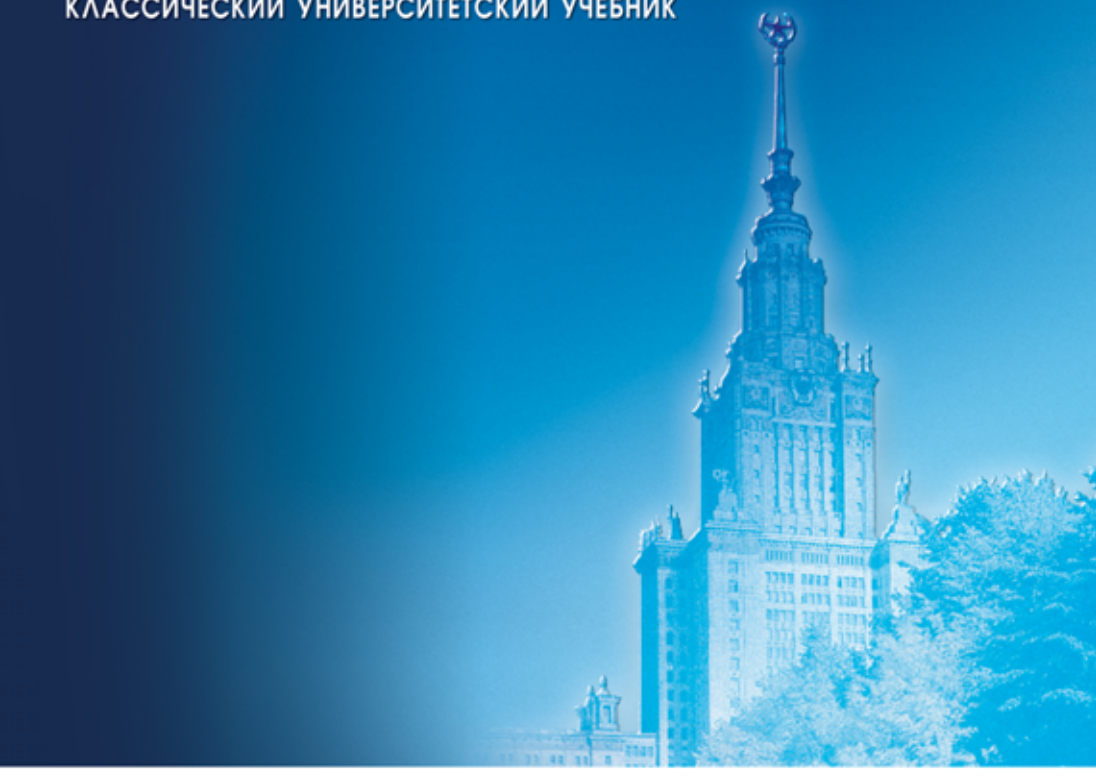
Московский государственный университет имени М.В. Ломоносова

КЛАССИЧЕСКИЙ УНИВЕРСИТЕТСКИЙ УЧЕБНИК



Н. С. Бахвалов, А. А. Корнев, Е. В. Чижонков

ЧИСЛЕННЫЕ МЕТОДЫ
РЕШЕНИЯ ЗАДАЧ И УПРАЖНЕНИЯ

Московский государственный университет имени М. В. Ломоносова

Н. С. Бахвалов, А. А. Корнев, Е. В. Чижонков

ЧИСЛЕННЫЕ МЕТОДЫ
РЕШЕНИЯ ЗАДАЧ И УПРАЖНЕНИЯ

2-е издание,
исправленное и дополненное
(электронное)

Допущено
УМО по классическому университетскому образованию
в качестве учебного пособия для студентов высших
учебных заведений, обучающихся по специальности
01.05.01 «Фундаментальные математика и механика»

Москва
Лаборатория знаний
2016

УДК 519.6(075.8)
ББК 22.193я73
Б30

Печатается
по решению Ученого совета
Московского государственного университета
имени М. В. Ломоносова

Бахвалов Н. С.

