^2 - 4y - 1 = 0;$
- д)  $2x - \cos y = 0, 2y - \exp x = 0.$

Доказать сходимость метода. Начальное приближение определить графически.

**IV.11.24.** Построить сходящийся метод простых итераций для поиска точек пересечения кривых

a)  $x^2 + y^2 = 25, \quad (x + 3)y = 1, \quad \text{б) } x - \operatorname{tg} y = 0, \quad (x - 1)y = 7.$

**IV.11.25.** Построить алгоритм метода Ньютона для решения системы уравнений. Сходится ли метод при выборе в качестве начального приближения точки  $M$ ?

- a)  $x - y^3 = 1, \quad (x + 3)(y - 1) = 5, \quad M = (0, 0).$
- б)  $x^2 - y^3 = 1, \quad 4x^2 + 25y^2 = 225, \quad M = (1, 1).$

**IV.11.26.** Предложить сходящийся итерационный метод (простых итераций) вычисления полуширины функции на полувысоте:

- a)  $x \in [0, +\infty) y(x) = 2xe^{-x};$
- б)  $x \in [0, +\infty) y(x) = x/(4 + x)^3;$
- в)  $x \in [0, +\infty) y(x) = x \exp(-(x + 1)^2).$

**IV.11.27.** Предложить метод простой итерации и определить область его сходимости для решения уравнения  $x = e^{2x} - 1$ .

**IV.11.28.** Определить порядок сходимости итерационного метода при вычислении квадратного корня  $x$  из числа  $a$  по формуле

$$x_{n+1} = x_n - (11x_n^4 - 4x_n^2a + a)(x_n^2 - a)/16x_n^5.$$

**IV.11.29.** Локализовать действительные корни алгебраического уравнения, выбрать точку начального приближения, написать итерационную формулу метода Ньютона для уточнения одного из действительных корней уравнения, проверить выполнения условий сходимости метода и привести оценку достижения заданной точности при вычислениях

$$8x^4 + 4x^3 - 14x^2 - x + 2 = 0.$$

**IV.11.30.\*** Уравнение зависит от времени  $t$ , причем при  $t = 0$  решения очевидны. Предложить итерационный алгоритм (продолжения по параметру) для отыскания положения этих корней в зависимости от  $t$  за время от  $t = 0$  до  $t = 1$ .

Выяснить, при каком значении  $t$  эволюция отрицательного корня заканчивается его исчезновением.

а)  $tx^3 + x^2 - 1 = 0$ , б)  $tx^4 + x^2 - 5x + 6 = 0$ .

**IV.11.31.** Для уравнения  $0.01 \sin(\pi x/2) + x^2 - 1 = 0$  получить выражение для приближенных значений корней, используя метод продолжения по параметру.

**IV.11.32.** Определить порядок сходимости итерационного метода для вычисления каждого из корней уравнения  $x^3 + 1.5x^2 - 0.5 = 0$ :

$$x^{(n+1)} = \frac{5}{9}x^{(n)} - \frac{2}{9} + \frac{1}{4x^{(n)}} + \frac{1}{72(x^{(n)})^2} - \frac{1}{72(x^{(n)})^3}.$$

## IV.12. Практические задачи

**IV.12.1.** Найти все действительные решения уравнения

$$0.001x^5 + x^2 - 1 = 0$$

с точностью а) до 0.1; б) до  $10^{-6}$ .

Указание. Грубое приближение найти, используя метод деления отрезка пополам. Более точное — с помощью метода Ньютона.

**IV.12.2.** Выписать формулы подходящего способа последовательных приближений для нахождения положительного корня уравнения

$$x - x^3 + 0.1 = 0.$$

Выбрав начальное приближение, оценить необходимое число итераций для достижения точности  $\varepsilon = 10^{-3}$ .

**IV.12.3.** Отделить корни следующих уравнений, а затем уточнить один из них с помощью подходящего итерационного процесса:

а)  $2x^2 + 5x - 3 = 0$ , б)  $3x + 4x^3 - 12x^2 - 5 = 0$ ,

в)  $(0.5)^x + 1 = (x - 1)^2$ , г)  $(x - 3)\cos x = 1$ ,  $-2\pi \leq x \leq 2\pi$ ,

д)  $\operatorname{arctg}(x - 1) + 2x = 0$ , е)  $x^2 - 20 \sin x = 0$ ,

ж)  $2\operatorname{tg} x - x/2 + 1 = 0$ , з)  $2\lg x - x/2 + 1 = 0$ ,

и)  $x^2 - e^x/5 = 0$ , к)  $\ln x + (x - 1)^3 = 0$ ,

л)  $x^{2^x} = 1$ , м)  $(x + 1)^{0.5} = 1/x$ .

**IV.12.4.** Вычислить с точностью  $\varepsilon = 10^{-3}$  координаты точек пересечения кривых

- a)  $\sin(x+1) - y = 1.2, \quad 2x + \cos y = 2,$
- б)  $\operatorname{tg}(xy + 0.4) = x^2, \quad 0.6x^2 + 2y^2 = 1,$
- в)  $\cos(x-1) + y = 0.5, \quad x - \cos y = 3,$
- г)  $\sin(x+2) - y = 1.5, \quad x + \cos(y-2) = 0.5.$

**IV.12.5.** Задана система уравнений, зависящих от времени  $t$ . Найти координаты точек пересечения при  $t = 0$  и координаты точек пересечения, полученные при эволюции этих точек при изменении  $t$  от  $t = 0$  до  $t = 1$ .

- а)  $x + y + 0.01t x^3 y^3 = 1, \quad x - y - 10^{-3}t \cos(x^2 y) = 2;$
- б)  $x + 2y + 10^{-3}t e^{xy} = 1, \quad 2x + y + 10^{-2}t x^2 y^3 = -1;$
- в)  $2x - y + 10^{-2}t \sin(xy^2) = 5, \quad 5x + y + 10^{-2}t x^4 y^2 = 2.$

Указание. Воспользоваться методом продолжения по параметру, см., например, [3].

**IV.12.6.** Отыскать с точностью до  $\varepsilon = 10^{-5}$  все точки пересечения следующих линий:

- а)  $2x^2 - xy - 5x + 1 = 0, \quad x + 3\lg x - y^2 = 0;$
- б)  $(x - 1.4)^2 - (y - 0.6)^2 = 1, \quad 4.2x^2 + 8.8y^2 = 1.42;$
- в)  $x^2 y^2 - 3x^3 + 6y^3 + 8 = 0, \quad x^4 - 9y + 2 = 0;$
- г)  $\sin x - y = 1.32, \quad \cos y - x = -0.85;$
- д)  $x^7 - 5x^2 y^4 + 1510 = 0, \quad y^3 - 3x^4 y - 105 = 0.$

**IV.12.7.** Методом простой итерации найти ширину функции на полувысоте с точностью  $10^{-3}$ :

- а)  $f(x) = x \exp(1/(x-2)), \quad 0 \leq x \leq 2;$
- б)  $f(x) = x \exp(-x^2), \quad x \geq 0;$
- в)  $f(x) = \ln x / x^2, \quad x \geq 1;$
- г)  $f(x) = \ln x / x, \quad x \geq 1;$
- д)  $f(x) = \sqrt{x} \exp(-x), \quad x \geq 0.$

**IV.12.8.** Профиль лазерного импульса по времени аппроксимируется формулой

$$f(t) = (t / t_m)^{2n} \exp(-n(t / t_m)^2),$$

в которой  $t_m$  имеет смысл времени максимума лазерного импульса и (в силу большой степени касания в нуле) неточно измеряется в эксперименте. Известно, что эта величина примерно равна 400 пс. Найти целое  $n$ , наиболее хорошо описывающее экспериментальный профиль, если известно, что ширина импульса на полуысоте, измеряемая точно, равна 320 пс.

### IV.13. Библиографическая справка

Итерационным методам решения нелинейных уравнений и систем посвящена обширная литература. Для выполнения работы вполне достаточно ознакомиться с основными идеями и теоремами по книгам [2–4]. Более полные сведения о методах можно получить из [7–9, 20], см. также [22] и библиографию в ней.

С итерационными методами решения нелинейных систем тесно связаны различные дискретные отображения. О них лучше прочитать в [24], а на более серьезном уровне в [23].

Основные идеи метода продолжения по параметру см в [3]. Часть задач в разделе взята из [8, 20, 22].

# V. ЗАДАЧА ПОИСКА ЭКСТРЕМУМА ФУНКЦИИ

## V.1. Основные понятия

**Определение.** Пусть на множестве  $U$ , состоящем из элементов и линейного метрического пространства, определена скалярная функция  $\Phi(u)$ .

1. Говорят, что  $\Phi(u)$  имеет локальный минимум на элементе  $u^*$ , если существует его конечная  $\varepsilon$ -окрестность, в которой выполнено

$$\Phi(u^*) \leq \Phi(u), \quad \|u - u^*\| \leq \varepsilon. \quad (1.1)$$

2.  $\Phi(u)$  достигает глобального минимума в  $U$  на элементе  $u^*$  (строгий, абсолютный минимум), если имеет место равенство

$$\Phi(u^*) = \inf_U \Phi(u). \quad (1.2)$$

**Замечание.** Если  $U$  — числовая ось, решается задача на нахождение минимума функции одного переменного, если  $U$  —  $n$ -мерное векторное пространство, имеется задача на нахождение минимума функции  $n$  переменных, если  $U$  — функциональное пространство, то решается задача на отыскание функции, доставляющей минимум функционалу (задача оптимального управления или динамического программирования).

Если к (1.1) или (1.2) добавляются условия

$$u_k^- \leq u_k \leq u_k^+, \quad k = 1, \dots, K$$

$$F_i^- \leq \Phi_i(u) \leq F_i^+, \quad i = 1, \dots, I,$$

где  $u_k^\pm$ ,  $F_i^\pm$  — числа,  $\Phi_i$  — заданные функции, то это задача на отыскание условного минимума. Если подобные ограничения отсутствуют, то это задача на отыскание безусловного минимума. Причем, если функции  $\Phi_i(u)$  линейны, задача поиска условного минимума называется задачей линейного программирования, если хотя бы одна из этих функций нелинейна, то имеется задача нелинейного программирования. Обе эти задачи вместе с задачей динамического программирования в теории оптимального управления называются задачами математического программирования. Задачи линейного программирования просты с идейной точки зрения, их сложность определяется количеством независимых переменных линейного функционала, количеством дополнительных условий типа равенств и неравенств, определяющим геометрические свойства многомерного выпуклого

многогранника, в котором ищется минимум. Минимум достигается либо в вершине, либо на ребре, либо на грани и т.д., и задача состоит в том, чтобы экономно дойти до нужной точки. В экономических вузах курс линейного программирования длится семестр, этому разделу посвящены отдельные учебные пособия, например, [26], поэтому этот раздел здесь опущен.

Без ограничения общности мы будем говорить об отыскании минимума функции, так как максимум функции  $\Phi(\mathbf{u})$  является минимумом функции  $-\Phi(\mathbf{u})$ .

В задачах отыскания экстремума  $\Phi(\mathbf{u})$  называют целевой функцией.

Отметим связь между задачами вычисления корней системы нелинейных алгебраических уравнений (СНАУ) и минимизации.

Пусть на множестве  $U \in L^n$  решается система нелинейных уравнений

$$f_1(u_1, \dots, u_n) = 0,$$

...

$$f_n(u_1, \dots, u_n) = 0.$$

Определим целевую функцию следующим образом

$$\Phi(u_1, \dots, u_n) = \sum_{k=1}^n f_k^2(u_1, \dots, u_n).$$

В области  $U$  справедливо  $\Phi(\mathbf{u}) \geq 0$ , причем минимальное значение  $\Phi(\mathbf{u})$  имеет при  $\mathbf{u} = \mathbf{u}^*$ , где  $\mathbf{u}^*$  — корень рассмотренной системы. Поэтому ее решение эквивалентно поиску минимума  $\Phi(\mathbf{u})$  в  $U$ . Если  $\Phi(\mathbf{u})$  строго больше нуля, то система решений не имеет.

Теперь положим, что необходимо найти минимум целевой функции  $\Phi(\mathbf{u})$ , у которой существуют первые производные. В этом случае задача сводится к решению СНАУ:

$$\frac{\partial \Phi(u_1, \dots, u_n)}{\partial u_1} = 0,$$

...

$$\frac{\partial \Phi(u_1, \dots, u_n)}{\partial u_n} = 0.$$

Точка, являющаяся решением указанной СНАУ, называется стационарной. Однако не всякая стационарная точка может быть точкой локального минимума целевой функции.

Следующую теорему приведем без доказательства.

**Теорема.** Пусть функция  $\Phi(u)$  дважды непрерывно дифференцируема. Тогда достаточным условием того, чтобы стационарная точка  $u^*$  была точкой локального минимума, является положительная определенность матрицы Гессе

$$\mathbf{G}(\mathbf{u}^*) = \begin{pmatrix} \frac{\partial^2 \Phi}{\partial u_1^2} & \cdots & \frac{\partial^2 \Phi}{\partial u_1 \partial u_m} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 \Phi}{\partial u_m \partial u_1} & \cdots & \frac{\partial^2 \Phi}{\partial u_m^2} \end{pmatrix}.$$

Отметим, что методы отыскания минимума  $\Phi(\mathbf{u})$  нередко оказываются более эффективными, чем методы численного решения СЛАУ.

## V.2. Метод перебора

Пусть  $U = [a, b]$ , т.е. отрезок числовой оси. Разобьем его на  $n$  равных частей узлами в точках  $u_i = a + i(b - a) / n$ ;  $i = 0, 1, \dots, n$ .

Вычислив значение  $\Phi(u)$  в этих точках, путем сравнения найдем точку  $u^*$ , в которой

$$\Phi(u^*) = \min_{0 \leq i \leq n} \Phi(u_i).$$

Далее полагаем:  $u^* \approx u_{\min}$ ,  $\Phi^* \approx \Phi(u^*)$ . Погрешность в определении  $u^*$  этого простейшего метода не превосходит числа

$$\varepsilon_n = \frac{b - a}{n}.$$

Этот метод прост, но неэкономичен, особенно для минимума функции многих переменных. Пусть число арифметических действий, необходимое для вычисления значений  $\Phi(u)$  в каждой точке, требует тысячи арифметических операций. Задача решается в гиперкубе  $U = \{0 \leq u_i \leq 1, 1 \leq i \leq 10\}$  с разбиением каждого из отрезков (по каждой координате) на 10 частей. Тогда при использовании компьютера с быстродействием  $10^6$  операций в секунду потребуется около  $10^7$  с (примерно 4 месяца) для поиска  $\min_U \Phi(u)$ .

Этот метод можно сделать более эффективным, если сначала определить минимум с грубым шагом, затем уточнять его значение с меньшим шагом на том из отрезков  $[x_i, x_{i+1}]$ , на котором предполагается его наличие. Можно и далее также уточнять решение задачи.

Дальнейшим усовершенствованием этого метода в случае поиска минимума функции одного переменного являются методы исключения отрезков, а именно дихотомии (деления пополам) и золотого сечения.

### V.3. Поиск минимума функции одного переменного

Для функции одного переменного можно строить эффективные алгоритмы нахождения минимума при условии, что минимум локализован на некотором отрезке  $[a, b]$ . Если на данном отрезке локализации расположен не один локальный минимум, а несколько, то алгоритмы нахождения минимума найдут один какой-нибудь, при этом не обязательно будет выполнено условие (1.2). Для того чтобы найденный минимум был глобальным минимумом на отрезке локализации, необходимо, чтобы функция была *унимодальной*, т.е. была монотонной по обе стороны от точки минимума (непрерывность функции при этом не требуется).

#### V.3.1. Метод деления отрезка пополам (метод дихотомии)

В этом методе отрезок  $[a, b]$  делится на три части выбором внутри отрезка точек  $u_1, u_2$ , в которых вычисляются значения целевой функции. Сравнив ее значения в этих точках можно сократить отрезок поиска точки минимума, перейдя к отрезку  $[a, u_2]$ , если  $\Phi(u_1) \leq \Phi(u_2)$  или  $[u_1, b]$ , если  $\Phi(u_1) \geq \Phi(u_2)$ . Этую процедуру можно продолжить.

Если вычисление значения функции стоит недорого, имеет смысл сокращать отрезок как можно сильнее, и тогда в методе дихотомии точки  $u_1, u_2$  выбираются близко к середине отрезка  $u_1 = (b + a - \Delta)/2$ ,  $u_2 = (b + a + \Delta)/2$ , где  $\Delta$  достаточно мало. Поскольку отношение  $\frac{b - u_1}{b - a}, \frac{u_2 - a}{b - a}$  близко к  $1/2$ , такой выбор объясняется стремлением обеспечить максимальное относительное уменьшение отрезков.

В конце вычисления, в качестве приближенного значения  $u^*$  берется середина последнего отрезка. В результате  $n$  итераций длина отрезка будет

$$\Delta_n = \frac{b - a}{2^n} + \left( \frac{1}{2^n} + \frac{1}{2^{n-1}} + \dots + \frac{1}{2} \right) \Delta = \frac{b - a}{2^n} + \left( 1 - \frac{1}{2^n} \right) \Delta,$$

т.е. точность определения  $u^*$  составляет  $\varepsilon_n = \Delta_n/2$ .

Найдя  $n$  из условия:  $\varepsilon_n \leq \varepsilon$ , получим число итераций, необходимое для достижения данной точности

$$n \geq \log_2 \frac{b - a - \Delta}{2\varepsilon - \Delta}.$$

Если в предыдущем неравенстве положить  $\Delta$  малой, то

$$\varepsilon_n \approx (b-a)/2^{n+1}.$$

### V.3.2. Метод золотого сечения

В случае, когда операция вычисления значения функции является дорогостоящей, имеет смысл сокращать отрезки локализации минимума так, чтобы наиболее эффективно использовать значение в пробной точке, оставшееся от предыдущего шага вычислений на новом отрезке локализации. Найдем расположение точек  $u_1, u_2$  на  $[a, b]$ , для чего рассмотрим отрезок  $[0, 1]$  и, для определенности, положим, что при его уменьшении исключается его правая часть. Из соображений симметрии точки  $u_1, u_2$  должны быть расположены симметрично относительно середины отрезка (рис. 3.1).

Пусть  $u_2 = \tau$ , тогда симметрично расположенная относительно центра отрезка точка имеет координату  $u_1 = 1 - \tau$ . Пробная точка  $u_1$  отрезка  $[0, 1]$  перейдет в пробную точку  $u_2^1 = 1 - \tau$  нового отрезка  $[0, \tau]$ . Условием деления отрезков  $[0, 1]$  и  $[0, \tau]$  в одном и том же отношении точками  $u_2 = \tau$  и  $u_2^1 = 1 - \tau$  является равенство  $\frac{1}{\tau} = \frac{\tau}{1-\tau}$ , или  $\tau^2 + \tau - 1 = 0$ , откуда положительный корень

$$\tau = (\sqrt{5} - 1)/2 \approx 0,61803\dots,$$

$$\text{т.е. } u_1 = 1 - \tau = (3 - \sqrt{5})/2, \quad u_2 = \tau = (\sqrt{5} - 1)/2.$$

Для отрезка  $[a, b]$

$$u_1 = a + (3 - \sqrt{5})(b-a)/2; \quad u_2 = a + (\sqrt{5} - 1)(b-a)/2.$$

#### Замечания

1. Точки  $u_1, u_2$  обладают следующим свойством: каждая из них делит отрезок  $[a, b]$  на две неравные части так, что отношение длины всего отрезка к длине его большей части равно отношению длин большей и

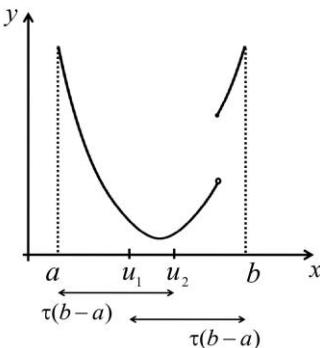


Рис. 3.1. Метод золотого сечения

меньшей части. Точки, обладающие таким свойством, называются точками золотого сечения, введенного Леонардо да Винчи.

2. На каждой итерации отрезок поиска минимума уменьшается в одном и том же отношении

$$\tau = (\sqrt{5} - 1)/2,$$

поэтому в результате  $n$  итераций длина становится равной

$$\Delta_n = \tau^n (b - a).$$

Следовательно, точность  $\varepsilon_n$  определения точки  $u^*$  после  $n$  итераций

$$\varepsilon_n = \Delta_n / 2 = \left( (\sqrt{5} - 1)/2 \right)^n (b - a) / 2,$$

а условие окончания вычислительного процесса будет  $\varepsilon_n \leq \varepsilon$ .

### V.3.3. Метод парабол

Методы, использующие исключение отрезков, основаны на сравнении значений функции в двух точках отрезка локализации, при этом учитываются лишь значения функции в этих точках.

Учесть информацию о значениях функции между точками позволяют методы полиномиальной аппроксимации. Их основная идея заключена в том, что функция  $\Phi(u)$  аппроксимируется полиномом, а точка его минимума служит приближением к  $u^*$ . Разумеется, в этом случае кроме свойства унимодальности, необходимо на  $\Phi(u)$  наложить и требования достаточной гладкости для ее полиномиальной аппроксимации.

Для повышения точности поиска  $u^*$  можно как увеличивать степень полинома, так и уменьшать пробный отрезок. Поскольку первый прием приводит к заметному увеличению вычислительной работы и появлению дополнительных экстремумов, обычно пользуются полиномами второй (метод парабол) или третьей (метод кубической интерполяции) степени.

Алгоритм поиска минимума состоит в следующем. Выбираем на отрезке локализации три точки  $u_1, u_2, u_3$  такие, что

$$u_1 < u_2 < u_3 \text{ и } u_1 \leq u^* \leq u_3.$$

По этим трем точкам построим параболу (квадратичный интерполяционный полином) в форме Ньютона

$$P_2(u) = \Delta_0 + \Delta_1(u - u_1) + \Delta_2(u - u_1)(u - u_2),$$

график которой проходит через точки  $(u_1, \Phi(u_1)), (u_2, \Phi(u_2)), (u_3, \Phi(u_3))$ .

Коэффициенты  $\Delta_k, k = 1, 2, 3$  находим из системы уравнений

$$P_2(u_1) = \Phi(u_1), \quad P_2(u_2) = \Phi(u_2), \quad P_2(u_3) = \Phi(u_3),$$

откуда находим

$$\Delta_0 = \Phi(u_1), \quad \Delta_1 = \frac{\Phi(u_2) - \Phi(u_1)}{u_2 - u_1},$$

$$\Delta_2 = \frac{1}{u_3 - u_1} \left[ \frac{\Phi(u_3) - \Phi(u_2)}{u_3 - u_2} - \frac{\Phi(u_2) - \Phi(u_1)}{u_2 - u_1} \right].$$

Точка  $\bar{u}$  минимума  $P_2(u)$  находится приравниваем его производной нулю:

$$\bar{u} = (u_1 + u_2 - \Delta_1 / \Delta_2) / 2.$$

Далее полагаем:  $u^* \approx \bar{u}$  (очередное приближение точки минимума). Этую процедуру можно продолжить до достижения необходимой точности, выбирая новые точки  $u_k, k = 1, 2, 3$ . Для этого можно применять методы исключения отрезков, используя в качестве двух пробных точек  $u_2$  и  $\bar{u}$ , таких, что  $u_2, \bar{u} \in [u_1, u_3]$ .

Иногда методом парабол называют метод нахождения нуля производной целевой функции, аналогичный методу секущих. При этом выбираются конкретные точки  $u_1 = u^{(s)} + h, u_2 = u^{(s)}, u_3 = u^{(s)} - h$ , и тогда новое приближение к положению минимума определяется формулой

$$u^{(s+1)} = u^{(s)} - \frac{h}{2} \cdot \frac{\Phi(u^{(s)} + h) - \Phi(u^{(s)} - h)}{\Phi(u^{(s)} + h) - 2\Phi(u^{(s)}) + \Phi(u^{(s)} - h)}.$$

Метод критичен к выбору начальной точки и значению третьей производной целевой функции в окрестности минимума: если эта производная мала, есть вероятность, что метод не сойдется вовсе. Теоретически у метода квадратичная сходимость, как у метода Ньютона.

### V.3.4. Модифицированный метод Брэндта

Для гарантированной сходимости не хуже линейной, а в окрестности минимума — квадратичной, прибегают к модифицированному методу Брэндта.

Начинают метод, например, с четырех точек золотого сечения на первоначальном отрезке локализации минимума  $[a, d]$ , расположенных в порядке возрастания аргумента:  $a < b < c < d$ , при этом

$$\Phi_a > \Phi_b, \quad \Phi_c < \Phi_d.$$

Построим две параболы так, что первая проходит через точки  $P_1 = (a, \Phi(a))$ ,  $P_2 = (b, \Phi(b))$ ,  $P_4 = (d, \Phi(d))$ , а вторая — через  $P_1 = (a, \Phi(a))$ ,  $P_3 = (c, \Phi(c))$ ,  $P_4 = (d, \Phi(d))$ . Пусть минимум первой параболы достигается в точке  $f$ , а второй в точке  $e$  (см. метод парабол). После того, как точки  $f$  и  $e$  найдены, проверяем их на приемлемость.

1 шаг. Проверяем, что выполнены два условия для каждой из точек:

$$e > \frac{a+b}{2} \text{ & } e < \frac{3c+d}{4}, \quad f > \frac{a+3b}{4} \text{ & } f < \frac{c+d}{2}.$$

При невыполнении условий помещаем точки на ближайшую границу разрешенного интервала.

2 шаг. Проверяем, что  $|f - e| > \varepsilon$ , где  $\varepsilon$  — заданная точность нахождения

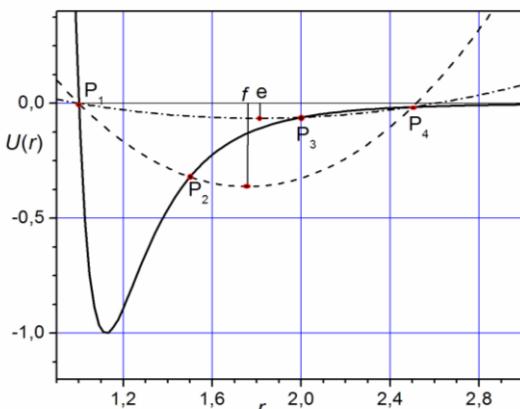
положения минимума. В противном случае полагаем, что  $f = e + \varepsilon$  или  $e = f + \varepsilon$ .

3 шаг. Упорядочиваем точки  $a, b, c, d, e, f$  в порядке возрастания, выбираем из них подряд четыре новые точки, для которых выполнено  $\Phi_i > \Phi_{i+1}, \Phi_{i+2} < \Phi_{i+3}$ . Процесс прекращается, когда отрезок локализации становится меньше  $2\varepsilon$ .

Трудности, с которыми сталкиваются методы полиномиальной аппроксимации

Рис. 3.2. Одна итерация метода Брендта в применении к потенциальному приближению  $a = d$ ,  $b = 1.5d$ ,  $c = 2d$ ,  $d = 2.5d$

при сильно асимметричной функции, проиллюстрированы на рис. 3.2 для поиска минимума потенциала Леннарда–Джонса



$U(r) = 4\epsilon \left( (d/r)^{12} - (d/r)^6 \right)$ . Параметры глубины потенциальной ямы  $\epsilon$  и диаметра молекулы  $d$  были взяты равными единице. Обе параболы метода Брендта в первой итерации дают близкие между собой результаты, весьма далекие от истинного положения минимума  $r_{\min} = 2^{1/6} d$ .

Замечание. Количество достижимых верных знаков при поиске корней уравнения  $\Phi(u) = 0$  — это почти количество верных разрядов в задании переменной. Количество верных знаков при поиске положения минимума вдвое меньше:

$$\Phi(u) = \Phi(u^*) + \frac{1}{2} \Phi''(u^*)(u - u^*)^2,$$

т.е. разницу в значениях функции можно увидеть только лишь при смещениях  $u - u^*$ , при которых заметна величина  $(u - u^*)^2$ .

## V.4. Поиск минимума функции многих переменных

### V.4.1. Методы спуска

Основная идея методов спуска состоит в том, чтобы построить алгоритм, позволяющий перейти из точки начального приближения  $\mathbf{u}^{(0)} = \{u_1^{(0)}, \dots, u_n^{(0)}\}$  в следующую точку  $\mathbf{u}^{(1)} = \{u_1^{(1)}, \dots, u_n^{(1)}\}$  таким образом, чтобы значение целевой функции приблизилось к минимальному. Одним из способов достижения этой цели является использование методов минимизации функции одного переменного. В качестве этого переменного выступают либо поочередно координаты, либо параметр движения вдоль определенного направления в пространстве переменных.

#### V.4.1.1. Метод покоординатного спуска

Этот метод является редукцией поиска функции многих переменных к методам поиска функции одной переменной. Пусть  $\mathbf{u}^{(0)} \in U$  — начальное приближение к расположению минимума  $\Phi(\mathbf{u})$ .

Рассмотрим  $\Phi(\mathbf{u}) = \Phi(u_1, \dots, u_n)$  как функцию одной переменной  $u_1$  при фиксированных  $u_2^{(0)}, \dots, u_n^{(0)}$  и находим одним из методов поиска минимума функции одной переменной  $\min_{u_1 \in U} \Phi(u_1, u_2^{(0)}, \dots, u_n^{(0)})$ . Полученное значение  $u_1$ , доставляющее минимум  $\Phi(u_1)$ , обозначим  $u_1^{(1)}$ ; при этом

$$\Phi(u_1^{(1)}, u_2^{(0)}, \dots, u_n^{(0)}) \leq \Phi(u_1^{(0)}, \dots, u_n^{(0)}).$$

Далее, при фиксированных значениях  $u_1^{(1)}, u_3^{(0)}, \dots, u_n^{(0)}$  ищем

$$\min_{u_2 \in U} \Phi(u_1^{(1)}, u_2, u_3^{(0)}, \dots, u_n^{(0)}),$$

как функции от  $u_2$ ; соответствующее значение  $u_2$  обозначим  $u_2^{(1)}$ ; при этом

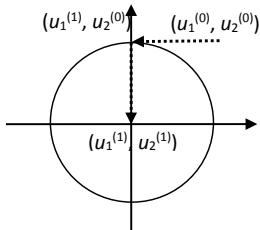


Рис. 4.1. Метод покоординатного спуска для целевой функции  $\Phi(\mathbf{u}) = u_1^2 + u_2^2$ . Окружностями изображены линии уровня целевой функции, пунктирумы стрелками — два шага метода покоординатного спуска

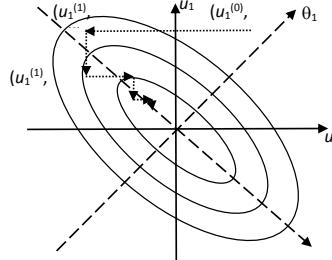


Рис. 4.2. Метод покоординатного спуска для целевой функции  $\Phi(\mathbf{u}) = 5u_1^2 + 5u_2^2 + 8u_1u_2$ . Эллипсами изображены линии уровня целевой функции, пунктирными стрелками — шаги метода покоординатного спуска

$$\Phi(u_1^{(1)}, u_2^{(1)}, u_3^{(0)}, \dots, u_n^{(0)}) \leq \Phi(u_1^{(1)}, u_2^{(0)}, \dots, u_n^{(0)}).$$

Этот процесс продолжаем аналогичным образом и для оставшихся координат; в результате получим

$$\Phi(u_1^{(1)}, \dots, u_n^{(1)}) \leq \Phi(u_1^{(0)}, \dots, u_n^{(0)}).$$

Таким образом, за цикл из  $n$  одномерных спусков переходим из точки  $\mathbf{u}^{(0)}$  в точку  $\mathbf{u}^{(1)}$ . Процесс повторяется до тех пор, пока не будет выполнено условие окончания вычислительного процесса, например,

$$|\Phi(\mathbf{u}^{(s+1)}) - \Phi(\mathbf{u}^{(s)})| \leq \varepsilon,$$

где  $\varepsilon > 0$  — заданная точность.

**Пример 1.** Найти минимум функции двух переменных

$$\Phi(\mathbf{u}) = u_1^2 + u_2^2.$$

Выбрав некоторую точку начального приближения, например  $u_0 = (2, 2)$ , получим минимум целевой функции за один цикл из двух шагов,

так как ее линии уровня — окружности с центром в начале координат (рис. 4.1).

Пример 2. Если же целевая функция будет  $\Phi(u) = 5u_1^2 + 5u_2^2 + 8u_1u_2$ , которая поворотом системы координат на угол  $-45^\circ$  и преобразованием

$$u_1 = \frac{v_1 + v_2}{\sqrt{2}}; u_2 = \frac{(-v_1 + v_2)}{\sqrt{2}}$$

приводится к виду  $\Phi'(v) = v_1^2 + 9v_2^2$ , то ее линиями уровня будут эллипсы  $v_1^2 / 9 + v_2^2 = c^2$ , поэтому спуск будет иметь иной характер (рис. 4.2).

#### V.4.2. Метод градиентного спуска

Напомним, что градиент функции

$$\text{grad } \Phi(\mathbf{u}) = \left( \frac{\partial \Phi}{\partial u_1}, \dots, \frac{\partial \Phi}{\partial u_n} \right)^T$$

есть вектор, ортогональный линиям уровня целевой функции, а его направление совпадает с направлением наибольшего роста  $\Phi(\mathbf{u})$  в данной точке. В точке минимума  $\text{grad } \Phi(\mathbf{u}) = 0$ .

Метод градиентного спуска основан на движении в направлении максимального убывания функции, т.е. в направлении  $-\text{grad } \Phi$ . Построим итерационный процесс следующим образом:

$$\mathbf{u}^{(s+1)} = \mathbf{u}^{(s)} - \tau \cdot \text{grad } \Phi^{(s)}, \quad \mathbf{u}^{(0)} = \mathbf{a},$$

где  $\tau$  — шаг спуска (итерационный параметр движения в направлении наиболее быстрого убывания функции). Итерации продолжим до выполнения заданного условия окончания процесса поиска минимума, например,

$$\|\text{grad } \Phi(\mathbf{u}^{(s+1)})\| \leq \varepsilon, \quad \varepsilon > 0.$$

Пример. Рассмотрим функцию двух переменных

$$\Phi(u_1, u_2) = u_1^2 / 4 + u_2^2.$$

В соответствии с методом градиентного спуска получим

$$u_1^{(s+1)} = u_1^{(s)} - \tau \frac{u_1^{(s)}}{2}, \quad u_2^{(s+1)} = u_2^{(s)} - \tau \cdot 2u_2^{(s)}.$$

Пусть начальное приближение  $\mathbf{u}^{(0)} = \{1, 1\}; \quad \tau = 0,1$ ; тогда  $\mathbf{u}^{(1)} = \{0.95, 0.80\}, \quad \mathbf{u}^{(2)} = \{0.9025, 0.6400\}, \quad \mathbf{u}^{(3)} = \{0.8574, 0.5120\}$ ,  $\Phi(\mathbf{u}^{(0)}) = 1.25, \Phi(\mathbf{u}^{(3)}) = 0.446$ . Если взять  $\tau = 2$ , то  $\mathbf{u}^{(1)} = \{0, -3\}$  и  $\Phi(\mathbf{u}^{(1)}) = 9$ , в то время, как  $\min_U \Phi(\mathbf{u}) = 0$ . Выбор шага оказывается существенным в этом методе, поэтому чаще используются методы с переменным шагом.

#### V.4.3. Метод наискорейшего спуска

Метод наискорейшего спуска основан на поиске минимума функции одного переменного ( $\tau$ ) в направлении максимального убывания функции, т.е. в направлении  $-\text{grad } \Phi$ . Для этого в методе градиентного спуска выберем шаг  $\tau$  так, чтобы функция  $\Phi(u)$  максимально уменьшала свое значение

$$\Phi(\mathbf{u}^{(s+1)}) = \min \Phi(\mathbf{u}^{(s)} - \tau \cdot \text{grad } \Phi(\mathbf{u}^{(s)})) \equiv \min_{\tau} \tilde{\Phi}(\tau, \mathbf{u}^{(s)}).$$

В предыдущем примере выбор шага в точке  $u^{(0)}$  сводится к задаче поиска минимума функции

$$\tilde{\Phi}(\tau, \mathbf{u}^{(0)}) = \frac{1}{4} \left(1 - \tau / 2\right)^2 + (1 - 2\tau)^2,$$

откуда  $\tau = 34/65$ , поскольку

$$\begin{aligned} \Phi(\mathbf{u}^{(1)}) &\equiv \tilde{\Phi}(\tau, \mathbf{u}^{(0)}) = \frac{1}{4} \left( u_1^{(0)} - \tau \frac{u_1^{(0)}}{2} \right)^2 + (u_2^{(0)} - 2\tau u_2^{(0)})^2 = \\ &= \left(1 - \tau / 2\right)^2 / 4 + (1 - 2\tau)^2, \quad u_1^{(0)} = u_2^{(0)} = 1. \end{aligned}$$

На следующих шагах  $\tau$  будет зависеть от  $u_i^{(s)}, s > 0, i = 1, 2$ .

Общий случай этого метода, а также метод сопряженных градиентов рассмотрены в части, посвященной численным методам решения систем линейных алгебраических уравнений.

Отметим следующее важное обстоятельство. Решение экстремальных задач в  $L^n$  зачастую сопряжено со значительными трудностями, особенно для многоэкстремальных задач. Некоторые из них исчезают, если ограничиться рассмотрением только выпуклых функций на выпуклых множествах.

**Определение.** Функция  $\Phi(u)$ , заданная на выпуклом множестве  $U \in L^n$ , называется выпуклой, если для любых точек  $u, v \in U$  и любого  $\alpha \in [0, 1]$  выполнено

$$\Phi[\alpha u + (1 - \alpha)v] \leq \alpha \Phi(u) + (1 - \alpha)\Phi(v).$$

Если для всех  $\alpha \in [0, 1]$  выполнено строгое неравенство

$$\Phi[\alpha u + (1-\alpha)v] < \alpha\Phi(u) + (1-\alpha)\Phi(v),$$

то функция  $\Phi(u)$  называется строгой выпуклой.

Это определение имеет наглядный геометрический смысл: график функции  $\Phi(u)$  на отрезке, соединяющем точки  $u, v$ , лежит ниже хорды, проходящей через точки  $\{u, \Phi(u)\}$  и  $\{v, \Phi(v)\}$ .

Для дважды непрерывно дифференцируемой функции  $\Phi(u)$  положительная определенность матрицы Гессе  $\Phi''_u(u)$  есть достаточное условие строгой выпуклости.

**Теорема.** Пусть  $\Phi(u)$  — выпуклая функция на выпуклом множестве  $U$ ,  $u \in U$ . Тогда любой ее локальный минимум на  $U$  является одновременно и глобальным.

Глобальный минимум строгой выпуклой функции  $\Phi(u)$  на выпуклом множестве  $U$  достигается в единственной точке.

**Доказательство.** Предположим противное, т.е.  $u_0$  — точка локального, а  $u^*$  — глобального минимума  $\Phi(u)$  на  $U$ ,  $u^* \neq u_0$  и  $\Phi(u_0) > \Phi(u^*)$ . Отсюда, с учетом выпуклости  $\Phi(u)$ , имеем

$$\Phi[\alpha u^* + (1-\alpha)u_0] \leq \alpha\Phi(u^*) + (1-\alpha)\Phi(u_0) < \Phi(u_0).$$

При  $\alpha \rightarrow +0$  точка  $u = \alpha u^* + (1-\alpha)u_0$  попадает в сколь угодно малую окрестность  $u_0$ . Поэтому полученное неравенство  $\Phi(u) < \Phi(u_0)$  противоречит предположению о том, что  $u_0$  — точка локального минимума (первая часть теоремы доказана).