Б30 Численные методы. Решения задач и упражнения [Электронный ресурс] : учебное пособие для вузов / Н. С. Бахвалов, А. А. Корнев, Е. В. Чижонков. — 2-е изд., испр. и доп. (эл.). — Электрон. текстовые дан. (1 файл pdf : 355 с.). — М. : Лаборатория знаний : Лаборатория Базовых Знаний, 2016. — (Классический университетский учебник). — Систем. требования: Adobe Reader XI ; экран 10".

ISBN 978-5-93208-205-8

Материал пособия соответствует программе курса «Численные методы», рекомендованной Министерством образования и науки РФ. Содержатся основные положения теории, большое количество подробно разобранных примеров, которые являются основой для компьютерного решения практических и учебных задач различного уровня сложности — от домашних упражнений до курсовых и дипломных работ. Включены упражнения для самостоятельной работы.

Книга такого типа по численным методам не имеет аналогов как в нашей стране, так и за рубежом.

Для студентов университетов, педагогических вузов, вузов с углубленным изучением математики, а также для студентов технических вузов, аспирантов и преподавателей, инженеров и научных работников, использующих в практической деятельности численные методы.

УДК 519.6(075.8)
ББК 22.193я73

Деривативное электронное издание на основе печатного аналога: Численные методы. Решения задач и упражнения : учебное пособие для вузов / Н. С. Бахвалов, А. А. Корнев, Е. В. Чижонков. — 2-е изд., испр. и доп. — М. : Лаборатория знаний, 2016. — 352 с. : ил. — (Классический университетский учебник). — ISBN 978-5-906828-04-0.

Подготовлено при участии
ООО «Лаборатория Базовых Знаний»

ISBN 978-5-93208-205-8

Оглавление

Предисловие

Учебное пособие написано на основе многолетнего опыта преподавания численных методов студентам механико-математического факультета и факультета вычислительной математики и кибернетики МГУ им. М. В. Ломоносова и полностью соответствует требованиям Государственного образовательного стандарта по математике, рекомендованного Министерством образования Российской Федерации.

Как правило, классический университетский курс, ориентированный на приближенное решение задач, состоит из теоретической (лекции) и практической (семинары) частей и сопровождается лабораторными работами. Поэтому учебная литература традиционно представлена теоретическими учебниками, сборниками задач и вычислительными практикумами. Предлагаемая вниманию читателя книга содержит в форме задач и упражнений наиболее ценные, по мнению авторов, сведения по численным методам из пособий всех указанных типов, и ее можно использовать не только в учебных, но и в справочных целях.

Пособие охватывает материал по разностным уравнениям, приближению функций, численному интегрированию и дифференцированию, интегральным уравнениям, задачам алгебры и решению нелинейных уравнений, приближенным методам решения дифференциальных уравнений как обыкновенных, так и с частными производными, а также по влиянию вычислительной погрешности в различных алгоритмах.

Главная цель пособия — помочь читателю глубоко и последовательно освоить предмет. Для этого материал разбит на крупные теоретические части — главы и, кроме того, на темы — параграфы, содержание которых структурировано специальным образом. Изучение каждой новой темы начинается со знакомства с основными определениями, формулировками фундаментальных теоретических результатов (теорем), полезными вспомогательными фактами и т. п., затем разбираются и анализируются типичные упражнения, отражающие специфику постановок задач и методы их решений. Первые задачи каждого параграфа решены подробно и сопровождаются комментариями. Сложность задач постепенно возрастает, поэтому нередки ссылки на уже разобранные примеры. Далее приводятся упражнения для самостоятельных занятий. Они, как правило, достаточно разнообразны и могут удовлетворить запросы большинства читателей. Затем содержатся наборы из нескольких упражнений, которые при одинаковом задании имеют различные условия. Это — образцы для контрольных работ по изучаемой теме, они сопровождены только ответами. В конце каждого параграфа имеются упражнения повышенной сложности, как правило, снабженные только указаниями и/или ответами. Их целесообразно использовать в качестве зачетных задач или как основу для небольших курсовых проектов.