Пусть  $u^{(1)}, u^{(2)}$  — две различные точки глобального минимума. Из строгой выпуклости  $\Phi(u)$  следует, что для всех  $\alpha \in [0, 1]$  выполняется строгое неравенство

$$\Phi[\alpha u^{(1)} + (1-\alpha)u^{(2)}] < \alpha\Phi(u^{(1)}) + (1-\alpha)\Phi(u^{(2)}) = \Phi^* = \min_U \Phi(u),$$

что противоречит предположению о том, что  $u^{(1)}, u^{(2)}$  — точки глобального минимума.

#### V.4.4. Метод наискорейшего спуска для решения систем нелинейных уравнений

Решение системы нелинейных уравнений можно свести к задаче нахождения минимума функционала:

$$\mathbf{F}(\mathbf{x}) = 0 \Leftrightarrow \Phi(\mathbf{x}) = \sum_{i=1}^n (f_i(\mathbf{x}))^2 = (\mathbf{F}(\mathbf{x}), \mathbf{F}(\mathbf{x})) \rightarrow \min .$$

В соответствии с методом градиентного спуска будем искать новое приближение в виде

$$\mathbf{x}^{(s+1)} = \mathbf{x}^{(s)} - \tau \cdot \text{grad } \Phi^{(s)},$$

$$\mathbf{x}^{(0)} = \mathbf{a},$$

где параметр спуска  $\tau$  в методе наискорейшего спуска определяется минимумом функционала в выбранном направлении, т.е. обращением в нуль производной

$$\frac{\partial}{\partial \tau} \Phi(\mathbf{x}^{(s)} - \tau \cdot \nabla \Phi(\mathbf{x}^{(s)})) = 0.$$

Рассмотрим скалярную функцию  $W(\tau) = \Phi(\mathbf{x}^{(s)} - \tau \cdot \nabla \Phi(\mathbf{x}^{(s)}))$ . Уравнение, определяющее оптимальный выбор параметра  $\tau$ ,

$$W'(\tau) = \frac{\partial}{\partial \tau} \Phi(\mathbf{x}^{(s)} - \tau \cdot \nabla \Phi(\mathbf{x}^{(s)})) = 0$$

необходимо, вообще говоря, решать численно, что довольно сложно. Поэтому укажем лишь приближенный метод нахождения оптимального параметра  $\tau_s$ . Предположим, что  $\tau$  — малая величина. Линеаризуем функцию  $W(\tau)$ :

$$\begin{aligned} W(\tau) &= \Phi(\mathbf{x}^{(s)} - \tau \cdot \nabla \Phi(\mathbf{x}^{(s)})) = \sum_{i=1}^n f_i^2(\mathbf{x}^{(s)} - \tau \cdot \nabla \Phi(\mathbf{x}^{(s)})) = \\ &= \sum_{i=1}^n (f_i(\mathbf{x}^{(s)}) - \tau \cdot \nabla f_i(\mathbf{x}^{(s)}) \cdot \nabla \Phi(\mathbf{x}^{(s)}))^2 \end{aligned}$$

Тогда

$$W'(\tau) = -2 \sum_{i=1}^n (f_i(\mathbf{x}^{(s)}) - \tau \cdot \nabla f_i(\mathbf{x}^{(s)}) \cdot \nabla \Phi(\mathbf{x}^{(s)})) \cdot \nabla f_i(\mathbf{x}^{(s)}) \cdot \nabla \Phi(\mathbf{x}^{(s)}) = 0.$$

Оптимальный параметр, определяемый этим уравнением, будет

$$\tau_s = \frac{\sum_{i=1}^n f_i(\mathbf{x}^{(s)}) \cdot \nabla f_i(\mathbf{x}^{(s)}) \cdot \nabla \Phi(\mathbf{x}^{(s)})}{\sum_{i=1}^n (\nabla f_i(\mathbf{x}^{(s)}) \cdot \nabla \Phi(\mathbf{x}^{(s)}))^2} = \frac{(\mathbf{F}, \mathbf{J} \nabla \Phi)}{(\mathbf{J} \nabla \Phi, \mathbf{J} \nabla \Phi)},$$

где  $\mathbf{J} = \begin{pmatrix} \frac{\partial f_i}{\partial x_j} \end{pmatrix}$  — матрица Якоби вектор-функции  $\mathbf{F}$ .

$$\frac{\partial \Phi}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{i=1}^n (f_i(\mathbf{x}))^2 = 2 \sum_{i=1}^n f_i(\mathbf{x}) \frac{\partial f_i(\mathbf{x})}{\partial x_j},$$

или

$$\nabla \Phi = 2\mathbf{J}^T \mathbf{F}.$$

Окончательно получаем

$$\tau_s = \frac{1}{2} \frac{(\mathbf{F}, \mathbf{J}\mathbf{J}^T \mathbf{F})^{(s)}}{(\mathbf{J}\mathbf{J}^T \mathbf{F}, \mathbf{J}\mathbf{J}^T \mathbf{F})^{(s)}}, \quad \mathbf{x}^{(s+1)} = \mathbf{x}^{(s)} - 2\tau_s \cdot \mathbf{J}^T (\mathbf{x}^{(s)}) \cdot \mathbf{F}(\mathbf{x}^{(s)}).$$

Для нелинейных систем большой размерности метод неэкономичен.

#### V.4.5. Динамический метод

Решение стационарной задачи нахождения минимума строится как установившееся решение динамической задачи. Рассмотрим систему дифференциальных уравнений

$$\frac{d\mathbf{u}}{dt} + \text{grad } \Phi(\mathbf{u}) = 0, \quad \text{grad } \Phi(\mathbf{u}) = \left( \frac{\partial \Phi}{\partial u_i} \right).$$

В этом случае вектор производных  $d\mathbf{u}/dt$  ортогонален линиям уровня  $\Phi(\mathbf{u})$  и направлен в сторону убывания значений  $\Phi(\mathbf{u})$ . Вдоль траектории  $\Phi(\mathbf{u})$  не возрастают. Формально справедливость этого утверждения следует из неравенства

$$\frac{d\Phi}{dt} = \left( \text{grad } \Phi(\mathbf{u}), \frac{d\mathbf{u}}{dt} \right) = -(\text{grad } \Phi(\mathbf{u}), \text{grad } \Phi(\mathbf{u})) \leq 0. \quad (5.1)$$

Другой нестационарный процесс, решение которого при весьма общих предположениях устанавливается к точке минимума  $\Phi(\mathbf{u})$ , описывается системой дифференциальных уравнений

$$\frac{d^2\mathbf{u}}{dt^2} + \gamma \frac{d\mathbf{u}}{dt} + \text{grad } \Phi(\mathbf{u}) = 0, \quad \gamma > 0.$$

Для решений этой системы имеем

$$\frac{d}{dt} \left( \frac{1}{2} \left( \frac{d\mathbf{u}}{dt}, \frac{d\mathbf{u}}{dt} \right) + \Phi(\mathbf{u}) \right) = \left( \frac{d\mathbf{u}}{dt}, \frac{d^2 \mathbf{u}}{dt^2} \right) + \left( \operatorname{grad} \Phi(\mathbf{u}), \frac{d\mathbf{u}}{dt} \right) = -\gamma \left( \frac{d\mathbf{u}}{dt}, \frac{d\mathbf{u}}{dt} \right) \leq 0. \quad (5.2)$$

Функцию  $\Phi(\mathbf{u})$  в первом случае и  $\frac{1}{2} \left( \frac{d\mathbf{u}}{dt}, \frac{d\mathbf{u}}{dt} \right) + \Phi(\mathbf{u})$  во втором можно

рассматривать как энергию материальной системы. Неравенства (5.1) и (5.2) показывают, что нестационарные процессы характеризуются диссириацией энергии. Материальная точка в поле сил с потенциалом  $\Phi(\mathbf{u})$  и трением  $-\gamma \frac{\partial \mathbf{u}}{\partial t}$  рано или поздно попадет в точку с минимумом  $\Phi(\mathbf{u})$ .

Для наших целей нужно определить разумный выбор коэффициента трения  $\gamma$ . Это проще всего сделать для простейшей модели квадратичной функции скалярного аргумента:

$$\Phi(x) = a^2 x^2 / 2.$$

Динамическое уравнение с потенциалом  $\Phi(x)$

$$x'' + \gamma x' + a^2 x = 0 \quad (5.3)$$

имеет характеристическое уравнение  $\lambda^2 + \gamma \lambda + a^2 = 0$  с корнями

$$\lambda_{1,2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - a^2}. \text{ Общее решение (5.3) тогда будет } x = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}.$$

Скорость убывания решения определяется величиной  $\sigma(\gamma) = \max(\operatorname{Re} \lambda_1, \operatorname{Re} \lambda_2)$ .

При  $\gamma \leq 2a$  имеем  $\gamma^2/4 - a^2 \leq 0$ , и поэтому  $\operatorname{Re} \lambda_1 = \operatorname{Re} \lambda_2 = \sigma(\gamma) = -\gamma/2 \geq -a$ . При  $\gamma > 2a$  величины  $\lambda_1$  и  $\lambda_2$  вещественны и  $\operatorname{Re} \lambda_1 = -\gamma/2 + \sqrt{\gamma^2/4 - a^2} > \operatorname{Re} \lambda_2 = -\gamma/2 - \sqrt{\gamma^2/4 - a^2}$ . Тогда

$$\sigma(\gamma) = \operatorname{Re} \lambda_1 = -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2} = -\frac{a^2}{\gamma/2 + \sqrt{\gamma^2/4 - a^2}} > -\frac{a^2}{\gamma/2} > -a.$$

Таким образом, график  $\sigma(\gamma)$  имеет вид, показанный на рис. 4.3.

Качественные выводы, которые можно сделать из такого рассмотрения:

- Если  $\gamma \ll 2a$ , то решение медленно устанавливается к положению равновесия, происходят колебания около положения равновесия.

- Если  $\gamma >> 2a$ , то решение тоже медленно устанавливается, т.к. при большом трении не развивается большой скорости движения к равновесию.
- Оптимальное значение  $\gamma$  лежит где-то посередине и зависит от свойств конкретной функции  $\Phi(x)$ .

С практической точки зрения можно предложить реализацию алгоритма нахождения минимума исходя из динамической системы уравнений (5.1) с помощью явного метода Эйлера (о чём речь подробнее пойдет в Главе VIII). Выберем шаг интегрирования  $\tau$  и коэффициент уменьшения скорости на каждом шаге  $\beta$ . Шаг выбирается максимально возможным, при котором алгоритм еще устойчив, а коэффициент

$\beta = 0.995$ . Выберем начальное приближение к положению равновесия (минимума энергии) и скорость материальной точки:  $\mathbf{x}_{(0)}, \mathbf{v}_{(0)} = 0$ .

1 шаг.  $\mathbf{v}_n^{1/2} = \mathbf{v}_{n-1} - \tau \cdot \nabla \Phi(\mathbf{x}_{n-1})$ .

2 шаг.  $\mathbf{v}_n = \beta \mathbf{v}_n^{1/2}$ . Настоящее трение приводит к связи  $\mathbf{v}_n = \mathbf{v}_n^{1/2} - \gamma \tau \mathbf{v}_n^{1/2}$ , т.е.  $\beta = 1 - \gamma \tau$ .

3 шаг.  $\mathbf{x}_n = \mathbf{x}_{n-1} + \tau \mathbf{v}_n$ .

Недостатком метода является наличие подгоночных параметров  $\tau$  и  $\beta$ . Хорошо, что алгоритм не заботится о значениях минимизируемой функции и о величине градиента. По ходу движения и то, и другое может возрастать. Например, при повороте оврага вбок материальная точка «заезжает» на стены оврага. В данном случае это достоинство, т.к. движение в основном идет в нужном направлении.

Напоследок несколько практических рекомендаций: не занулять скорость при возрастании значения функции и не менять резко величину  $\beta$  с 0.995 до 0.8 в окрестности минимума (хоть и хочется, чтобы колебания установились быстрее).

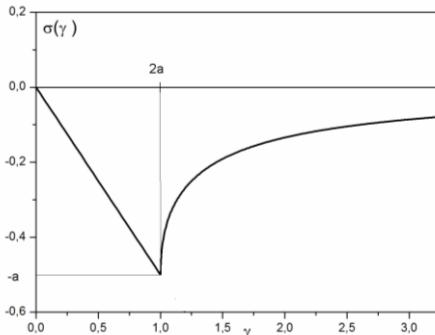


Рис. 4.3. Зависимость скорости затухания решения от коэффициента трения для квадратичной функции

## V.5. Задачи с решениями

**V.5.1.** Свести задачу решения системы нелинейных уравнений

$$\begin{cases} u - 5 \cdot 10^{-2} e^{uv} = 0, \\ v - 5 \cdot 10^{-2} e^{-(u+v)} = 0, \end{cases}$$

к вариационной задаче в области  $\Omega = \{|u - 0,1| \leq 0,1; |v - 0,1| \leq 0,1\}$ .

**Решение.** Решение системы сводится к нахождению условий минимума функционала

$$\Phi(u, v) = (u - 5 \cdot 10^{-2} e^{uv})^2 + (v - 5 \cdot 10^{-2} e^{-(u+v)})^2.$$

**V.5.2.** Свести задачу минимуме функции  $\Phi(u, v) = u^4 + v^4 - u^2 - v^2$  в области  $\Omega > \{|u| \leq 1; |v| \leq 1\}$  к решению системы алгебраических уравнений.

**Решение.** Задача о нахождении минимума функции  $\Phi(u, v)$  сводится к решению системы уравнений  $\frac{\partial \Phi}{\partial u} = 0, \frac{\partial \Phi}{\partial v} = 0$ .

**V.5.3.** Найти значения  $\{x, y\}$ , при которых достигается минимум функции  $f(x, y) = x^3 + y^3 - 3xy$ .

**Решение.** Вычислим частные производные  $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$  и приравняем их к нулю. Получим систему двух нелинейных уравнений

$$3x^2 - 3y = 0, \quad -3x + 3y^2 = 0.$$

Решениями этой системы являются пары  $\{0, 0\}, \{1, 1\}$ . Подстановкой убеждаемся, что вторая точка является точкой глобального минимума.

**V.5.4.** Методом деления отрезка пополам найти точку локального минимума для функции  $f(x) = x^3 + e^{-x} - x$  на отрезке  $[0, 1]$  с точностью  $\varepsilon = 10^{-2}$ .

**Решение.** Обозначим границы отрезка  $a_0 = 0, b_0 = 1$  и зададим  $\Delta/2 = \delta = 10^{-3}$ . Вычислим

$$f(0.5 \cdot (a_0 + b_0 - \Delta)) \approx 0.2324, \text{ и } f(0.5 \cdot (a_0 + b_0 + \Delta)) \approx 0.2307.$$

Так как второе значение меньше первого, то положим  $(a_1, b_1) = (0.4990, 1)$ . Далее получим  $(a_3, b_3) = (0.6238, 0.7505), (a_4, b_4) = (0.6861, 0.7505), (a_5, b_5) = (0.6861, 0.7191), (a_6, b_6) = (0.7016, 0.7191), (a_7, b_7) = (0.7101, 0.7198)$ .

**V.5.5.** С помощью метода Ньютона найти минимум функции

$$F(t) = \sin t - \cos t, \quad t_0 = -0.5.$$

**Решение.** Найдем точку минимума функции  $F(t)$  как корень уравнения  $F'(t) = 0$ . Для этого построим итерационный процесс Ньютона:

$$t^{(n+1)} = t^{(n)} - \frac{F'(t^{(n)})}{F''(t^{(n)})}, \quad t^{(0)} = -0.5.$$

При  $t^{(0)} = -0.5$  имеем  $F'(t^{(0)}) = \cos t + \sin t \approx 0.3982$ ,  
 $F''(t^{(0)}) = -\sin t^{(0)} + \cos t^{(0)} \approx 1.3570$ ,  $t^{(1)} = t^{(0)} - \frac{0.3982}{1.3570} \approx -0.7934$ . Дальнейшие вычисления дают  $t^{(2)} = -0.7854$ ,  $t^{(3)} = -0.7854$ .

**V.5.6.** Методом наискорейшего спуска приближенно вычислить корни системы:

$$x + x^2 - 2yz = 0.1,$$

$$y - y^2 + 3xz = -0.2,$$

$$z + z^2 + 2xy = 0.3,$$

расположенные в окрестности начала координат.

**Решение.** Выберем точку начального приближения в начале координат:  $\mathbf{u}^{(0)} = (0, 0, 0)^T$ . Нелинейная система уравнений записывается  $\mathbf{F}(\mathbf{u}) = 0$ , где

$$\mathbf{F} = \begin{pmatrix} x + x^2 - 2yz - 0.1 \\ y - y^2 + 3xz + 0.2 \\ z + z^2 + 2xy - 0.3 \end{pmatrix}, \text{ матрица Якоби } \mathbf{J} = \begin{pmatrix} 1+2x & -2z & -2y \\ 3z & 1-2y & 3x \\ 2y & 2x & 1+2z \end{pmatrix}.$$

$$\text{Для начального приближения } \mathbf{F}^{(0)} = \begin{pmatrix} -0.1 \\ 0.2 \\ -0.3 \end{pmatrix}, \quad \mathbf{J}^{(0)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \mathbf{E}.$$

$$\mathbf{u}^{(1)} = \mathbf{u}^{(0)} - 2\tau_0 \cdot \mathbf{J}^T(\mathbf{u}^{(0)}) \cdot \mathbf{F}(\mathbf{u}^{(0)}) = \mathbf{u}^{(0)} - 2 \frac{1}{2} \mathbf{E} \mathbf{F}^{(0)} = (0.1, -0.2, 0.3)^T.$$

Аналогичным образом находим второе приближение:

$$\mathbf{F}^{(1)} = \begin{pmatrix} 0.13 \\ 0.05 \\ 0.05 \end{pmatrix}, \quad \mathbf{J}^{(1)} = \begin{pmatrix} 1.2 & -0.6 & 0.4 \\ 0.9 & 1.4 & 0.3 \\ -0.4 & 0.2 & 1.6 \end{pmatrix},$$

$$\mathbf{J}^{(1)T} \mathbf{F}^{(1)} = \begin{pmatrix} 0.181 \\ 0.002 \\ 0.147 \end{pmatrix}, \quad \mathbf{J}^{(1)} \mathbf{J}^{(1)T} \mathbf{F}^{(1)} = \begin{pmatrix} 0.2748 \\ 0.2098 \\ 0.1632 \end{pmatrix}.$$

$$\tau_1 = \frac{1}{2} \frac{(\mathbf{F}^{(1)}, \mathbf{J}^{(1)} \mathbf{J}^{(1)T} \mathbf{F}^{(1)})}{2 (\mathbf{J}^{(1)} \mathbf{J}^{(1)T} \mathbf{F}^{(1)}, \mathbf{J} \mathbf{J}^T \mathbf{F}^{(1)})} = \frac{1}{2} \cdot 0.3719,$$

$$\mathbf{u}^{(2)} = \mathbf{u}^{(1)} - 2\tau_1 \cdot \mathbf{J}^{(1)T} \cdot \mathbf{F}^{(1)} = \begin{pmatrix} 0.1 \\ -0.2 \\ 0.3 \end{pmatrix} - 0.379 \begin{pmatrix} 0.181 \\ 0.002 \\ 0.147 \end{pmatrix} = \begin{pmatrix} 0.0327 \\ -0.2007 \\ 0.2453 \end{pmatrix}.$$

Для контроля точности вычислим невязку:

$$\mathbf{F}^{(2)} = \begin{pmatrix} 0.032 \\ -0.017 \\ -0.007 \end{pmatrix}.$$

## V.6. Теоретические задачи

**V.6.1.** Привести пример функции  $f(x)$ , заданной на множестве  $U = R$  и обладающей следующим свойством: а) глобальный минимум  $f(x)$  достигается на счетном множестве точек; б)  $f(x)$  имеет бесконечное число точек локального минимума, но глобальный минимум  $f(x)$  на  $U$  не достигается; в)  $f(x)$  ограничена снизу на  $U$ , но не достигает минимума.

**V.6.2.** Пользуясь только определением точки минимума, доказать, что линейная на отрезке  $[a, b]$  функция  $f(x) = \alpha x + \beta$ ,  $\alpha \neq 0$  может достигать минимального на этом отрезке значения лишь в точках  $x = a$  и  $x = b$ .

**V.6.3.** Привести примеры отрезков, на которых функции  $x^2$ ,  $-x^2$ ,  $\ln x$ ,  $\cos x$  унимодальны.

**V.6.4.** Доказать следующие свойства унимодальных функций:

- а) любая из точек локального минимума унимодальной функции является и точкой ее глобального минимума на отрезке  $[a, b]$ ;
- б) функция унимодальная на отрезке  $[a, b]$  является унимодальной и на любом меньшем отрезке  $[c, d] \subset [a, b]$ ;
- в) пусть  $f(x) \in Q[a, b]$  и  $a \leq x_1 < x_2 \leq b$ ; тогда если  $f(x_1) \leq f(x_2)$ , то  $x^* \in [a, x_2]$ , если  $f(x_1) > f(x_2)$ , то  $x^* \in [x_1, b]$ , где  $x^*$  — одна из точек минимума на отрезке  $[a, b]$ .

**V.6.5.** Доказать, что из выпуклости функции  $f(x)$  на отрезке  $[a, b]$  следует ее унимодальность на  $[a, b]$ , ограничиваясь только дифференцируемыми на  $[a, b]$  функциями. Верно ли обратное?

**V.6.6.** Привести пример унимодальной, но не выпуклой функции.

**V.6.7.** Пусть функция  $f(x)$  дифференцируема на множестве  $U = \mathbb{R}$  и производная  $f'(x)$  обращается в нуль в единственной точке  $\tilde{x}$ . Доказать, что если  $\lim_{|x| \rightarrow \infty} f(x) = +\infty$ , то  $\tilde{x}$  — точка глобального минимума  $f(x)$  на  $U$ .

**V.6.8.** Пусть  $f(x)$  — дифференцируемая унимодальная на отрезке  $[a, b]$  функция, причем  $|f'(x)| \leq M$ . Оценить точность  $\delta(N)$  определения минимального значения  $f^*$  методом перебора в результате  $N$  вычислений  $f(x)$ .

**V.6.9.** Может ли применение методов исключения отрезков привести к неверному определению точки минимума  $x^*$ , если  $f(x)$  не унимодальна? Ответ пояснить рисунком.

**V.6.10.** Зависит ли точность определения точки минимума  $x^*$ , которую гарантируют методы дихотомии и золотого сечения в результате  $N$  вычислений  $f(x)$ , от конкретной функции  $f(x)$ ?

**V.6.11.** Требуется найти точку минимума унимодальной функции на отрезке длины 1 с точностью  $\varepsilon = 0,02$ . Имеется возможность измерить не более 10 значений  $f(x)$ . Какой из прямых методов минимизации можно использовать для этого?

**V.6.12.** Указать класс функций, для точного определения точек минимума которых достаточно одной итерации метода парабол.

**V.6.13.** В окрестности точки минимума  $x^*$  график функции  $f(x)$  близок к симметричному относительно вертикальной оси, проходящей через точку  $x^*$ . А график функции  $g(x)$  заметно асимметричен. Для какой из этих функций следует ожидать более высокой скорости сходимости метода парабол?

## V.7. Практические задачи

**V.7.1.** Используя методы дихотомии и сведения вариационной задачи решения алгебраического уравнения, найти точку локального минимума функции

a)  $f(x) = 2x^2 - \ln x$ ,      б)  $f(t) = t^3 / 3 + t^2$ ,

в)  $f(t) = t^4 / 4 - 2t^2$ ,      г)  $f(t) = te^{-t^2/2}$ ,

д)  $f(t) = 3t^4 - 8t^3 + 6t^2$ ,    е)  $f(t) = (t-5)e^t$ ,

ж)  $f(t) = (t^2 - 3) / (t + 2)$ .

**V.7.2.** Найти минимум функции  $y = x^4 + e^{-x}$ ,  $x \in [0, 1]$  методом парабол.

**V.7.3.** Найти минимум функции  $y = x^x$  методом исключения отрезков.

**V.7.4.** Для нахождения минимума функции  $f(x)$  сделать один шаг модифицированного метода Брендта от заданного начального приближения:

a)  $f(x) = 10x(2 - \ln x)^2$

$x$	$x_1 = 6.$	$x_2 = 7.$	$x_3 = 8.$	$x_4 = 9.$
$f(x)$	2.60185	0.20480	0.50488	3.50078

б)  $f(x) = 1/x + e^x$

$x$	$x_1 = 0.5$	$x_2 = 0.75$	$x_3 = 1.0$	$x_4 = 1.25$
$f(x)$	3.64872	3.45033	3.71828	4.29034

в)  $f(x) = 1/x + \cos x^2$

$x$	$x_1 = 1.0$	$x_2 = 1.5$	$x_3 = 2.0$	$x_4 = 2.5$
$f(x)$	1.54030	0.03849	-0.15364	1.39945

г)  $f(x) = 1/x + 4\ln(x+1)$

$x$	$x_1 = 0.1$	$x_2 = 0.5$	$x_3 = 0.9$	$x_4 = 1.3$
$f(x)$	10.38124	3.62186	3.67853	4.10087

д)  $f(x) = \sqrt{x} + \sin x$

$x$	$x_1 = 3.5$	$x_2 = 4.0$	$x_3 = 4.5$	$x_4 = 5.$
$f(x)$	1.52005	1.24320	1.14379	1.27715

е)  $f(x) = (x-1)^2 + \operatorname{arctg} x$

$x$	$x_1 = 0.$	$x_2 = 0.5$	$x_3 = 1.0$	$x_4 = 1.5$
$f(x)$	1	0.71365	0.78540	1.23279

**V.7.5.** Методом покоординатного спуска и с помощью динамического метода найти точки локального минимума функций

а)  $f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2,$

б)  $f(x, y) = (x^2 + y^2 - 1)^2 + (y - x \sin x)^2,$

в)  $f(x, y) = x^3 + 8y^3 - 6xy + 1,$

г)  $f(x, y) = (x - 3)^2 + (y - 2)^2 + (x - y - 4)^2,$

д)  $f(x, y) = 2x^3 - xy^2 + 5x^2 + y^2.$

**V.7.6.** Найти точку локального минимума функции методом спуска и динамическим методом:

a)  $f(x, y) = 3x^2 - 2x\sqrt{y} + y - 8x + 8,$

б)  $f(x, y) = x^3 + 8y^3 - 6xy + 1,$

в)  $f(x, y) = x^2 + y^2 + xy + x - y + 1,$

г)  $f(x, y) = 2x^3 - xy^2 + 5x^2 + y^2.$

## V.8. Библиографический комментарий

Изложение теоретического материала здесь следует пособиям [3, 5, 24, 27]. Методам оптимизации, в частности численным методам оптимизации, посвящена обширная литература, например, [32]. Отдельно отметим книги авторов, в течение многих лет читавших соответствующие курсы на Физтехе [28–31]. Модифицированный метод Брендта подробно описан в [33, 34].

# VI. ТАБЛИЧНОЕ ЗАДАНИЕ И ИНТЕРПОЛИРОВАНИЕ ФУНКЦИЙ

## VI.1. Задача интерполяции

Задача интерполяции состоит в нахождении обобщенного многочлена

$$P_n(x) = \sum_{k=0}^n c_k \varphi_k(x), \quad (1.1)$$

где  $\varphi_k(x)$  — фиксированные функции, а значения коэффициентов определяются из условия равенства со значением приближаемой функции в узлах интерполяции

$$P_n(x_k) = f_k, \quad k = 0, 1, \dots, n. \quad (1.2)$$

Набор точек  $x_j$  на интервале  $[a, b]$ ,  $a \leq x_0 < x_1 < \dots < x_n \leq b$ , в которых заданы значения функции  $f(x_j)$  называют *сеткой*. Множество точек  $x_j$  иногда также называют *узлами сетки* или *узлами интерполяции*.

Сетка называется *равномерной*, если

$$x_{j+1} - x_j = h = \text{const}, \quad j = 0, 1, \dots, n-1; \quad a = x_0, \quad b = x_n.$$

Если  $\varphi_k(x) = x^k$ , то соответствующая интерполяция называется *алгебраической*, если  $\varphi_k$  — тригонометрические функции, то говорят о *тригонометрической интерполяции*.

Если построенный обобщенный полином (1.1) используется для восстановления функции на всем отрезке  $[a, b]$ , то говорят о *глобальной* интерполяции. Если же для восстановления функции между каждыми двумя соседними узлами строится многочлен заданной невысокой степени, то говорят о *кусочно-многочленной* интерполяции.

Если значения функции  $f(x)$  заданы в узлах  $x_j$  на интервале  $[a, b]$ ,  $a \leq x_0 < x_1 < \dots < x_n \leq b$ , то говорят, что функция  $f(x)$  задана *таблицей*.

## VI.2. Алгебраическая интерполяция

Теорема 1. Пусть заданы  $n + 1$  узел  $x_0, x_1, \dots, x_n$ , среди которых нет совпадающих, и значения функции в этих узлах  $f(x_0), f(x_1), \dots, f(x_n)$ . Тогда существует один и только один многочлен  $P_n(x) = P_n(x, f, x_0, x_1, \dots, x_n)$  степени не выше  $n$ , принимающий в узлах  $x_k$  заданные значения  $f(x_k)$ .

Интерполяционный многочлен можно записать (и соответственно вычислить) различными способами, представляя его в виде разложения по степеням  $x$  (в форме Лагранжа и в форме Ньютона) или в виде *разложения по ортогональным многочленам*.

### VI.2.1. Непосредственное вычисление коэффициентов интерполяционного полинома

Полином степени  $n$  можно записать в виде

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n, \quad (2.1)$$

где  $a_0, \dots, a_n$  — неопределенные коэффициенты. Их можно определить из  $n+1$  условия:

$$\begin{aligned} a_0 + a_1 x_0 + a_2 x_0^2 + \dots + a_n x_0^n &= f(x_0), \\ a_0 + a_1 x_1 + a_2 x_1^2 + \dots + a_n x_1^n &= f(x_1), \\ \dots & \\ a_0 + a_1 x_n + a_2 x_n^2 + \dots + a_n x_n^n &= f(x_n). \end{aligned} \quad (2.2)$$

Определитель системы (2.2) есть детерминант Вандермонда, известный из курса линейной алгебры. Его значение в случае, когда выполняются условия теоремы 1, отлично от нуля, что доказывает существование и единственность полинома. Эта линейная система во многих случаях является плохо обусловленной. Последнее связано с тем, что последовательные степени  $1, x, x^2, \dots, x^n$  «почти линейно зависимы» на интервале  $0 < x < 1$ .

### VI.2.2. Интерполяционный полином в форме Лагранжа. Интерполяционный полином в форме Ньютона

Введем вспомогательные многочлены

$$l_k = \frac{(x-x_0) \dots (x-x_{k-1})(x-x_{k+1}) \dots (x-x_n)}{(x_k-x_0) \dots (x_k-x_{k-1})(x_k-x_{k+1}) \dots (x_k-x_n)}. \quad (2.3)$$

Многочлен  $P_n(x)$ , заданный равенством

$$P_n(x) = P_n(x, f, x_0, \dots, x_n) = f(x_0) l_0(x) + f(x_1) l_1(x) + \dots + f(x_n) l_n(x), \quad (2.4)$$

есть интерполяционный многочлен в форме Лагранжа.

Употребляются и другие виды записи интерполяционного многочлена (в форме Ньютона, Лебега или Стильеса). Для практического применения обычно используют запись в *форме Ньютона*.

Рекурсивно определим *разделенные разности*.

Разделенной разностью нулевого порядка  $f(x_k)$  функции  $f(x)$  в точке  $x_k$  назовем значение функции в этой точке

$$f(x_k) = f(x_k).$$

Разделенной разностью первого порядка  $f(x_k, x_t)$  функции  $f(x)$  для произвольной пары точек  $x_k$ , и  $x_t$  определим через разделенные разности нулевого порядка:

$$f(x_k, x_t) = \frac{f(x_t) - f(x_k)}{x_t - x_k}.$$

Разделенную разность  $f(x_0, x_1, \dots, x_n)$  порядка  $n$  определим через разделенную разность порядка  $n - 1$ , положив

$$f(x_0, x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n) - f(x_0, \dots, x_{n-1})}{x_n - x_0}.$$

Интерполяционный многочлен в *форме Ньютона*  $P_n(x, x_0, x_1, \dots, x_n)$  может быть записан через разделенные разности следующим образом:

$$\begin{aligned} P_n(x, f, x_0, x_1, \dots, x_n) &= f(x_0) + f(x_0, x_1)(x - x_0) + \\ &+ f(x_0, x_1, x_2)(x - x_0)(x - x_1) + \dots + \\ &+ f(x_0, x_1, \dots, x_n)(x - x_0) \dots (x - x_{n-1}). \end{aligned} \quad (2.5)$$

Если  $x_0 < x_1 < \dots < x_n$ , то соответствующую интерполяцию называют *интерполяцией вперед*; в случае  $x_0 > x_1 > \dots > x_n$  интерполяцию называют *интерполяцией назад*.

**Теорема 2.** *Интерполяционный многочлен в форме Ньютона* (2.5) эквивалентен *интерполяционному многочлену в форме Лагранжа* (2.4).

### VI.2.3. Формула погрешности алгебраической интерполяции

Оценим погрешность  $R_s(x) = f(x) - P_s(x, f)$ ,  $x_k < x < x_{k+1}$ , возникающую при приближенной замене  $f(x)$  алгебраическим многочленом  $P_s(x, f)$ . В основе оценки лежит следующая общая теорема о формуле погрешности.

**Теорема 3.** *Пусть  $f(x)$  определена на некотором отрезке  $a \leq x \leq \beta$  и имеет производные до некоторого порядка  $s + 1$  включительно.*

Пусть  $x_0, x_1, \dots, x_s$  — произвольный набор попарно различных точек отрезка  $[a, b]$ ,  $f(x_0), f(x_1), \dots, f(x_s)$  — значения функции  $f(x)$  в этих точках;  $P_s(x)$  — интерполяционный многочлен степени не выше  $s$ , построенный по этим значениям. Тогда погрешность интерполяции  $R_s(x) = f(x) - P_s(x)$  представляется формулой

$$R_s(x) = \frac{1}{(s+1)!} f^{(s+1)}(z)(x-x_0)(x-x_1)\dots(x-x_s), \quad (2.6)$$

где  $z = z(x)$  — некоторая точка отрезка  $[a, b]$ .

Величина, определенная равенством (2.10), называется *остаточным членом интерполяции*, по аналогии с остаточным членом в форме Лагранжа при представлении функции по формуле Тейлора. Заметим, что интерполяционные формулы в вычислительной математике играют ту же роль, что формула Тейлора в математическом анализе.

Из (2.6) следует оценка

$$\|R_s(x)\| \leq \frac{1}{(s+1)!} \|f^{(s+1)}(x)\|_C \cdot \|(x-x_0)(x-x_1)\dots(x-x_s)\|, \quad (2.7)$$

где под нормой функции понимается максимальное значение ее абсолютной величины на отрезке.

Вопрос об оптимальном выборе узлов интерполяции связан с минимизацией нормы функции

$$\omega(x) = (x-x_0)(x-x_1)\dots(x-x_s).$$

На отрезке  $[-1, 1]$  минимальна норма приведенного многочлена Чебышёва (см. приложение).

#### VI.2.4. О сходимости интерполяционного процесса

На отрезке  $a \leq x \leq b$  будем рассматривать бесконечную последовательность узлов интерполяции

$$\begin{aligned} &x_1^1 \\ &x_1^2, x_2^2 \\ &x_1^3, x_2^3, x_3^3 \\ &\dots \\ &x_1^n, x_2^n, x_3^n, \dots, x_n^n \end{aligned}$$

и соответствующую последовательность интерполяционных многочленов

$P_n(x, f)$ , построенную для некоторой функции  $f(x)$ , принимающей конечные значения во всех узлах интерполяции.

**Теорема 4. (Фабера).** *Какова бы ни была последовательность узлов интерполяции, существует непрерывная функция  $f$ , для которой последовательность интерполяционных многочленов расходится.*

**Теорема 5.** *Для каждой функции  $f$ , непрерывной на конечном отрезке, существует такая последовательность узлов интерполяции, что соответствующий ей интерполяционный процесс равномерно сходится к  $f$ .*

**Теорема 6.** *Не существует последовательности узлов, для которой интерполяционный процесс был бы равномерно сходящимся для всякой непрерывной на отрезке функции.*

**Теорема 7.** *Если функция  $f$  имеет ограниченную производную на отрезке, то интерполяционный процесс, в котором за узлы принимаются корни многочленов Чебышёва, сходится равномерно к  $f$ .*

Интерполирующую функцию иногда называют *интерполянтом*. Гладкий кусочно-многочленный интерполянт называется *сплайном*.

### VI.2.5. Обусловленность задачи интерполяции

Интерполяционный многочлен

$$P_n(x) = \sum_{k=0}^n c_k \varphi_k(x), \quad (2.8)$$

где  $\varphi_k(x)$  — фиксированные функции, а значения коэффициентов  $c_k$  определяются из условия совпадения со значениями приближаемой функции в узлах интерполяции, можно записать в виде (ср. с (2.4))

$$P_n(x) = \sum_{k=0}^n f_k l_k(x), \quad (2.9)$$

для  $l_k(x_i) = 0, k \neq i, l_k(x_k) = 1; k, i = 0, 1, \dots, n$ . Многочлены  $l_k(x)$  иногда называют **фундаментальными полиномами**.

Придадим значениям функции  $f(x_j)$  возмущения  $\delta f(x_j)$ . Интерполяционный многочлен  $P_n(x, f)$  заменится многочленом  $P_n(x, f + \delta f)$ .

Так как  $P_n(x, f + \delta f) = P_n(x, f) + P_n(x, \delta f)$  в силу линейности (2.14) по  $f$ , то возмущение  $P_n(x, \delta f)$ , которое претерпевает интерполяционный многочлен, можно оценить как:

$$\|P_n(x, \delta f)\| \leq \|\delta f(x)\| \sum_{k=0}^n |l_k(x)|. \quad (2.10)$$

Здесь, как и в (2.7), под нормой функции понимается ее максимальное значение на отрезке.

Это возмущение при заданных узлах интерполяции и фиксированных базисных функциях  $l_k(x)$  зависит только от  $\delta f$ .

Введем в рассмотрение функцию  $L_n(x) = \sum_{k=0}^n |l_k(x)|$ , которая называется *функцией Лебега*.

За меру чувствительности интерполяционного многочлена к возмущениям задания функции в узлах  $\delta f$  принимается наименьшее число  $L_n$ , при котором для каждого  $\delta f$  выполнено неравенство

$$\max_{a < x < b} |P_n(x, \delta f)| \leq L_n \|\delta f(x)\|. \quad (2.11)$$

Числа  $L_n$ , зависящие от  $a = x_0, x_1, \dots, x_n = b$ , называют *константами Лебега*. Эти числа растут с ростом  $n$ .

Очевидно, что  $L_n = \max_{a < x < b} L_n(x)$ .

Для алгебраической интерполяции ( $\varphi_k(x) = x^k$ ) в случае равномерно расположенных узлов доказана оценка

$$\frac{2^{n-3}}{(n-3/2)\sqrt{n-1}} < L_n < 2^{n-1}, \quad n \geq 4,$$

т.е. чувствительность результата интерполяции к погрешностям задания функции в узлах резко возрастает с ростом  $n$ . Такие погрешности неизбежны как при получении табличных значений в результате измерений, так и в результате округлений.

Если узлами интерполяции являются корни полинома Чебышёва, то

$$L_n = \frac{2}{\pi} \ln n + 1 - \theta_n, \quad 0 \leq \theta_n \leq 1/4,$$

т.е. с ростом  $n$  константы Лебега растут очень медленно. Говорят, что узлы в нулях многочлена Чебышёва являются асимптотически оптимальными. В этом случае вычислительная неустойчивость не является препятствием для использования интерполяционных многочленов высокой степени. Аналогичная оценка константы Лебега справедлива и при выборе точек интерполяции в точках экстремума многочленов Чебышёва.

## VI.3. Тригонометрическая интерполяция

### VI.3.1 Постановка задачи

Задача (линейной) тригонометрической интерполяции состоит в построении тригонометрического многочлена вида

$$\begin{aligned} Q_n \left( \cos \frac{2\pi(x - x_0)}{L}, \sin \frac{2\pi(x - x_0)}{L} \right) &= \\ &= \sum_{k=0}^n a_k \cos \frac{2\pi k(x - x_0)}{L} + \sum_{k=1}^n b_k \sin \frac{2\pi k(x - x_0)}{L}. \end{aligned}$$

Здесь  $k$  и  $n$  — натуральные числа,  $[x_0, x_n]$  — отрезок интерполяции,  $L = x_n - x_0$  — положительное число (длина отрезка интерполяции),  $a_k$  и  $b_k$  — числовые коэффициенты.

Теорема 8. (Первый вариант задания узлов интерполяции). Пусть  $N = 2n + 1$ ,  $n$  — натуральное число. При произвольном задании значений функции  $f_m$ , периодической с периодом  $L$ , в узлах сетки

$$x_m = \frac{Lm}{N} + \frac{L}{2N}, \quad m = 0, 1, \dots, N-1,$$

существует один и только один интерполяционный тригонометрический многочлен

$$Q_n \left( \cos \frac{2\pi x}{L}, \sin \frac{2\pi x}{L}, f \right) = \sum_{k=0}^n a_k \cos \frac{2\pi kx}{L} + \sum_{k=1}^{n+1} b_k \sin \frac{2\pi kx}{L},$$

удовлетворяющий равенствам  $Q_n(x_m) = f_m$ ,  $m = 0, \dots, N-1$ .

Коэффициенты этого многочлена задаются формулами

$$a_0 = \frac{1}{N} \sum_{m=0}^{N-1} f_m, \quad b_{n+1} = \frac{1}{N} \sum_{m=0}^{N-1} (-1)^m f_m,$$

$$a_k = \frac{2}{N} \sum_{m=0}^{N-1} f_m \cos k \left( \frac{2\pi m}{N} + \frac{\pi}{N} \right), \quad k = 1, 2, \dots, n,$$

$$b_k = \frac{2}{N} \sum_{m=1}^{N-1} f_m \sin k \left( \frac{2\pi m}{N} + \frac{\pi}{N} \right), \quad k = 1, 2, \dots, n.$$

**Теорема 9.** (Второй вариант задания узлов интерполяции) Пусть  $N = 2n$ ,  $n$  — натуральное число. При произвольном задании значений функции  $f_m$ , периодической с периодом  $L$ , в узлах сетки

$$x_m = Lm/N, \quad m = 0, \pm 1, \dots, \pm (n - 1),$$

существует единственный интерполяционный тригонометрический многочлен

$$Q_n \left( \cos \frac{2\pi x}{L}, \sin \frac{2\pi x}{L}, f \right) = \sum_{k=0}^n a_k \cos \frac{2\pi kx}{L} + \sum_{k=1}^{n-1} b_k \sin \frac{2\pi kx}{L},$$

удовлетворяющий равенствам  $Q_n(x_m) = f_m$ ,  $m = 0, \pm 1, \pm 2, \dots, \pm (n - 1)$ .

Коэффициенты этого многочлена задаются формулами

$$a_0 = \frac{1}{n} \sum_{m=0}^n f_m,$$

$$a_k = \frac{2}{N} \sum_{m=0}^{N-1} f_m \cos \frac{2\pi km}{N}, \quad k = 1, 2, \dots, n-1.$$

$$b_k = \frac{2}{N} \sum_{m=1}^{N-1} f_m \sin \frac{2\pi km}{N}, \quad k = 1, 2, \dots, n-1.$$

$$a_n = \frac{1}{N} \sum_{m=0}^{N-1} (-1)^m f_m,$$

**Теорема 10.** (Произвольное расположение узлов интерполяции на отрезке периодичности) Пусть заданы значения  $f_i$ :  $i = 1, \dots, N$  периодической с периодом  $L$  функции  $f(x)$  в  $N = 2n$  несовпадающих точках  $x_i$ :  $i = 1, \dots, N$ , принадлежащих отрезку  $[a, b]$ ,  $b - a = L$ ,  $f(a) = f_1 = f_N = f(b)$ .

Тогда существует один и только один интерполяционный тригонометрический многочлен

$$Q_n \left( \cos \frac{2\pi (x-a)}{L}, \sin \frac{2\pi (x-a)}{L}, f \right) = \sum_{k=0}^{N-1} f_k \cdot l_k(x),$$

$$l_k(x) = \frac{\sin \frac{\pi (x-x_1-a)}{L} \dots \sin \frac{\pi (x-x_i-a)}{L} \dots \sin \frac{\pi (x-x_{N-1}-a)}{L}}{\sin \frac{\pi (x_k-x_1-a)}{L} \dots \sin \frac{\pi (x_k-x_i-a)}{L} \dots \sin \frac{\pi (x_k-x_{N-1}-a)}{L}}.$$

В произведениях отсутствует сомножитель, соответствующий  $i = k$ , так что  $l_k(x_k) = 1$ .

Тригонометрические многочлены обладают определенными преимуществами перед алгебраическим многочленом. Во-первых, при  $N \rightarrow \infty$  погрешность тригонометрической интерполяции

$$R_N(x, f) = f(x) - Q\left(\cos \frac{2\pi x}{L}, \sin \frac{2\pi x}{L}, f\right)$$

равномерно стремится к нулю, если  $f(x)$  имеет хотя бы вторую производную, причем скорость убывания погрешности автоматически учитывает гладкость  $f(x)$ , т. е. возрастает с ростом числа  $(r+1)$  производных:

$$\max_x |R_N(x)| = O\left(M_{r+1} \frac{\ln N}{N^r}\right), \quad M_{r+1} = \max_x \left| \frac{d^{r+1} f(x)}{dx^{r+1}} \right|.$$

Во-вторых, чувствительность тригонометрического интерполяционного многочлена к погрешности задания значений  $f_m$  в узлах с ростом числа узлов «почти не возрастает».

Эти два положительных свойства тригонометрической интерполяции, а именно, возрастание точности при увеличении гладкости и вычислительную устойчивость, можно придать и алгебраической интерполяции функций на отрезке за счет специального выбора узлов интерполяции и использования алгебраических многочленов Чебышева, обладающих многими замечательными свойствами.

### VI.3.2. Обусловленность тригонометрической интерполяции

Чувствительность интерполяционного тригонометрического многочлена к погрешности задания значений  $f_m$  оценивается следующим образом. Пусть вместо  $f = [f_m]$  задана сеточная функция  $f + \delta f = \{f_m + \delta f_m\}$ . Тогда возникающая погрешность

$$\delta Q_n = Q_n\left(\cos \frac{2\pi x}{L}, \sin \frac{2\pi x}{L}, \delta f\right)$$

и, следовательно, мерой чувствительности интерполяционного тригонометрического многочлена к возмущению  $\delta f$  входных данных могут служить числа  $L_n$ , называемые *константами Лебега* (2.16):

$$\max_x |\delta Q_n| \leq L_n \max_m |f_m|.$$

*Теорема 11. Константы Лебега тригонометрического интерполяционного многочлена удовлетворяют оценке  $L_n \leq 2n$ .*

Интерполяционный полином на сетке из нулей или экстремумов полиномов Чебышёва наследует от тригонометрической интерполяции слабый рост константы Лебега при увеличении  $n$ .

## VI.4. Классическая кусочно-многочленная интерполяция

Рост константы Лебега при увеличении количества узлов сетки, а также пример простой функции (функция Рунге – см. задачи на доказательство), для которой алгебраический интерполяционный процесс на равномерной сетке не сходится, являются причинами того, что для функций, заданных таблично на большом числе точек, не используют интерполянты высоких степеней. Обычно в таких случаях переходят к построению сплайнов. Рассмотрим вначале более простую задачу кусочно-многочленной интерполяции.

Пусть функция  $f(x)$  задана таблицей. Для восстановления функции между узлами можно воспользоваться функцией, которая между каждыми двумя соседними узлами является многочленом заданной невысокой степени, например, первой, второй, третьей и т. д.

Соответствующая интерполяция называется *кусочно-линейной*, *кусочно-квадратичной* и т. п. Такую интерполяцию можно строить локально, не привлекая данные с отдаленных участков задания функции, а можно строить интерполяцию, пытаясь сохранить максимальную гладкость построенной функции. В последнем случае все коэффициенты локальной интерполяции (кроме кусочно-линейной) оказываются связанными между собой.

### VI.4.1. Оценка неустранимой погрешности при интерполяции

Пусть функция  $f(x)$  определена на отрезке  $[0, \pi]$  и пусть заданы ее значения в узлах равномерной сетки  $x_k = k\pi/n$ ,  $k = 0, 1, \dots, n$ . По таблице  $f(x_0)$ ,  $f(x_1)$ , ...,  $f(x_n)$  в принципе, нельзя восстановить функцию  $f(x)$  точно, потому что значения различных функций могут совпадать в точках  $x_k$ ,  $k = 1, \dots, n$ , т. е. различные функции могут иметь одинаковую таблицу.

Если о функции известно лишь то, что она непрерывна, то ее нельзя восстановить в точке  $x \neq x_k$ ,  $k = 0, 1, \dots, n$ , ни с какой гарантированной точностью.

Укажем две функции:

$$f_{(I)}(x) = \frac{\sin nx}{n^{s+1}}, \quad f_{(II)}(x) = -\frac{\sin nx}{n^{s+1}},$$

для которых таблицы  $f_{(I)}(x_k) = f_{(II)}(x_k)$ ,  $k = 0, 1, \dots, n$ , совпадают (обе таблицы содержат лишь нули). Эти функции уклоняются друг от друга на величину

порядка  $h^{s+1}$ .

$$\max_x |f_{(I)}(x) - f_{(II)}(x)| = 2 \max_x \left| \frac{\sin nx}{n^{s+1}} \right| = 2h^{s+1}.$$

Таким образом, зная лишь оценку  $s+1$  производной, в принципе нельзя восстановить табличную функцию с точностью, большей, чем  $O(h^{s+1})$ . Данная погрешность неустранима.

#### VI.4.2. Насыщаемость (гладкостью) кусочно-многочленной интерполяции

Пусть функция  $f(x)$  определена на отрезке  $[a, b]$ , и задана ее таблица  $f(x_k)$  в равноотстоящих узлах  $x_k$ ,  $k = 0, 1, \dots, n$ ; с шагом  $h = (b - a)/n$ .

Погрешность кусочно-многочленной интерполяции степени  $s$  (с помощью интерполяционных многочленов  $P_s(x, f_{kj})$  на отрезке  $x_{k-j} \leq x \leq x_k$ ) в случае, если на  $[a, b]$  существует и ограничена  $f^{(s+1)}(x)$ , имеет порядок  $O(h^{s+1})$ .

Если о функции  $f(x)$  известно лишь, что она имеет ограниченную производную до некоторого порядка  $q$ ,  $q < s$ , то неустранимая погрешность при ее восстановлении по таблице есть  $O(h^{q+1})$ . Можно показать, что при интерполяции с помощью  $P_s(x, f_{kj})$  порядок  $O(h^{q+1})$  достигается.

Если  $f(x)$  имеет ограниченную производную порядка  $q+1$ ,  $q > s$ , то погрешность интерполяции с помощью  $P_s(x, f_{kj})$  остается  $O(h^{q+1})$ , т. е. порядок погрешности не реагирует на дополнительную, сверх  $s+1$  производную, гладкость функции  $f(x)$ . Это свойство кусочно-многочленной интерполяции называют свойством насыщаемости (гладкостью).

#### VI.4.3. Нелокальная гладкая кусочно-многочленная интерполяция

Пусть задана таблица функции на некотором отрезке. Поставим задачу найти на каждом отрезке разбиения  $x_k \leq x \leq x_{k+1}$ ,  $k = 0, 1, \dots, n-1$ , кубический многочлен  $P_3(x, k)$  так, чтобы возникающая при этом на отрезке  $a \leq x \leq b$  кусочно-многочленная функция совпадала с заданной сеточной функцией в узлах и имела непрерывные производные до порядка  $s=2$ . Разность между степенью интерполяционного полинома и глобальной гладкостью сплайна называется *дефектом сплайна*. Точки, в которых заданы значения функции, называются *узлами интерполяции*, а точки, в которых сшиваются интерполяционные многочлены, — *узлами сплайна*. Вообще говоря, узлы интерполяции и узлы сплайна могут не совпадать.

Будем строить кубический сплайн дефекта 1 при условии совпадения узлов интерполяции и узлов сплайна. Кубический многочлен задается четырьмя коэффициентами на каждом интервале, всего  $4n$  коэффициентов. Условие интерполяции на каждом из отрезков разбиения слева и справа даст  $2n$  условий, требование непрерывности первой и второй производной во внутренних узлах приведет еще к  $2(n - 1)$  условию, всего  $4n - 2$  условия для нахождения  $4n$  коэффициентов.

Недостающие условия можно задавать различными способами. Наиболее употребляемыми являются следующие:

1) «свободный» сплайн, соответствующий минимуму потенциальной энергии упругой линейки, поставленной на ребро, и закрепленной так, чтобы она проходила через заданные точки:

$$\frac{d^2 P_3(x, 0)}{dx^2} = \frac{d^2 P_3(x, n)}{dx^2} = 0; \quad (4.1)$$

2) кубический сплайн Шёнберга

$$\frac{d^3 P_3(x, 0)}{dx^3} = \frac{d^3 c_0}{dx^3}, \quad \frac{d^3 P_3(x, n)}{dx^3} = \frac{d^3 c_n}{dx^3}. \quad (4.2)$$

Здесь  $c_0(x), c_n(x)$  — единственные кубические кривые, которые проходят соответственно через четыре первые и четыре последние заданные точки.

3) эмпирический сплайн, требующий непрерывности третьей производной в первой и предпоследней точке интервала («not a knot»). В стандартных пакетах для построения сплайнов именно этот случай реализуется чаще всего.

4) естественный сплайн, обеспечивающий минимизацию разрывов последней существующей производной.

*Теорема 10. Интерполяционный кубический сплайн  $S(x)$  дефекта 1, удовлетворяющий одному из краевых условий (4.1–4.2), существует и единственен.*

Рассмотрим построение кубического сплайна в общем случае на неравномерной сетке  $x_n - x_{n-1} = h_{n-1}$ ,  $x_{n+1} - x_n = h_n$ . Пусть  $m_n$  — значение второй производной в точке  $x_n$  (пока неизвестное!). На каждом отрезке  $[x_n, x_{n+1}]$  для второй производной кусочно-кубического сплайна имеем:

$$S''_{xx} = \frac{1}{h_n} (m_n (x_{n+1} - x) + m_{n+1} (x - x_n)).$$

Так как сплайн — полином третьей степени, то его вторая производная — линейная функция. Интегрируем выражение для второй производной

сплайна, получаем на отрезке  $[x_n, x_{n+1}]$ :

$$S'_x = \frac{1}{h_n} \left( m_{n+1} \frac{(x_{n+1}-x)^2}{2} - m_n \frac{(x-x_n)^2}{2} \right) + A_n.$$

$A_n$  — константа интегрирования. Интегрируя последнее равенство еще раз, получаем

$$S(x) = \frac{1}{6h_n} \left( m_n(x_{n+1}-x)^3 + m_{n+1}(x-x_n)^3 \right) + A_n x + B_n.$$

Линейное слагаемое в этом равенстве, включающее в себя две произвольные константы, перепишем в виде

$$A_n x + B_n = (\beta_n - \alpha_n)x + \alpha_n x_{n+1} - \beta_n x_n,$$

т. е. вместо двух констант интегрирования  $A_n, B_n$  введем две новые константы, более удобные для дальнейших выкладок. Тогда

$$\begin{aligned} S(x) = & \frac{1}{6h_n} \left( m_n(x_{n+1}-x)^3 + m_{n+1}(x-x_n)^3 \right) + \\ & + \alpha_n (x_{n+1}-x) + \beta_n (x-x_n). \end{aligned}$$

Из условий интерполяции  $S(x_n) = f_n, S(x_{n+1}) = f_{n+1}$  получаем

$$\begin{aligned} f_n = & \frac{m_n h_n^2}{6} + \alpha_n h_n \Rightarrow & \alpha_n = & \frac{f_n}{h_n} - \frac{m_n h_n}{6}, \\ f_{n+1} = & \frac{m_{n+1} h_n^2}{6} + \beta_n h_n \Rightarrow & \beta_n = & \frac{f_{n+1}}{h_n} - \frac{m_{n+1} h_n}{6}, \end{aligned}$$

$$A_n = \frac{f_{n+1} - f_n}{h_n} - \frac{(m_{n+1} - m_n)h_n}{6}.$$

Приравняем первые производные в каждом узле сетки справа и слева (кроме граничных):  $S'_x(x_n + 0) = S'_x(x_n - 0)$ , получим систему уравнений для определения коэффициентов сплайна:

$$\begin{aligned} \frac{m_n h_{n-1}}{2} - \frac{m_{n-1} h_{n-1}}{2} + \frac{f_n - f_{n-1}}{h_{n-1}} - \frac{(m_n - m_{n-1})h_{n-1}}{6} = \\ = \frac{m_{n+1} h_n}{2} - \frac{m_n h_n}{2} + \frac{f_{n+1} - f_n}{h_n} - \frac{(m_{n+1} - m_n)h_n}{6}, \end{aligned} \quad (4.3)$$

Эта система дополняется соответствующими граничными условиями. В случае свободного сплайна  $m_0 = m_N = 0$ .

Для определения коэффициентов  $m_n$  получена система линейных уравнений с трехдиагональной матрицей. Матрица этой системы симметрична, имеет свойство диагонального преобладания и, как можно показать, положительно определена, а, следовательно, неособенная. Значит, решение рассматриваемой СЛАУ существует и единствено. Следовательно, и задача о построении кубического сплайна имеет единственное решение. Для других типов краевых условий доказательство проводится аналогично.

Коэффициенты  $m_i$  называются моментами кубического сплайна.

Теорема 11. Пусть  $S_3(x)$  кубический сплайн дефекта 1, интерполирующий на системе узлов  $a = x_0 < x_1 < \dots < x_n = b$  четырежды непрерывно дифференцируемую на  $[a, b]$  функцию  $f(x)$ . Тогда  $\forall n \exists c > 0: \forall x \in [a, b]$  справедливо неравенство

$$|f(x) - S_3(x)| \leq c\Delta^4, \text{ где } \Delta = \max_{i=0, \dots, n-1} h_i.$$

Порядок аппроксимации производной для каждой следующей производной уменьшается на единицу.

## VI.5. Дробно-полиномиальные аппроксимации

Многочленная интерполяция не всегда удобна на бесконечных интервалах или для функций с полюсами. Кроме того, в ряде случаев мы можем строить приближение функции по ее поведению в окрестности отдельных точек или на бесконечности, не используя точек интерполяции вообще. В этом случае построенные приближения не будут интерполяциями в строгом смысле этого слова. Однако такие приближения зачастую бывают исключительно полезны. В этом разделе мы будем использовать для них слово аппроксимация, хотя строго этот термин будет введен только в VIII главе.

### VI.5.1. Рациональная интерполяция

В ряде случаев большую точность приближения можно получить, используя рациональную интерполяцию. Это важно в тех случаях, когда интервал аппроксимации бесконечен и поведение функции на бесконечности не описывается степенной зависимостью или при наличии у функции полюсов.

При заданных значениях функции в узлах  $f(x_1), \dots, f(x_n)$  приближение к  $f(x)$  ищется в виде

$$R(x) = \frac{a_0 + a_1 x + \dots + a_p x^p}{b_0 + b_1 x + \dots + b_q x^q}, \quad n = p + q + 1. \quad (5.1)$$

Коэффициенты  $a_i, b_i$  определены с точностью до общего множителя, и находятся из условий  $R(x_j) = f(x_j)$ ,  $j = 1, \dots, n$ , или

$$\sum_{k=0}^p a_k x_j^k - f(x_j) \sum_{k=0}^q b_k x_j^k = 0, \quad j = 1, \dots, n. \quad (5.2)$$

Система (5.2) представляет собой систему  $n$  уравнений относительно  $n + 1$  неизвестного. Превышение количества неизвестных над количеством уравнений является естественным, т.к. числитель и знаменатель определены с точностью до общего множителя.

Существует рекурсивный алгоритм вычисления функция  $R(x)$  в случаях, когда  $n$  — нечетно и  $p = q$ , и когда  $n$  — четно и  $p = q + 1$ .

Для начала определим два набора величин:

$n$  величин  $R_{1,0} = 0 \quad R_{2,1} = 0 \quad \dots \quad R_{n,n-1} = 0$ ,

$n$  величин  $R_{1,1} = f_1 \quad R_{2,2} = f_2 \quad \dots \quad R_{n,n} = f_n$ .

После этого с помощью формулы

$$R_{i,k} = R_{i+1,k} + \frac{R_{i+1,k} - R_{i,k-1}}{\left( \frac{x - x_i}{x - x_k} \right) \left( 1 - \frac{R_{i+1,k} - R_{i,k-1}}{R_{i+1,k} - R_{i+1,k-1}} \right) - 1}$$

последовательно вычисляются

$n$  величин  $R_{1,2}, \quad R_{2,3} \quad \dots \quad R_{n-1,n}$ ,

$n - 1$  величин  $R_{1,2}, \quad R_{2,3} \quad \dots \quad R_{n-1,n}$

и так далее до вычисления

двух величин  $R_{1,n-1}$  и  $R_{2,n}$ ,

одной величины  $R_{1,n}$ ,

которая и является ответом.

Все деления должны сопровождаться проверками на нуль в знаменателе, например, при  $x = x_k$  вся большая дробь в последнем выражении должна быть положена равной нулю.

Как и для полиномиальной интерполяции, алгоритм позволяет оценить погрешность по последней поправке — разнице между  $R_{1,n}$  и  $R_{1,n-1}$  или  $R_{2,n}$ .

Пример. Рассмотрим построение рациональной интерполяции для функции  $f(x) = \operatorname{arctg} x$  на интервале  $0 \leq x < \infty$ . Существует предел  $f(+\infty) = \pi/2$ . Это означает, что в формуле (5.2) нужно выбрать  $p = q$ . Кроме

того, при  $x \rightarrow 0$  функция  $f(x) \approx x$  и раскладывается в ряд по нечетным степеням  $x$ . Для того чтобы удовлетворить обоим требованиям при не слишком большой степени полиномов, удобнее аппроксимировать не  $f(x)$ , а ее квадрат:

$$f^2(x) = \arctg^2 x \approx P_p(x^2) / Q_p(x^2).$$

Для удовлетворения граничных условий надо полагать

$$a_0 = 0, \quad a_1/b_0 = 1, \quad a_p = (\pi/2)^2, \quad b_p = 1.$$

В этом случае свободными остаются  $2p - 2$  параметров. Столько же нужно брать узлов интерполяции на  $[0, +\infty]$ . При  $p = 1$  дополнительных условий уже не требуется, так что после извлечения корня и небольших преобразований имеем без единой точки интерполяции

$$\arctg x = \frac{x}{\sqrt{1 + (2x/\pi)^2}}.$$

Даже такое грубое приближение дает погрешность не хуже 12%. При введении узлов интерполяции точность быстро возрастает (при предлагаемом способе аппроксимации узлы интерполяции нужно добавлять парами).

### VI.5.2. Аппроксимация Паде

Пусть задан степенной ряд

$$f(z) = \sum_{k=0}^{\infty} c_k z^k, \quad (5.3)$$

определяющий функцию  $f(z)$ .

Аппроксимация Паде — рациональная функция вида

$$[L/M] = \frac{a_0 + a_1 z + \dots + a_L z^L}{b_0 + b_1 z + \dots + b_M z^M}, \quad (5.4)$$

разложение которой в ряд Тейлора (с центром в нуле) совпадает с разложением (5.3) до тех членов, пока это возможно. Как и в случае рациональной аппроксимации, весь набор из  $L + M + 2$  коэффициентов определен с точностью до общего множителя. Мы для определенности положим  $b_0 = 1$ . Теперь мы имеем  $L + M + 1$  свободных параметров.

Коэффициенты разложения функции  $[L/M]$  в ряд совпадают с коэффициентами при степенях  $1, z, z^2, \dots, z^{L+M}$  в (5.3). Должно выполняться

$$\sum_{k=0}^{\infty} c_k z^k = \frac{a_0 + a_1 z + \dots + a_L z^L}{1 + b_1 z + \dots + b_M z^M} + O(z^{L+M+1}). \quad (5.5)$$

Если (5.3) сходится к  $f(z)$  в круге  $|z| < R$ , где  $0 < R < \infty$ , то последовательность аппроксимаций Паде (5.4) может сходиться в более широкой области.

Чтобы продемонстрировать, насколько хорошо применимы аппроксимации Паде в естественных ситуациях, рассмотрим следующий пример:

$$f(z) = \sqrt{\frac{1+z/2}{1+2z}} = 1 - \frac{3}{4}z + \frac{39}{32}z^2 - \dots \quad (5.6)$$

Вычислим аппроксимацию Паде [1 / 1] этой функции из условий (5.5):

$$\frac{a_0 + a_1 z}{1 + b_1 z} = 1 - \frac{3}{4}z + \frac{39}{32}z^2 + O(z^3),$$

или, умножая на знаменатель,

$$a_0 + a_1 z + O(z^3) = (1 + b_1 z) \left( 1 - \frac{3}{4}z + \frac{39}{32}z^2 \right).$$

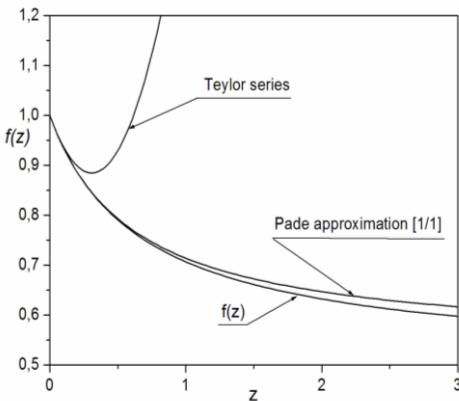


Рис. 5.1. Функция (5.6), сумма трех членов ее разложения в ряд Тейлора и аппроксимация Паде [1 / 1] этой функции

Приравнивая коэффициенты при одинаковых степенях  $z$ , получим систему уравнений для коэффициентов Паде-аппроксимации. В нашем примере приравнивая коэффициент при  $z^2$  слева и справа, получим  $b_1 = 13/8$ . Свободные члены слева и справа дают  $a_0 = 1$ . И, наконец, из равенства коэффициентов при  $z$  можно получить  $a_1 = 7/8$ . Таким образом, мы нашли

$$[1/1] = \frac{1 + 7z/8}{1 + 13z/8}.$$

На рис. 5.1 приведены для сравнения графики функций  $f(z)$ , первые три члена разложения в ряд Тейлора и  $[1 / 1](z)$  при  $z \geq 0$ . В частности, имеем

$$f(\infty) = 0.5, \quad [1/1](\infty) = \frac{7}{13} = 0.54\dots$$

Этот пример показывает замечательную точность, достигаемую аппроксимацией Паде, которая использует всего три члена разложения.

В таблице 1 представлены несколько Паде-аппроксимаций экспоненты, большая часть которых встретится в дальнейшем в качестве функций устойчивости неявных методов Рунге–Кутты.

Таблица 1

**Аппроксимация Паде функции  $\exp(z)$**

M \ L	0	1	2
0	$\frac{1}{1}$	$\frac{1+z}{1}$	$\frac{1+z+z^2/2}{1}$
1	$\frac{1}{1-z}$	$\frac{1+z/2}{1-z/2}$	$\frac{1+2z/3+z^2/6}{1-z/3}$
2	$\frac{1}{1-z+z^2/2}$	$\frac{1+z/3}{1-2z/3+z^2/6}$	$\frac{1+z/2+z^2/12}{1-z/2+z^2/12}$

## VI.6. Задачи с решениями

**VI.6.1.** Вывести расчетные формулы для построения свободного сплайна методом неопределенных коэффициентов.

**Решение.** Для построения сплайна можно искать решение на каждом отрезке  $[x_k, x_{k+1}]$  методом неопределенных коэффициентов в виде

$$S_3(x, k) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3.$$

Общее число неопределенных коэффициентов —  $4n$ , где  $n$  — количество отрезков, на которые разбит отрезок  $[a, b]$ . Условия интерполяции на левом и правом концах отрезка  $[x_k, x_{k+1}]$  при

$$h_k = x_{k+1} - x_k$$

дают

$$a_k = f(x_k), \quad k = 0, 1, \dots, n-1 \quad (6.1)$$

$$a_k + b_k h_k + c_k h_k^2 + d_k h_k^3 = f(x_{k+1}), \quad k = 0, 1, \dots, n-1. \quad (6.2)$$

Непрерывность первой производной в точке  $x_k$  слева и справа приведет к

$$b_{k-1} + 2c_{k-1}h_{k-1} + d_{k-1}h_{k-1}^2 = b_k, \quad k = 1, 2, \dots, n-1. \quad (6.3)$$

Непрерывность второй производной

$$2c_{k-1} + 6d_{k-1}h_{k-1} = 2c_k, \quad k = 1, \dots, n-1. \quad (6.4)$$

Дополнительные условия свободного сплайна приводят к условиям на коэффициенты

$$2c_0 = 0, \quad (6.5)$$

$$2c_{n-1} + 6d_{n-1}h_{n-1} = 0. \quad (6.6)$$

В задаче нет  $n$ -го отрезка. Искусственно введем коэффициент  $c_n$ , положив

$$c_n = 0, \quad (6.7)$$

а граничное условие (6.6) приведем к виду условия (6.4).

Из условия (6.4) со сдвигом индексов имеем

$$d_k = (c_{k+1} - c_k) / (3h_k). \quad (6.8)$$

Подстановка (6.8) и (6.1) в (6.2) приводит к

$$f(x_k) + b_k h_k + c_k h_k^2 + h_k^3(c_{k+1} - c_k) / (3h_k) = f(x_{k+1}),$$

или

$$b_k = (f_{k+1} - f_k) / h_k - h_k(c_{k+1} + 2c_k) / 3. \quad (6.9)$$

Подставляя в (6.3) выражение (6.9), для соседних интервалов после приведения подобных членов получим систему линейных уравнений относительно коэффициентов  $c_k$ :

$$\begin{aligned} h_{k-1}c_{k-1} + 2(h_{k-1} + h_k)c_k + h_kc_{k+1} &= \\ = 3((f_{k+1} - f_k) / h_k - (f_k - f_{k-1}) / h_{k-1}). \end{aligned} \quad (6.10)$$

Это система с диагональным преобладанием и граничными условиями (6.5) и (6.7), решение этой системы существует и единственno. Для решения систем с матрицей трехдиагональной структуры существует эффективный алгоритм прогонки. После нахождения коэффициентов  $c_k$  остальные находятся с помощью формул (6.1), (6.9), (6.8).

**VI.6.2.** (Т.К. Старожилова). При исследовании некоторой химической реакции через каждые 10 минут измерялась концентрация одного из реагентов. Результаты измерений представлены в таблице.

$t$ , мин	15	25	35	45	55	65	75
$C$ , моль/литр	700	370	160	130	220	220	220

С помощью интерполяции найти минимальную концентрацию реагента. Метод интерполяции выберите самостоятельно, обоснуйте его выбор.

**Решение.** Пронумеруем измерения  $t_n = 15 + n\tau$ ,  $\tau = 10$ ,  $n = 0, 1, \dots, 6$  и построим таблицу разностей:

$N$	0	1	2	3	4	5	6
$C$	700	370	160	130	220	220	220
$(\Delta C)_n = C_n - C_{n-1}$	-330	-210	-30	90	0	0	
$(\Delta^2 C)_n = (\Delta C)_{n+1} - (\Delta C)_n$	120	180	120	-90	0		
$(\Delta^3 C)_n = (\Delta^2 C)_{n+1} - (\Delta^2 C)_n$	60	-60	-210	90			

Изменение концентрации должно описываться гладкой функцией. Решение может быть функцией с затухающими осцилляциями (если положение равновесия, к которому стремится система есть устойчивый фокус). Но информацию о решении мы имеем неполную – есть только сеточная функция. Ее таблица такова, что выход на постоянное значение может быть истолкован как проекция на сетку разрывной функции. В окрестности разрыва могут возникнуть осцилляции.

В качестве интерполянта выберем кубический сплайн

$$S''_{tt}(t) = m_n(t_{n+1} - t) + m_{n+1}(t - t_n), \quad t \in [t_n, t_{n+1}]$$

со свободными граничными условиями  $S''_{tt}(t_0) = S''_{tt}(t_6) = 0$ .

Тогда уравнения для определения  $m_n$  в случае равномерной сетки запишутся в виде:

$$m_0 = m_6 = 0,$$

$$m_{n-1} + 4m_n + m_{n+1} = \frac{6}{\tau^2} (C_{n-1} - 2C_n + C_{n+1}) = \frac{6}{\tau^2} (\Delta^2 C)_n, \quad n = 1, \dots, 5.$$

Перепишем систему уравнений для моментов сплайна

$$m_0 = m_6 = 0,$$

$$4m_1 + m_2 = 7.2,$$

$$m_1 + 4m_2 + m_3 = 10.8,$$

$$m_2 + 4m_3 + m_4 = 7.2,$$

$$m_3 + 4m_4 + m_5 = -5.4,$$

$$m_4 + 4m_5 = 0.$$

Исключения в методе Гаусса будем вести одновременно сверху и

снизу, получим

$$m_0 = m_6 = 0, \quad m_1 = -\frac{1}{4} m_2 + 1.8,$$

$$m_2 = -\frac{4}{15} m_3 + \frac{4}{15} (10.8 - 1.8), \quad \frac{52}{15} m_3 = 7.2 - \frac{4}{15} (9 - 5.4) = \frac{26}{15} 3.6,$$

$$m_4 = -\frac{4}{15} m_5 + \frac{4}{15} (-5.4), \quad m_5 = -\frac{1}{4} m_4.$$

Находим отсюда все моменты сплайна. Результаты для удобства занесем в таблицу.

$n$	0	1	2	3	4	5	6
$m_n$	0	1.32	1.92	1.8	-1.92	0.48	0
C	700	370	160	130	220	220	220
$(\Delta C)_n = C_n - C_{n-1}$	-330	-210	-30	90	0	0	

Для того чтобы найти минимум  $S(t)$ , необходимо проанализировать знаки  $S'_t(t)$ . Заметим, что вторая производная  $S''_t(t)$  неотрицательная на отрезке  $[t_0, t_3]$ , так как  $m_n \geq 0$  при  $n = 0, 1, 2, 3$ . Следовательно,  $S'_t(t)$  — неубывающая на отрезке  $[t_0, t_3]$ . Вычислим  $S'_t(t_3)$  и  $S'_t(t_2)$ :

$$S'_t(t_3) = \frac{m_3(t_3 - t_2)^2 - m_2(t_3 - t_2)^2}{2\tau} + \frac{C_3 - C_2}{\tau} - \frac{m_3 - m_2}{6}\tau = \frac{(\Delta C)_3}{\tau} + \frac{2m_3 + m_2}{6}\tau$$

$$S'_t(t_2) = \frac{m_3(t_2 - t_1)^2 - m_2(t_2 - t_1)^2}{2\tau} + \frac{C_3 - C_2}{\tau} - \frac{m_3 - m_2}{6}\tau = \frac{(\Delta C)_3}{\tau} - \frac{m_3 + 2m_2}{6}\tau,$$

$$S'_t(t_3) = \frac{-30}{10} + \frac{36 + 19.2}{6} = -3 + 6 + 3.2 = 6.2 > 0,$$

$$S'_t(t_2) = \frac{-30}{10} - \frac{18 + 38.4}{6} = -3 - 3 - 6.4 = -12.4 < 0.$$

Следовательно, на отрезке  $[t_0, t_2]$ , функция  $S'_t(t)$  отрицательная, а на отрезке

$[t_2, t_3]$  меняет знак. Оценим  $S_t'(t)$  на отрезке  $[t_3, t_4]$ .

$$\begin{aligned} S_t'(t) &= \frac{1}{2\tau} (m_4(t-t_3)^2 - m_3(t-t_4)^2) + \frac{C_4 - C_3}{\tau} - \frac{m_4 - m_3}{6}\tau = \\ &= \frac{(\Delta C)_4}{\tau} - \frac{m_4 - m_3}{6}\tau - \frac{m_3(t-t_4)^2 - m_4(t-t_3)^2}{2\tau} = \\ &= 15.2 - \frac{1.8(t-t_4)^2 + 1.92(t-t_3)^2}{20} \geq 15.2 - \frac{1.92 \cdot 100}{20} = 15.2 - 9.6 = 5.6 > 0. \end{aligned}$$

Значит, на отрезке  $[t_3, t_4]$  функция  $S_t'(t)$  положительная. Рассмотрим производную сплайна на отрезке  $[t_4, t_5]$ :

$$S_t'(t) = \frac{1}{2\tau} (m_5(t-t_4)^2 - m_4(t_5-t)^2) + \frac{C_5 - C_4}{\tau} - \frac{m_5 - m_4}{6}\tau.$$

Тогда

$$\begin{aligned} S_t'(t_5) &= \frac{1}{2\tau} (m_5(t_5-t_4)^2 - m_4(t_5-t_5)^2) + \frac{C_5 - C_4}{\tau} - \frac{m_5 - m_4}{6}\tau = \\ &= 0 - \frac{4,8+19,2}{6} + \frac{0,48(t_5-t_4)^2 + 1,92(t_5-t_5)^2}{20} = \\ &= 0 - \frac{4,8+19,2}{6} + \frac{48}{20} = 0 - 4 + 2,4 = -1,6 < 0. \end{aligned}$$

На отрезке  $[t_5, t_6]$   $S_{tt}'(t)$  неотрицательная, значит,  $S_t'(t)$  не убывает, то есть  $S_t'(t) > S_t'(t_5) > 11.4 > 0$ . В итоге получаем, что минимумы интерполирующего сплайна (функции  $S(t)$ ) находятся на отрезках  $[t_2, t_3]$ , и  $[t_5, t_6]$ , что следует из знаков второй производной. Эти экстремумы находятся из условия  $S_t'(t) = 0$ .

$$\begin{aligned} \frac{1}{2\tau} (m_3(t-t_2)^2 - m_2(t-t_3)^2) + \frac{C_3 - C_2}{\tau} - \frac{m_3 - m_2}{6}\tau &= 0, \\ \frac{1}{20} (180z^2 - 192(z-1)^2) + \frac{-30}{10} - \frac{18-19.2}{6} &= 0, \end{aligned}$$

где  $z = \frac{1}{\tau}(t-t_2)$ . Приведя подобные слагаемые, получим  $3z^2 - 96z + 62 = 0$ .