Важная методическая особенность пособия — расположение решений, указаний и ответов непосредственно за условиями задач и упражнений, а не в конце книги, как это принято в задачниках. Тщательный отбор и подача материала в такой форме способствуют эффективному усвоению численных методов даже при самостоятельной работе. Поэтому данное учебное пособие рекомендуется студентам, аспирантам и преподавателям высших учебных заведений с углубленным изучением математики и всем, кто по роду своей деятельности сталкивается с приближенным решением задач, допускающих математическую формулировку. Даже специалист в области вычислительной математики может найти сформулированные в виде упражнений необычные формулы, факты, утверждения, неизвестные ему ранее. Например, различные численные аспекты решения уравнения Шрёдингера и задачи Стокса.

Критические замечания и предложения по совершенствованию книги просьба сообщать авторам на кафедру вычислительной математики механико-математического факультета МГУ им. М. В. Ломоносова.

Авторы.

Глава 1

Погрешность решения задачи

Если a — точное значение некоторой величины, a^* — известное приближение к нему, то *абсолютной погрешностью* приближенного значения a^* обычно называют некоторую величину $\Delta(a^*)$, про которую известно, что

$$|a^* - a| \leqslant \Delta(a^*)\,.$$

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

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

Относительную погрешность часто выражают в процентах.

В этой главе на модельных упражнениях показано принципиальное отличие между математически точными вычислениями и вычислениями с произвольно высокой, но конечной точностью. Приведены примеры *катастрофического* накопления вычислительной погрешности в стандартных алгоритмах, рассмотрены методы возможного улучшения исследуемых алгоритмов.

1.1. Вычислительная погрешность

Наиболее распространенная форма представления действительных чисел в компьютерах — *числа с плавающей точкой*. Множество F чисел с плавающей точкой характеризуется четырьмя параметрами: основанием системы счисления p, разрядностью t и интервалом показателей $[L, U]$. Каждое число x, принадлежащее F, представимо в виде

$$x = \pm\left(\frac{d_1}{p} + \frac{d_2}{p^2} + \cdots + \frac{d_t}{p^t}\right) p^{\alpha}\,,$$

где целые числа $p, \alpha, d_1, \ldots, d_t$ удовлетворяют неравенствам $0 \leqslant d_i \leqslant p-1$, $i = 1, \ldots, t$; $L \leqslant \alpha \leqslant U$. Часто d_i называют *разрядами*, t — *длиной мантиссы*, α — *порядком числа*. *Мантиссой* (дробной частью) x называют число в скобках. Множество F называют *нормализованным*, если для каждого $x \neq 0$ справедливо условие $d_1 \neq 0$.

Удобно определить, что округление с точностью ε — это некоторое отображение fl действительных чисел $\mathbf{R}$ на множество F чисел с плавающей точкой, удовлетворяющее следующим аксиомам.

1) Для произвольного $y \in \mathbf{R}$ такого, что результат отображения $fl(y) \in F$, имеет место равенство при $fl(y) \neq 0$

$$fl(y) = y\,(1+\eta), \quad |\eta| \leqslant \varepsilon\,.$$

2) Обозначим результат арифметической операции $*$ с числами $a, b \in F$ через $fl(a * b)$. Если $fl(a * b) \neq 0$, то

$$fl(a * b) = (a * b)(1+\eta), \quad |\eta| \leqslant \varepsilon\,.$$

Приведенные соотношения позволяют изучать влияние ошибок округления в различных алгоритмах.

Если результат округления не принадлежит F, то его обычно называют *переполнением* и обозначают ∞.

Будем считать, что ε — точная верхняя грань для $|\eta|$. При традиционном способе округления чисел имеем $\varepsilon = \frac{1}{2}\,p^{1-t}$, при округлении отбрасыванием разрядов $\varepsilon = p^{1-t}$. Величину ε часто называют *машинной точностью*.