Решив квадратное уравнение, выберем корень из условия  $0 \leq z \leq 1$ :

$$z = 16 - \frac{1}{3} \sqrt{48^2 - 186} = 16 \left( 1 - \sqrt{1 - \frac{186}{48^2}} \right) \approx 16 \frac{186}{2 \cdot 48^2} = \frac{31}{38}.$$

Следовательно,  $t^* = t_2 + z\tau = t_2 + \frac{31}{38}\tau$  — точка минимума функции  $S(t)$ . Вычислим сам минимум

$$\begin{aligned} S(t^*) &= \frac{1}{6\tau} (m_3(t^* - t_2)^3 + m_2(t_3 - t^*)^3) + \left( \frac{C_2}{\tau} - \frac{m_2\tau}{6} \right) (t_3 - t^*) + \\ &\quad + \left( \frac{C_3}{\tau} - \frac{m_3\tau}{6} \right) (t^* - t_2) = \\ &= \frac{\tau^2}{6} (m_3 z^3 + m_2 (1-z)^3) + \left( C_2 - \frac{m_2\tau^2}{6} \right) (1-z) + \left( C_3 - \frac{m_3\tau^2}{6} \right) z \approx \\ &\approx \frac{1}{6} \left( 180 \left( \frac{31}{48} \right)^3 + 192 \left( \frac{17}{48} \right)^3 \right) + \left( 160 - \frac{192}{6} \right) \frac{17}{48} + \left( 130 - \frac{180}{6} \right) \frac{31}{48} = \\ &= \frac{6}{48^3} (5 \cdot 31^3 + 6 \cdot 17^3) + \frac{1}{48} (128 \cdot 17 + 3100) = \frac{178433}{2304 \cdot 8} + \frac{1319}{12} \approx 119.6. \end{aligned}$$

Оценим теперь второй минимум функции на другом отрезке.

$$\begin{aligned} \frac{1}{2\tau} (m_6(t-t_5)^2 - m_5(t-t_6)^2) + \frac{C_6 - C_5}{\tau} - \frac{m_6 - m_5}{6} \tau &= 0, \\ \frac{1}{20} (-48(z-1)^2) + 0 - \frac{-4.8}{6} &= 0, \end{aligned}$$

подходящим корнем уравнения является  $z = 1 - \sqrt{10/3}$ . Вычисляем значение минимума в этой точке:

$$\begin{aligned} S(t^{**}) &= \frac{1}{6\tau} m_5(t_6 - t^*)^3 + \left( \frac{C_5}{\tau} - \frac{m_5\tau}{6} \right) (t_6 - t^*) + \left( \frac{C_6}{\tau} \right) (t^* - t_5) = \\ &= 220 - \frac{4.8}{6} - \frac{0.48}{60} \left( \frac{10\sqrt{10}}{3} \right)^3 > 209. \end{aligned}$$

**Ответ:**  $\approx 119.6$  моль/литр.

**VI.6.3.** Построить минимальную дробно-рациональную аппроксимацию функции  $f(x) = \operatorname{tg} x$  на интервале  $0 \leq x < \pi/2$ .

Решение. Функция имеет полюс первого порядка при  $x = \pi/2$ , причем  $f(x) \approx (\pi/2 - 1)^{-1}$ ; при  $x \rightarrow 0$  функция  $f(x) \approx x$  и раскладывается в ряд по нечетным степеням  $x$ . Аналогично разобранному в п. 5.1 примеру будем аппроксимировать  $f^2(x)$ . Для передачи особенности типа полюса будем искать аппроксимацию в виде

$$f^2(x) = \operatorname{tg}^2(x) \approx \frac{x^2}{\left(1 - 4x^2 / \pi^2\right)^2} \frac{P_p(x^2)}{Q_q(x^2)}; \quad p+q \geq 1.$$

Соотношение степеней этих многочленов может быть произвольным, но должно выполняться

$$\frac{P_p(0)}{Q_q(0)} = 1, \quad \frac{P_p(\pi^2/4)}{Q_q(\pi^2/4)} = \frac{64}{\pi^4}.$$

Даже не привлекая ни одного узла интерполяции, можно взять

$$P_1(x^2) = 1 - \left(4/\pi^2 - 256/\pi^6\right)x^2, \quad Q_0(x^2) = 1.$$

Это уже дает точность не хуже 0,2% на всем отрезке интерполирования.

## VI.7. Задачи на доказательство

**VI.7.1.** Докажите, что при выборе в качестве узлов интерполяции нулей полинома Чебышева  $n + 1$  степени, алгебраический интерполяционный полином можно записать в виде

$$P_n(x, f) = \sum_{k=0}^n a_k T_k(x),$$

$$\text{где } a_0 = \frac{1}{n+1} \sum_{m=0}^n f_m T_0(x_m), \quad a_k = \frac{2}{n+1} \sum_{m=0}^n f_m T_k(x_m).$$

Здесь  $n + 1$  — число узлов интерполяции.

**VI.7.2.** Докажите, что при выборе в качестве узлов интерполяции экстремумов полинома Чебышева алгебраический интерполяционный полином можно записать в виде

$$P_n(x, f) = \sum_{k=0}^n a_k T_k(x),$$

где коэффициенты

$$a_0 = \frac{1}{2n} (f_0 + f_n) + \frac{1}{n} \sum_{m=1}^{n-1} f_m T_0(x_m), \quad a_k = \frac{1}{n} (f_0 + (-1)^k f_n) + \frac{2}{n} \sum_{m=1}^{n-1} f_m T_k(x_m),$$

$$a_n = \frac{1}{2n} (f_0 + (-1)^n f_n) + \frac{1}{n} \sum_{m=1}^{n-1} f_m (-1)^m T_n(x_m).$$

Здесь  $n + 1$  — число узлов интерполяции.

**VI.7.3.** Доказать, что если узлы интерполяции расположены симметрично относительно некоторой точки  $c$ , а значения интерполируемой функции в симметричных узлах равные, то интерполяционный многочлен в форме Лагранжа — четная функция аргумента  $x - c$ .

**VI.7.4.** Разности функции  $f(x)$  с постоянным шагом  $h$  в точке  $x$  определяются для  $k = 0, 1, 2, \dots$  по следующим формулам:

$$\Delta^0 f(x) = f(x),$$

$$\Delta^k f(x) = \Delta^{k-1} f(x+h) - \Delta^{k-1} f(x), \quad k = 1, 2, \dots$$

Показать, что

$$\Delta f(x) = f(x+h) - f(x),$$

$$\Delta^2 f(x) = f(x+2h) - 2f(x+h) + f(x),$$

...

$$\Delta^k f(x) = \sum_{j=0}^k (-1)^{k-j} C_k^j f(x+jh).$$

**VI.7.5.** Пусть  $f(x)$  есть многочлен степени  $n$ . Доказать, что тогда разность (не разделенная!)  $\Delta^k f(x)$  порядка  $k \leq n$  есть многочлен степени  $n - k$ . В частности, при  $n = k$   $\Delta^n f(x) = \text{const}$ .

Указание.  $\Delta^0 f(x) = f(x)$ ,  $\Delta^k f(x) = \Delta^{k-1} f(x+h) - \Delta^{k-1} f(x)$ ,  $k = 1, 2, \dots$

**VI.7.6.** В этой задаче разделенной разностью порядка  $n$  с шагом  $h$  функции  $f(x)$  на равномерной сетке назовем величину  $\frac{1}{h^n} \Delta^n f(x)$ . Показать, что для многочлена  $P_2(x) = a_0 x^n + a_1 x^{n-1} + \dots + a_n$  разделенная разность порядка  $n$  совпадает с производной порядка  $n$ :  $\frac{1}{h^n} \Delta^n P_n(x) \equiv \frac{d^n P_n(x)}{dx^n} = a_0 n!$

**VI.7.7.** Доказать, что погрешность алгебраической интерполяции на сетке  $a = x_0 < x_1 < \dots < x_n = b$  может быть выражена как  
 $f(x) - P_n(x) = f(x_0, x_1, \dots, x_n, x)\omega_{n+1}(x)$ , где  $\omega_{n+1}(x) = (x - x_0)(x - x_1)\dots(x - x_n)$ .

**VI.7.8.** Пусть разделянная разность  $n$ -го порядка  $f(x_0, x_1, \dots, x_n) = 0$  для любых  $a = x_0 < x_1 < \dots < x_n = b$ . Доказать, что тогда  $f(x)$  на отрезке  $[a, b]$  есть алгебраический полином степени не больше  $n - 1$ .

Указание. Воспользуйтесь результатом задачи VI.7.7.

**VI.7.9.** Доказать методом математической индукции, что разделянная разность  $n$ -го порядка  $f(x_0, x_1, \dots, x_n)$  может быть представлена в виде

$$f(x_0, x_1, \dots, x_n) = \sum_{k=0}^n \frac{f(x_k)}{\prod_{j \neq k} (x_k - x_j)}.$$

**VI.7.10.** Функция  $e^x$  приближается на  $[0, 1]$  интерполяционным многочленом степени 3 с чебышевским набором узлов интерполяции:

$$x_k = \frac{1}{2} + \frac{1}{2} \cos \frac{(2k-1)\pi}{8}, \quad k = 1 \div 4.$$

Доказать, что погрешность интерполяции в равномерной норме не превосходит величины  $10^{-3}$ .

**VI.7.11.** При единственном  $n$ -кратном узле интерполяции  $x = (a + b)/2$  интерполяционный многочлен Лагранжа совпадает с отрезком ряда Тейлора:

$$P_{n-1}(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}((a+b)/2)}{k!} (x - (a+b)/2)^k.$$

Доказать, что погрешность интерполяции с единственным кратным узлом в  $2^{(n-1)}$  раз больше погрешности интерполяции с оптимальным (каким?) выбором узлов интерполяции.

**VI.7.12.** Показать, что норма многочлена  $\omega_{n+1}(x) = (x - x_0)(x - x_1)\dots(x - x_n)$  на равномерной сетке с шагом  $h = (x_n - x_0)/n$  может быть оценена как

$$\|\omega_{n+1}(x)\| \approx n! h^{n+1}.$$

**VI.7.13.** Докажите, что если  $f(x)$  есть ограничение на отрезок некоторой аналитической во всей комплексной плоскости функции, принимающей вещественные значения на вещественной оси и имеющей особенности только в бесконечности, то для произвольного числа  $q > 0$  найдется  $A(q) > 0$ , такая, что для любой системы узлов интерполяции выполнена оценка  $\|f(x) - P_n(x_0, \dots, x_n, f)\| \leq Aq^n$ .

Указание. Воспользоваться интегральным представлением Коши для оценки  $f^{(n+1)}(\zeta)$ .

**VI.7.14.** Докажите, что на равномерной сетке на отрезке  $[-1, 1]$  интерполяционный процесс, примененный к функции Рунге  $R(x) = \frac{1}{1 + 25x^2}$ , расходится.

Указание. Воспользуйтесь доказательством, которое получено при решении предыдущей задачи с соответствующими модификациями.

Что можно сказать о сходимости интерполяционного процесса для этой функции на равномерной сетке на  $[0, 2]$ ?

Что можно сказать о сходимости интерполяционного процесса для этой функции на сетке из нулей многочленов Чебышёва?

**VI.7.15.** Докажите, что если система узлов интерполяции  $a = x_0 < x_1 < \dots < x_n = b$  линейной заменой переменных переводится в систему узлов  $t_i = \alpha x_i + \beta$ ,  $i = 0, \dots, n$ , то константа Лебега не изменится (константа Лебега не зависит от длины отрезка интерполяции, а зависит только от взаимного расположения точек на отрезке).

## VI.8. Теоретические задачи

**VI.8.1.** С каким шагом надо составить таблицу значений функции  $y = f(x)$ , чтобы при использовании линейной интерполяции на заданном интервале погрешность не превосходила  $10^{-3}$ :

а)  $f(x) = \sin x$ ,  $-\infty < x < \infty$ , б)  $f(x) = \ln x$ ,  $x \geq 1$ , в)  $f(x) = e^x$ ,  $0 \leq x \leq 1$ ?

**VI.8.2.** С каким шагом надо составить таблицу значений функции  $y = f(x)$ , чтобы погрешность квадратичной интерполяции на заданном интервале не превосходила  $10^{-3}$ :

а)  $f(x) = \sin x$ ,  $-\infty < x < \infty$ , б)  $f(x) = \ln x$ ,  $x \geq 1$  в)  $f(x) = e^x$ ,  $0 \leq x \leq 1$ ?

Сравнить с результатом предыдущей задачи.

**VI.8.3.** С какой точностью можно вычислить  $\sin 5^\circ$  по известным значениям  $\sin 0^\circ$ ,  $\sin 30^\circ$ ,  $\sin 45^\circ$ ,  $\sin 60^\circ$ , используя интерполяцию: а) линейную, б) квадратичную, в) кубическую?

**VI.8.4.** Построить тригонометрический интерполяционный полином второй степени

$$T_2(x) = a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x,$$

удовлетворяющий условиям  $T_2(0) = 0$ ,  $T_2(\pi/4) = 1$ ,  $T_2(\pi/2) = 1$ ,  $T_2(3\pi/4) = 1$ ,  $T_2(\pi) = 1$ .

**VI.8.5.** Пусть  $x_i = a + \frac{b-a}{n-1}(i-1)$ ,  $i = 1 \div n$ . Вычислить  $\|\omega_n(x)\|$  при  $n = 2, 3, 4$ .

Указание:  $\omega_n(x) = \prod_{i=1}^n (x - x_i)$ .

**VI.8.6.** С какой точностью имеет смысл задавать значения таблицы  $f(x) = \sin x$  на отрезке  $[0,1]$ , если шаг таблицы  $h = 0.1$  и предлагается использовать линейную интерполяцию?

**VI.8.7.** С какой точностью можно извлечь кубический корень из 1200, интерполируя функцию  $f(x) = \sqrt[3]{x}$  между узлами  $x_0 = 10^3$ ,  $x_1 = 11^3 = 1331$ ,  $x_2 = 12^3 = 1728$ ?

**VI.8.8.** Данна таблица значений  $y(x)$ . Построить интерполяционный многочлен степени не выше третьей, записав его в форме Лагранжа, в форме Ньютона и в форме  $P_4(x) = a_0x^4 + a_1x^3 + \dots + a_4$ :

а)

$x$	-3.	-2.	-1.	0.
$y$	16.	7.	4.	1.

б)

$x$	-1.	-2.	0.	2.
$y$	7.	19.	3.	-5.

**VI.8.9.** Оценить погрешность приближения функции  $e^x$  интерполяционным многочленом в форме Лагранжа  $L_2(x)$ , построенным по узлам  $x_0 = 0.0$ ,  $x_1 = 0.1$ ,  $x_2 = 0.2$  в точке а)  $x = 0.05$ , б)  $x = 0.15$ .

**VI.8.10.** Функция  $f(x) = 1/(A^2 - x)$  приближается на  $[-4, -1]$  многочленом Лагранжа по узлам  $x_0 = -4$ ,  $x_1 = -3$ ,  $x_2 = -2$ ,  $x_3 = -1$ . При каких значениях  $A$  оценка погрешности в равномерной норме не превосходит  $10^{-5}$ ?

**VI.8.11.** Составляется таблица значений функции  $y = \sin x$  на неравномерной сетке  $x_0 < x_1 < x_2 < \dots < x_n$ ,  $\max_j(x_{j+1} - x_j) = h = \text{const}$ .

а) При каком  $h$  линейная интерполяция позволяет восстановить  $\sin x$  с точностью  $10^{-4}$  между узлами?

б) Ответить на тот же вопрос для квадратичной интерполяции.

в) Какова в обоих случаях допустимая неточность в задании табличных значений, увеличивающая погрешность полученной интерполяции не более чем вдвое, т.е. до  $2 \cdot 10^{-4}$ ?

Ответить на последний вопрос при дополнительном предположении, что шаг сетки  $x_{j+1} - x_j = h$  постоянен, и без этого дополнительного предположения.

**VI.8.12.** Задана табличная функция

$x$	0	$\pi/6$	$\pi/4$	$\pi/3$
$\sin(x)$	0	0.5	0.71	0.87

С какой точностью можно восстановить значение в точке  $x = \pi/5$ , если известно, что функция в узлах задана с абсолютной погрешностью, не превосходящей  $10^{-2}$ ?

**VI.8.13.** Задана табличная функция

$x$	0	$\pi/6$	$\pi/4$	$\pi/3$
$\cos(x)$	1.	0.87	0.71	0.5

а) С какой точностью можно восстановить значение в точке  $x = \pi/5$ , если известно, что функция в узлах задана с абсолютной погрешностью, не превосходящей  $10^{-2}$ ?

б) С какой точностью можно восстановить значение в точке  $x = 7\pi/24$ , если известно, что функция в узлах задана с абсолютной погрешностью, не превосходящей  $10^{-2}$ ?

**VI.8.14.** Задана табличная функция

$x$	0.	$\pi/6$	$\pi/4$	$\pi/3$
$\sin(x)$	0.	0.5	0.71	0.87

С какой точностью можно восстановить значение в точке  $x = 7\pi/24$ , если известно, что функция в узлах задана с абсолютной погрешностью не превосходящей  $10^{-2}$ ?

**VI.8.15.** Пусть в узлах интерполяции  $\{x_m\}$  заданы не только узловые значения функции  $y(x)$ , но и узловые значения ее первой производной. Найти базисные функции  $H$  и  $G$  в формуле кубической эрмитовой интерполяции вида

$$P_3(p, q) = H^R(p, q)y_{m+1} + H^L(p, q)y_m + G^R(p, q)y'_{m+1}h + G^L(p, q)y'_mh,$$

где  $p = (x - x_m)/h$ ,  $q = (x_{m+1} - x)/h$ ,  $p + q = 1$ ,  $h = x_{m+1} - x_m$ .

Найти остаточный член такой интерполяции.

**VI.8.16.** Значения интерполируемой функции  $f(t)$  заданы в узлах  $t_1, t_2, t_3$ .

Постройте дробно-рациональную функцию (интерполянт):

$$F(t) = (a_0 + a_1t)/(d_0 + t), \text{ так чтобы } f(t_i) = F(t_i), i = 1, 2, 3.$$

**VI.8.17** (В.Б. Пирогов). Для функции  $f(x)$  на интервале  $x \in [a, b]$  по ее точным значениям в четырех чебышёвских узлах интерполяции построен интерполяционный полином третьей степени. Известны точные значения этого полинома в трех точках:  $x_0, x_1 = x_0 + c_1 h, x_2 = x_1 + c_2 h$ .

Оценить, с какой точностью можно восстановить с помощью квадратичной интерполяции значения функции  $f(x)$  в любой точке  $x \in [x_0, x_2]$ , если известно, что  $\max_{x \in [a, b]} |f^{(4)}(x)| \leq M_4$ , а  $|f^{(3)}(x_3)| = M_3$ .

- a)  $a = 2, b = 3, x_0 = 2.5, h = 0.1, c_1 = 1, c_2 = 2, M_4 = 3, M_3 = 2, x_3 = x_1$ ;
- б)  $a = 3, b = 4, x_0 = 3.3, h = 0.1, c_1 = 3, c_2 = 2, M_4 = 6, M_3 = 4, x_3 = 3$ ;
- в)  $a = 0.5, b = 1.5, x_0 = 0.6, h = 0.1, c_1 = 3, c_2 = 1, M_4 = 3, M_3 = 1, x_3 = b$ ;
- г)  $a = 1, b = 3, x_0 = 1.5, h = 0.2, c_1 = 2, c_2 = 3, M_4 = 2, M_3 = 3, x_3 = a$ ;
- д)  $a = 1.5, b = 3, x_0 = 2, h = 0.2, c_1 = 2, c_2 = 1, M_4 = 6, M_3 = 2, x_3 = b$ ;
- е)  $a = 2, b = 3.5, x_0 = 2.4, h = 0.2, c_1 = 2, c_2 = 1, M_4 = 3, M_3 = 3, x_3 = a$ .

Указание 1. Для оценки  $\max_{x \in [x_0, x_2]} |(x - x_0)(x - x_1)(x - x_2)|$  можно воспользоваться менее точной, но более простой оценкой:

$$\max_{x \in [x_0, x_2]} |(x - x_0)(x - x_1)(x - x_2)| < \max \begin{cases} \max_{x \in [x_0, x_1]} |(x - x_0)(x - x_1)| \cdot \max_{x \in [x_0, x_1]} |(x - x_2)|, \\ \max_{x \in [x_1, x_2]} |(x - x_1)(x - x_2)| \cdot \max_{x \in [x_1, x_2]} |(x - x_0)|. \end{cases}$$

Указание 2. Имеет смысл изобразить графически модули вспомогательных многочленов Лагранжа и их сумму.

**VI.8.18.** а) Пусть на плоскости заданы координаты вершин треугольника  $L, M, N: \{1, 1\}, \{3, 3\}, \{2, 4\}$  и значения интерполируемой функции  $f_i = f(x_i, y_i), i = 1, 2, 3$  (в этих точках соответственно). Постройте по этим данным линейную функцию — интерполянт  $F(x, y)$  для точек, находящихся внутри треугольника.

б) Как будет выглядеть линейный интерполянт  $F(t)$ , если значения интерполируемой функции  $f(t) = \{f_{n,m}, f_{n+1,m}, f_{n,m+1}, f_{n+1,m+1}\}$  заданы в точках — вершинах прямоугольника  $\{x_n, y_m\}, \{x_{n+1}, y_m\}, \{x_n, y_{m+1}\}, \{x_{n+1}, y_{m+1}\}$ ?

**VI.8.19.** Выбрать вид аппроксимации Паде и найти коэффициенты аппроксимации для функций:

- а)  $f(x) = x \exp(-x)$ ;
- б)  $f(x) = x^2 \exp(-x)$ ;
- в)  $f(x) = x^2 \exp(-x^2)$ .

**VI.8.20.** Вычислить определитель

$$W(x_1, x_2, \dots, x_{2n+1}) =$$

$$= \begin{vmatrix} 1 & \sin x_1 & \cos x_1 & \sin 2x_1 & \cos 2x_1 & \dots & \cos nx_1 \\ 1 & \sin x_2 & \cos x_2 & \sin 2x_2 & \cos 2x_2 & \dots & \cos nx_2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & \sin x_{2n+1} & \cos x_{2n+1} & \sin 2x_{2n+1} & \cos 2x_{2n+1} & \dots & \cos nx_{2n+1} \end{vmatrix}.$$

Указание. Воспользоваться представлением Эйлера  $e^{ix} = \cos x + i \cdot \sin x$ .

**VI.8.21.** Как ведет себя оценка погрешности в задаче экстраполяции при удалении точки  $t$  от интервала  $[t_0, t_N]$ ? Использована равномерная сетка. Рассмотреть случаи  $t \in [t_N, t_N + \tau]$ ,  $t \in [t_N + \tau, t_N + 2\tau]$ ,  $t \in [t_N + 2\tau, t_N + 3\tau]$ .

## VI.9. Практические задачи

**VI.9.1.** Методом обратной интерполяции найти корень нелинейного уравнения, используя приведенные таблицы. Оценить точность полученного решения.

a)	$x^2 + \ln x - 4 = 0$	$x$	1.5	1.6	1.9	2.
		$f(x)$	-1.345	-0.970	0.252	0.693

б)	$(x-1)^2 - \frac{e^x}{2} = 0$	$x$	0.2	0.25	0.27	0.3
		$f(x)$	0.029	-0.080	-0.122	-0.185

в)	$4x - \cos x = 0$	$x$	0.	0.1	0.3	0.5
		$f(x)$	-1.	-0.595	0.245	1.122

г)	$\sqrt{4-x^2} - e^x = 0$	$x$	0.5	0.6	0.7	0.8
		$f(x)$	0.288	0.086	-0.140	-0.393

д)	$x^2 - \lg(x+2) = 0$	$x$	0.5	0.6	0.8	1.
		$f(x)$	-0.148	-0.055	0.193	0.523

е)	$x - \cos x = 0$	$x$	0.5	0.6	0.8	1.
		$f(x)$	-0.378	-0.225	0.103	0.460

ж)	$x^2 - \sin x = 0$	$x$	0.5	0.6	0.8	1.
		$f(x)$	-0.229	-0.205	-0.077	0.159

**VI.9.2.** Для функции, заданной таблично, найти значение первой производной в указанной точке с максимальной возможной точностью 1) с помощью интерполяции, 2) методом неопределенных коэффициентов.

а)

$f'(5) = ?$	$x$	0.	1.	3.	4.	5.
	$f(x)$	0.5	0.3	0.3	0.2	0.1

б)

$f'(0.3) = ?$	$x$	0.	0.1	0.2	0.3	0.4
	$f(x)$	5.	2.5	3.	-2.5	-0.2

в)

$f'(3.) = ?$	$x$	0.	1.	2.	3.	4.
	$f(x)$	4.	2.5	1.	-1.	-2.

г)

$f'(2.) = ?$	$x$	0.	1.	2.	5.	7.
	$f(x)$	1.	0.5	0.3	0.2	0.1

д)

$f'(3.) = ?$	$X$	0.	2.	3.	5.	7.
	$f(x)$	-1.	0.	2.	3.	5.

**VI.9.3.** Найти значение  $x$ , при котором  $y(x) = 1$ . Функция  $y(x)$  задана таблицей

$x$	-3.	0.	1.	5.
$y$	-7.	2.	4.	7.

**VI.9.4.** Найти значение  $x$ , при котором  $y(x) = 0$ . Функция  $y(x)$  задана таблицей

$x$	-5.	-3.	2.	6.
$y$	-22.	-4.	6.	50.

**VI.9.5.** На отрезке  $[2, 5]$  для дробно-линейной функции  $y(x) = \frac{x - 3.5}{x + \sqrt{3} - 3.5}$

построить

а) квадратичный интерполяционный полином в форме Ньютона на сетке из нулей полинома Чебышёва. Оценить погрешность данного метода интерполяции;

б) интерполяционный полином в форме Лагранжа на сетке из экстремумов полинома Чебышёва с тремя нулями на данном отрезке. Оценить погрешность этого метода интерполяции.

**VI.9.6.** В области определения функции  $f(x) = \arcsin(x)$  построить ее интерполянт с максимальной точностью

а) по 4 точкам б) по 5 точкам.

Какую сетку Вам следует использовать? Однаковы ли принципы выбора сеток в пунктах задачи?

**VI.9.7.** Для четной функции, заданной таблично, построить тригонометрическую интерполяцию с максимальной точностью

$x$	0	1/4	1/6	1/2	1	5/4	7/6	3/2
$f$	1	0	-2	-5	1	0	-2	-5

**VI.9.8.** Для нечетной функции с периодом  $T = 1$ , заданной таблично, построить тригонометрическую интерполяцию с максимальной точностью:

$x$	1/8	1/6	1/3
$F$	1	2	2

**VI.9.9.** Для функции, заданной таблично,

$x$	3	4	5
$f$	0,25	-0,2	-0,1

вычислить корень уравнения  $f(x) = 0$  с использованием свободного сплайна. Для решения кубического уравнения использовать метод простых итераций. Доказать его сходимость.

**VI.9.10.** Функция  $f(x)$  задана следующей таблицей. Определить значение аргумента, при котором функция обращается в ноль.

a)

$x$	0.1	0.2	0.3	0.4
$f(x)$	-0.9	-0.6	0.3	0.6

б)

$x$	-1	0	1	2
$f(x)$	0.3	0.1	-0.1	-0.2

Указание. Использовать обратную интерполяцию.

**VI.9.11.** Про функцию известно, что она имеет максимум при  $x = 1$ , и ее значение в этой точке равно 1. В точке  $x = 2$  ее значение равно 0, а первая производная равна 3. Приблизить функцию интерполяционным полиномом на отрезке  $[1, 2]$ .

**VI.9.12.** Про функцию известно, что она имеет минимум при  $x = 1$ , и ее значение в этой точке равно 1. В точке  $x = 0$  ее значение равно 5, а первая производная равна -3. Приблизить функцию интерполяционным полиномом на  $[0, 1]$ .

**VI.9.13.** Для функции  $\sin(x)$ , заданной в девяти узлах  $0; \pm\pi/6, \pm\pi/4, \pm\pi/3, \pm\pi/2$ , построить интерполяционный многочлен, позволяющий вычислить

значение  $\sin(\pi/9)$  с точностью  $\varepsilon = 10^{-2}$ , записав его в виде алгебраического многочлена.

**VI.9.14.** Среди всех многочленов вида  $3x^3 - a_2x^2 + a_1x + a_0$  найти многочлен, наименее уклоняющийся от нуля на  $[0,2]$ .

**VI.9.15.** Зависимость  $y = f(x)$  задана таблицей

$x$	1	2	3	4	5	6	7
$y$	3	7	13	21	31	43	57

Вычислить  $f(x)$ , используя линейную, квадратичную и кубическую интерполяцию при следующих значениях  $x$ :

а)  $x = 2.1$ ; б)  $x = 2.9$ ; в)  $x = 3.1$ ; г)  $x = 3.8$ ; д)  $x = 5.8$ .

**VI.9.16.** Пусть за узлы интерполяции приняты нули многочлена Чебышева  $T_{n+1}(x)$ , т.е. точки  $x_k = \cos((\pi + 2\pi k)/(2n + 2))$ ,  $k = 0, 1, 2, \dots, n$ .

а) По данной таблице функций

$x$	$x_0$	$x_1$	...	$x_{n-1}$	$x_n$
$y$	$y_0$	$y_1$	...	$y_{n-1}$	$y_n$

построить интерполяционный полином  $P_n(x)$  в форме  $P_n(x) = \sum_{k=0}^n c_k T_k(x)$ ,

т.е. выписать формулы для вычисления  $c_k$ .

б) Осуществить вычисления и привести многочлен  $P_n(x) = \sum_{k=0}^n c_k T_k(x)$  к виду  $P_n(x) = a_0x^n + a_1x^{n-1} + \dots + a_n$  в случае  $n = 2$  для функции

$X$	$x_0$	$x_1$	$x_2$
$Y$	2	1	3

Полиномы Чебышёва описаны в Приложении.

**VI.9.17.** Пусть в узлах  $x_k = k / (n + 1)$ ,  $k = 1, 2, \dots, n$ , заданы значения периодической с периодом 1 функции  $y = f(x)$ .

Пусть заданная периодическая с периодом 1 функция  $f(x)$  нечетная. Какие упрощения при этом возникнут?

**VI.9.18.** Функция  $y = f(x)$  задана таблицей

$x$	0.2050	0.2052	0.2060	0.2065	0.2069	0.2075
$y$	0.20792	0.20813	0.20896	0.20949	0.20990	0.21053

Вычислить  $f(0.2062)$ , пользуясь линейной, квадратичной и кубической интерполяциями. В какой форме — Лагранжа или Ньютона — удобнее записывать интерполяционные многочлены?

Составить представление о погрешности, используя остаточный член интерполяции  $R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\dots(x - x_n)$  и приближенное равенство  $\max_x |f^{(k)}(x)| \approx k! \max_{x_j} |f(x_j, x_{j+1}, \dots, x_{j+k+1})|$ , где  $f(x_j, x_{j+1}, \dots, x_{j+k+1})$  — разделенная разность порядка  $k$ .

**VI.9.19.** (Т.К. Старожилова). При исследовании некоторой химической реакции через каждые 5 минут определялось количество вещества, оставшегося в системе. Результаты измерений указаны в таблице

$T$	7	12	17	22	27	32	27
$A$	83.7	72.9	63.2	54.7	47.5	41.4	36.3

Определить количество вещества в системе по истечении 25 минут после начала реакции.

Указание. Составить таблицу разделенных разностей. Из этой таблицы видно, что уже третьи разности теряют регулярный характер. Поэтому воспользуемся квадратичной интерполяцией.

**VI.9.20.** По заданным значениям функции

$x$	1	2	2.5	3
$y$	-6	-1	15.6	16

найти значение  $x$ , при котором  $y = 0$ .

**VI.9.21.** Используя таблицу значений функции  $y = \operatorname{sh} x$ ,

$x$	2.2	2.4	2.6	2.8	3.0
$y$	4.457	5.466	6.695	8.198	10.019

найти значение  $x$ , при котором  $\operatorname{sh} x = 5$ .

**VI.9.22.** Используя значение функции  $y = \lg x$ , указанные в таблице,

$x$	20	25	30
$y$	1.3010	1.3979	1.4771

найти значение  $x$ , при котором  $\lg x = 1.35$ .

**VI.9.23.** Вычислить положительный корень уравнения  $z^7 + 28z^4 - 480 = 0$  посредством обратного интерполирования.

Указание. Вычислить  $y = z^7 + 28z^4 - 480$  при  $z = 1.90, 1.91, 1.92, 1.93, 1.94$ . Убедиться, что корень лежит между 1.92 и 1.93.

**VI.9.24.** В таблице приведены результаты измерения плотности воды ( $D$ ) в

интервале температур между 20° и 25 С. Вычислить плотность воды при температуре 22.7 С.

$t$	20 С	21 С	22 С	23 С	24 С	25 С
$D$	0.998230	0.998019	0.997797	0.997565	0.997323	0.997071

Указание. Составив таблицу конечных разностей, увидим, что третьи разности практически равны нулю. Следовательно, естественно воспользоваться квадратичной интерполяцией.

**VI.9.25.** Функция  $f(x)$  задана таблицей своих значений в узлах интерполяции. а) Построить кубический сплайн для этой функции, предполагая, что сплайн имеет нулевую кривизну при  $x = x_0$  и  $x = x_4$ . Вычислить приближенное значение функции в точке  $x^*$ .

а)  $x^* = 1.5$

0.	1.	2.	3.	4.
0.00000	0.50000	0.86603	1.00000	0.86603

б)  $x^* = 0.8$

0.1	0.5	0.9	1.3	1.7
-2.3026	-0.69315	-0.10536	0.26236	0.53063

в)  $x^*=3.0$

0.	1.7	3.4	5.1	6.8
0.00000	1.3038	1.8439	2.2583	2.6077

г)  $x^*=0.1$

-0.4	-0.1	0.2	0.5	0.8
1.9823	1.6710	1.3694	1.0472	0.64350

д)  $x^*=1.5$

0.	1.	2.0	3.0	4.0
1.00000	1.5403	1.5839	2.0100	3.3464

**VI.9.26.** При исследовании некоторой химической реакции через каждые 10 минут измерялась концентрация образующегося в ходе реакции вещества. Результаты измерений представлены в таблице.

$t$ , мин	10	20	30	40	50	60	70
$C$ , моль/литр	10	340	550	580	490	490	490

С помощью интерполяции найти максимальную концентрацию вещества. Метод интерполяции выберите самостоятельно, обоснуйте выбор метода интерполяции.

**VI.9.27.** При исследовании некоторой химической реакции через каждые две секунды измерялась температура смеси. Результаты измерений представлены в таблице.

$t, \text{с}$	5	7	9	11	13	15	17	19	21
$T, \text{К}$	296	520	744	982	1248	1570	2256	2256	2256

С помощью интерполяции найти  $t^*$ , при котором производная  $dT/dt$  максимальна.

**VI.9.28.** У пациента больницы через каждые полчаса измерялась температура тела. Результаты измерений представлены в таблице.

$t, \text{ ч}$	1	1,5	2,0	2,3	3,0	3,5	4,0	4,5	5,0
$T, \text{ С}$	37,3	37,58	37,86	38,21	38,70	39,26	40,17	40,17	40,17

С помощью интерполяции найти  $t^*$ , при котором производная  $dT/dt$  максимальна.

**VI.9.29.** Согласно переписи население США менялось следующим образом:

1910 – 92 228 496 человек,  
1920 – 106 021 537,  
1930 – 123 202 624,  
1940 – 132 164 569,  
1950 – 151 325 798,  
1960 – 179 323 175,  
1970 – 203 211 926,  
1980 – 226 545 805,  
1990 – 248 709 873,  
2000 – 281 421 906.

- По приведенным данным построить интерполянт в форме Ньютона. Вычислить экстраполированное значение численности населения США в 2010 году и сравнить с точным значением 308 745 538 человек.
- По этим же данным построить сплайн-аппроксимацию, экстраполировать данные на 2010 год, сравнить с точным значением. Какие дополнительные условия для построения сплайна нужно поставить в этом случае?
- Какой из результатов оказывается более точным?

## **VI.10. Библиографический комментарий**

Интерполяция и восполнение функций — обширный раздел вычислительной математики. С разными подходами к теме можно познакомиться по учебникам [2, 3, 8, 27]. Очень глубокое изложение теории алгебраической и тригонометрической интерполяции дано в [35]. Из [35] взята задача 7.12 данного раздела.

Сплайн-интерполяции посвящена обширная литература. Многие приложения сплайнов описаны в [36–38]. В частности, о B-сплайнах см. [37]. О задачах интерполяции, сплайнах и эрмитовой интерполяции см. [38]. О приближении поверхностей с помощью кривой Безье можно прочитать в [22]. Об аппроксимации Паде см. [39].

## VII. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

Задача приближенного вычисления определенного интеграла в зависимости от полноты входящих данных может ставиться по-разному.

Первая постановка связана с вычислением интеграла от таблично заданной функции (значения которой, например, получены в результате эксперимента). В этом случае отсутствует априорная оценка гладкости подынтегральной функции, возможности выбора узлов интегрирования весьма ограничены. В этом случае наиболее эффективными будут интегральные формулы интерполяционного типа (формулы Ньютона–Котеса).

Вторая постановка связана с вычислением определенного интеграла от известной функции. Наиболее дорог с вычислительной точки зрения в этом случае подсчет значений функции. В этом случае желательно построить метод, обеспечивающий как можно более высокую точность при минимальных вычислениях. Так как свобода выбора узлов квадратурной формулы в этом случае в руках вычислителя, то наиболее эффективными будут квадратурные формулы Гаусса.

### VII.1. Квадратурные формулы Ньютона–Котеса (интерполяционного типа)

Пусть необходимо вычислить определенный интеграл

$$I = \int_a^b f(x)dx. \quad (1.1)$$

Пусть имеется некоторое разбиение отрезка  $[a,b]$ :

$$a = x_0 < x_1 < \dots < x_n = b. \quad (1.2)$$

Тогда в качестве приближения значения интеграла можно использовать интегральную сумму

$$I = \sum_{k=0}^{n-1} (x_{k+1} - x_k) f(\xi_k), \quad x_k \leq \xi_k \leq x_{k+1}. \quad (1.3)$$

В этом случае точки  $\xi_k$  называются узлами квадратурной формулы, а величины  $(x_{k+1} - x_k)$  весами квадратурной формулы.

Формула (1.3) точна в том случае, если  $f(x)$  постоянна на каждом отрезке разбиения (т.е. на многочленах степени ноль). Оказывается, если в качестве узлов квадратурной формулы взять центральную точку интервала

$$\xi_k = 0.5(x_{k+1} + x_k),$$

то квадратурная формула (1.3) будет точна и для линейной функции на каждом отрезке разбиения. При любом выборе узлов формула (1.3) носит название *формулы прямоугольников*.

К вычислению интеграла (1.1) можно подойти как к вычислению площади под кривой, и тогда в предположении кусочно-линейной интерполяции подынтегральной функции на каждом отрезке разбиения получается *формула трапеций* для приближенного вычисления интеграла (1.1):

$$I \approx \sum_{k=0}^{n-1} \frac{h_k}{2} (f(x_{k+1}) + f(x_k)), \quad h_k = x_{k+1} - x_k. \quad (1.4)$$

*Формула Симпсона* получается с помощью построения квадратичной интерполяции подынтегральной функции по трем равноотстоящим узлам  $x_i, x_{i+1/2}, x_{i+1}$  на каждом отрезке разбиения

$$f(x) \approx P_2(x) = f(x_k) + f(x_k, x_{k+1})(x - x_k) + f(x_k, x_{k+1}, x_{k+1/2})(x - x_k)(x - x_{k+1}), \quad (1.5)$$

что приводит к квадратурной формуле

$$I \approx \sum_{k=0}^{n-1} \frac{h_k}{6} (f(x_{k+1}) + 4f(x_{k+1/2}) + f(x_k)), \quad h_k = x_{k+1} - x_k. \quad (1.6)$$

При разбиении полного отрезка интегрирования на сдвоенные интервалы равной длины можно получить формулу Симпсона в несколько другой форме:

$$I \approx \sum_{k=0}^{[n/2]} \frac{h_{2k}}{3} (f(x_{2k}) + 4f(x_{2k+1}) + f(x_{2k+2})). \quad (1.7)$$

Аналогично, интерполируя подынтегральную функцию полиномом третьей степени по четырем равноотстоящим точкам, получим выражение интеграла (1.1) в соответствии с «правилом 3/8»:

$$I \approx \sum_{i=0}^{n-1} \frac{h_k}{8} (f(x_k) + 3f(x_{k+1/3}) + 3f(x_{k+2/3}) + f(x_{k+1})). \quad (1.8)$$

Квадратурные формулы интерполяционного типа более высокого порядка рассматриваются редко по двум важным взаимосвязанным причинам. Заметим, что все приведенные выше формулы являются *правильными*,

т.к. все веса квадратурных формул были положительными. При использовании интерполянтов более высоких степеней получающиеся квадратурные формулы перестают быть правильными: для полиномов степени выше седьмой среди весов квадратурных формул появляются отрицательные величины. Более того, Д. Пойа показал, что

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n |\alpha_{nk}| = \infty, \quad (1.9)$$

в то время как для самой суммы выполняется очевидное равенство

$$\sum_{k=0}^n \alpha_{nk} = b - a. \quad (1.10)$$

Здесь  $\alpha_{nk}$  — веса квадратурных формул, получающихся при замене  $f(x)$  интерполянтом степени  $n$ . Такое увеличение суммы абсолютных значений весов квадратурных формул связано с быстрым ростом постоянной Лебега при алгебраической интерполяции на равномерной сетке.