1.1. Построить нормализованное множество F с параметрами $p = 2$, $t = 3$, $L = -1$, $U = 2$.

$\lhd$ Каждый элемент $x \in F$ имеет вид

$$x = \pm\left(\frac{d_1}{2} + \frac{d_2}{4} + \frac{d_3}{8}\right) 2^{\alpha}\,, \text{ где } \alpha \in \{-1, 0, 1, 2\},\ d_i \in \{0, 1\}$$

и $d_1 \neq 0$ для $x \neq 0$.

Зафиксируем различные значения мантисс m_i для ненулевых элементов множества:

$$\frac{1}{2}\,, \quad \frac{1}{2} + \frac{1}{8} = \frac{5}{8}\,, \quad \frac{1}{2} + \frac{1}{4} = \frac{3}{4}\,, \quad \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}\,,$$

или $m_i \in \left\{\frac{1}{2}\,,\ \frac{5}{8}\,,\ \frac{3}{4}\,,\ \frac{7}{8}\right\}$. Далее, умножая m_i на 2^{α} с $\alpha \in \{-1, 0, 1, 2\}$ и добавляя знаки $\pm$, получим все ненулевые элементы множества F: $\pm\frac{1}{4}$, $\pm\frac{5}{16}$, $\pm\frac{3}{8}$, $\pm\frac{7}{16}$, $\pm\frac{1}{2}$, $\pm\frac{5}{8}$, $\pm\frac{3}{4}$, $\pm\frac{7}{8}$, ± 1, $\pm\frac{5}{4}$, $\pm\frac{3}{2}$, $\pm\frac{7}{4}$, ± 2, $\pm\frac{5}{2}$, ± 3, $\pm\frac{7}{2}$. После добавления к ним числа *нуль* имеем искомую модель системы действительных чисел с плавающей точкой. $\rhd$

1.2. Сколько элементов содержит нормализованное множество F с параметрами p, t, L, U?

Ответ: $2\,(p-1)\,p^{t-1}(U - L + 1) + 1$.

1.3. Каков результат операций $fl(x)$ при использовании модельной системы из 1.1 для следующих значений x:

$$\frac{23}{32}\ ,\ \frac{1}{8}\ ,\ 4\,,\ \frac{1}{2} + \frac{3}{4}\ ,\ \frac{3}{8} + \frac{5}{4}\ ,\ 3 + \frac{7}{2}\ ,\ \frac{7}{16} - \frac{3}{8}\ ,\ \frac{1}{4} \cdot \frac{5}{16}\ .$$

Ответ: $\frac{3}{4}$, 0, ∞ $\left(x > \frac{7}{2}\right)$, $\frac{5}{4}$, $\frac{3}{2}$ или $\frac{7}{4}$, ∞, 0, 0.

1.4. Верно ли, что всегда $fl\left(\frac{a+b}{2}\right) \in [a, b]$?

Ответ: нет (см. 1.3).

1.5. Пусть отыскивается наименьший корень уравнения $y^2-140y+1=0$. Вычисления производятся в десятичной системе счисления, причем в мантиссе числа после округления удерживается четыре разряда. Какая из формул $y=70-\sqrt{4899}$ или $y=\dfrac{1}{70+\sqrt{4899}}$ дает более точный результат?

$\lhd$ Воспользуемся первой формулой. Так как $\sqrt{4899}=69,992...$, то после округления получаем $\sqrt{4899}\approx 69,99$, $\quad y_1\approx 70-69,99=0,\underline{01}$.

Вторая формула представляет собой результат «избавления от иррациональности в числителе» первой формулы. Последовательно вычисляя, получаем $70+69,99=139,99\approx 140,0$, $\frac{1}{140}=0,00714285\ldots$. Наконец, после последнего округления имеем $y_2=0,00\underline{7143}$.

Если произвести вычисления с большим количеством разрядов, то можно проверить, что в y_1 и y_2 все подчеркнутые цифры результата верные; однако во втором случае точность результата значительно выше. В первом случае пришлось вычитать близкие числа, что привело к эффекту *пропадания значащих цифр*, часто существенно искажающему конечный результат вычислений. Увеличение абсолютной погрешности также может происходить в результате деления на малое (умножение на большое) число. Еще одна опасность — выход за диапазон допустимых значений в промежуточных вычислениях, например после умножения исходного уравнения на достаточно большое число. $\rhd$

1.6. Пусть приближенное значение производной функции $f(x)$ определяется при $h\ll 1$ по формуле $f'(x)\approx\dfrac{f(x+h)-f(x-h)}{2h}$, а самими значения $f(x)$ вычисляются с абсолютной погрешностью Δ. Какую погрешность можно ожидать при вычислении производной, если $\left|f^{(k)}(x)\right|\leqslant M_k$, $\quad k=0,1,\ldots$?

$\lhd$ В данном случае имеется два источника погрешности: *погрешность метода* и *вычислительная погрешность*. Первая связана с неточностью формулы в правой части при отсутствии ошибок округления. Разложим функцию $f(x\pm h)$ в ряд Тейлора в точке x:

$$f(x\pm h)=f(x)\pm h\,f'(x)+\frac{h^2}{2}\,f''(x)\pm\frac{h^3}{6}\,f'''(x_\pm).$$

Подставляя полученные разложения в правую часть приближенного равенства, получим

$$\frac{f(x+h)-f(x-h)}{2h}=f'(x)+\frac{h^2}{6}\left[\frac{f'''(x_+)+f'''(x_-)}{2}\right].$$

Ограничиваясь главным членом в разложении по степеням h, имеем оценку для погрешности метода

$$\left|\frac{f(x+h)-f(x-h)}{2h}-f'(x)\right|\leqslant\frac{h^2}{6}\,M_3\,.$$

С другой стороны, в силу наличия ошибок округления в вычислениях участвуют не точные значения $f(x\pm h)$, а их приближения $f^*(x\pm h)$ с заданной абсолютной погрешностью. Поэтому полная погрешность выглядит так:

$$Err = \left| \frac{f^*(x+h) - f^*(x-h)}{2h} - f'(x) \right| .$$

Добавляя в числитель дроби $\pm f(x+h)$ и $\pm f(x-h)$, после перегруппировки слагаемых получим

$$Err \leqslant \left| \frac{f^*(x+h)-f(x+h)}{2h} - \frac{f^*(x-h)-f(x-h)}{2h} \right| + \left| \frac{f(x+h)-f(x-h)}{2h} - f'(x) \right| .$$

Оценка вычислительной погрешности для каждого из двух первых слагаемых имеет вид $\frac{\Delta}{2\,h}$, а погрешность метода в предположении ограниченности третьей производной получена выше. Окончательно имеем $Err \leqslant \frac{\Delta}{h} + \frac{h^2}{6}\, M_3$.

Зависимость такого рода при малых h наблюдается при численных экспериментах: при уменьшении h сначала погрешность квадратично убывает, а затем линейно растет; начиная с некоторого h ошибка может стать больше, чем сама производная $f'(x)$. Здесь эффект пропадания значащих цифр (см. 1.5) усиливается за счет деления на малую величину. $\triangleright$

Ответ: $Err \leqslant \frac{\Delta}{h} + \frac{h^2}{6}\, M_3$.

1.7. Найти абсолютную погрешность вычисления суммы $S = \sum\limits_{j=1}^{n} x_j$, где все x_j — числа одного знака.