### VII.1.1. Оценка погрешности квадратурных формул

Оценка погрешности квадратурных формул Ньютона–Котеса производится через оценку остаточного члена погрешности интерполяции подынтегральной функции.

**1.** Формула прямоугольников при произвольном выборе узла интерполяции на отрезке разбиения дает погрешность для отдельного интервала разбиения:

$$\begin{aligned} \varepsilon_k &\leq \max_{x_k \leq x \leq x_{k+1}} |f'(x)| \int_{x_k}^{x_{k+1}} |x - \xi_k| dx \leq \frac{1}{2} \max_{x_k \leq x \leq x_{k+1}} |f'(x)|(x_{k+1} - x_k)^2 = \\ &= \frac{1}{2} \max_{x_k \leq x \leq x_{k+1}} |f'(x)| h_k^2, \end{aligned} \quad (1.11)$$

а при суммировании по всем отрезкам равной длины с учетом равенства  $\sum_k h_k = b - a$  будем иметь формулу первого порядка аппроксимации

$$\varepsilon \leq \frac{1}{2} \max_{a \leq x \leq b} |f'(x)|(b - a)h. \quad (1.12)$$

Для формулы прямоугольников с центральной точкой для получения более точной оценки погрешности интегрирования интерполяцию по значению в средней точке нужно рассматривать как интерполяцию с кратным

узлом, в котором известно не только значение функции, но и производная. В этом случае интерполяционный многочлен совпадает с отрезком ряда Тейлора из двух членов, а погрешность интегрирования на отдельном интервале выражается через остаточный член:

$$\varepsilon_k \leq \frac{1}{2} \max_{x_k \leq x \leq x_{k+1}} |f''(x)| \int_{x_k}^{x_{k+1}} (x - x_{k+1/2})^2 dx = \frac{1}{24} \max_{x_k \leq x \leq x_{k+1}} |f''(x)| h_k^3, \quad (1.13)$$

и для полной погрешности

$$\varepsilon \leq \frac{1}{24} \max_{a \leq x \leq b} |f''(x)|(b-a)h^2. \quad (1.14)$$

**2.** Оценка погрешности формула трапеций получается интегрированием остаточного члена погрешности линейной интерполяции для отдельного интервала:

$$\varepsilon_k \leq \frac{1}{2} \max_{x_k \leq x \leq x_{k+1}} |f''(x)| \int_{x_k}^{x_{k+1}} |(x - x_k)(x - x_{k+1})| dx = \frac{1}{12} \max_{x_k \leq x \leq x_{k+1}} |f''(x)| h_k^3, \quad (1.15)$$

Суммарная погрешность

$$\varepsilon \leq \frac{1}{12} \max_{a \leq x \leq b} |f''(x)|(b-a)h^2. \quad (1.16)$$

Как видим, формула прямоугольников со средней точкой и формула трапеций являются формулами второго порядка аппроксимации, однако константа погрешности у формулы прямоугольников в два раза меньше.

**3.** Оценка погрешности формулы Симпсона, как и в случае формулы прямоугольников со средней точкой, может быть повышена на единицу, если рассматривать среднюю точку интервала как кратный узел интерполяции с известной производной, тогда интерполяционный многочлен Эрмита представляется в виде суммы обычного многочлена второго порядка (1.5), построенного по трем точкам, и добавочного члена, обусловленного кратностью узла интерполяции:

$$P_3(x) = P_2(x) + f(x_k, x_{k+1}, x_{k+1/2}, x_{k+1/2})(x - x_k)(x - x_{k+1})(x - x_{k+1/2}), \quad (1.17)$$

где под разделенной разностью с кратными узлом понимается предел

$$f(x_k, x_{k+1}, x_{k+1/2}, x_{k+1/2}) = \lim_{\varepsilon \rightarrow 0} f(x_k, x_{k+1}, x_{k+1/2} - \varepsilon, x_{k+1/2} + \varepsilon).$$

Для дифференцируемой функции этот предел существует и конечен.

Добавочный член к  $P_2(x)$  является нечетной функцией относительно середины интервала интерполяции, поэтому интеграл от него по отрезку  $[x_k, x_{k+1}]$  обращается в нуль, а остаточный член интерполяции позволяет найти ошибку интегрирования формулы Симпсона на элементарном отрезке

$$\varepsilon_k \leq \frac{1}{24} \max_{x_k \leq x \leq x_{k+1}} \left| f^{IV}(x) \int_{x_k}^{x_{k+1}} (x - x_k)(x - x_{k+1/2})^2 (x - x_{k+1}) dx \right| = \quad (1.18)$$

$$= \frac{1}{2880} \max_{x_k \leq x \leq x_{k+1}} \left| f^{IV}(x) h_k^5 \right|$$

и на всем интервале интегрирования:

$$\varepsilon \leq \frac{b-a}{2880} \max_{a \leq x \leq b} \left| f^{IV}(x) h^4 \right|. \quad (1.19)$$

Аналогично, для формулы (1.7) со сдвоенными интервалами интерполяции будем иметь оценку погрешности:

$$\varepsilon \leq \frac{b-a}{180} \max_{a \leq x \leq b} \left| f^{IV}(x) h^4 \right|. \quad (1.20)$$

**4.** Хотя формула «правило 3/8» строится по большему числу узлов по сравнению с формулой Симпсона, но, как и для формулы трапеций, для нее невозможно улучшить оценку. Это формула также четвертого порядка аппроксимации:

$$\varepsilon \leq \frac{b-a}{6480} \max_{a \leq x \leq b} \left| f^{IV}(x) h^4 \right|. \quad (1.21)$$

### VII.1.2. Связь между формулами прямоугольников, трапеций и Симпсона

Рассмотрим приближенные вычисления интеграла на отрезке  $[x_k - h, x_k + h]$  по формулам прямоугольников, трапеций и Симпсона:

$$I_k^H = y_k \cdot 2h, \quad (1.22)$$

$$I_k^T(2h) = \frac{1}{2} (y_{k-1} + y_{k+1}) \cdot 2h = (y_{k-1} + y_{k+1}) \cdot h, \quad (1.23)$$

$$I_k^S(h) = \frac{h}{2} (y_{k-1} + 2y_k + y_{k+1}), \quad (1.24)$$

$$I_k^C(h) = \frac{h}{3}(y_{k-1} + 4y_k + y_{k+1}). \quad (1.25)$$

Полусумма (1.22) и (1.23) дает (1.24), т.е.

$$I_k^T(h) = \frac{1}{2}(I_k^{\Pi}(2h) + I_k^T(2h)). \quad (1.26)$$

Аналогично, две третьих (1.22) и одна треть (1.23) в сумме дадут

$$I_k^C(h) = \frac{1}{3}(2I_k^{\Pi}(2h) + I_k^T(2h)). \quad (1.27)$$

Можно комбинировать формулы из одного класса с разными шагами для увеличения точности, например:

$$I_k^C(h) = I_k^T(h) + \frac{1}{3}(I_k^T(h) - I_k^T(2h)). \quad (1.28)$$

Последнее слагаемое в (1.28) называется поправкой Ричардсона.

## VII.2. Экстраполяция Ричардсона. Правило Рунге практического оценивания погрешности. Алгоритм Ромберга

Пусть для приближенного вычисления значения  $I$  интеграла применяется некая квадратурная формула  $p$ -го порядка аппроксимации  $I^p$  из семейства составных формул Ньютона–Котеса. При условии, что подынтегральная функция является  $p$  раз непрерывно дифференцируемой, тогда существует константа  $c$  такая, что выполняется условие

$$I = I^p(h) + c \cdot h^p. \quad (2.1)$$

При уменьшении шага вдвое будем иметь

$$I = I^p(h/2) + c_1 \cdot (h/2)^p. \quad (2.2)$$

При малых  $h$  постоянные  $c$  и  $c_1$  близки (их отличие есть величина  $O(h)$ ), тогда имеем

$$I^p(h) + c \cdot h^p = I^p(h/2) + c_1 \cdot (h/2)^p \approx I^p(h/2) + c \cdot (h/2)^p.$$

Отсюда можем получить оценку константы в (2.1), (2.2):

$$c \approx c_1 \approx \left( I^p(h/2) - I^p(h) \right) / \left( h^p - (h/2)^p \right). \quad (2.3)$$

Тогда можно получить уточненное значение интеграла:

$$I \approx I^p(h/2) + (I^p(h/2) - I^p(h)) / (2^p - 1).$$

К полученному равенству можно относиться двояко:

1) для контроля точности (*принцип Рунге* практического оценивания погрешности):

$$|I - I^p(h/2)| \approx |(I^p(h/2) - I^p(h)) / (2^p - 1)|. \quad (2.4)$$

Его применение считается правомочным, если выполнено неравенство

$$|(2^p(I^p(h/2) - I^p(h)) / (I^p(h) - I^p(2h))) - 1| < 0.1.$$

2) для повышения порядка точности вычисления интеграла (*экстраполяция Ричардсона*)

$$I^{p+1} = I^p(h/2) + (I^p(h/2) - I^p(h)) / (2^p - 1) + O(h^{p+1}). \quad (2.5)$$

Применение экстраполяции Ричардсона к формуле трапеций приводит к формуле Симпсона (ср. (1.28)), это означает, что происходит увеличение порядка аппроксимации не на единицу, как следовало ожидать, а на две. Это свойство является характерной чертой формул Ньютона–Котеса четных порядков аппроксимации. Применяя данную процедуру к формулам Симпсона, можно получить вычисление интеграла с шестым порядком аппроксимации:

$$I = I^4(h/2) + (I^4(h/2) - I^4(h)) / 15 + O(h^6). \quad (2.6)$$

Повторяя эту процедуру дальше, получим *алгоритм Ромберга* уточнения вычисления интеграла. Этот алгоритм может рассматриваться как экстраполяция вычисленных значений интегралов на последовательности вдвое сгущающихся сеток в нулевой шаг интегрирования. Важно, что экстраполяция должна производиться не по величине шага  $h$ , а по величине  $h^2$ .

### VII.3. Квадратурные формулы Гаусса

Если свобода выбора узлов интегральной формулы в руках вычислителя, то можно ставить задачу следующим образом: выбрать узлы и веса квадратурной формулы для обеспечения алгебраической точности данной квадратуры как можно более высоко порядка. Иными словами, мы будем подбирать узлы квадратурной формулы  $x_k$  и веса квадратурной формулы  $c_k$ ,

чтобы квадратурная формула с  $2n$  параметрами

$$\int_{-1}^1 f(x)dx = \sum_{i=1}^n c_i f(x_i) \quad (3.1)$$

была точна для многочленов как можно более высокой степени. Для формул Гаусса нам удобнее начинать нумерацию узлов не с нуля, а с единицы. Для выбранного интервала интегрирования эта задача решается наиболее просто, в других случаях нужно делать линейную замену переменных.

**Теорема.** *Если в качестве узлов квадратурной формулы (3.1) взять нули полиномов Лежандра  $q_n(x)$ :  $q_n(x_i) = 0$ ,  $i = 1, \dots, n$ , а в качестве весов интегралы от базисных функций многочленов Лагранжа  $l_i(x)$  степени  $n - 1$ , а именно*

$$c_i = \int_{-1}^1 \frac{(x-x_1)(x-x_2)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_1)(x_i-x_2)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)} dx, \quad (3.2)$$

*то квадратурная формула (3.1) будет точна при подстановке в нее вместо  $f(x)$  любого многочлена степени не выше  $2n - 1$ .*

Приведем узлы и веса квадратур для нескольких первых квадратур Гаусса.

$$n = 2, \quad x_1 = \sqrt{3}/3, \quad x_2 = -\sqrt{3}/3, \quad c_1 = c_2 = 1;$$

$$n = 3, \quad x_1 = -x_3 = \sqrt{3}/5, \quad x_2 = 0, \quad c_1 = c_3 = 5/9, \quad c_2 = 8/9;$$

$$n = 4, \quad x_1 = -x_4 = 0.861136, \quad x_2 = -x_3 = 0.339981,$$

$$c_1 = c_4 = 0.347855, \quad c_2 = c_3 = 0.652145.$$

Для вычисления интеграла по произвольному отрезку  $[a, b]$  система ортогональных многочленов на заданном отрезке получается из многочленов Лежандра после линейной замены:

$$t \in [-1, 1] \rightarrow x = \frac{b+a}{2} + \frac{b-a}{2}t, \quad x \in [a, b]. \quad (3.3)$$

Для оценки остаточного члена интегрирования по формулам Гаусса можно воспользоваться оценкой

$$r_n^{\Gamma} = \frac{(b-a)^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3} f^{(2n)}(\xi), \quad \xi \in [a, b]. \quad (3.4)$$

Приведем вычисленные значения коэффициента перед производной в оценке остаточного члена для нескольких первых квадратурных формул

Гаусса при интегрировании по отрезку  $[-1, 1]$ :

$$\begin{aligned} r_2^{\Gamma} &= \frac{1}{135} f^{(4)}(\xi), & r_3^{\Gamma} &= \frac{1}{15750} f^{(6)}(\xi), \\ r_4^{\Gamma} &= \frac{2}{3472875} f^{(8)}(\xi), & r_5^{\Gamma} &= \frac{13}{1237732650} f^{(10)}(\xi), \dots \end{aligned} \quad (3.5)$$

Видим быстрое затухание ошибки при достаточной гладкости интегрируемой функции.

### VII.3.1. Квадратурные формулы Гаусса–Кристоффеля

В некоторых приложениях возникает необходимость вычисления интегралов с заданной весовой функцией  $p(x)$ :  $\int_a^b f(x)p(x)dx$ . При некотором

наборе весовых функций могут оказаться полезными квадратурные формулы Гаусса–Кристоффеля вида

$$\int_a^b f(x)p(x)dx = \sum_{i=1}^n c_i f(x_i). \quad (3.6)$$

**Теорема.** *Квадратурная формула (3.6) точна для многочленов степени не выше  $2n - 1$ , если ее узлами  $x_i$  являются корни многочлена  $Q_n(x)$  из семейства многочленов, ортогональных на промежутке интегрирования  $(a, b)$  с весом  $p(x)$ , а весовыми коэффициентами являются числа*

$$c_i = \int_{-1}^1 \frac{(x - x_1)(x - x_2)\dots(x - x_{i-1})(x - x_{i+1})\dots(x - x_n)}{(x_i - x_1)(x_i - x_2)\dots(x_i - x_{i-1})(x_i - x_{i+1})\dots(x_i - x_n)} p(x)dx. \quad (3.7)$$

## VII.4. Приемы вычисления несобственных интегралов

Рассмотрим сходящиеся интегралы следующих двух типов:

$$\int_a^b f(x)dx, \text{ причем } f(x) \rightarrow \infty \text{ при } x \rightarrow a \text{ (первый тип);} \quad (4.1)$$

$$\int_a^{\infty} f(x)dx, \text{ (второй тип).} \quad (4.2)$$

**Замечание.** Второй интеграл, вообще говоря, может быть сведен к первому заменой переменной интегрирования:  $t = 1/x$ . Поэтому пока будем говорить об интегралах первого типа.

Очевидно, непосредственное использование квадратурных формул трапеций и Симпсона для вычисления таких интегралов невозможно (так как точка  $x = a$  является для этих формул узлом интегрирования). По методу прямоугольников вычисления формально провести можно, но результат будет сомнительным, так как оценка погрешности теряет смысл (производные подынтегральной функции не ограничены).

Продемонстрируем приемы, которые позволяют получать в подобных случаях надежные результаты, на примере интеграла

$$I = \int_0^1 \frac{\cos x}{\sqrt{x}} dx.$$

а) Иногда подходящая замена *переменной* интегрирования позволяет вообще избавиться от особенности.

В рассматриваемом примере после замены  $x = t^2$  получаем

$$I = 2 \int_0^1 \cos t^2 dt,$$

и интеграл вычисляется с требуемой точностью по любой из квадратурных формул.

б) Та же цель (избавление от особенности) достигается иногда предварительным *интегрированием по частям*:

$$I = \int_0^1 \frac{\cos x}{\sqrt{x}} dx = 2\sqrt{x} \cos x \Big|_0^1 + 2 \int_0^1 \sqrt{x} \sin x dx.$$

Последний интеграл формально может быть вычислен стандартным образом, но оценка погрешности для любой квадратурной формулы будет иметь лишь первый порядок, так как при  $x = 0$  не существует вторая производная от подынтегральной функции. Проводя еще раз интегрирование по частям, придем к интегралу от дважды непрерывно дифференцируемой функции, который с гарантированной точностью может быть вычислен по формулам трапеций или прямоугольников.

в) Если упомянутыми простыми средствами избавиться от особенности не удается, то прибегают к универсальному *методу выделения особенности*. В рассматриваемом случае представим интеграл в виде суммы двух интегралов:

$$I = I_1 + I_2, \quad I_1 = \int_0^\delta \frac{\cos x}{\sqrt{x}} dx, \quad I_2 = \int_\delta^q \frac{\cos x}{\sqrt{x}} dx.$$

Второй интеграл особенности не содержит и вычисляется по любой квадратурной формуле. Вопрос о выборе величины  $\delta$  обсуждается ниже.

Первый интеграл с требуемой точностью вычисляем аналитически, используя представление подынтегральной функции в окрестности особой точки ( $x = 0$ ) в виде отрезка ряда по степеням  $x$ , который получим после замены  $\cos x$  соответствующим рядом Тейлора:

$$I_1 = \int_0^\delta \frac{1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \dots + (-1)^m \frac{x^{2m}}{(2m)!}}{\sqrt{x}} dx = 2\sqrt{\delta} - \frac{1}{2!} \frac{2}{5} \delta^{5/2} + \\ + \frac{1}{4!} \frac{2}{9} \delta^{9/2} + \dots + (-1)^m \frac{1}{(2m)!} \frac{1}{2m+1/2} \delta^{2m+1/2}.$$

Важно, что подобное аналитическое представление в малой окрестности особой точки можно получить практически во всех конкретных случаях. Как это сделать — зависит от квалификации вычислителя.

Допустим, что мы решили ограничиться в полученном представлении первыми  $m$  слагаемыми. При этом для данного примера мы допускаем погрешность, которая не превосходит последнего приведенного в записи для  $I_1$  слагаемого в силу того, что ряд для  $\delta$  — знакопеременный. Следовательно, для выбора двух параметров ( $\delta$  и  $m$ ) имеем следующий критерий:

$$\frac{1}{(2m)!} \frac{1}{(2m+\frac{1}{2})} \delta^{2m+\frac{1}{2}} \leq \frac{\varepsilon}{2}, \quad (4.3)$$

а  $\varepsilon/2$  выступает в качестве допустимого уровня погрешности при вычислении  $I_2$ . Таким образом, один из параметров ( $m$  или  $\delta$ ) можно задавать по своему усмотрению, второй — определяется из неравенства (4.3). При этом нужно принять в расчет следующее соображение.

Если  $\delta \ll 1$ , то существенно ухудшается оценка погрешности для любой квадратурной формулы, которую мы предполагаем использовать для вычисления  $I_2$ , так как в качестве коэффициента при  $h^p$  (где  $p$  — порядок точности выбранной формулы) фигурирует максимальное на отрезке  $[\delta, 1]$  значение  $p$ -й производной от подынтегральной функции, которое при  $x = \delta$  в рассматриваемом случае имеет порядок  $\delta^{-(p+0.5)}$ .

Кроме того, при вычислении интеграла  $I_2$  придется вычислять подынтегральную функцию  $f(x)$  от аргумента либо равного  $\delta$  (для формул трапеций и Симпсона), либо очень близкого к  $\delta$  (для формулы прямоугольников с центральной точкой). Но значение  $f(\delta)$  при малом  $\delta$  может быть настолько боль-

шим (в рассматриваемом случае  $f(\delta) \sim 1/\sqrt{|\delta|}$ ), что абсолютная погрешность функции  $f(\delta)$  не позволит вычислить интеграл с требуемой точностью при заданной длине мантиссы и выбранном шаге интегрирования.

Следовательно, целесообразно задать «не слишком малое»  $\delta$  (например,  $\delta = 0.1$ ), а затем  $t$  найти из условия (4.3).

**Замечание.** Разумеется, если поиск последовательных членов разложения подынтегральной функции затруднителен, то приходится ограничиваться доступными членами. В этом случае из условия типа (4.3) находится параметр  $\delta$ .

Аккуратное вычисление интеграла с особенностью может быть выполнено гораздо более экономичными средствами. Это достигается с помощью приема *регуляризации* (метод Канторовича), или *выделения особенности*. Поясним его в более общей ситуации. Пусть требуется вычислить интеграл

$$\int_0^1 \frac{f(t)}{\sqrt{t}} dt, \text{ где } f(t) — \text{гладкая функция. Регуляризация заключается в том,}$$

что проделывается тождественное преобразование:

$$\int_0^1 \frac{f(t)}{\sqrt{t}} dt = \int_0^1 [f(t) - \varphi(t)] t^{-1/2} dt + \int_0^1 \varphi(t) t^{-1/2} dt.$$

Функция  $\varphi(t)$  выбирается такой, чтобы первый интеграл правой части не содержал особенности и при небольшом объеме вычислений достаточно точно определялся хотя бы по формуле Симпсона. Второй интеграл особенность содержит, но вычисляется аналитически. В данном случае цель будет достигнута, если в качестве  $\varphi(t)$  взять отрезок ряда Тейлора  $f(t)$  в точке  $t = 0$ . Это приводит к вычислению

$$\int_0^1 \frac{f(t) - f(0) - tf'(0)}{\sqrt{t}} dt + f(0) \int_0^1 \frac{1}{\sqrt{t}} dt + f'(0) \int_0^1 \sqrt{t} dt.$$

В примере, с которого мы начинали  $f(t) = \cos t$ , приходим к вычислению

$$\int_0^1 (\cos t - 1) t^{-1/2} dt + \int_0^1 t^{-1/2} dt.$$

Второе слагаемое есть 2, первое вычислим по формуле Симпсона: сначала с шагом 0.5, что даст значение 1.807967, а затем с шагом 0.25, что даст значение 1.808850. Эти вычисления «стоили» всего четырех вычислений

подынтегральной функции. Поучительно сравнить их с тем, сколько вычислений этой функции потребуется при «студенческом» рецепте (для достижения такой же точности).

Рассмотрим пример вычисления интеграла второго типа  $\int_0^\infty e^{-x^2} dx$ .

Можно, как уже отмечалось, свести его к интегралу первого типа. Но мы воспользуемся универсальным приемом выделения особенности. Особенность состоит в том, что верхний предел интегрирования — бесконечность. Представим интеграл в виде суммы двух интегралов:  $I = I_1 + I_2$  где  $I_1$  — интеграл по конечному отрезку  $[a, A]$ ;  $I_2$  — интеграл по  $[A, \infty]$ . Вычисление  $I_1$  при заданном  $A$  затруднений не вызывает.

Выберем теперь  $A$  так, чтобы в пределах допустимой погрешности вторым интегралом можно было пренебречь, т. е. так, чтобы  $|I_2| \leq \varepsilon / 2$ . Например, учитывая, что при  $A \geq 1$

$$\int_A^\infty e^{-x^2} dx \leq \int_A^\infty x e^{-x^2} dx = \frac{1}{2} e^{-A^2},$$

и требуя, чтобы выполнялось условие

$$\frac{1}{2} e^{-A^2} \leq \frac{1}{2} \varepsilon,$$

найдем  $A \geq \sqrt{|\ln \varepsilon|}$ .

## VII.5. Вычисление интегралов от быстроосциллирующих функций

Начнем с простого примера. Пусть требуется вычислить  $\int_0^1 e^{-t} \sin kt dt$

при большом значении  $k$ , например,  $k = 100$ .

Интегралы вида  $\int_0^1 f(t) \sin kt dt$ , где  $f(t)$  — гладкая функция, часто при-

ходится вычислять в некоторых разделах физики. Сложность задачи состоит в том, что подынтегральная функция совершает большое число колебаний. Вычисление интеграла по стандартной формуле Симпсона, конечно, возможно, но требует сетки с очень малым шагом: каждая волна должна быть описана некоторым числом узлов сетки, а волн много.

Дело осложняется еще и тем, что вычисление должно проводиться с высокой точностью. Так как результат есть сумма большого числа близких величин с противоположными знаками (интегралов от отдельных волн подынтегральной функции), происходит сильное сокращение знаков. Для обеспечения точности остатка (результата) отдельные слагаемые должны вычисляться с существенно более высокой точностью. Для вычисления подобных интегралов используется следующий прием. Гладкая функция  $f(t)$  аппроксимируется некоторой другой гладкой функцией  $f^*(t)$ , такой, чтобы интеграл от  $f^*(t) \sin kt$  вычислялся аналитически.

Итак, все сводится к тождественному преобразованию:

$$\int_0^\pi f(t) \sin kt dt = \int_0^\pi f^*(t) \sin kt dt + \int_0^\pi [f(t) - f^*(t)] \sin kt dt.$$

Второе слагаемое является малым и отбрасывается. Правда, если оценить отбрасываемую величину, опираясь только на оценку типа  $|f(t) - f^*(t)| \leq \varepsilon$ , т.е. в данном случае величиной  $\pi\varepsilon$ , ничего хорошего (даже если  $\varepsilon$  – точная оценка погрешности аппроксимации) не получится, так как величина  $\pi\varepsilon$  может оказаться значительно большей интересующего нас интеграла. На самом деле погрешность существенно меньше. Это ведь интеграл от гладкой функции, не превосходящей  $\varepsilon$ , умноженной на быстроосциллирующую функцию. Естественно ожидать, что погрешность будет во столько раз меньше результата, во сколько раз  $|f - f^*|$  меньше  $f$ . При  $f(t) = e^{-t}$  интеграл вычисляется аналитически.

## VII.6. Задачи на доказательство

**VII.6.1.** Доказать, что вычисление интеграла от строго выпуклой функции  $f'' > 0$  по формулам прямоугольников со средней точкой дает заниженное значение интеграла.

**VII.6.2.** Доказать, что вычисление интеграла от строго выпуклой функции  $f'' > 0$  методом трапеций дает завышенное значение интеграла.

**VII.6.3.** Доказать, что для погрешности квадратурной формулы трапеций справедливо представление

$$\int_a^b f(x) dx - \frac{b-a}{2} (f(a) + f(b)) = \frac{1}{2} \int_a^b (x-a)(x-b) f''(x) dx.$$

**VII.6.4.** Доказать, что при применении правила Рунге к формуле трапеций получается формула Симпсона. Насколько при этом повышается порядок точности метода?

**VII.6.5.** Пусть для вычисления интеграла  $\int\limits_{-a}^a f(x)p(x)dx$  с четной весовой функцией  $p(x)$  используется симметричный относительно нуля набор узлов  $x_{n+1-k} = -x_k$ ,  $k = 1, \dots, n$ . Доказать, что веса, соответствующие симметричным узлам, будут равны:  $c_{n+1-k} = c_k$ ,  $k = 1, \dots, n$ .

**VII.6.6.** Показать, что процедура уточнения значения интеграла по формуле (2.5)

$$I^{p+1} = I^p(h/2) + (I^p(h/2) - I_p(h)) / (2^p - 1)$$

является экстраполяцией, т.е. при  $I_p(h/2) \neq I_p(h)$  величина  $I_{p+1}$  всегда лежит вне отрезка с концами  $I_p(h/2)$  и  $I_p(h)$ .

**VII.6.7.** Пусть  $C_q = \int\limits_a^b |f^{(q)}(x)| dx < \infty$ ,  $q \leq 4$ . Получить оценку погрешности формулы Симпсона  $|R_4(f)| \leq k C_q h^4$ , где  $k$  — абсолютная постоянная,  $h$  — шаг интегрированной.

## VII.7. Задачи с решениями

**VII.7.1.** (Пример применения алгоритма Рунге–Ромберга.) Вычислить интеграл  $\int\limits_0^{1.1} \frac{\sqrt{1-x^2}}{1.1-x} dx$ , используя таблицу значений подынтегральной функции:

$x$	0.	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1.
$f(x)$	0.909091	1.007266	1.139113	1.278655	1.443376	1.643421	1.889822	2.151657	0.00

**Решение.** Специально выбрана функция, имеющая максимум ближе к одному из концов интегрирования (максимум в точке  $x = 1./1.1$ ), и обращающаяся в нуль на том же конце. Поведение функции на правом краю интервала интегрирования сеточной функцией никак не прописано. Этот интеграл можно вычислить точно, и его значение равно

$$I_{ex} = 1.1\pi + 1 - \frac{0.42}{\sqrt{0.21}} \left[ \operatorname{arctg} \frac{0.1}{\sqrt{0.21}} + \operatorname{arctg} \frac{1}{\sqrt{0.21}} \right] \approx 1.485130.$$

Применение метода трапеций дает приближенное значение интеграла  $I_h = 1.375982$ , вычисление с двойным шагом  $I_{2h} = 1.231714$ , уточнение экстраполяцией Ричардсона совпадает с применением метода Симпсона

$I_S = I_R = 1.424071$ . Оценка точности любого из методов в данном случае будет некорректна, т.к. максимальное значение производных на правом конце равно бесконечности. Для функций с конечными значениями производной уточнение экстраполяцией дает очень хороший результат, но не для данного случая.

Построим решение по алгоритму Рунге–Ромберга на основе экстраполяции вычисленных интегралов в нулевой шаг интегрирования (по  $h^2$ ) для метода трапеций и метода Симпсона.

Метод трапеций:

$h^2$	$I$	$\Delta^1$	$\Delta^2$
$(0.5)^2 = 0.25$	0.948961		
		-1.508016	
$(0.25)^2 = 0.0625$	1.231714		6.697392
		-3.0777173	
$(0.125)^2 = 0.015625$	1.375982		

Интерполяция в нулевой шаг интегрирования дает  $I = 1.430612$ , ошибка довольно велика (3.67%).

Метод Симпсона:

$h^2$	$I$	$\Delta^1$	$\Delta^2$
$(0.5)^2 = 0.25$	1.113765833		
		-1.131730222	
$(0.25)^2 = 0.0625$	1.32596525		4.1011788
		-2.092944	
$(0.125)^2 = 0.015625$	1.424072		

Уточнение по Рунге двух последних значений (метод шестого порядка аппроксимации) дает  $I = 1.430612$ , что совпадает с результатом интерполяции в нулевой шаг интегрирования по трем сеткам метода трапеций, что естественно.

Интерполяция в нулевой шаг интегрирования по трем значениям метода Симпсона (метод восьмого порядка) дает  $I = 1.460779$ , ошибка вдвое меньше (1.64%), но тоже велика.

Измельчим шаг интегрирования еще вдвое, при этом впервые появляется единственная точка на убывающем участке функции:

$h^2$	$I$	$\Delta^1$	$\Delta^2$	$\Delta^3$
$(0.25)^2 = 0.0625$	1.32596525			
		-2.092944		-79.845679
$(0.125)^2 = 0.015625$	1.424072		23.7507015	
		-3.48458(6)		
$(0.0625)^2 = 0.00390625$	1.464907			

Интерполяция в нулевой шаг интегрирования по трем последним сеткам дает  $I = 1.479968$ , ошибка 0.35%.

Интерполяция в нулевой шаг интегрирования по четырем последним шагам дает  $I = 1.480273$ , ошибка 0.33%. Заметим, что вычисление методом Симпсона даже для самой мелкой сетки дает значительно большие отличия: 1.36%.

Интересно сравнить этот результат с интегрированием по Симпсону, когда поведение функции на правом конце учтено еще лучше:

$h^2$	$I$	$\Delta^1$	$\Delta^2$
$(0.1)^2 = 0.01$	1.442913		
		-3.7344000	
$(0.05)^2 = 0.0025$	1.470921		137.78488888
		-5.02613333	
$(0.025)^2 = 0.000625$	1.480345		

Интерполяция в нулевой шаг интегрирования дает  $I = 1.483702$ , погрешность (0.096%).

**VII.7.2.** Определить фактический порядок точности методов трапеций и Симпсона вычисления интеграла из задачи VII.7.1  $\int_0^1 \frac{\sqrt{1-x^2}}{1.1-x} dx$ . Здесь подынтегральная функция недифференцируема на правой границе.

Решение. Так как в Задаче VII.7.1 фиксирована таблица функции, будем не «сгущать» сетку, а «разрежать» ее. Для метода трапеций на последовательности трех сеток с шагами 0.5, 0.25 и 0.125 получим из таблиц этой задачи:  $2^P = \frac{1.231714 - 0.948961}{1.375982 - 1.231714} = 1.959915$ .

На этих же сетках для метода Симпсона имеем:  
 $2^P = \frac{1.32596525 - 1.113765833}{1.424072 - 1.32596525} = 2.16294$ , а на трех сетках 0.25, 0.125 и 0.0625 для метода Симпсона:  $2^P = \frac{1.424072 - 1.32596525}{1.464907 - 1.424072} = 2.40252$ .

Построим теперь новые таблицы подынтегральной функции с более мелкими шагами и проведем новые серии вычислений. Для более подробной последовательности вдвое сгущающихся сеток с шагами 0.1, 0.05, 0.025 для метода Симпсона:

$$2^P = \frac{1.470921 - 1.442913}{1.480345 - 1.470921} = 2.9720.$$

Как видим, предельный реальный порядок сходимости как метода трапеций, так и метода Симпсона, оказываются значительно хуже теоретического. Если для метода трапеций происходит понижение порядка на единицу, то для метода Симпсона реальный порядок сходимости  $p = \log_2 2.9720 \approx 1.57$ . Это очень далеко от теоретической четверки.

Теперь применим формулу (2.10) для уточнения результата. Для удобства сделаем таблицу сравнения результата экстраполяции в нулевой шаг метода Симпсона и процесса Эйткена уточнения результата на одних и тех же последовательностях из трех сеток. Например, уточнение метода Эйткена на последних трех подробных сетках с шагами 0.1, 0.05, 0.025 приводит к результату

$$I \approx I_h + ch^p = 1.442913 + \frac{(1.470921 - 1.442913)^2}{2 \cdot 1.470921 - 1.480345 - 1.470921} = 1.485124.$$

Напомним, что точное значение интеграла — 1.485130.

Шаги сетки	Экстраполяция метода Симпсона из VII.7.1		Процесс Эйткена	
	Значение интеграла	Абсолютная ошибка	Значение интеграла	Абс. ошибка
0.5, 0.25, 0.125	1.460779	0.024351	1.508433	-0.023303
0.25, 0.125, 0.0625	1.479968	0.005162	1.494022	-0.008892
0.1, 0.05, 0.025	1.483702	0.001428	1.485124	0.000006

Из приведенных в таблице численных результатов следует, что процесс уточнения, исходя из реального порядка сходимости (Эйткена), дает примерно такие же значения, как и экстраполяция в нулевой шаг при больших шагах. Так продолжается до тех пор, пока в проблемной области поведения функции не появляется несколько точек на самой мелкой сетке. После этого у процесса уточнения результата по Эйткену конкурентов нет. Заметим, что для первого значения шага 0.1 мелкой сетки табличная функция «не чувствует» особенность подынтегральной функции вблизи правой границы отрезка интегрирования.

Приведенные в таблице численные результаты показывают важность учета, с одной стороны, реального порядка точности метода и минимальной информации об особенностях поведения подынтегральной функции — с другой стороны.

**VII.7.3.** Вычислить  $\int_{-1}^1 \ln(1+x^2/3)dx$  по формулам Гаусса с точностью  $\varepsilon = 10^{-3}$ .

Решение. Воспользуемся четностью подынтегральной функции:

$$\int_{-1}^1 \ln(1+x^2/3)dx = 2 \int_0^1 \ln(1+x^2/3)dx$$

Оценим количество узлов интегрирования по квадратурным формулам Гаусса, необходимое для достижения заданной точности. Для этого нам нужны оценки сверху величин производных подынтегральной функции  $f(x) = \ln(1+x^2/3)$ :

$$f'(x) = \frac{2x}{3+x^2} = \frac{x+\sqrt{3}i + x-\sqrt{3}i}{(x+\sqrt{3}i)(x-\sqrt{3}i)} = \frac{1}{x+\sqrt{3}i} + \frac{1}{x-\sqrt{3}i},$$

$$f''(x) = -\left(x+\sqrt{3}i\right)^{-2} - \left(x-\sqrt{3}i\right)^{-2},$$

$$f'''(x) = 2\left(x+\sqrt{3}i\right)^{-3} + 2\left(x-\sqrt{3}i\right)^{-3},$$

$$f''''(x) = -6\left(x+\sqrt{3}i\right)^{-4} - 6\left(x-\sqrt{3}i\right)^{-4},$$

$$\max_{[0,1]} |f''''(x)| \leq 6(\sqrt{3})^{-4} + 6(\sqrt{3})^{-4} = \frac{4}{3},$$

тогда ошибка интегрирования на двух узлах на интервале единичной длины в соответствии с (3.4) – (3.5) будет оценена как

$$r_2^I = \frac{1}{2^5} \frac{1}{135} f^{(4)}(\xi) \leq \frac{1}{4320} \frac{4}{3} \approx 3 \cdot 10^{-4} < \varepsilon. \quad (7.1)$$

Таким образом, для достижения заданной точности нам достаточно двух узлов для вычисления интеграла по отрезку  $[0, 1]$ :

$$\int_{-1}^1 \ln(1+x^2/3)dx = 2(c_1 \ln(1+x_1^2/3) + c_2 \ln(1+x_2^2/3)). \quad (7.2)$$

Здесь  $c_1$  и  $c_2$  — веса квадратурной формулы на отрезке  $[0, 1]$ , а  $x_1$  и  $x_2$  — узлы квадратуры на этом интервале.

На  $[-1, 1]$  узлы и веса  $\tilde{c}_1 = \tilde{c}_2 = 1$ ,  $t_1 = -t_2 = \sqrt{3}/3$ . Выполняя линейную замену (3.3), для весов и узлов квадратурной формулы получим:  $c_1 = c_2 = 1/2$ ,  $x_{1,2} = 1/2 \pm \sqrt{3}/6$ . Подставляя это в (7.2), получим

$$\begin{aligned} \int_{-1}^1 \ln(1+x^2/3) dx &= 2 \left[ \frac{1}{2} \ln \left( 1 + \frac{1}{3} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right)^2 \right) + \frac{1}{2} \ln \left( 1 + \frac{1}{3} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right)^2 \right) \right] = \\ &= \ln \left( 1 + \frac{73}{9 \cdot 36} \right) = 0.2031928. \end{aligned}$$

Сравнивая результат с точным значением  $2 \ln \frac{4}{3} - 4 + \frac{2\sqrt{3}\pi}{3} \approx 0.2029629$ , видим, что точность оценки (7.1) оказалась высокой.

**VII.7.4.** Предложить метод вычисления несобственного интеграла  $\int_0^\infty \frac{1}{1+x^5} dx$  с точностью  $\varepsilon = 10^{-4}$ .

Решение. Разобъем интеграл на два:

$$\int_0^\infty \frac{1}{1+x^5} dx = \int_0^M \frac{1}{1+x^5} dx + \int_M^\infty \frac{1}{1+x^5} dx \quad (7.3)$$

Параметр  $M$  можно выбирать из условия, что значение второго интеграла в правой части (7.3) окажется меньше половины требуемой точности, а вторую половину допустимой погрешности можно было бы использовать для приближенного вычисления первого интеграла, например, методом трапеций или прямоугольников.

Поступим иначе. При больших  $M$  значение второго интеграла мало отличается от интеграла, вычисляемого точно:

$$\int_M^\infty \frac{1}{x^5} dx = \frac{1}{4M^4}. \quad (7.4)$$

Выберем  $M$  из условия, что разность этих двух интегралов по модулю не превосходит половины заданной точности:

$$\left| \int_M^\infty \frac{1}{1+x^5} dx - \int_M^\infty \frac{1}{x^5} dx \right| = \left| \int_M^\infty \frac{1}{x^5(1+x^5)} dx \right| \leq \int_M^\infty \frac{1}{x^{10}} dx = \frac{1}{9M^9} \leq \frac{1}{2} \varepsilon = \frac{1}{2} \cdot 10^{-4}.$$

Это неравенство позволяет определить  $M \geq 2.3534$ . Возьмем  $M = 2.5$ , тогда второй интеграл в (7.3) с нужной точностью будет равен 0.0064 в соответствии с оценкой (7.4). Для вычисления первого интеграла в (7.3), например, методом прямоугольников со средней точкой, нужно оценить шаг численного интегрирования, обеспечивающий нужную точность. Оценим вторую производную подынтегральной функции на отрезке  $[0, 2.5]$ .

$$f'(x) = -\frac{5x^4}{(1+x^5)^2}, \quad f''(x) = \frac{30x^8 - 20x^3}{(1+x^5)^3},$$

$$f'''(x) = \frac{30x^2(-7x^{10} + 16x^5 - 2)}{(1+x^5)^3}.$$

Точки экстремума второй производной  $x_{1,2} = \sqrt[5]{\frac{8 \pm \sqrt{50}}{7}}$ , максималь-

ное значение второй производной  $|f''(x)| \leq 3.5$ . Тогда для обеспечения заданной точности формулы прямоугольников со средней точкой нужно выбрать шаг интегрирования, удовлетворяющий неравенству

$$\frac{1}{24} h^2 \cdot (2.5 - 0) \cdot 3.5 \leq \frac{1}{2} \varepsilon = \frac{1}{2} 10^{-4},$$

откуда  $h < 0.012$ . Возьмем, например,  $h = 0.01$ . Тогда  $x_k = kh$ ,  $k = 0, \dots, 250$ ,

$$\int_0^{2.5} \frac{1}{1+x^5} dx \approx \sum_{k=0}^{249} h \cdot f(x_k + h/2) = 0.01 \cdot \sum_{k=0}^{249} \frac{1}{1+(x_k + h/2)^5}.$$

## VII.8. Теоретические задачи

**VII.8.1.** Описать алгоритм автоматического выбора шага интегрирования, основанный на экстраполяции по Ричардсону.

**VII.8.2.** Оценить погрешность квадратурной формулы

$$\int_a^b f(x) dx \equiv (b-a) \sum_{k=1}^n c_k f(x_k),$$

порожденную погрешностями в таблице значений  $f(x_k)$ .

**VII.8.3.** Пусть  $\Delta ABC$  — треугольник в плоскости  $(x,y)$ , точки  $M, N, K$  — середины его сторон. Показать, что кубатурная формула

$$\iint_{\Delta ABC} f(x, y) dx dy = \frac{1}{3} S_{\Delta ABC} (f(M) + f(N) + f(K))$$

точна для всех полиномов  $f(x, y) = a_{11}x^2 + a_{12}xy + a_{22}y^2 + b_1x + b_2y + c$ .

**VII.8.4.** Показать, что квадратурная формула  $I_n(f) = \frac{\omega}{n} \sum_{k=0}^{n-1} f\left(\frac{k\omega}{n}\right)$  для

вычисления интегралов вида  $I(f) = \int_a^b f(x) dx$  точна для всех тригонометрических многочленов с периодом  $\omega$  степени не выше  $n - 1$ .

**VII.8.5.** Для вычисления  $\int_0^1 f(x) dx$  применяется составная формула трапеций. Оценить минимальное число разбиений  $N$ , обеспечивающее точность  $10^{-3}$  на двух классах функций:

$$1) \|f''\|_C \leq 1, \quad 2) \|f''\|_{L_1} = \int_0^1 |f''(x)| dx \leq 1.$$

Указание: воспользоваться результатом задачи 7.6.3 на каждом элементарном отрезке интегрирования.

**П.8.6.** Получить формулу Симпсона методом неопределенных коэффициентов.

**VII.8.7.** Доказать, что формула Эйлера–Маклорена:

$$\int_a^{a+h} f(x) dx \approx \frac{h}{2} (f(a+h) + f(a)) - \frac{h^2}{12} (f'(a+h) - f'(a))$$

имеет порядок аппроксимации не ниже четвертого. Как на этой основе построить уточнение составной квадратурной формулы трапеций с постоянным шагом, если известны значения производных подынтегральной функции на концах отрезка интегрирования?

**VII.8.8.** Оценить минимальное число разбиений отрезка для вычисления интеграла по составной формуле трапеций, обеспечивающее точность  $10^{-4}$ .

$$a) I = \int_0^1 e^{-x^2} dx, \quad b) \int_0^1 \sin(x^2) dx.$$

**VII.8.9.** Предложить метод вычисления интеграла  $\int_{-1}^2 e^{|x-1|(x-1)} dx$  с заданной точностью  $\varepsilon$  по формуле трапеций.

**VII.8.10.** Показать, что квадратурная формула Гаусса с двумя узлами на  $[-1, 1]$  точна для полиномов третьей степени.

**VII.8.11.** Показать, что квадратурная формула Гаусса–Кристоффеля

$$\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx \approx \frac{\pi}{3} \left( f\left(-\frac{\sqrt{3}}{2}\right) + f(0) + f\left(\frac{\sqrt{3}}{2}\right) \right)$$

точна для многочленов пятой степени.

**VII.8.12.** Построить квадратуру Гаусса–Кристоффеля с двумя узлами для вычисления интеграла:

а)  $I(f) = \int_{-1}^1 x^2 f(x) dx$ ,      б)  $I(f) = \int_{-1}^1 \frac{f(x) dx}{\sqrt{1-x^2}}$ ,

в)  $I(f) = \int_{-\pi/2}^{\pi/2} \cos x f(x) dx$ ,      г)  $I(f) = \int_0^\infty \exp(-x) f(x) dx$ ,

д)  $I(f) = \int_0^\pi \sin x f(x) dx$ ,      ж)  $I = \int_{-\infty}^{+\infty} \exp(-x^2) f(x) dx$ ,

з)  $\int_{-1}^1 \sqrt{1-x^2} f(x) dx$ .

**VII.8.13.** Построить квадратурную формулу Гаусса–Кристоффеля с двумя узлами для вычисления интеграла  $\int_0^1 p(x) f(x) dx$ ,  $p(x)$  — весовая функция:

а)  $p(x) = x$ ;      б)  $p(x) = \sin \pi x$ ;

в)  $p(x) = \exp(x)$ ;      г)  $p(x) = \ln(1+x)$ ;

д)  $p(x) = 1-x$ ;      е)  $p(x) = \exp(-x)$ .

**VII.8.14.** Вычислить приближенное значение интеграла, используя квадратурную формулу Гаусса с двумя узлами и оценить ее погрешность.

а)  $I = \int_0^1 \sin(x^2) dx$ ,

б)  $\int_0^4 \frac{x dx}{(3x+4)^2}$ .

**VII.8.15.** Применяя метод Канторовича выделения особенностей, предложить алгоритм приближенного вычисления интеграла с точностью до  $10^{-5}$ :

$$\int_0^1 \frac{dx}{\sqrt{x}(1+x)}.$$

**VII.8.16.** Предложить способ приближенного вычисления интеграла  $\int_0^{1/2} \ln\left(\ln\left(\frac{1}{x}\right)\right) dx$  с точностью не менее  $10^{-4}$  с помощью метода устранения особенности. Количество шагов интегрирования не должно превышать  $10^4$ .

**VII.8.17.** Предложить способ вычисления несобственного интеграла с заданной точностью. Оценить величину шага интегрирования.

а)  $\int_0^1 \frac{1}{\sqrt{x(1-x)}} dx, \varepsilon = 10^{-4},$     б)  $\int_0^1 \cos \frac{\pi}{x} dx, \varepsilon = 5 \cdot 10^{-5},$

в)  $\int_0^1 \frac{\ln x}{\sqrt{1-x}} dx, \varepsilon = 10^{-3},$     г)  $\int_0^1 x \sin \frac{1}{x} dx, \varepsilon = 5 \cdot 10^{-5},$

д)  $\int_1^\infty \frac{1}{x^2} \sin \frac{1}{x^2} dx, \varepsilon = 10^{-4},$     е)  $\int_1^\infty \frac{\cos(x^2)}{x} dx, \varepsilon = 10^{-4},$

ж)  $\int_0^1 \frac{dx}{\sqrt{x}\sqrt{x+1}}, \varepsilon = 10^{-3}.$

**VII.8.18.** Предложить алгоритм вычисления интеграла с заданной точностью  $\varepsilon$ , используя метод регуляризации подынтегральной функции:

а)  $\int_0^2 \frac{\cos(x)}{\sqrt{2x-x^2}} dx,$  б)  $\int_0^3 \frac{\sin(\sqrt{x})}{\sqrt{3x^2-x^3}} dx,$  в)  $\int_0^4 \frac{e^{-\sqrt{x}}-1}{\sqrt[4]{4x^2-x^3}} dx,$  г)  $\int_0^5 \frac{1-e^x}{x\sqrt[5]{5x-x^2}} dx,$

д)  $\int_0^{0.5} \frac{\sqrt{2} \sin(x)}{x\sqrt{x-2x^2}} dx,$  е)  $\int_0^3 \frac{e^{-x}}{\sqrt[3]{x}\sqrt{3-x}} dx,$  ж)  $\int_0^2 \frac{e^x}{\sqrt[2]{2x-x^2}} dx.$

**VII.8.19.** Предложить алгоритм вычисления интеграла  $\int_a^b f(x) \cos \omega x dx,$

если  $f(x)$  задана как сеточная функция на отрезке  $[a, b].$

## VII.9. Практические задачи

**VII.9.1.** Для функции, заданной таблично, вычислить значение интеграла с использованием указанной формулы. Уточнить полученное значение интеграла с помощью экстраполяции Ричардсона.

а) формула Симпсона

$x$	-1.	-0.75	-0.5	-0.25	0.	0.25	0.5	0.75	1.
$f(x)$	-1	-0.14	-0.032	0.01	0	0.002	0.003	0.0031	0.0029

б) формула трапеций

$x$	0.	0.25	0.5	0.75	1.	1.25	1.5	1.75	2.
$f(x)$	0	0.028	0.054	0.078	0.1	0.2	0.133	0.145	0.154

**VII.9.2.** Для функции, заданной таблично,

$x$	0.	0.25	0.5	0.75	1.	1.25	1.5	1.75	2.
$f(x)$	0	0.004	0.015	0.034	0.059	0.089	0.123	0.3	0.2

вычислить значение интеграла с использованием формулы Симпсона. Оценить погрешность по правилу Рунге.

**VII.9.3.** Для функции, заданной таблично,

$x$	-1	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	1
$f(x)$	-0.333	0	-0.125	-0.056	0	0.046	0.083	0.115	0.143

вычислить значение интеграла с использованием формулы трапеций. Оценить погрешность по правилу Рунге.

**VII.9.4.** Для функции, заданной таблично,

$x$	0	0,5	1	1,5	2
$f$	0,5	0,25	0,25	0,1	0,1

вычислить значение интеграла с использованием формулы Симпсона. Оценить погрешность вычисления этого интеграла.

**VII.9.5.** Для функции, заданной таблично, вычислить значение определенного интеграла методом трапеций, сделать уточнение результата

по правилу Рунге. Сравнить уточненный результат с результатом вычислений по формуле Симпсона.

а)		б)		в)	
$x$	$f(x)$	$x$	$f(x)$	$x$	$f(x)$
0.	1.000000	0.	0.000000	0.	1.000000
0.25	0.989616	0.125	0.021470	0.15	1.007568
0.5	0.958851	0.25	0.293050	0.3	1.031121
0.75	0.908852	0.375	0.494105	0.45	1.073456
1.0	0.841471	0.5	0.541341	0.6	1.140228
1.25	0.759188	0.625	0.516855	0.75	1.242129
1.5	0.664997	0.75	0.468617	0.9	1.400176
1.75	0.562278	0.875	0.416531	1.05	1.660300
2.	0.454649	1.	0.367879	1.2	2.143460

г)		д)	
$x$	$f(x)$	$x$	$f(x)$
0.	1.000000	0.	0.000000
0.25	0.979915	0.125	0.124670
0.5	0.927295	0.25	0.247234
0.75	0.858001	0.375	0.364902
1.0	0.785398	0.5	0.473112
1.25	0.716844	0.625	0.563209
1.5	0.655196	0.75	0.616193
1.75	0.600943	0.875	0.579699
2.	0.553574	1.	0.000000

**VII.9.6.** Оценить точность вычисления интегралов в VII.9.5 методом трапеций и методом Симпсона по правилу Рунге.

**VII.9.7.** С помощью разложения в степенные ряды вычислить интегралы с точностью до  $10^{-4}$ :

$$\text{а) } I = \int_0^1 e^{-x^2} dx, \quad \text{б) } I = \int_0^1 \sin(x^2) dx.$$

**VII.9.8.** Вычислить несобственный интеграл с точностью  $10^{-4}$

$$\text{а) } \int_0^1 \frac{dx}{(1+x)\sqrt{x}}, \quad \text{б) } \int_0^1 \frac{\cos x}{\sqrt{x}} dx, \quad \text{в) } \int_0^\infty \frac{\sin x}{1+x^2} dx,$$

$$\text{г) } \int_0^1 \frac{\ln(x^2 + 1)}{\sqrt{x}} dx, \quad \text{д) } \int_0^{1.5} \frac{e^x}{x^2} dx - \int_0^{1.5} \frac{1+x}{x^2} dx, \quad \text{е) } \int_1^\infty \frac{\operatorname{arctg} x}{x^2} dx.$$

Сравните различные приемы для решения каждой задачи.

**VII.9.9.** Вычислить несобственный интеграл с точностью  $10^{-3}$ :

$$\text{а) } \int_1^\infty \frac{dx}{(1+x)\sqrt{x}}; \quad \text{б) } \int_1^\infty e^{-x^2} \sin x dx; \quad \text{в) } \int_0^\infty \frac{(1-\cos x)dx}{x\sqrt{x}}.$$

**VII.9.10.** Вычислить интеграл от быстро осциллирующей функции с точностью  $10^{-6}$ :

$$\text{а) } \int_0^1 \frac{\sin 100x dx}{1+x}, \quad \text{б) } \int_1^2 \cos 100x \ln x dx.$$

**VII.9.11.** Функция  $f(x)$  задана своими сеточными значениями. Найти  $\int_a^b f(x) \sin kx dx$  построением сплайна для аппроксимации  $f(x)$ .

а)  $k = 50$

$x_i$	0.	1.	2.	3.	4.
$f_i$	0.00000	0.50000	0.86603	1.00000	0.86603

б)  $k = 80$

$x_i$	0.1	0.5	0.9	1.3	1.7
$f_i$	-2.3026	-0.69315	-0.10536	0.26236	0.53063

в)  $k = 40$

$x_i$	0.	1.7	3.4	5.1	6.8
$f_i$	0.	1.3038	1.8439	2.2583	2.6077

г)  $k = 100$

$x_i$	-0.4	-0.1	0.2	0.5	0.8
$f_i$	1.9823	1.6710	1.3694	1.0472	0.64360

д)  $k = 30$

$x_i$	0.	1.	2.	3.	4.
$f_i$	1.	1.5403	1.5839	2.0100	3.3464

## VII.10. Библиографический комментарий

О методах численного интегрирования, квадратурных и кубатурных формулах на начальном уровне можно прочитать в [2, 4, 5, 7, 8, 27, 35]. Об

оценке функционала погрешности для формул интерполяционного типа, о свойстве правильности и о некоторых других вопросах численного интегрирования см. [35]. О вычислении интегралов на квазиравномерных сетках с гарантированной точностью см. [38, 40].

# VIII. ЗАДАЧА КОШИ ДЛЯ СИСТЕМ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

## VIII.1. Аппроксимация, устойчивость, сходимость

Будем рассматривать численные методы решения задачи Коши для обыкновенных дифференциальных уравнений (ОДУ):

$$\frac{du}{dt} = f(t, u), \quad t > 0, \quad u(0) = u_0, \quad (1.1)$$

а также систем ОДУ:

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(t, \mathbf{u}), \quad \mathbf{u}(0) = \mathbf{u}_0,$$

где  $\mathbf{u}, \mathbf{f}$  — векторы-столбцы искомых функций и правых частей соответственно.

К аналогичной форме приводится задача Коши для обыкновенного дифференциального уравнения (системы уравнений) порядка выше первого

$$\frac{d^m u}{dt^m} = g\left(t, u, \frac{du}{dt}, \dots, \frac{d^{m-1}u}{dt^{m-1}}\right), \quad t > 0, \quad u(0) = a_0,$$

$$\frac{du}{dt}(0) = a_1, \dots, \frac{d^{m-1}u}{dt^{m-1}}(0) = a_{m-1},$$

если положить

$$u_1 = u, \quad u_2 = du_1/dt, \quad u_3 = du_2/dt, \dots, \quad u_m = du_{m-1}/dt,$$

$$\frac{du_m}{dt} = g(t, u_1, u_2, \dots, u_m),$$

$$u_i(0) = a_{i-1}, \quad i = 1, 2, \dots, m.$$

Введем в расчетной области  $t \in [0, T]$  точки (узлы расчетной сетки)  $\{t_n = nt, n = 0, 1, \dots, N\}$ , в которых вычисляется искомое решение. Совокупность узлов называется *расчетной сеткой* (сеточной областью),  $\tau$  — шагом

*интегрирования*. Здесь для простоты введена равномерная сетка. В реальных расчетах применяются и неравномерные сетки.

Введем сеточную функцию  $\mathbf{y}$ , определенную в узлах сетки и представляющую собой совокупность приближенных значений искомой функции,  $\mathbf{U}^\tau$  — проекцию точного решения искомой задачи на сетку.

Будем использовать также операторное обозначение дифференциальной задачи

$$\mathbf{L}(\mathbf{u}) = \mathbf{F}, \quad (1.2)$$

где

$$\mathbf{L}(\mathbf{u}) = \begin{cases} \frac{du}{dt} - f(t, u), & t > 0; \\ u(0), & t = 0; \end{cases}, \quad \mathbf{F} = \begin{cases} 0, & t > 0; \\ u_0, & t = 0; \end{cases}$$

и аппроксимирующей разностной задачи

$$\mathbf{L}_\tau(\mathbf{y}) = \mathbf{F}_\tau, \quad (1.3)$$

где  $\mathbf{L}_\tau$  — обозначения разностного оператора,  $\mathbf{F}_\tau$  — проекция  $\mathbf{F}$  на расчетную сетку. Заметим, что  $u$  и  $\mathbf{y}$  являются элементами разных пространств.

**Определение.** Решение задачи (1.3) *сходится* при  $\tau \rightarrow 0$  к решению исходной задачи (1.2), если  $\|\mathbf{y} - \mathbf{U}^\tau\| \rightarrow 0$  при  $\tau \rightarrow 0$ . Если при этом имеет место оценка  $\|\mathbf{y} - \mathbf{U}^\tau\| \leq C \tau^p$ , причем постоянная в правой части не зависит от сеточных параметров, то имеет место сходимость порядка  $p$ .

**Определение.** Говорят, что задача (1.3) *аппроксимирует* задачу (1.2) на ее решении, если невязка  $\|\mathbf{r}_\tau\| \rightarrow 0$  при  $\tau \rightarrow 0$ , где  $\mathbf{r}_\tau \equiv \mathbf{L}_\tau(\mathbf{U}^\tau) - \mathbf{F}_\tau$ . При этом если имеет место оценка  $\|\mathbf{r}_\tau\| \leq C_1 \tau^q$ , причем постоянная в правой части не зависит от сеточных параметров, то имеет место сходимость порядка  $q$ .

**Определение.** Задача (1.3) *устойчива*, если существуют параметры  $\varepsilon$ ,  $\tau_0$ :  $\forall \tau < \tau_0$  и  $\forall \xi, \eta : \|\xi\| \leq \varepsilon, \|\eta\| \leq \varepsilon$  две близкие возмущенные задачи одновременно однозначно разрешимы и из равенств  $\mathbf{L}_\tau(\mathbf{y}_1) = \mathbf{F}_\tau + \xi_\tau$ ,  $\mathbf{L}_\tau(\mathbf{y}_2) = \mathbf{F}_\tau + \eta_\tau$  следует

$$\|\mathbf{y}_1 - \mathbf{y}_2\| \leq C_2 (\|\xi_\tau\| + \|\eta_\tau\|),$$

причем  $C_2$  не зависит от сеточных параметров.

**Теорема 1.**(А.Ф.Филлипова – В.С.Рябенького, П. Лакса – Р. Рихтмайера). *Решение задачи (1.3) сходится к решению исходной задачи*

(1.2), если задача (1.3) устойчива и аппроксимирует задачу (1.2) на ее решении; если аппроксимация имеет порядок  $p$ , то сходимость также имеет порядок  $p$ .

Приведем примеры простейших разностных уравнений, аппроксимирующих (1.1):

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n), \quad 0 \leq n \leq N-1,$$

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+1}, y_{n+1}), \quad 0 \leq n \leq N-1,$$

$$\frac{y_{n+1} - y_{n-1}}{2\tau} = f(t_n, y_n), \quad 1 \leq n \leq N-1.$$

Первая из схем называется *явной* (явная схема Эйлера), вторая — *неявной* (неявная схема Эйлера). Алгоритмическая реализация первой схемы — «бегущий счет» (рекуррентная формула), второй — решение нелинейного алгебраического уравнения на каждом временном шаге.

Для реализации третьей схемы (*метод Эйлера с центральной точкой*) необходимо задание функции  $u_n$  в двух точках  $t_0$  и  $t_1$ . Один из возможных вариантов — решение на первом шаге нелинейного уравнения (метод трапеций):

$$\frac{y_1 - y_0}{\tau} = \frac{1}{2} [f(0, y_0) + f(\tau, y_1)].$$

В данном случае мы сталкиваемся с явлением несовпадения формальных порядков дифференциального и разностного уравнений (дифференциальное уравнение первого порядка, разностное — второго). Теория таких методов (их называют многошаговыми) опирается на общую теорию линейных разностных уравнений.

Рассмотрим еще один способ получения простейших одношаговых расчетных схем для численного решения уравнения (1.1). Решение (1.1) может быть записано как

$$u(t + \tau) = u(t) + \int_0^\tau u'(t + \xi) d\xi.$$

После аппроксимации интеграла в правой части по формуле прямоугольников и замене его на величину  $\tau u'(t)$  получим

$$u(t + \tau) = u(t) + \tau u'(t) + O(\tau^2),$$

или

$$u(t + \tau) = u(t) + \tau f(t, u) + O(\tau^2),$$

поскольку  $u'(t) = f(t, u)$ .

Опуская член  $O(\tau^2)$  и обозначая  $t = t_n$ ,  $t + \tau = t_{n+1}$ ,  $u(t) = y_n$ ,  $u(t + \tau) = y_{n+1}$ , получим явный метод Эйлера.

Если интеграл в правой части приблизить формулой трапеций, то

$$u(t + \tau) = u(t) + \frac{\tau}{2} [u'(t) + u'(t + \tau)] + O(\tau^3),$$

откуда имеем

$$y_{n+1} = y_n + 0.5\tau [f(t_n, y_n) + f(t_{n+1}, y_{n+1})].$$

Этот метод называется *неявным методом трапеций*.

## VIII.2. Исследование устойчивости разностных схем для ОДУ

Исследование устойчивости разностных схем напрямую с использованием определения возможно лишь для малого класса уравнений и схем. В этом смысле ситуация близка к прямому исследованию сходимости разностных схем.

Устойчивость исследуется исходя из канонической формы записи разностной схемы (способ приведения к канонической форме записи может быть неединственным)

$$y_{n+1} = \mathbf{R}_\tau y_n + \tau \rho_n. \quad (2.1)$$

Оператор  $\mathbf{R}_\tau$  называется оператором перехода или разрешающим оператором. Уравнение (2.1) является линейным, т.е. при переходе к канонической записи разностной схемы мы пользуемся процедурой линеаризации.

Теорема. (Достаточное условие устойчивости.) Пусть разностная схема  $\mathbf{L}_\tau \mathbf{y} = \mathbf{f}_\tau$  приведена к каноническому виду (2.1), и пусть выполнены неравенства

$$\|y_0\| \leq C_2 \|\mathbf{f}\|, \quad \|\rho_n\| \leq C_2 \|\mathbf{f}\|.$$

Тогда для устойчивости

$$\|y_n\| \leq C_2 \|\mathbf{f}\| \quad (2.2)$$

достаточно, чтобы нормы степеней оператора  $\|\mathbf{R}_h^m\|$  были равномерно по  $\tau$  ограничены, т.е. чтобы выполнялась оценка

$$\|\mathbf{R}_\tau^m\| \leq C_3, \quad m = 1, \dots, N.$$

При этом в качестве числа  $C$  в оценке (2.2) может быть взята величина

$$C = (1 + T) C_2 C_3. \quad (2.3)$$

Доказательство следует из цепочки равенств

$$y_1 = \mathbf{R}_\tau y_0 + \tau \rho_0,$$

$$y_2 = \mathbf{R}_\tau y_1 + \tau \rho_1 = \mathbf{R}_\tau^2 y_0 + \tau (\mathbf{R}_\tau \rho_0 + \rho_1),$$

...

$$y_n = \mathbf{R}_\tau y_{n-1} + \tau \rho_n = \mathbf{R}_\tau^n y_0 + \tau (\mathbf{R}_\tau^{n-1} \rho_0 + \mathbf{R}_\tau^{n-2} \rho_1 + \dots + \rho_{n-1}).$$

и следующего отсюда неравенства

$$\max_n |y_n| \leq \max_n \|\mathbf{R}_\tau^n\| \left( |y_0| + N \tau \max_n |\rho_n| \right).$$

С учетом  $N\tau = T$  из последнего неравенства следует (2.2) с константой  $C$ , определяемой выражением (2.3).

Для проверки выполнения требования

$$\|\mathbf{R}_\tau^m\| \leq C_3, \quad m = 1, \dots, N \quad (2.4)$$

можно воспользоваться необходимым спектральным признаком устойчивости. Для ОДУ он заключается в следующем. Для нормы оператора справедливо неравенство

$$\|\mathbf{R}_\tau\| \geq \max_i |\lambda_i|.$$

Тогда для выполнения требования (2.4) необходимо, чтобы все собственные значения оператора послойного перехода лежали в круге

$$|\lambda_i| \leq 1 + c\tau, \quad (2.5)$$

при этом постоянная  $c$  не должна зависеть от сеточных параметров. Тогда

$$|1 + c\tau|^m \leq e^{cT}, \quad m = 1, \dots, N. \quad (2.6)$$

При выполнении условия (2.5) будем называть устойчивость нестрогой, а в случае, когда выполнено более сильное условие

$$|\lambda_i| \leq 1, \quad (2.7)$$

будем говорить о *строгой устойчивости*. Требование строгой устойчивости кажется оправданным в том случае, когда расчет ведется не до заранее определенной правой границы интервала расчета по  $t$ , а до выполнения каких-либо условий для решения.

Пример 1. Исследуем устойчивость явной схемы Эйлера решения задачи Коши для ОДУ  $u' = f(t, u)$ ,  $u(0) = u_0$ :

$$\frac{y_{n+1} - y_n}{\tau} = f(t_n, y_n), \quad y_0 = u_0.$$

Запишем разностную схему в виде

$$y_{n+1} = y_n + \tau f(t_n, y_n).$$

Для приведения к каноническому виду линеаризуем разностную схему в окрестности некоторой гладкой траектории  $\varphi(t)$ , тогда получим

$$y_{n+1} = y_n + \tau \left( f(t_n, \varphi(t)) + \frac{\partial f}{\partial u} (y_n - \varphi(t)) \right) + \dots$$

$$y_{n+1} = \mathbf{R}_h y_n + \tau \rho_n, \quad \mathbf{R}_h = 1 + \tau \frac{\partial f}{\partial u}, \quad \rho_n = f(t_n, \varphi(t)) - \frac{\partial f}{\partial u} \varphi(t) + \dots$$

Для нестрогой устойчивости (2.5) явной схемы Эйлера достаточно, чтобы была ограничена производная правой части по решению  $\left| \frac{\partial f}{\partial u} \right| \leq c$ .

Для строгой устойчивости ограничения на спектр оператора перехода более жесткие:

$$-1 \leq 1 + \tau \frac{\partial f}{\partial u} \leq 1,$$

что приводит к двум условиям  $\frac{\partial f}{\partial u} \leq 0$ ,  $\tau \leq 2 \left| \frac{\partial f}{\partial u} \right|^{-1}$ .

Пример 2. Исследуем устойчивость неявной схемы Эйлера решения задачи Коши для ОДУ  $u' = f(t, u)$ ,  $u(0) = u_0$ . Неявная схема Эйлера записывается как

$$\frac{y_{n+1} - y_n}{\tau} = f(t_{n+1}, y_{n+1}), \quad y_0 = u_0.$$

Снова представим разностную схему в виде

$$y_{n+1} = y_n + \tau f(t_{n+1}, y_{n+1})$$

Линеаризуем разностную схему в окрестности некоторой гладкой траектории  $\varphi(t)$ , тогда получим

$$y_{n+1} = y_n + \tau \left( f(t_{n+1}, \varphi(t)) + \frac{\partial f}{\partial u}(y_{n+1} - \varphi(t)) \right) + \dots$$

$$y_{n+1} = \mathbf{R}_\tau y_n + \tau \rho_n, \quad \mathbf{R}_\tau = \left( 1 - \tau \frac{\partial f}{\partial u} \right)^{-1}, \quad \rho_n = f(t_{n+1}, \varphi(t)) - \frac{\partial f}{\partial u} \varphi(t) + \dots$$

Поэтому для нестрогой устойчивости (2.5) неявной схемы Эйлера достаточно, как и для явной схемы, чтобы была ограничена производная правой части по решению  $\left| \frac{\partial f}{\partial u} \right| \leq c$ , а для строгой устойчивости требуем

$$-1 \leq \left( 1 - \tau \frac{\partial f}{\partial u} \right)^{-1} \leq 1.$$

Для  $\frac{\partial f}{\partial u} \leq 0$  оба эти условия выполнены всегда, а в случае  $\frac{\partial f}{\partial u} > 0$  возникает требование  $\tau > 2 \left| \frac{\partial f}{\partial u} \right|^{-1}$ .

Как видим, область строгой устойчивости неявного метода Эйлера значительно больше, чем у явного.

В дальнейшем для жестких систем ОДУ понятие строгой устойчивости будет обобщено на понятие А-устойчивости метода. Неявный метод Эйлера обладает самой большой областью устойчивости.

Отметим также, что для неявных методов строгая устойчивость иногда приводит к нежелательным последствиям — численное решение оказывается «устойчивым» и стремится к положению равновесия, а точное решение задачи устроено совершенно иначе. Для того чтобы понять, почему этот эффект нежелателен, достаточно рассмотреть систему линейных дифференциальных уравнений, описывающую малые колебания маятника (Задача VIII.7.10). Замечательный советский математик Л. А. Чудов в своих лекциях иногда называл это явление «сверхустойчивостью неявной схемы».

Пример 3. Исследуем устойчивость схемы с центральной разностью для решения ОДУ  $u' = f(t, u)$  вида

$$\frac{y_{n+1} - y_{n-1}}{2\tau} = f(t_n, y_n).$$

Это пример схемы, для которой формальный разностный порядок схемы не совпадает с порядком дифференциального уравнения, т.е. для нахождения значения  $y_{n+1}$  нам необходимо знать не только  $y_n$ , но и  $y_{n-1}$ .

В этом случае порядок аппроксимации определяется не только порядком аппроксимации уравнения, но и порядком аппроксимации при вычислении  $y_1$ .

Замечание 1. Как мы видели на простейших примерах, при рассмотрении устойчивости определяющую роль играет производная правой части по решению. В случае систем уравнений — матрица Якоби правой части системы.

Замечание 2. Для трехслойных схем привести их непосредственно к виду (2.1) достаточно сложно. Мы не будем этого делать, а предположим, что для собственных значений оператора перехода  $\lambda$  последовательные значения  $y$  связаны равенством

$$y_{n+1} = \lambda y_n, \quad y_n = \lambda y_{n-1}.$$

С учетом этих двух замечаний для определения собственных значений оператора перехода будем иметь

$$\lambda^2 - 2\tau \frac{\partial f}{\partial u} \lambda - 1 = 0.$$

По теореме Виета имеем  $\lambda_1 \lambda_2 = -1$ , что означает, что при действительном дискриминанте  $D = 4\tau^2(f_u')^2 + 4$  разностная схема строгой устойчивостью обладать не может. При достаточно малом шаге имеем:  $|\lambda_{1,2}| \leq 1 + c\tau$ ,

$c = 2 \left| \frac{\partial f}{\partial u} \right|$ , т.е. условие нестрогой устойчивости выполнено при ограниченности производной правой части по решению.

### VIII.3. Явные методы Рунге–Кутты

Сформулируем теперь идею построения численных методов для решения ОДУ. Используем формулу Ньютона–Лейбница:

$$u(t + \tau) - u(t) = \int_t^{t+\tau} f(t, u) dt.$$

Разделим правую и левую части на длину отрезка интегрирования, получим

$$\frac{u(t + \tau) - u(t)}{\tau} = \frac{1}{\tau} \int_t^{t+\tau} f(t, u) dt.$$

Правую часть этого равенства обозначим как  $\Phi$  — она называется функцией приращения. Ее смысл — интегральное среднее правой части по отрезку интегрирования:

$$\frac{u(t + \tau) - u(t)}{\tau} = \Phi(t, u).$$

Заметим, что хотя в левой части стоит конечная разность, приближающая первую производную с первым порядком аппроксимации, равенство является точным при точном вычислении интегрального среднего в правой части. К сожалению, для большинства задач точно вычислить его невозможно, и для этого применяются те или иные приближенные формулы. В зависимости от качества приближения мы получим методы, обладающие разными свойствами.

Очевидный способ построения численных методов такой. Введем на отрезке  $[t, t + \tau]$  набор узлов интегрирования:  $t_j = t + c_j \tau$ . Дополнительно потребуем, чтобы  $c_1 = 0$ . Набор узлов интегрирования однозначно определяет вес квадратурной формулы  $b_j$ . Для определения значения функции в узлах квадратуры воспользуемся какой-либо экстраполяцией в эти узлы. Если мы используем только одно значение  $u(t)$ , то мы получим семейство явных методов Рунге–Кутты.

Их принято представлять в следующей форме.

**Определение.** *S-стадийный одношаговый явный метод для численного решения задачи Коши для обыкновенного дифференциального уравнения (1.1):*

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + c_2 \tau, y_n + \tau a_{21} k_1), \\ k_3 &= f(t_n + c_3 \tau, y_n + \tau(a_{31} k_1 + a_{32} k_2)), \dots, \\ k_s &= f(t_n + c_s \tau, y_n + \tau(a_{s1} k_1 + \dots + a_{s,s-1} k_{s-1})), \\ y_{n+1} &= y_n + \tau(b_1 k_1 + \dots + b_s k_s), \end{aligned} \tag{3.1}$$

где  $k_i$  — промежуточные вспомогательные величины.

Напомним, что явным называется такой численный метод, который позволяет найти значение в следующей точке разностной области, используя только известные значения без необходимости решать нелинейные системы алгебраических уравнений.

Коэффициенты, определяющие конкретный метод, могут быть представлены в виде *таблицы Бутчера* (табл. 1).

Таблица 1

0					
$c_2$	$a_{21}$				
$c_3$	$a_{31}$	$a_{32}$			
$\dots$	$\dots$	$\dots$	$\dots$		
$c_s$	$a_{s1}$	$a_{s2}$	$\dots$	$a_{ss-1}$	
	$b_1$	$b_2$	$\dots$	$b_{s-1}$	$b_s$

Обычно также используют условие, предложенное Куттой и не являющееся обязательным, однако упрощающее вывод условий порядка аппроксимации для многостадийных методов:

$$c_n = \sum_j a_{nj}.$$

Явные методы Рунге–Кутты являются одношаговыми — для построения решения на данном шаге необходимо знать только искомые значения на предыдущем.

#### VIII.4. Устойчивость явных методов Рунге–Кутты

Для исследования устойчивости методов Рунге–Кутты для численного решения задачи

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(t, \mathbf{u}), \quad \mathbf{u}(0) = \mathbf{u}_0,$$

представим ее дискретный аналог в виде

$$\frac{\mathbf{y}_{n+1} - \mathbf{y}_n}{\tau} = \mathbf{F}(t_n, \mathbf{y}_n),$$

здесь  $\mathbf{F}(t, \mathbf{y})$  — функция приращения метода Рунге–Кутты, она, конечно, связана с функцией правой части системы ОДУ.

Сформулируем вначале следующую лемму.

**Лемма 1.** Пусть  $C$  — постоянная Липшица для функции правых частей системы  $\mathbf{f}(t, \mathbf{u})$ , тогда функция приращения  $\mathbf{F}(t, \mathbf{u})$  метода (3.1) удовлетворяет неравенству  $\|\mathbf{F}(t, \mathbf{u}_n) - \mathbf{F}(t, \mathbf{v}_n)\| \leq C_2 \|\mathbf{u}_n - \mathbf{v}_n\|$ , где

$$C_2 = C \left( \sum_i |b_i| + \tau C \sum_{i,j} |b_i a_{ij}| + \tau^2 C^2 \sum_{i,j,k} |b_i a_{ij} a_{jk}| + \dots \right).$$

В некоторых важных частных случаях эту оценку можно улучшить, рассматривая более тонкие свойства рассматриваемой функции. Для *правильных* (все коэффициенты неотрицательны) методов Рунге–Кутты  $C_2 \approx (e^{C\tau} - 1)/\tau$ .

**Теорема 2.** (Устойчивость методов Рунге–Кутты). *Пусть правая часть системы ОДУ  $\mathbf{f}(t, \mathbf{u})$  удовлетворяет условиям Липшица по аргументу  $\mathbf{u}$  с постоянной  $C$ :*

$$\|\mathbf{f}(t, \mathbf{u}) - \mathbf{f}(t, \mathbf{v})\| \leq C \|\mathbf{u} - \mathbf{v}\|$$

(эта оценка не зависит от сеточного параметра). Пусть также  $C_2\tau \ll 1$ .  $C_2$  оценено в Лемме 1. Тогда метод Рунге–Кутты устойчив, и имеет место оценка

$$\|y_n - v_n\| \leq e^{C_2 T} \|y_0 - v_0\| + 2\varepsilon \frac{e^{C_2 T}}{C_2}.$$

Здесь  $\varepsilon$  — максимальная ошибка округления на данной ЭВМ,  $\{y_n\}$  — «точное» сеточное решение задачи,  $\{v_n\}$  — решение возмущенной задачи,  $T$  — длина отрезка интегрирования.

**Замечание.** Данный вывод не зависит от числа стадий метода Рунге–Кутты. Более тонкие оценки получаются с учетом информации о характере решения.

Пусть матрица

$$\mathbf{A}(\mathbf{u}) = \frac{1}{2} (\mathbf{f}_{\mathbf{u}}(\mathbf{u}) + \mathbf{f}_{\mathbf{u}}^*(\mathbf{u}))$$

строго отрицательна, т. е.  $(\mathbf{A}(\mathbf{u})\xi, \xi) \leq -a(\xi, \xi)$  для любых  $\xi, \mathbf{u}$  и  $a > 0$  (траектория, в окрестности которой выполняется это условие, называется *устойчивой*). Тогда при интегрировании правильным методом Рунге–Кутты  $k$ -го порядка аппроксимации погрешность приближенного решения есть  $O(\tau^k)$  при любом  $t > 0$  при выполнении условий  $a\tau \ll 1$ .

При численном интегрировании устойчивой траектории методом Рунге–Кутты порядка  $k$  при всех  $t > 0$  погрешность метода есть  $O(\tau^k)$ .

Пусть теперь  $(\mathbf{Ay}, \mathbf{y}) \leq 0$  для любого вектора  $\mathbf{y}$ . Такие траектории называются «*неустойчивыми*» (нейтральными).