$\triangleleft$ Используя аксиому

$$fl(a+b) = (a+b)(1+\eta), \quad |\eta| \leqslant \tfrac{1}{2}\, p^{1-t} ,$$

имеем

$$fl\,(S) = (\ldots((x_1+x_2)(1+\eta_2)+x_3)(1+\eta_3)+\ldots+x_n)(1+\eta_n) =$$
$$= (x_1+x_2)\prod_{j=1}^{n-1}(1+\eta_{j+1}) + x_3\prod_{j=2}^{n-1}(1+\eta_{j+1}) + \ldots + x_n\prod_{j=n-1}^{n-1}(1+\eta_{j+1}) .$$

Перепишем полученное выражение в виде

$$fl\,(S) = \sum_{j=1}^{n} x_j\,(1+E_j) ,$$

где для модулей E_j справедливы равенства

$$|E_1| = \tfrac{n-1}{2}\, p^{1-t} + O(p^{2(1-t)}) ,$$

$$|E_i| = \left| \prod_{j=i-1}^{n-1}(1+\eta_{j+1}) \right| = \tfrac{n+1-i}{2}\, p^{1-t} + O(p^{2(1-t)})$$

при $2 \leqslant i \leqslant n$.

Найденное представление означает, что суммирование чисел на компьютере в режиме с плавающей точкой эквивалентно точному суммированию с относительным возмущением E_j в слагаемом x_j. При этом относительные возмущения неодинаковы: они максимальны в первых слагаемых и минимальны в последних. Абсолютная погрешность Δ вычисления суммы равна $\Delta = \sum\limits_{j=1}^{n} |x_j|\,|E_j|$. Оценки E_j не зависят от x_j, поэтому в общем случае погрешность Δ будет наименьшей, если числа суммировать в порядке возрастания их абсолютных значений начиная с наименьшего. $\triangleright$

Ответ: $\Delta = \sum\limits_{j=1}^{n} |x_j|\,|E_j|$.

1.8. Пусть вычисляется сумма $\sum\limits_{j=1}^{10^6} \frac{1}{j^2}$. Какой алгоритм $S_0 = 0$, $S_n = S_{n-1} + \frac{1}{n^2}$, $n = 1, \ldots, 10^6$, или $R_{10^6+1} = 0$, $R_{n-1} = R_n + \frac{1}{n^2}$, $n = 10^6, \ldots, 1$, $\tilde{S}_{10^6} = R_0$, следует использовать, чтобы суммарная вычислительная погрешность была меньше?

Ответ: следует воспользоваться вторым алгоритмом (см. решение 1.7).

1.9. Можно ли непосредственными вычислениями проверить, что ряд $\sum\limits_{j=1}^{\infty} \frac{1}{j}$ расходится?

1.10. Предложить способ вычисления суммы, состоящей из слагаемых одного знака, минимизирующий влияние вычислительной погрешности.

$\triangleleft$ Рассмотрим оценки величин E_j из 1.7. Имеем

$$|E_1| = \frac{n-1}{2}\, p^{1-t} + O(p^{2(1-t)})\,,$$

$$|E_i| = \frac{n+1-i}{2}\, p^{1-t} + O(p^{2(1-t)})\,,\ 2 \leqslant i \leqslant n\,.$$

Из этих оценок следует, что $\left|\frac{E_1}{E_n}\right| \approx n$, т. е. первое слагаемое вносит возмущение примерно в n раз большее, чем последнее. Неравноправие слагаемых объясняется тем, что в образовании погрешностей каждое слагаемое участвует столько раз, сколько суммируются зависящие от него частичные суммы.

Влияние всех слагаемых можно уравнять с помощью следующего приема. Пусть для простоты количество слагаемых равно $n = 2^k$. На первом этапе разобьем близкие слагаемые x_j на пары и сложим каждую из них. При этом в каждое слагаемое вносится относительное возмущение одного порядка. Далее будем складывать уже полученные суммы. Для этого повторяем процесс разбиения и попарного суммирования до тех пор, пока получающиеся суммы не превратятся в одно число (степень двойки 2^k

нужна только здесь). Абсолютная погрешность по-прежнему имеет вид $\Delta = \sum\limits_{j=1}^{n} |x_j|\,|\tilde{E}_j|$, но теперь для всех $\tilde{E}_j$ справедлива оценка

$$\left|\tilde{E}_j\right| = \frac{1+\log_2 n}{2}\; p^{1-t} + O(p^{2(1-t)})\,,\; 1 \leqslant j \leqslant n\,.$$

Таким образом, меняя только порядок суммирования можно уменьшить оценку погрешности примерно в $\frac{n}{\log_2 n}$ раз. Значения $\tilde{E}_j$ отличаются от E_j в силу другого порядка суммирования. $\triangleright$

1.11. Предложить способ вычисления знакопеременной суммы, минимизирующий влияние вычислительной погрешности.