При численном интегрировании нейтральной системы методом Рунге–Кутты порядка  $k \geq 2$  точность метода падает на порядок при  $t = O(1/\tau)$ .

Такие случаи возникают, когда имеется необходимость проводить численные расчеты при исследовании процессов с большим количеством колебаний, вращений и т. д. Важно отметить, что оценки погрешности численного решения получены с использованием более сильных, чем условие

липшиц-непрерывности, свойств правых частей рассматриваемых дифференциальных уравнений.

### VIII.5. Методы Адамса

Для решения ОДУ или систем ОДУ существуют *методы Адамса*. Эти методы являются *многошаговыми и одностадийными*. СтРОЯтся они следующим образом.

Пусть нам известно приближенное решение в некоторых узлах расчетной сетки:  $t_n, t_{n-1}, \dots, t_{n-m}$ . В окрестности этих узлов заменим  $f(t, u)$  интерполяционным полиномом, записанным в форме Ньютона:

$$\begin{aligned} f(t) = & f(t_n) + f(t_n, t_{n-1})(t - t_n) + \\ & + f(t_n, t_{n-1}, t_{n-2})(t - t_n)(t - t_{n-1}) + \\ & + f(t_n, t_{n-1}, t_{n-2}, t_{n-3})(t - t_n)(t - t_{n-1})(t - t_{n-2}) + \dots \end{aligned} \quad (5.1)$$

Чтобы вычислить решение в точке  $n+1$ , запишем его в интегральном виде:

$$u_{n+1} = u_n + \int_{t_n}^{t_{n+1}} f(t, u(t)) dt = \int_{t_n}^{t_{n+1}} f(t) dt. \quad (5.2)$$

Подставим в это интегральное представление интерполяционный полином:

$$\begin{aligned} u_{n+1} = & u_n + \tau_n f(t_n) + \frac{\tau_n^2}{2} f(t_n, t_{n-1}) + \\ & + \frac{\tau_n^2}{6} (2\tau_n + 3\tau_{n-1}) f(t_n, t_{n-1}, t_{n-2}) + \frac{\tau_n^2}{12} (2\tau_n^2 + 8\tau_n\tau_{n-1} + \\ & + 4\tau_n\tau_{n-2} + 6\tau_{n-1}^2 + 6\tau_n\tau_{n-2}) f(t_n, t_{n-1}, t_{n-2}, t_{n-3}). \end{aligned} \quad (5.3)$$

Здесь  $\tau_n = t_{n+1} - t_n$ ,  $f(t_n, t_{n-1})$ ,  $f(t_n, t_{n-1}, t_{n-2})$  и т.д. — разделенные разности.

Формула (5.3) является формулой четвертого порядка аппроксимации. Если опустить последнее слагаемое, то получим формулу третьего порядка, если опустить еще и предпоследнее, то — второго. Если же положить  $\tau_n = \text{const}$ , то формула значительно упростится:

$$u_{n+1} = u_n + \tau f_n + \frac{\tau^2}{2} \Delta_1 f_n + \frac{5}{12} \tau^3 \Delta_2 f_n + \frac{3}{8} \tau^4 \Delta_3 f_n,$$

где  $\Delta_k f_n$  —  $k$ -я конечная разность назад от правой части. На интервале  $[t_n, t_{n+1}]$  представление (5.1) является экстраполяцией, что обуславливает небольшую область устойчивости явных методов Адамса.

Явные методы Адамса, от первого до четвертого порядка аппроксимации, на равномерной сетке представляются в виде

$$y_{n+1} = y_n + \tau f_n,$$

$$y_{n+1} = y_n + \tau \left( \frac{3}{2} f_n - \frac{1}{2} f_{n-1} \right),$$

$$y_{n+1} = y_n + \tau \left( \frac{23}{12} f_n - \frac{16}{12} f_{n-1} + \frac{5}{12} f_{n-2} \right),$$

$$y_{n+1} = y_n + \tau \left( \frac{55}{24} f_n - \frac{59}{24} f_{n-1} + \frac{37}{24} f_{n-2} - \frac{9}{24} f_{n-3} \right).$$

В общем виде методы Адамса могут быть представлены следующим образом:

$$u_{n+1} = u_n + \tau \sum_{j=0}^{k-1} \eta_j \Delta^j f_j.$$

Если в интерполяционном многочлене использовать значение  $f(t_{n+1})$ , то аналогичным образом строятся *неявные* методы Адамса. Неявные методы Адамса требуют решения нелинейного уравнения для нахождения  $y_{n+1}$ .

## VIII.6. Экстраполяция Ричардсона

Пусть в точке  $t$  известно значение решения  $u(t)$ . Пусть методом Рунге–Кутты порядка аппроксимации  $p$  в результате выполнения численного интегрирования на двух шагах величины  $\tau$  найдено численное значение  $y$  в точке  $t + 2\tau$ , а в результате выполнения одного шага  $2\tau$  получено значение  $y_{2\tau}$  (в той же точке). Тогда выражение

$$y' = y_\tau + \frac{y_\tau - y_{2\tau}}{2^p - 1}$$

аппроксирует величину  $u(t + 2\tau)$  с порядком  $p+1$ . Другими словами, экстраполяция Ричардсона позволяет увеличивать на единицу точность метода.

### VIII.7. Задачи на доказательство

VIII.7.1. Задача Коши для системы ОДУ  $\dot{x} = -2y, \dot{y} = 2x, x(0) = 0, y(0) = 1$  решается с помощью метода Рунге–Кутты с таблицей Бутчера

0	
1	1
1/2	1/2

Доказать, что решение разностной задачи сходится к решению дифференциальной.

VIII.7.2. Задача Коши для системы ОДУ  $\dot{x} = -2y, \dot{y} = -2x, x(0) = 1, y(0) = 1$  решается с помощью метода Рунге–Кутты с таблицей Бутчера

0	
1	1/2
0	1

Используя определение сходимости доказать, что решение разностной задачи сходится к решению дифференциальной.

VIII.7.3. Нелинейное автономное дифференциальное уравнение решается с помощью явного метода Рунге–Кутты второго порядка аппроксимации с числом стадий, равным двум. Указать коэффициенты метода, имеющего минимальную погрешность аппроксимации для данного класса задач.

VIII.7.4. На основе определения сходимости доказать сходимость и определить ее порядок для явной и неявной схем Эйлера и метода трапеций применительно к решению дифференциальной задачи  $\dot{y} = ay, y(0) = 1$ .

VIII.7.5. Вывести условия правдака (до 3 включительно) для явного метода Рунге–Кутты с произвольным числом стадий.

VIII.7.6. Доказать Лемму 1.

VIII.7.7. (Л.А. Чудов) Система уравнений колебаний маятника

$$\frac{dx}{dt} = -y, \frac{dy}{dt} = x$$

с начальными условиями  $x(0) = 0, y(0) = 1$  решается с помощью неявного метода Эйлера. Доказать, что для решения данной задачи  $\|R_\tau\| < 1$ , что приводит к нефизической диссипации энергии. Какой закон сохранения существует для дифференциальной задачи? Насколько на каждом шаге по времени нарушается закон сохранения для разностной задачи?

**VIII.7.8.** Задача Коши для системы ОДУ  $\dot{x} = -10y$ ,  $\dot{y} = 10x$ ,  $x(0) = 0$ ,  $y(0) = 1$  решается с помощью метода трапеций

$$\frac{\mathbf{y}^{n+1} - \mathbf{y}^n}{\tau} = \frac{\mathbf{f}^{n+1} + \mathbf{f}^n}{2}.$$

Доказать, что метод консервативен (в разностной задаче выполняется тот же закон сохранения, что в дифференциальной задаче).

**VIII.7.9.** Задача Коши для системы ОДУ  $\dot{x} = -10y$ ,  $\dot{y} = 10x$ ,  $x(0) = 0$ ,  $y(0) = 1$  решается с помощью методов Рунге–Кутты. Доказать, что в разностной задаче нет закона сохранения, аналогичного закону сохранения энергии для решения дифференциальной задачи колебаний маятника. Таблицы Бутчера методов следующие:

0		
1	1	
	1/2	1/2

0		
1	1/2	
	0	1

**VIII.7.10.** Доказать, что «классический» четырехстадийный метод Рунге–Кутты представляет собой обобщение формулы Симпсона численного интегрирования для случая ОДУ.

## VIII.8. Задачи с решениями

**VIII.8.1.** Задача Коши для системы ОДУ  $\dot{x} = -2y$ ,  $\dot{y} = 2x$ ,  $x(0) = 0$ ,  $y(0) = 1$  решается с помощью метода трапеций:  $\frac{\mathbf{y}^{n+1} - \mathbf{y}^n}{\tau} = \frac{\mathbf{f}^{n+1} + \mathbf{f}^n}{2}$ . Доказать, что решение разностной задачи сходится к решению дифференциальной.

Решение: Решение дифференциальной задачи легко ищется.

Рассмотрим решение разностной задачи.

Имеем:  $\frac{x^{n+1} - x^n}{\tau} = -y^{n+1} - y^n$ ,  $\frac{y^{n+1} - y^n}{\tau} = x^{n+1} + x^n$ , откуда

$$\begin{pmatrix} x \\ y \end{pmatrix}^{n+1} = \frac{1}{1+\tau^2} \begin{pmatrix} 1-\tau^2 & -2\tau \\ 2\tau & 1-\tau^2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}^n = \begin{pmatrix} \cos \xi & -\sin \xi \\ \sin \xi & \cos \xi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}^n,$$

где  $\cos \xi = \frac{1 - \tau^2}{1 + \tau^2}$ .

$$\text{Тогда } \begin{pmatrix} x \\ y \end{pmatrix}^n = \begin{pmatrix} \cos \xi & -\sin \xi \\ \sin \xi & \cos \xi \end{pmatrix}^{T/\tau} \begin{pmatrix} x \\ y \end{pmatrix}^0 = \begin{pmatrix} \cos \frac{T}{\tau} \xi & -\sin \frac{T}{\tau} \xi \\ \sin \frac{T}{\tau} \xi & \cos \frac{T}{\tau} \xi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}^0.$$

Теперь необходимо вычислить предел

$$\begin{aligned} & \lim_{\tau \rightarrow 0} \begin{pmatrix} \cos \frac{T}{\tau} \xi & -\sin \frac{T}{\tau} \xi \\ \sin \frac{T}{\tau} \xi & \cos \frac{T}{\tau} \xi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}^0 = \\ & = \lim_{\tau \rightarrow 0} \begin{pmatrix} \cos \frac{T}{\tau} \arcsin \frac{2\tau}{1+\tau^2} & -\sin \frac{T}{\tau} \arcsin \frac{2\tau}{1+\tau^2} \\ \sin \frac{T}{\tau} \arcsin \frac{2\tau}{1+\tau^2} & \cos \frac{T}{\tau} \arcsin \frac{2\tau}{1+\tau^2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}^0 = \begin{pmatrix} \cos 2T & -\sin 2T \\ \sin 2T & \cos 2T \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}^0. \end{aligned}$$

Отсюда следует сходимость.

**VIII.8.2.** Получить все явные методы Рунге–Кутты с числом стадий от 1 до 3.

**Решение.** Получим простейшие методы Рунге–Кутты. Для этого рассмотрим погрешность

$$\xi(\tau) = u(t + \tau) - \left[ u(t) + \sum_{j=0}^r b_j k_j \right]$$

и представим ее в виде разложения в ряд Маклорена:

$$\xi(\tau) = \sum_{i=0}^p \frac{\xi(0)}{i!} \tau^i + \frac{\xi^{(p+1)}(\theta\tau)}{(p+1)!} \tau^{p+1},$$

где  $\frac{\xi^{(p+1)}(\theta\tau)}{(p+1)!} \tau^{p+1}$  — остаточный член ряда;  $0 < \theta < 1$ . Будем полагать (что

можно сделать соответствующим выбором коэффициентов)

$$\xi(0) = \xi'(0) = \dots = \xi^{(p)}(0) = 0.$$

В таком случае разложение для  $\xi(\tau)$  имеет более простой вид:

$$\xi(\tau) = \frac{\xi^{(p+1)}(\theta\tau)}{(p+1)!} \tau^{p+1},$$

где  $p$  — порядок аппроксимации метода.

**1.** Пусть  $p = 1, s = 1$ . Тогда  $\xi(\tau) = u(t + \tau) - u(t) - \tau b_1 f(t, u)$ ,

$$\xi(0) = 0,$$

$$\xi'(0) \equiv [u'(t + \tau) - b_1 f(t, u)]|_{t=0} = f(t, u)(1 - b_1),$$

отсюда

$$\xi''(0) = u''(t + \tau).$$

Видно, что условие  $\xi'(0) = 0$  выполняется лишь при  $b_1 = 1$ , что соответствует методу Эйлера, при этом

$$\frac{u(t + \tau) - u(t)}{\tau} - f(t, u) = \frac{\xi''(t + \theta\tau)}{2} \tau = R_\tau,$$

здесь  $R_\tau$  — невязка, имеющая первый порядок малости по  $\tau$ .

**2.** Более сложный случай:  $p = 2, s = 2$ . Тогда

$$\xi(\tau) = u(t + \tau) - u(t) - \tau b_1 f(t, u) - \tau b_2 f(t + c_2 \tau, u + a_{21} \tau f(t, u)).$$

Вводя обозначения  $\tilde{t} = t + c_2 \tau$ ,  $\tilde{u} = u + a_{21} \tau f(t, u)$ , получим следующие выражения для производных погрешности  $\xi$  по аргументу  $\tau$ :

$$\xi'(\tau) = u'(t + \tau) - b_1 f(t, u) - b_2 f(\tilde{t}, \tilde{u}) -$$

$$- \tau b_2 [c_2 f'_t(\tilde{t}, \tilde{u}) + a_{21} f'_u(\tilde{t}, \tilde{u}) f(t, u)],$$

$$\xi''(\tau) = u''(t + \tau) - 2b_2 [c_2 f'_t(\tilde{t}, \tilde{u}) + a_{21} f'_u(\tilde{t}, \tilde{u}) f(t, u)] -$$

$$- \tau b_2 [c_2^2 f''_{tt}(\tilde{t}, \tilde{u}) + 2c_2 a_{21} f''_{tu}(\tilde{t}, \tilde{u}) f(t, u) +$$

$$+ a_{21}^2 f''_{uu}(\tilde{t}, \tilde{u}) f^2(t, u)],$$

$$\xi'''(\tau) = u'''(t + \tau) - 3b_2 [c_2^2 f''_{tt}(\tilde{t}, \tilde{u}) +$$

$$+ 2c_2 a_{21} f''_{tu}(\tilde{t}, \tilde{u}) f(t, u) + a_{21}^2 f''_{uu}(\tilde{t}, \tilde{u}) f^2(t, u)] + o(\tau).$$

Подставим в эти выражения следующие следствия исходного дифференциального уравнения:

$$u' = f, \quad u'' = f'_t + f'_u f, \quad u''' = f''_{tt} + 2f''_{tu} f + f''_{uu} f^2 + f'_u u''.$$

С учетом этих следствий получим

$$\begin{aligned}\xi(0) &= 0, & \xi'(0) &= (1 - b_1 - b_2) f(t, u), \\ \xi''(0) &= (1 - 2b_2 c_2) f'_t(t, u) + (1 - 2b_2 a_{21}) f'_u(t, u) f(t, u), \\ \xi'''(0) &= (1 - 3b_2 c_2^2) f''_{tt}(t, u) + \\ &\quad + (2 - 6b_2 c_2 b_{21}) f''_{tu}(t, u) f(t, u) + \\ &\quad + (1 - 3b_2 a_{21}^2) f''_{uu}(t, u) f^2(t, u) + f'_u(t, u) u''(t).\end{aligned}$$

Второе из полученных соотношений выполняется при  $b_1 + b_2 = 1$ , третье — при  $1 - 2b_2 c_2 = 0$ ,  $1 - 2b_2 a_{21} = 0$ .

Таким образом, мы имеем три алгебраических уравнения и четыре параметра. Они определяют однопараметрическое семейство схем. Задавая один из параметров, можно получать различные методы Рунге–Кутты с аппроксимацией 2-го порядка. При формально одинаковом порядке аппроксимации они будут обладать различными свойствами (реальной погрешностью).

Так, при  $b_1 = 1/2$ , имеем:  $b_2 = 1/2$ ,  $c_2 = 1$ ,  $a_{21} = 1$ ; метод будет выглядеть следующим образом:

$$u_{n+1} = u_n + \frac{\tau}{2} [f(t_n, u_n) + f(t_{n+1}, \tilde{u}_{n+1})].$$

Положив  $b_1 = 0$ , имеем:  $b_2 = 1$ ,  $c_2 = 1/2$ ,  $a_{21} = 1/2$ ; соответствующий метод:

$$\begin{aligned}u_{n+1/2} &= u_n + \frac{\tau}{2} f(t_n, u_n), \\ u_{n+1} &= u_n + f\left(t_n + \frac{\tau}{2}, u_{n+1/2}\right).\end{aligned}$$

**3.** При  $p = 3$ ,  $s = 3$  аналогично получаем систему уравнений:

$$\begin{aligned}c_2 &= a_{21}, & c_3 &= a_{31} + a_{32}, & b_3 a_{32} c_2 &= 1/6, \\ b_2 c_2 + b_3 c_3 &= 1/2, & b_1 + b_2 + b_3 &= 1, & b_2 c_2^2 + b_3 c_3^2 &= 1/3,\end{aligned}$$

которая также имеет бесконечное множество решений. Все условия порядка приводят к четырем независимым уравнениям для шести неизвестных:

$$\begin{aligned}b_3 a_{32} a_{21} &= 1/6, & b_2 a_{21} + b_3 (a_{31} + a_{32}) &= 1/2, & b_1 + b_2 + b_3 &= 1, \\ b_2 a_{21}^2 + b_3 (a_{31} + a_{32})^2 &= 1/3.\end{aligned}$$

Это приводит к двухпараметрическому семейству:  $c_2 = u$ ,  $c_3 = v$ ,  
 $b_2 = \frac{2-3v}{6u(u-v)}$ ,  $b_3 = \frac{2-3u}{6v(v-u)}$ ,  $a_{32} = \frac{v(v-u)}{u(2-3u)}$ ,  $b_1 = 1 - b_2 - b_3$ ,  $a_{21} = u$ ,  
 $a_{31} = v - a_{32}$ .

Вспомогательные векторы  $k_i$  одного из возможных методов будут

$$k_1 = \tau f(t_n, u_n), \quad k_2 = \tau f\left(t_n + \frac{\tau}{2}, u_n + \frac{k_1}{2}\right),$$

$$k_3 = \tau f(t_n + \tau, u_n - k_1 + 2k_2), \quad u_{n+1} = u_n + \frac{k_1 + 4k_2 + k_3}{6}.$$

В случае  $p = 4$ ,  $s = 4$  имеем двухпараметрическое семейство методов Рунге–Кутты. Соотношения порядка (условия на коэффициенты, когда метод имеет порядок аппроксимации 4) получаются аналогичным образом, предлагаем читателям вывести систему соотношений порядка самостоятельно. Из семейства явных методов четвертого порядка наиболее известен следующий «классический» метод:

$$k_1 = \tau f(t_n, u_n), \quad k_2 = \tau f\left(t_n + \frac{\tau}{2}, u_n + \frac{k_1}{2}\right),$$

$$k_3 = \tau f\left(t_n + \frac{\tau}{2}, u_n + \frac{k_2}{2}\right), \quad k_4 = \tau f(t_n + \tau, u_n + k_3),$$

$$u_{n+1} = u_n + \frac{k_1 + 2k_2 + 2k_3 + k_4}{6}.$$

### VIII.9. Теоретические задачи

**VIII.9.1.** Исследовать на устойчивость метод трапеций:  
 $\frac{y_{n+1} - y_n}{\tau} = \frac{1}{2}(f(t_n, y_n) + f(t_{n+1}, y_{n+1}))$ ,  $y_0 = u_0$ . Каковы условия строгой устойчивости этого метода?

**VIII.9.2.** Исследовать на устойчивость метод Эйлера с пересчетом:  
 $v = y_n + \frac{\tau}{2} f(t_n, y_n)$ ,  $\frac{y_{n+1} - y_n}{\tau} = f\left(t_n + \frac{\tau}{2}, v\right)$ ,  $y_0 = u_0$ . Каковы условия строгой устойчивости этого метода?

**VIII.9.3.** Построить сеточный аналог общего решения обыкновенного дифференциального уравнения:

$$\frac{d^2u}{dx^2} + xu = x,$$

на равномерной сетке  $D_h = \{x_m: x_m = hm; m = 0, \pm 1, \dots, \pm\infty\}$ .

**VIII.9.4.** Построить семейство явных трехстадийных методов Рунге–Кутты, основанных на квадратурной формуле Гаусса

$$\int_{-1}^1 f(x) dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

Исследовать его на аппроксимацию для случая автономного нелинейного уравнения.

**VIII.9.5.** Для решения системы ОДУ  $\dot{y} = f(t, y)$  используется метод

0		$\sqrt{2}$		$\sqrt{2}$		$\frac{2\sqrt{2}-1}{2\sqrt{2}}$		$\frac{1}{2\sqrt{2}}$	
---	--	------------	--	------------	--	---------------------------------	--	-----------------------	--

Исследовать его на аппроксимацию. Если возможно повышение порядка аппроксимации, внесите изменения в таблицу Бутчера.

**VIII.9.6.** Для решения системы ОДУ  $\dot{y} = f(t, y)$  используется метод

0		$3/2$		$3/2$		$3/4$		$1/4$	
---	--	-------	--	-------	--	-------	--	-------	--

Исследовать его на аппроксимацию. Если возможно повышение порядка аппроксимации, внесите изменения в таблицу Бутчера.

**VIII.9.7.** Приближенное решение задачи Коши:

$$\frac{dx}{dt} = ax; \quad x(0) = X_0; \quad 0 \leq t \leq T$$

вычисляется по разностной схеме:

$$1) \frac{x_{n+1} + 4x_n - 5x_{n-1}}{2\tau} = a(2x_n + x_{n-1}); \quad x_0 = X_0,$$

$$2) \frac{x_{n+1} - \frac{4}{3}x_n + \frac{1}{3}x_{n-1}}{2\tau} = \frac{1}{3}ax_{n+1}; \quad x_0 = X_0,$$

$$3) \frac{x_{n+1} - x_n}{\tau} = \frac{1}{2}a(3x_n - x_{n-1}); \quad x_0 = X_0,$$

$$4) \frac{x_{n+1} - x_{n-1}}{\tau} = \frac{1}{3}a(3x_{n+1} + 4x_n + x_{n-1}); \quad x_0 = X_0.$$

Способ задания дополнительного краевого условия  $x_1$  предложить самостоятельно.

Найти порядок аппроксимации разностной схемы. Исследовать влияние способа задания  $x_1$  на порядок аппроксимации. Исследовать на устойчивость разностную схему. Найти точное решение разностной задачи. Исследовать его сходимость к точному решению дифференциальной задачи.

### VIII.9.8. Для решения задачи Коши

$$y'' + 6y' + 5y = 0, \quad y(0) = 0, \quad y'_x(0) = 2.$$

на  $(0, 1)$  предложена разностная схема

$$(y_{l+1} - 2y_l + y_{l-1})/h^2 + 6(y_{l+1} - y_{l-1})/2h + 5y_l = 0, \quad l = 1 \dots L-1,$$

$$y_0 = 0; \quad y_1 = 2h - 6h^2.$$

Исследовать разностную задачу на аппроксимацию и определить порядок сходимости ее решения к решению дифференциальной задачи.

VIII.9.9. Для решения задачи Коши на отрезке  $[0, T]$  введена равномерная сетка с шагом  $\tau$ . Рассматривается дифференциальное уравнение, которому в соответствие ставится численный метод:

a)  $y' - y = 0, \quad 0 \leq x \leq 1, \quad y(0) = 1, \quad (y_{n+1} - y_n)/h - (3y_{n+1} - y_n)/2 = 0, \quad y_0 = 1.$

б)  $y' - 2y = 0, \quad 0 \leq x \leq 1, \quad y(0) = 1, \quad (y_{n+1} - y_n)/h + y_{n+1} - 3y_n = 0, \quad y_0 = 1.$

в)  $y' - 7y = 0, \quad 0 \leq x \leq 1, \quad y(0) = 1$ , метод задан таблицей Бутчера:

1/2	0	
	1/2	0
	1/2	1/2

г)  $y' - 3y = 0, \quad 0 \leq x \leq 1, \quad y(0) = 1$ , метод задан таблицей Бутчера:

-1/2	0	
	-1/2	0
	2	-1

д)  $y' - 3y = 0, \quad 0 \leq x \leq 1, \quad y(0) = 1,$

$$\begin{array}{c|cc} & 0 \\ \hline 1/4 & 1/4 & 0 \\ & -1 & 2 \end{array}$$

Выписать общее решение разностного уравнения и исследовать сходимость численного решения к решению дифференциальной задачи на основе определения сходимости.

Пусть по данной схеме проводятся вычисления на компьютере, где для хранения мантиссы отводится 10 бит. Оценить максимальную ошибку округления для данной задачи.

**VIII.9.10.** Для решения задачи Коши использована разностная схема на равномерной сетке:

а)  $y' + \sin y = 0, \quad y(0) = 1, \quad t \in [0, 20\pi],$

$$\frac{y^{n+1} - y^n}{\tau} + \sin(y^n + 0,5\tau \sin y^n) \cos(0,5\tau \sin y^n) = 0, \quad y^0 = 1.$$

б)  $y' - \cos y = 0, \quad y(0) = 0, \quad t \in [0, 25\pi],$

$$\frac{y^{n+1} - y^n}{\tau} - \cos(y^n) \cos(0,5\tau \cos y^n) + \sin(y^n) \sin(0,5\tau \cos y^n) = 0, \quad y^0 = 0$$

в)  $y' = y^2(2.5 - t), \quad y(0) = 1/3, \quad t \in [0, 2],$

$$\frac{y^{n+1} - y^n}{\tau} = (y^{n+1})^2 (2.5 - (n+1)\tau), \quad y^0 = \frac{1}{3}.$$

г)  $y' = -2te^{-y}, \quad y(0) = \ln 4, \quad t \in [0, 2],$

$$\frac{y^{n+1} - y^n}{\tau} = -t_n e^{-y_n} - t_{n+1} e^{-y_{n+1}}, \quad y^0 = \ln 4.$$

д)  $y' = 4y(1 - y), \quad y(0) = \alpha \quad (0 < \alpha < 1), \quad t \in [0, 100],$

$$\frac{y^{n+1} - y^n}{\tau} = 2(y^{n+1} + y^n)(1 - y^n), \quad y^0 = \alpha \quad (0 < \alpha < 1).$$

$$и) \frac{z^{n+1} - z^n}{\tau} = (z^{n+1} + z^n)(2 - z^n - z^{n+1}), \quad z^0 = \alpha \quad (0 < \alpha < 1).$$

С каким порядком разностная задача аппроксимирует дифференциальную на ее решении?

**VIII.9.11.** Простейший метод Рунге–Кутты 2-го порядка можно получить с помощью применения метода Эйлера первого порядка с последующей экспоненциальной интерполяцией по Ричардсону. Записать таблицу Бутчера метода.

**VIII.9.12.** С помощью ЯМРК порядка  $p$  решается задача Коши:

$$\dot{x} = \sin y, \quad \dot{y} = -x \quad x(0) = 0, y(0) = 1 \text{ на отрезке } t \in [0, 1000].$$

Исследовать метод на устойчивость на траектории, получить оценку константы в условии устойчивости.

**VIII.9.13.** Выяснить порядок аппроксимации метода

0			
1	1		
1/2	0	1/2	
	1/6	1/6	4/6

При необходимости исправить таблицу Бутчера так, чтобы метод имел наивысший порядок для данного числа стадий.

**VIII.9.14.** Система ОДУ решается методом Рунге–Кутты:

0			
1/3	1/3		
2/3	1/3	1/3	
	1/4	0	3/4

При необходимости исправить таблицу Бутчера так, чтобы метод имел наивысший порядок для данного числа стадий.

**VIII.9.15.** Построить все трехстадийные методы с порядком аппроксимации 2, имеющие вид

0			
$c_2$	$c_2$		
$c_3$	0	$c_3$	
	0	0	1

**VIII.9.16.** Аппроксимирует ли разностная схема

$$\frac{3y_m - 4y_{m-1} + y_{m-2}}{2h} + y_m = x_m^2 + 2x_m,$$

$$m = 2, \dots, M, \quad hM = 1, \quad y_0 = 0, \quad y_1 = h$$

дифференциальную задачу Коши

$$\frac{dy}{dx} + y = x^2 + 2x, \quad x \in (0, 1), \quad y(0) = 0$$

со вторым порядком по  $h$ ? Если нет, то видоизменить разностную схему так, чтобы она имела второй порядок аппроксимации.

**VIII.9.17.** Построить четырехстадийный явный метод Рунге–Кутты, основанный на квадратурной формуле «правило 3/8».

## VIII.10. Практические задачи

**VIII.10.1.** Для решения задач Коши

$$\text{a) } y' = x + \cos y; \quad y(1) = 30; \quad 1 \leq x \leq 2,$$

$$\text{б) } y' = x^2 + y^2; \quad y(2) = 1; \quad 1 \leq x \leq 2,$$

используется метод Эйлера с центральной точкой. Провести вычисления с шагом  $h = 0,01$ . Использовать также метод Эйлера с пересчетом. Сравнить и объяснить результаты.

**VIII.10.2.** Получить численно решение систем ОДУ с использованием методов Рунге–Кутты порядка 1, 2, 3, 4. Исследовать сеточную сходимость методов при измельчении сетки. Объяснить полученные результаты.

$$\text{а) } \frac{d v}{d t} = v + w,$$

$$\frac{d w}{d t} = v^2 - w^2,$$

$$v(0) = 1, \quad w(0) = 2, \quad 0 \leq t \leq 1;$$

$$\text{б) } \frac{d v}{d t} = v \cdot w,$$

$$\frac{d w}{d t} = v + w,$$

$$v(1) = 2, \quad w(1) = 3, \quad 1 \leq t \leq 2.$$

**VIII.10.3.** Приближенно решить задачу Коши:

$$\frac{d^2 y}{d t^2} = y \sin t,$$

$$y(0) = 0; \quad y'(0) = 1; \quad 0 \leq t \leq 1.$$

а) Описать алгоритм, основанный на переходе к системе двух уравнений первого порядка с последующим решением этой системы.

б) Описать алгоритм, основанный на замене уравнения  $y'' = y \sin x$  разностным уравнением второго порядка.

в) Решить задачу любым из этих методов.

**VIII.10.4.** Известно, что система нелинейных ОДУ второго порядка

$$\frac{d^2 x}{dt^2} = x(y^2 - 1),$$

$$\frac{d^2 y}{dt^2} = y(x^2 - 1),$$

с начальными условиями  $x(0) = \alpha$ ,  $y(0) = 1$ ,  $x'(0) = 0$ ,  $y'(0) = 0$  описывает некоторые замкнутые траектории в плоскости  $(x, y)$ . Получить численно эти траектории с использованием методов Рунге–Кутты порядка 1, 2, 3, 4. Считать, что  $0 < t < 20$ . Исследовать сеточную сходимость методов при измельчении сетки. Объяснить полученные результаты.

\*Методы Штёрмера (см., например, [48]) предназначены специально для решения систем ОДУ второго порядка. Реализовать методы Штёрмера порядка 2 и 4. Провести вычисления, сравнить результаты с полученными с помощью методов Рунге–Кутты соответствующего порядка.

**VIII.10.5.** Функция  $\beta = \beta(\alpha)$ , задана как результат решения задач Коши для нелинейного обыкновенного дифференциального уравнения второго порядка

$$\frac{d^2 x}{dt^2} + x^3 = 1 - t^2, \quad 0 \leq t \leq 1.$$

с начальными условиями  $x(0) = \alpha$ ,  $x'(0) = 0$ ,  $-1 < \alpha < 0$ , при этом функция определена как  $\beta = x(t, \alpha)|_{t=1}$ .

Качественно исследовать свойства функции. Для этого воспользоваться методами фазовой плоскости. На основе численных расчетов построить табличное представление функции. Непрерывна ли функция, таблицу которой Вы получили? Предложить аппроксимацию функции  $\beta(\alpha)$ , принимая во внимание свойства этой функции.

## **VIII.11. Устойчивость методов Рунге–Кутты на различных типах траекторий и практические задачи**

Решение задач из данного раздела предполагает самостоятельную реализацию численных методов на компьютере с использованием любого языка программирования.

**VIII.11.1.** Получите численное решение системы двух ОДУ:

$$u' = A + u^2 v - (B+1) u, \quad u(0) = 1,$$

$$v' = B u - u^2 v, \quad v(0) = 1,$$

$$A = 1, \quad B \in [1, 5]$$

явными методами Рунге–Кутты первого и четвертого порядка. Изучите фазовые портреты. Удалось ли Вам получить предельные циклы и бифуркацию Хопфа (при которой предельный цикл вырождается в точку; при этом  $B \rightarrow A \cdot (A+1)$ )?

Эта система — модель Лефевра–Пригожина «брюсселятор».

**VIII.11.2.** Изучите поведение численного решения ОДУ второго порядка (уравнения Ван-дер-Поля):

$$y'' + e(y^2 - 1)y' + y = 0, \quad e > 0,$$

представленного в виде системы двух ОДУ первого порядка

$$x' = z, \quad z' = e(1-x^2)z - x,$$

или в представлении Льенара:

$$z' = -y, \quad y' = z - e\left(\frac{y^3}{3} - y\right); \quad e > 0,$$

$$x(0) = 2, \quad z(0) = 0, \quad 0 < t \leq 100$$

в зависимости от изменения параметра  $e$  ( $0,01 < e \leq 1$ ). Использовать методы Рунге–Кутты порядка 1, 2, 3, 4 и методы Адамса порядка 2, 3, 4.

**VIII.11.3.** Исследуйте поведение фазовых траекторий для системы ОДУ

$$x' = y, \quad y' = x^2 - 1$$

вблизи особых точек  $(1, 0)$  и  $(-1, 0)$  с помощью двух методов Рунге–Кутты (первого и четвертого порядка аппроксимации). Объясните их поведение. Значения  $x(0)$  и  $y(0)$  изменяйте самостоятельно.

**VIII.11.4.** Получите траекторию движения спутника вокруг планеты, пройдя численное решение задачи двух тел

$$x' = z,$$

$$y' = u,$$

$$z' = -\frac{x}{(x^2 + y^2)^{3/2}},$$

$$u' = -\frac{y}{(x^2 + y^2)^{3/2}},$$

$$x(0) = 0,5; \quad y(0) = z(0) = 0, \quad u(0) = \sqrt{3} \approx 1.73,$$

при  $0 \leq t \leq 20$  методами Рунге–Кутты первого, второго, третьего и четвертого порядков аппроксимации. Исследуйте зависимость численного решения от шага интегрирования.

**VIII.11.5.** Методами Рунге–Кутты разных порядков аппроксимации численно решить систему Лоренца:

$$x' = -\sigma(x - y),$$

$$y' = -xz + rx - y,$$

$$z' = xy - bz,$$

с начальными условиями  $x(0) = y(0) = z(0) = 1$  при  $b = 8/3$ ,  $\sigma = 10$ ,  $r = 28$ .

Считаем, что  $0 \leq t \leq 50$ . Объяснить полученные результаты.

## VIII.12. Библиографический комментарий

Теорема 1 была установлена, доказана и опубликована в 1956 году независимо В.С. Рябеньким и А. Ф. Филипповым в СССР [41], П. Лаксом и Р. Рихтмайером в США [42]. Отметим, что ранее близкие теоремы были установлены Л.В. Канторовичем, некоторые из них нашли отражение в [49]. Об исследовании на устойчивость линейных разностных уравнений с использованием разрешающего оператора подробнее см. в [43].

Простейшие численные методы решения задач Коши для систем ОДУ подробно рассмотрены в большом количестве учебников и специальных изданий. Самое подробное изложение различных численных методов можно найти в [44]. Теоремы об устойчивости явных методов Рунге–Кутты впервые были опубликованы в первом издании книги [3] в 1984 году (Федоренко Р.П. . Введение в вычислительную физику. — Москва : Изд-во МФТИ, 1994. — 528 с.).

## Приложение. Некоторые системы ортогональных многочленов

### Многочлены Чебышёва

Одним из самых известных семейств многочленов являются многочлены Чебышёва (первого рода), обозначаемые в литературе  $T_n(x)$ . Многочлены Чебышёва удовлетворяют рекуррентному соотношению

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad (1)$$

которое позволяет, зная первые два многочлена, определить все остальные:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1,$$

$$T_3(x) = 4x^3 - 3x, \quad T_4(x) = 8x^4 - 8x^2 + 1,$$

$$T_5(x) = 16x^5 - 20x^3 + 5x, \dots$$

Для многочленов Чебышёва известно много способов представления, одним из самых удобных оказывается тригонометрическое:

$$T_n(x) = \cos(n \arccos x).$$

Действительно, при любом угле  $\Theta$  справедливо тождество

$$\begin{aligned} \cos((n+1)\Theta) &= \cos(n\Theta + \Theta) = \cos \Theta \cos(n\Theta) - \sin \Theta \sin(n\Theta) = \\ &= 2 \cos \Theta \cos(n\Theta) - (\cos \Theta \cos(n\Theta) + \sin \Theta \sin(n\Theta)) = \\ &= 2 \cos \Theta \cos(n\Theta) - \cos((n-1)\Theta). \end{aligned}$$

Полагая в этом тождестве  $\Theta = \arccos x$ , получим

$$\cos((n+1) \arccos x) = 2x \cos(n \arccos x) - \cos((n-1) \arccos x). \quad (2)$$

Сравнивая (1) и (2) убеждаемся, что функция  $\cos(n \arccos x)$  удовлетворяет тому же разностному уравнению, что и  $T_n(x)$  при начальных условиях

$$\cos(0 \cdot \arccos x) = 1 = T_0(x), \quad \cos(1 \cdot \arccos x) = x = T_1(x),$$

это означает, что две функции совпадают при любом  $n$ .

Рекуррентное соотношение (1) является разностным уравнением по  $n$  при  $x$  — параметре с характеристическим уравнением

$$\lambda^2 - 2x\lambda + 1 = 0, \quad \lambda_{1,2} = x \pm \sqrt{x^2 - 1}.$$

При  $x \neq \pm 1$  корни простые, следовательно,  $T_n(x) = c_1 \lambda_1^n + c_2 \lambda_2^n$ . Из начальных условий  $T_0(x) = 1$ ,  $T_1(x) = x$  будем иметь  $c_1 = c_2 = 1/2$ . Тогда

$$T_n(x) = \left( (x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n \right) / 2.$$

При комплексных значениях  $x = z$  функция  $\sqrt{z^2 - 1}$  не является однозначной. Под  $\sqrt{z^2 - 1}$  будем понимать ветвь, которая при действительных  $z > 1$  принимает положительные значения и определена в комплексной плоскости с разрезом  $[-1, 1]$ .

Многочлены Чебышёва можно представить в виде конечного ряда

$$T_n(x) = \frac{n}{2} \sum_{k=0}^{[n/2]} \frac{(-1)^k (n-k-1)!}{k!(n-2k)!} (2x)^{n-2k}.$$

**Нули многочленов Чебышёва.** Из уравнения

$$\cos(n \arccos x) = 0$$

получаем, что корни многочлена Чебышёва степени  $n$  суть

$$x_m = \cos\left(\frac{\pi(2m+1)}{2n}\right), \quad m = 0, 1, \dots, n-1.$$

**Точками экстремума**  $T_n(x)$  будут точки, в которых  $|T_n(x)| = 1$ , т.е.

$$x_{(m)} = \cos\left(\frac{\pi m}{n}\right), \quad m = 0, 1, \dots, n, \quad T_n(x_{(m)}) = \cos \pi m = (-1)^m.$$

Из (1) следует, что коэффициент при старшей степени у многочлена  $T_n(x)$  растет как  $2^{n-1}$ . Приведенный многочлен  $\bar{T}_n(x) = 2^{1-n} T_n(x)$  с коэффициентом при старшей степени, равным единице, называют *многочленом, наименее уклоняющимся от нуля в силу* следующей леммы:

**Лемма.** *Если  $P_n(x)$  — многочлен степени  $n$  со старшим коэффициентом, равным единице, то справедливо неравенство:*

$$\max_{[-1,1]} |P_n(x)| \geq \max_{[-1,1]} |\bar{T}_n(x)| = 2^{1-n}.$$

Иногда рассматривают многочлены Чебышёва второго рода  $U_n(x)$ .

$$U_n(x) = \frac{1}{n+1} \frac{d}{dx} T_{n+1}(x),$$

или, в тригонометрическом виде:  $U_n(x) = \frac{\sin((n+1)\Theta)}{\sin \Theta}$ ,  $\cos \Theta = x$ .

Нули многочленов Чебышёва второго рода совпадают с точками экстремума многочленов Чебышёва первого рода степени на единицу выше. Многочлен Чебышёва второго рода характеризуется как многочлен с коэффициентом при старшей степени  $2^n$ , интеграл от которого по отрезку  $[-1, 1]$  принимает наименьшее возможное значение.

Рекуррентное соотношение для полиномов Чебышёва 2 рода такое же, как и для многочленов первого рода:

$$U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x),$$

$$U_0(x) = 1, \quad U_1(x) = 2x, \quad U_2(x) = 4x^2 - 1,$$

$$U_3(x) = 8x^3 - 4x, \dots$$

$$U_n(x) = \left( \left( x + \sqrt{x^2 - 1} \right)^{n+1} - \left( x - \sqrt{x^2 - 1} \right)^{n+1} \right) / \left( 2\sqrt{x^2 - 1} \right),$$

$$U_n(x) = \sum_{k=0}^{[n/2]} \frac{(-1)^k (n-k)!}{k!(n-2k)!} (2x)^{n-2k}.$$

Связь многочленов Чебышёва первого и второго рода может быть представлена также в виде:

$$(x^2 - 1)U_{n-1}(x) = xT_n(x) - T_{n-1}(x).$$

Для многочленов Чебышёва первого и второго рода имеет место ортогональность с весом:

$$\int_{-1}^1 T_n(x)T_m(x) \frac{dx}{\sqrt{1-x^2}} = \begin{cases} 0, & n \neq m, \\ \pi/2, & n = m \neq 0, \\ \pi, & n = m = 0. \end{cases}$$

$$\int_{-1}^1 U_n(x)U_m(x) \sqrt{1-x^2} dx = \begin{cases} 0, & n \neq m, \\ \pi/2, & n = m \neq 0. \end{cases}$$

Это свойство роднит многочлены Чебышёва с другими системами ортогональных многочленов (см. ниже).

Поведение многочленов Чебышёва вне отрезка ортогональности не менее интересно и тоже экстремально. Среди всех многочленов данной степени, значения которых на отрезке  $[-1, 1]$  не превосходят по модулю единицы, многочлены Чебышёва имеют наибольшее значение в любой точке вне этого отрезка, т.е. функции  $T_n(x)$  очень сильно растут вне интервала ортогональности. Для примера рассмотрим значения  $T_n(1+\varepsilon)$  при нескольких значениях  $n$  и  $\varepsilon$ :

$n \backslash \varepsilon$	$10^{-4}$	$10^{-3}$	$10^{-2}$
10	1.0	1.1	2.2
100	2.2	44	$6.9 \cdot 10^5$
200	8.5	$3.8 \cdot 10^3$	$9.4 \cdot 10^{11}$
1000	$6.9 \cdot 10^5$	$1.3 \cdot 10^{19}$	$1.2 \cdot 10^{61}$

Для практических приложений очень важным свойством многочленов Чебышёва является их ортогональность в смысле скалярного произведения, определенного для векторов, заданных как значения этих многочленов в выделенном наборе точек. Оказывается, что на множестве нулей многочлена Чебышёва степени  $n+1$  все многочлены меньшей степени образуют ортогональную систему. Это означает, что если  $\{x_m\}$  — множество нулей многочлена степени  $n+1$ , т.е.  $T_{n+1}(x_m) = 0$ ,  $m = 0, 1, \dots, n$ , то

$$\frac{2}{n+1} \sum_{m=0}^n T_k(x_m) T_l(x_m) = \delta_{kl}, \quad k, l = 0, 1, \dots, n$$

## Некоторые другие системы ортогональных многочленов

**I. Многочлены Лежандра**  $q_n(x)$  — наиболее употребительные из классических ортогональных многочленов и единственны, для которых условие их ортогональности на отрезке  $[-1, 1]$  выполняется «в чистом виде» без введения весовой функции в скалярное произведение, а именно:

$$\int_{-1}^1 q_n(x) q_m(x) dx = \begin{cases} 0, & n \neq m, \\ \frac{2}{2n+1}, & n = m. \end{cases}$$

Для многочленов Лежандра известна явная, порождающая их формула Родрига:

$$q_n(x) = \frac{1}{n! 2^n} \frac{d^n}{dx^n} (x^2 - 1)^n.$$

Многочлены Лежандра удовлетворяют рекуррентной формуле

$$(n+1)q_{n+1}(x) = (2n+1)x q_n(x) - nq_{n-1}(x),$$

которая позволяет, зная нулевой и первый многочлены, легко определять все последующие:

$$q_0(x) = 1, \quad q_1(x) = x, \quad q_2(x) = (3x^2 - 1)/2,$$

$$q_3(x) = (5x^3 - 3x)/2, \quad q_4(x) = (35x^4 - 30x^2 + 3)/8,$$

$$q_5(x) = (63x^5 - 70x^3 + 15x)/8, \dots$$

Сравнивая многочлены Лежандра с многочленами Чебышёва  $T_n(x)$ , видим ряд общих свойств, в частности, на  $[-1, 1]$  обе системы функций имеют  $n$  различных действительных корней.

**П. Многочлены Лагерра**  $L_n(x)$  определяются требованием их ортогональности на луче  $[0, +\infty)$  с весовой функцией  $\exp(-x)$ , обеспечивающей интегрируемость любого многочлена на луче:

$$\int_0^{+\infty} L_n(x)L_m(x)\exp(-x)dx = \begin{cases} 0, & n \neq m, \\ (n!)^2, & n = m. \end{cases}$$

Многочлены Лагерра также могут быть определены с помощью рекуррентного соотношения

$$L_{n+1}(x) = (2n+1-x)L_n(x) - n^2L_{n-1}(x).$$

Это позволяет найти все старшие многочлены после нулевого и первого:

$$L_0(x) = 1, \quad L_1(x) = -x + 1, \quad L_2(x) = x^2 - 4x + 2,$$

$$L_3(x) = -x^3 + 9x^2 - 18x + 6, \quad L_4(x) = x^4 - 16x^3 + 72x^2 - 96x + 24,$$

$$L_5(x) = -x^5 + 25x^4 - 200x^3 + 600x^2 - 600x + 120, \dots$$

Заметим, что под тем же обозначением  $L_n(x)$  в литературе может встречаться многочлен  $\frac{1}{n!}L_n(x)$ .

Существуют ортогональные многочлены Лагерра более общего вида, их ортогональность может пониматься как ортогональность с весовой функцией

$$p(x) = x^\alpha \exp(-x), \quad \alpha > -1.$$

**III. Многочлены Эрмита**  $H_n(x)$  ортогональны на всей числовой оси с весом  $p(x) = \exp(-x^2)$ :

$$\int_{-\infty}^{+\infty} H_n(x)H_m(x) \exp(-x^2) dx = \begin{cases} 0, & n \neq m, \\ 2^n n! \sqrt{\pi}, & n = m. \end{cases}$$

Многочлены Эрмита удовлетворяют рекуррентному равенству

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x).$$

Приведем несколько полиномов Эрмита:

$$H_0(x) = 1, \quad H_1(x) = 2x, \quad H_2(x) = 4x^2 - 2,$$

$$H_3(x) = 8x^3 - 12x, \quad H_4(x) = 16x^4 - 48x^2 + 12,$$

$$H_5(x) = 32x^5 - 160x^3 + 120x, \dots$$

Ортогональные многочлены Якоби (под которыми понимают многочлены Лежандра, а также многочлены Чебышёва первого и второго рода), Лагерра, Эрмита являются решениями одного семейства ОДУ 2-го порядка. Все ортогональные многочлены  $P_n(x)$  одной системы линейно независимы и удовлетворяют рекуррентному соотношению вида:

$$\alpha_n P_{n+1}(x) + (\beta_n - x) P_n(x) + \gamma_n P_{n-1}(x) = 0.$$

Общим свойством всех приведенных систем ортогональных многочленов  $P_n(x)$  является то, что все  $n$  корней — простые и находятся на промежутке ортогональности, при этом корни  $P_n(x)$  разделяются корнями  $P_{n-1}(x)$ . Наименее уклоняются от нуля на отрезке  $[-1, 1]$  в норме С многочлены Чебышёва первого рода, в норме  $L_1$  — многочлены Чебышёва второго рода, в норме  $L_2$  — многочлены Лежандра.

Ортогональные многочлены встречаются в задачах интерполяции, интегрирования, ускорения итерационных процессов решения систем линейных уравнений и т.д.

## Ответы

**I.6.5.** Указание: Рассмотреть представление произвольного действительного числа в виде бесконечной двоичной дроби:

$$a = \text{sign } a \cdot 2^q \cdot \left( \frac{a_1}{2} + \frac{a_2}{2^2} + \dots + \frac{a_p}{2^p} + \frac{a_{p+1}}{2^{p+1}} + \dots \right),$$

где  $a_i$  равно 0 или 1, и соответствующее ему округленное представление:

$$a = \text{sign } a \cdot 2^q \cdot \left( \frac{a_1}{2} + \frac{a_2}{2^2} + \dots + \frac{a_p}{2^p} \right).$$

Ответ:  $2^{-p}$ .

**I.6.6.** 1)  $f'(x_0) = \frac{-3f(x_0) + 4f(x_1) - f(x_2)}{2h}$ . Полная погрешность

$$\frac{M_3 h^2}{3} + \frac{4M_0 \varepsilon_M}{h}. \text{ Оптимальный шаг } \sqrt[3]{\frac{6M_0 \varepsilon_M}{M_3}}.$$

2)  $f'(x_0) = \frac{-3f(x_N) + 4f(x_{N-1}) - f(x_{N-2})}{2h}$ . Полная погрешность

$$\frac{M_3 h^2}{3} + \frac{4M_0 \varepsilon_M}{h}. \text{ Оптимальный шаг } \sqrt[3]{\frac{6M_0 \varepsilon_M}{M_3}}.$$

3)  $f'(x_0) = \frac{-11f(x_0) + 18f(x_1) - 9f(x_2) + 2f(x_3)}{6h}$ . Полная погрешность

$$\frac{M_4 h^3}{4} + \frac{20M_0 \varepsilon_M}{3h}. \text{ Оптимальный шаг } \sqrt[4]{\frac{80M_0 \varepsilon_M}{9M_3}} \text{ I.8.16. а) } 10^{-4}. \text{ б) } 10^{-3}.$$

в)  $10^{-3}$ . I.8.20. а) 18 шагов. б)  $\delta = 1.4 \cdot 10^{-7}$ . в)  $\alpha \in [10-3\pi, 11-3\pi]$ . I.8.28. Пол-

ная погрешность  $\frac{M_4 h^3}{4} + \frac{20M_0 \varepsilon_M}{3h}$ . Оптимальный шаг  $\sqrt[4]{\frac{80M_0 \varepsilon_M}{9M_3}}$ .

**II.9.2.** Задача об ошибке правой части.

$$\begin{cases} \mathbf{A}(\mathbf{x} + \Delta \mathbf{x}) = \mathbf{b} + \Delta \mathbf{b} \\ \mathbf{A}\mathbf{x} = \mathbf{b} \end{cases} \Rightarrow \mathbf{A}\Delta \mathbf{x} = \Delta \mathbf{b} \Leftrightarrow \Delta \mathbf{x} = \mathbf{A}^{-1}\Delta \mathbf{b} \Rightarrow \|\Delta \mathbf{x}\| = \|\mathbf{A}^{-1}\Delta \mathbf{b}\| \leq \|\mathbf{A}^{-1}\| \|\Delta \mathbf{b}\|$$

Эта последняя оценка становится точным равенством, когда вектор  $\Delta\mathbf{b}$  реализует максимум матричной (подчиненной нормы)  $\mathbf{A}^{-1}$ . При выборе первой векторной нормы соответствующий максимум для подчиненной матричной нормы реализуется на векторе сигнатуры элементов той строки  $\mathbf{A}^{-1}$ , где достигается максимум строчных сумм модулей.

$$\mathbf{A}^{-1} = \frac{1}{122 \cdot 101 - 110^2} \begin{pmatrix} 122 & -110 \\ -110 & 101 \end{pmatrix} = \frac{1}{222} \begin{pmatrix} 122 & -110 \\ -110 & 101 \end{pmatrix},$$

поэтому равенство будет иметь место для всех  $\Delta\mathbf{b}$ , пропорциональных вектору  $(1, -1)^T$ . Так как  $\|\mathbf{b}\|_1 = 342$ , а  $\frac{\|\Delta\mathbf{b}\|_1}{\|\mathbf{b}\|_1} = 0.01$ , то  $\|\Delta\mathbf{b}\|_1 = 0.01 \cdot 342 \cdot (1, -1)^T = (3.42, -3.42)$ .

Теперь точное решение системы  $\mathbf{x} = (2, 1)^T$ , при заданной правой части мы имеем оценку

$$\frac{\|\Delta\mathbf{x}\|_1}{\|\mathbf{x}\|_1} \leq v(\mathbf{b}) \frac{\|\Delta\mathbf{b}\|_1}{\|\mathbf{b}\|_1}, \quad v(\mathbf{b}) = \|\mathbf{A}^{-1}\|_1 \frac{\|\mathbf{b}\|_1}{\|\mathbf{x}\|_1} = \frac{122 + 110}{222} \cdot \frac{342}{2} = \frac{232}{222} \cdot 171 \approx 178.7$$

Тогда

$$\frac{\|\Delta\mathbf{x}\|_1}{\|\mathbf{x}\|_1} \leq 178.7 \cdot 0.01 = 1.787.$$

Найдем минимум относительной погрешности решения при фиксированной относительной погрешности правой части.

$$\begin{aligned} \left( \frac{\|\Delta\mathbf{x}\|_1}{\|\mathbf{x}\|_1} \right)_{\min} &= \frac{\min_{\|\Delta\mathbf{b}\|=\text{const}} \|\mathbf{A}^{-1}\Delta\mathbf{b}\|}{\|\mathbf{x}\|_1} = \frac{\min_{\|\Delta\mathbf{b}\|=\text{const}} \frac{\|\mathbf{A}^{-1}\Delta\mathbf{b}\|}{\|\Delta\mathbf{b}\|} \|\Delta\mathbf{b}\|}{\|\mathbf{x}\|_1} = \frac{\|\Delta\mathbf{b}\| \min_{\Delta\mathbf{b}} \frac{\|\mathbf{A}^{-1}\Delta\mathbf{b}\|}{\|\Delta\mathbf{b}\|}}{\|\mathbf{x}\|_1} = \\ &= \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{x}\|} \left( \max_{\Delta\mathbf{b}} \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{A}^{-1}\Delta\mathbf{b}\|} \right)^{-1} = \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{x}\|} \left( \max_{\Delta\mathbf{b}} \frac{\|\mathbf{A}\Delta\mathbf{b}\|}{\|\Delta\mathbf{b}\|} \right)^{-1} = \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{x}\| \|\mathbf{A}\|}. \end{aligned}$$

Реализуется равенство на векторе с нужной нормой:  $\Delta\mathbf{b} = \text{const} \cdot \mathbf{A}\Delta\mathbf{x}$ , где  $\Delta\mathbf{x}$  реализует норму  $\mathbf{A}$ .

Ответы: а)  $\|\Delta\mathbf{x}\|_1 / \|\mathbf{x}\|_1 \leq 1.787$ , равенство реализуется для

$$\Delta\mathbf{b} = (3.42, -3.42)^T; (\|\Delta\mathbf{x}\|_1 / \|\mathbf{x}\|_1)_{\min} = 0.00737 \text{ реализуется на}$$

$$\Delta\mathbf{b} = (3.11, 3.42)^T$$

б)  $\|\Delta\mathbf{x}\|_2 / \|\mathbf{x}\|_2 \leq 0.662$ ; равенство реализуется для  $\Delta\mathbf{b} = (1.9, 0)^T$ ,

$$(\|\Delta\mathbf{x}\|_2 / \|\mathbf{x}\|_2)_{\min} = 0.00273 \text{ реализуется на } \Delta\mathbf{b} = (0.901, -0.999)^T$$

в)  $\|\Delta\mathbf{x}\|_2 / \|\mathbf{x}\|_2 \leq 0.735$ ; равенство реализуется для  $\Delta\mathbf{b} = (0, 2.1)^T$ ,

$$(\|\Delta\mathbf{x}\|_2 / \|\mathbf{x}\|_2)_{\min} = 0.00366 \text{ реализуется на } \Delta\mathbf{b} = (1.110, -0.990)^T$$

г)  $\|\Delta\mathbf{x}\|_3 / \|\mathbf{x}\|_3 \leq 1.458$ ; равенство реализуется для  $\Delta\mathbf{b} = (0.747, -0.664)^T$ ,

$$(\|\Delta\mathbf{x}\|_3 / \|\mathbf{x}\|_3)_{\min} = 0.001 \text{ реализуется на } \Delta\mathbf{b} = (0.664, 0.747)^T$$

д)  $\|\Delta\mathbf{x}\|_3 / \|\mathbf{x}\|_3 \leq 1.477$ ; равенство реализуется для  $\Delta\mathbf{b} = (1.711, -1.198)^T$ ,

$$(\|\Delta\mathbf{x}\|_3 / \|\mathbf{x}\|_3)_{\min} = 0.00985 \text{ реализуется на } \Delta\mathbf{b} = (1.198, 1.711)^T.$$

**П.9.7.** а) 4/1.001; б) ~0.0025; в) метод Гаусса – с 4 знаками, метод Гаусса с выбором главного элемента — с 1. П.9.13. 19 шагов в любом случае.

**П.9.22.** а) 1.  $\tau_{\text{опт}} = 2/7$ , 2.  $\tau_{\text{опт}} = 2/23$ ; б) 1.  $\tau_{\text{опт}} = 2/5$ , 2.  $\tau_{\text{опт}} = 2/19$ ; в) 1.  $\tau_{\text{опт}} = 1/8$ , 2.  $\tau_{\text{опт}} = 2/25$ . П.9.25.  $u^1 = 1/5(8, -7, 19)$ ,  $u^2 = 1/5(-1, 0, 21)$

**П.9.26.** 1. а)  $\tau \in (0, 1/6)$ ,  $\tau_{\text{опт}} = 1/7$ ; б)  $\tau' \in (0, 1/6)$ ,  $\tau'_{\text{опт}} = 1/7$ ;

2. а)  $\tau \in (0, 1/6)$ ,  $\tau_{\text{опт}} = 2/23$ ; б)  $\tau' \in (0, 1/6)$ ,  $\tau_{\text{опт}} = 2/23$ ; 3. а)  $\tau = 1/11$ ;

б)  $\tau = 1/11$ ; 4. а)  $\tau \in (0, 2/11)$ ,  $\tau_{\text{опт}} = 2/13$ ; б)  $\tau' \in (0, 1/6)$ ;

5. а)  $\tau \in (0, 2/7)$ ,  $\tau_{\text{опт}} = 1/4$ ; б)  $\tau' \in (0, 2/25)$ ,  $\tau'_{\text{опт}} < 2/25$ .

**П.9.28.** а)  $\omega_{\text{опт}} = 1.36847$ , расчет можно вести с  $\omega = 1.4$ , вычисление сначала  $x$ , потом  $y$  и  $z$ , в)  $\omega_{\text{опт}} = 1.29461$ , расчет можно вести с  $\omega = 1.3$ , вычисление сначала  $y$ , потом  $x$  и  $z$ .

**IV.11.1.** а)  $\sqrt{a^2 + b^2} < 1$ , с — любое; б)  $|ab| < 1$ , с — любое; в)  $|a - b| < 1$ , с — любое; г)  $|ab| < 1$ , с — любое; д)  $|ab| < \sqrt{e/2}$ , с — любое. IV.11.7 Третий.

**IV.11.25.**  $x^{(k+1)} = \operatorname{arctg} x^{(k)} + \pi l$ ,  $l$  — номер корня. IV.11.37 третий порядок для корня  $x = 0.5$  и первый для кратного корня  $x = -1$ .

**IV.12.8.** а) Т. Max  $x_{\max} = 1$ ,  $f_{\max} = 1/e$ ,  $f(x) = f_{\max}/2$ :  $x_1 = 0.335$ ,  $x_2 = 1.528$ ,  $\Delta_{1/2} = x_2 - x_1 = 1.193$ , б) Т. Max  $x_{\max} = 1/\sqrt{2}$ ,  $f_{\max} = 1/\sqrt{2} \exp(-1/2)$ ,  $f(x) = f_{\max}/2$ :  $x_1 = 0.226$ ,  $x_2 = 1.359$ ,  $\Delta_{1/2} = x_2 - x_1 = 1.133$ , в) Т. Max  $x_{\max} = \sqrt{e}$ ,  $f_{\max} = 1/(2e)$ ,  $f(x) = f_{\max}/2$ :  $x_1 = 1.123$ ,  $x_2 = 3.816$ ,  $\Delta_{1/2} = x_2 - x_1 = 2.693$ , г) Т. Max  $x_{\max} = e$ ,  $f_{\max} = 1/e$ ,  $f(x) = f_{\max}/2$ :  $x_1 = 1.261$ ,  $x_2 = 14.561$ ,  $\Delta_{1/2} = x_2 - x_1 = 13.300$ . д) Т. Max  $x_{\max} = 1/2$ ,  $f_{\max} = 1/\sqrt{2} \exp(-1/2)$ ,  $f(x) = f_{\max}/2$ :  $x_1 = 0.051$ ,  $x_2 = 1.846$ ,  $\Delta_{1/2} = x_2 - x_1 = 1.795$ .

**IV.12.9.** Для  $n = 1$   $t_m = 277$  пс, для  $n = 2$   $t_m = 388$  пс, для  $n = 3$   $t_m = 474$  пс. Поэтому выбираем  $n = 2$ .

**V.7.4.** а) Парабола по точкам 1–2–4 имеет минимум в точке  $x_5 = 7.38888$ ,  $y_5 = 4.197 \cdot 10^{-8}$ , а по точкам 1–3–4 — в точке  $x_6 = 7.38752$ ,  $y_6 = 3.193 \cdot 10^{-6}$ . В следующем приближении остаются точки 2–6–5–3.

б) Парабола по точкам 1–2–4 имеет минимум в точке  $x_5 = 0.74530$ , а по точкам 1–3–4 — в точке  $x_6 = 0.72572$ . Сносов точек 5 и 6 нет,  $y_5 = 3.448815$ ,  $y_6 = 3.44416$ . В следующем приближении остаются точки 1–6–5–2.

в) Парабола по точкам 1–2–4 имеет минимум в точке  $x_5 = 1.76613$ , а по точкам 1–3–4 — в точке  $x_6 = 2.06572$ . Сносов точек 5 и 6 нет,  $y_5 = -0.43354$ ,  $y_6 = 0.05346$ . В следующем приближении остаются точки 2–5–3–6.

г) Парабола по точкам 1–2–4 имеет минимум в точке  $x_5 = 0.87947$ , а по точкам 1–3–4 — в точке  $x_6 = 1.03285$ . Сносов точек 5 и 6 нет,  $y_5 = 3.66101$ ,  $y_6 = 3.80595$ . В следующем приближении остаются точки 1–2–5–3.

д) Парабола по точкам 1–2–4 имеет минимум в точке  $x_5 = 4.36467$ , а по точкам 1–3–4 — в точке  $x_6 = 4.43883$ . Сносов точек 5 и 6 нет,  $y_5 = 1.14903$ ,  $y_6 = 1.14404$ . В следующем приближении остаются точки 5–6–3–4.

е) Параобра по точкам 1–2–4 имеет минимум в точке  $x_5 = 0.64340$ , а по точкам 1–3–4 — в точке  $x_6 = 0.64508$ . Сносов точек 5 и 6 нет,  $y_5 = 0.69888$ ,  $y_6 = 0.69887$ . В следующем приближении остаются точки 2–5–6–3.

**VI.8.1.** а) 0.045, б) 0.045, в) 0.027. VI.8.9. а)  $10^{-5}$ , б)  $10^{-4}$ .

**VI.8.15.**  $H^R = p(p + 2qp)$ ,  $G^R = -p \cdot qp$ ,  $H^L = q(q + 2qp)$ ,  $G^L = q \cdot qp$ .

**VI.8.20.**  $W(x_1, x_2, \dots, x_{2n+1}) = 2^{2n^2} \prod_{1 \leq i < j \leq 2n+1} \sin \frac{x_j - x_i}{2}$ .

**VI.8.21.** При  $t \in [t_N, t_N + \tau]$   $|R_N(t)| \leq \tau^{N+1} \max_{\xi \in [t_0, t_N + \tau]} |f^{(N+1)}(\xi)|$ ,

при  $t \in [t_N + \tau, t_N + 2\tau]$   $|R_N(t)| \leq (N+2)\tau^{N+1} \max_{\xi \in [t_0, t_N + 2\tau]} |f^{(N+1)}(\xi)|$ .

При  $t \in [t_N + 2\tau, t_N + 3\tau]$   $|R_N(t)| \leq \frac{(N+2)(N+3)}{2!} \tau^{N+1} \max_{\xi \in [t_0, t_N + 3\tau]} |f^{(N+1)}(\xi)|$ .

**VI.9.1.** а)  $x = 1.841$ , б)  $x = 0.213$ , в)  $x = 0.243$ , г)  $x = 0.639$ , д)  $x = 0.739$ , е)  $x = 0.877$ . VI.9.2. а)  $f'(5) = -0.085$ , б)  $f'(0.3) = -58$ , в)  $f'(3) = -23/12 \approx -1.9$ , г)  $f'(2) = -2.18$ , д)  $f'(3) = 1.6$ .

**VI.9.4.** 4.244.

**VI.9.18.** 49.968.

**VI.9.25.** а) Отрезок  $[1, 2]$   $a = 0.5$ ,  $b = 0.451808$ ,  $c = -0.0722874$ ,  $d = -0.134956$ ,  $f(x^*) = 0.706145$ . б) Отрезок  $[0.5, 0.9]$ ,  $a = -0.693147$ ,  $b = 2.72502$ ,  $c = -4.86964$ ,  $d = 4.32685$ ,  $f(x^*) = -0.197082$ . в) Отрезок  $[1.7, 3.4]$ ,  $a = 1.30384$ ,  $b = 0.535474$ ,  $c = -0.204257$ ,  $d = 0.0447927$ ,  $f(x^*) = 1.75317$ . г) Отрезок  $[-0.1, 0.2]$ ,  $a = 1.67096$ ,  $b = -1.02012$ ,  $c = 0.885522$ ,  $d = -0.128102$ ,  $f(x^*) = 1.46946$ . д) Отрезок  $[1., 2.]$ ,  $a = 1.54030$ ,  $b = 0.252035$ ,  $c = -0.432401$ ,  $d = 0.223917$ ,  $f(x^*) = 1.58621$ .

**VII.8.8.** 41.

**VIII.9.1.** Метод строго устойчив при  $\frac{\partial f}{\partial u} \leq 0$  для любых значений сеточного параметра.

**VIII.9.2.** Метод строго устойчив при  $\frac{\partial f}{\partial u} \leq 0$  и  $\tau \left| \frac{\partial f}{\partial u} \right| < 2$ .

	0			
	$\frac{3-\sqrt{3}}{6}$	$\frac{3-\sqrt{3}}{6}$		
VIII.9.4.	$\frac{3+\sqrt{3}}{6}$	$\gamma$	$\frac{3+\sqrt{3}}{6}-\gamma$	
	0	1/2	1/2	

**VIII.9.5.** Второй порядок аппроксимации. VIII.9.6. Первый порядок. Для порядка 2 — вместо  $3/2$  в таблице Бутчера должна быть 2.

**VIII.9.7.** 1) Порядок аппроксимации  $O(\tau)$ . Свойства схемы можно улучшить, если в правой части взять  $a(x_n + 2x_{n-1})$ .

2) Порядок аппроксимации  $O(\tau^3)$ . Общее решение разностного уравнения

$$y^n = C_1 \left( \frac{2 + \sqrt{1+2\tau a}}{3 - 2\tau a} \right)^n + C_2 \left( \frac{2 - \sqrt{1+2\tau a}}{3 - 2\tau a} \right)^n.$$

3) Порядок аппроксимации  $O(\tau^2)$ . Общее решение разностного уравнения

$$y^n = C_1 \left( -2(1 - \tau a) + \sqrt{9 - 6\tau a + \frac{\tau^2 a^2}{4}} \right)^n + C_2 \left( -2(1 - \tau a) - \sqrt{9 - 6\tau a + \frac{\tau^2 a^2}{4}} \right)^n.$$

4) Схема не аппроксимирует задачу. Для аппроксимации следует взять

$$\frac{x_{n+1} - x_{n-1}}{\tau} = \frac{1}{3} a(3x_{n+1} + 4x_n - x_{n-1}).$$

**VIII.9.8.** Порядок аппроксимации второй во внутренних точках. Начальное условие аппроксимируется со вторым порядком с учетом самого дифференциального уравнения. По основной теореме вычислительной математики порядок сходимости второй.

**VIII.9.10.** а) Заметим, что

$$\sin(y^n + 0.5\tau \sin y^n) \cos(0.5\tau \sin y^n) = \frac{1}{2} (\sin(y^n + \tau \sin y^n) + \sin y^n).$$

Тогда метод перепишется в виде  $k_1 = \sin y^n$ ,  $k_2 = \sin(y^n + \tau k_1)$ ,  $y^{n+1} = y^n + \frac{\tau}{2}(k_1 + k_2)$ , то есть использован метод Рунге–Кутты второго порядка аппроксимации. Читателям предлагается самим проанализиро-

вать, как будет эволюционировать погрешность округления при реализации метода Рунге–Кутты в его «обычном» виде и в том виде, как он был записан в задаче.

г) Вычислим точное решение задачи — проинтегрируем уравнение с разделяющимися переменными. Получим  $y = \ln(4 - x^2)$ . Видно, что решение на всем отрезке интегрирования не существует, стало быть, не приходится говорить о порядке аппроксимации.

**VII.9.17.** Порядок аппроксимации — первый из-за аппроксимации граничного условия. Для второго порядка необходимо положить  $y_1 = 2h$ .

## **Литература**

1. Сборник задач по основам вычислительной математики / под ред. О.М. Белоцерковского. — Москва : МФТИ, 1974. — 148 с.
2. Рябенький В.С. Введение в вычислительную математику. — Москва : Физматлит, 2007. — 288 с.
3. Федоренко Р.П. Введение в вычислительную физику. 2-е изд. — Долгопрудный : Издательский дом «Интеллект», 2009. — 504 с.
4. Косарев В.И. 12 лекций по вычислительной математике. — Москва : Физматкнига, 2013. — 240 с.
5. Петров И.Б., Лобанов А.И. Лекции по вычислительной математике. — Москва : БИНОМ, 2009. — 533 с.
6. Упражнения и задачи контрольных работ по вычислительной математике. Часть 1 / под ред. В.В. Демченко. — Москва : МФТИ, 2019. — 143 с.
7. Галанин М.П., Савенков Е.Б. Методы численного анализа математических моделей. — Москва : Изд-во МГТУ им. Н.Э. Баумана, 2010. — 591 с.
8. Бахвалов Н.С., Лапин А.В., Чижонков Б.В. Численные методы в задачах и упражнениях. учеб. пособие / под ред. В.А. Садовничего. — Москва : Высшая школа, 2000. — 190 с.
9. Лабораторный практикум по курсу Основы вычислительной математики. / В.Д. Иванов, В.И. Косарев и др. — Москва : МЗ Пресс, 2003. — 196 с.
10. <https://standards.ieee.org/content/ieee-standards/en/standard/754-2019.html> (доступ 3 февраля 2020)
11. Воеводин В.В. Вычислительные основы линейной алгебры. — Москва : Наука, 1977. — 303 с.
12. Голуб Дж., Ван Лоун Ч. Матричные вычисления. — Москва : Мир, 1999. — 548 с.
13. Деммель Дж. Вычислительная линейная алгебра. Теория и приложения. — Москва : Мир, 2001. — 429 с.
14. Фадеев А.К., Фадеева В.Н. Вычислительные методы линейной алгебры. — Санкт-Петербург : Лань, 2002. — 736 с.

15. *Воеводин В.В., Кузнецов Ю.А.* Матрицы и вычисления. — Москва : Наука, 1984. — 320 с.
16. *Беклемишев Д.В.* Курс аналитической геометрии и линейной алгебры. 17 изд., стереотипное. — Санкт-Петербург : Лань, 2020. — 448 с.
17. *Коновалов А.Н.* Введение в вычислительные методы линейной алгебры. — Новосибирск : Наука, 1993. — 158 с.
18. *Годунов С.К.* Лекции по современным аспектам линейной алгебры. — Новосибирск : Научная книга (ИДМИ), 2002. — 216 с. — (Университетская серия. Т. 12)
19. *Амосов А.А., Дубинский Ю.А., Копченова Н.В.* Вычислительные методы для инженеров. — Москва : Высшая школа, 1994. — 544 с.
20. *Вержбицкий В.М.* Численные методы. Линейная алгебра и нелинейные уравнения. — Москва : Высшая школа, 2000. — 266 с.
21. *Бахвалов Н.В., Жидков Н.П., Кобельков Г.М.* Численные методы. — Москва : Лаборатория Базовых Знаний, 2002. — 632 с.
22. *Каханер Д., Моулер К., Нэш С.* Численные методы и программное обеспечение. — Москва : Мир, 1998. — 575 с.
23. *Шарковский Ю.Н., Майстренко Ю.А., Романенко Е.Ю.* Разностные уравнения и их приложения. — Киев : Наукова думка, 1986. — 279 с.
24. *Лобанов А.И.* Математическое моделирование нелинейных процессов: учебник для академического бакалавриата / А.И. Лобанов, И.Б. Петров. — Москва : Издательство Юрайт, 2019. — 255 с. — (Бакалавр. Академический курс.) — ISBN 978-5-9916-8897-0. — Текст : электронный // ЭБС Юрайт [сайт]. — URL: <https://urait.ru/bcode/437003> (дата обращения: 03.02.2020).
25. *Васильев Ф.П., Иваницкий А.Ю.* Линейное программирование. — Москва : Изд-во «Факториал», 1998. — 176 с.
26. *Шананин А.А., Обросова Н.К.* Экономическая интерпретация двойственности в задачах линейного программирования. — Москва : Изд-во РУДН, 2007. — 36 с.
27. *Калиткин Н.Н.* Численные методы. — Москва : Наука, 1978. — 512 с.; 2 изд. Санкт-Петербург : БВХ-Петербург, 2014. — 592 с.
28. *Бирюков С.И.* Оптимизация. Введение в теорию. Численные методы. — Москва : МЗ-пресс, 2003. — 244 с.
29. *Бирюков А.Г.* Методы оптимизации. Условия оптимальности в экстремальных задачах. — Москва : МФТИ, 2010. — 244 с.

30. Жадан В.Г. Методы оптимизации Ч. 1: Введение в выпуклый анализ и теорию оптимизации. — Москва : МФТИ, 2014. — 270 с.
31. Жадан В.Г. Методы оптимизации Ч. 2: Численные алгоритмы. — Москва : МФТИ, 2015. — 319 с.
32. Мусеев Н.Н., Иванилов Ю.П., Столлярова Е.М. Методы оптимизации. — Москва : Наука, 1978. — 351 с.
33. Ильина В.А., Силаев П.К. Численные методы для физиков-теоретиков. Т.1. — Москва–Ижевск : Институт компьютерных исследований, 2003. — 133 с.
34. Ильина В.А., Силаев П.К. Численные методы для физиков-теоретиков. Т. 2. — Москва–Ижевск : Институт компьютерных исследований, 2003. — 118 с.
35. Гавриков М.Б., Локуциевский О.В. Начала численного анализа. — Москва : ТОО «Янус», 1995. — 580 с.
36. Завьялов Ю.С., Квасов Б.И., Мирошниченко В.Л. Методы сплайн-функций. — Москва : Наука, 1980. — 355 с.
37. Завьялов Ю.С., Леус В.А., Скоропоселов В.А. Сплайны в инженерной геометрии. — Москва : Машиностроение, 1985. — 224 с.
38. Калиткин Н.Н., Альшина Е.А. Численные методы. Книга 1. Численный анализ. — Москва : Академия, 2013. — 299с.
39. Бейкер Дж., Грейвс-Моррис П. Аппроксимации Паде. — Москва : Мир, 1986. — 502 с.
40. Калиткин Н.Н., Альшин А.Б., Альшина Е.А., Рогов Б.В. Вычисления на квазиравномерных сетках. — Москва : Наука–Физматлит, 2005. — 224 с.
41. Рябенький В.С., Филиппов А.Ф. Об устойчивости разностных уравнений. — Москва : Гостехиздат, 1956. — 172 с.
42. Lax P.D., Richtmyer R.D. Survey of the stability of linear finite difference equations // Comm. Pure Appl. Math. 9 (1956), 267–293.
43. Годунов С.К., Рябенький В.С. Разностные схемы. Введение в теорию. — Москва : Наука, 1977. — 440 с.
44. Хайрер Э., Нерсетт С., Ваннер Г. Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. — Москва : Мир, 1990. — 512 с.
45. Лобанов А.И., Мещеряков М.В., Чудов Л.А. Задачи для самостоятельного исследования в курсе вычислительной математики: учебное пособие. — Москва : МФТИ, 2001. — 76 с.

46. *Малинецкий Г.Г.* Задачи по курсу нелинейной динамики. Сер. Синергетика — от прошлого к будущему, № 148. Изд. 2, испр. — Москва:URSS. 2018. — 136 с.
47. *Пинни Э.* Обыкновенные дифференциально-разностные уравнения. — Москва : ИЛ, 1961. — 248 с.
48. *Самарский А.А.* Введение в численные методы. 3-е изд. — Санкт-Петербург: Лань, 2005 — 288 с.
49. *Канторович Л.В., Акилов Г.П.* Функциональный анализ. 4-е изд. — Москва : ВНВ, 2004. — 816 с.

Учебное издание

**Аристова Елена Николаевна  
Завьялова Наталья Александровна  
Лобанов Алексей Иванович**

**ПРАКТИЧЕСКИЕ ЗАНЯТИЯ  
ПО ВЫЧИСЛИТЕЛЬНОЙ  
МАТЕМАТИКЕ В МФТИ**

**Часть 1**

Издание второе, исправленное и дополненное

Редактор *O. П. Котова*.  
Корректор *I. A. Волкова*  
Дизайн обложки *E. A. Березина*

Подписано в печать 11.06.2021. Формат 60 × 84 1/16. Усл. печ. л. 15,12.  
Уч.-изд. л. 14,0. Тираж 150 экз. Заказ № 56.

Федеральное государственное автономное образовательное  
учреждение высшего образования «Московский физико-  
технический институт (национальный исследовательский университет)»  
141700, Московская обл., г. Долгопрудный, Институтский пер., 9  
Тел. (495) 408-58-22, e-mail: [rio@mipt.ru](mailto:rio@mipt.ru)

---

Отпечатано в полном соответствии с предоставленным оригиналом макетом  
ООО «Печатный салон ШАНС» 127412. Москва, ул. Ижорская, д. 13, стр. 2  
Тел. (495) 484 26-55.

Е. Н. Аристова, Н. А. Завьялова, А. И. Лобанов

Практические занятия по вычислительной математике в МФТИ. Часть I



ISBN 978-5-7417-0774-6



9 785741 707746 >