1.12. Пусть значение многочлена $P_n(x) = a_0 + a_1 x + ... + a_n x^n$ вычисляется в точке $x = 1$ по схеме Горнера:

$$P_n(x) = a_0 + x(a_1 + x(...(a_{n-1} + a_n x)...))\,.$$

Какую погрешность можно ожидать в результате, если коэффициенты округлены с погрешностью η?

У к а з а н и е. Воспользоваться решением 1.7, учитывая незнакоопределенность a_i, и с точностью до слагаемых $O\left(\eta^2\right)$ получить

$$|P_n(1) - P_n^*(1)| \leqslant n\,\eta\,(|a_0| + |a_1| + ... + |a_n|)\,.$$

1.13. Оценить погрешность вычисления скалярного произведения двух векторов $S = \sum\limits_{j=1}^{n} x_j y_j$, если их компоненты округлены с погрешностью η.

О т в е т: с точностью до слагаемых $O\left(\eta^2\right)$ имеем $|S - S^*| \leqslant n\,\eta\,\|x\|_2\|y\|_2$, где $\|z\|_2^2 = \sum\limits_{j=1}^{n} z_j^2$.

1.14. Пусть вычисляется величина $S = a_1 x_1 + ... + a_n x_n$, где коэффициенты a_i округлены с погрешностью η. Оценить погрешность вычисления S при условии, что $x_1^2 + ... + x_n^2 = 1$.

О т в е т: с точностью до слагаемых $O\left(\eta^2\right)$ имеем $|S - S^*| \leqslant n\,\eta\,\|a\|_2$, где $\|a\|_2^2 = \sum\limits_{j=1}^{n} a_j^2$.

1.15. Для элементов последовательности

$$I_n = \int\limits_0^1 x^n \mathrm{e}^{x-1} dx$$

справедливо точное рекуррентное соотношение $I_n = 1 - n\,I_{n-1}$, $I_1 = \frac{1}{\mathrm{e}}$. Можно ли его использовать для приближенного вычисления интегралов, считая, что ошибка округления допускается только при вычислении I_1?

$\lhd$ Пусть в результате округления значения I_1 получено значение I_1^*, использование которого приводит к величинам $I_n^* = 1 - n\,I_{n-1}^*$. Для погрешности $\Delta_n = I_n - I_n^*$ имеем соотношение $\Delta_n = -n\,\Delta_{n-1}$, откуда следует $\Delta_n = (-1)^{n+1} n!\,\Delta_1$. Полученная формула гарантирует факториальный рост погрешности и ее знакопеременность. Учитывая, что точные значения удовлетворяют неравенству

$$0 < I_n < \int\limits_0^1 x^n dx = \frac{1}{n+1}\,,$$

получим, что начиная с некоторого n величина погрешности существенно больше искомого результата. Алгоритмы такого рода называются *неустойчивыми*. $\rhd$

1.16. Можно ли использовать для приближенного вычисления интегралов

$$I_n = \int\limits_0^1 x^n \mathrm{e}^{x-1} dx$$

точное рекуррентное соотношение $I_{n-1} = \dfrac{1 - I_n}{n}$ (в обратную сторону по сравнению с 1.15), считая, что ошибка округления допускается только при вычислении стартового значения I_N? Как выбрать это значение?

О т в е т: да (см. решение 1.15), $I_N \approx 0$ при достаточно больших N.

1.17. Пусть вычисления ведутся по формуле

$$y_{n+1} = 2y_n - y_{n-1} + h^2 f_n\,,$$

где $n = 1, 2, \ldots$; y_0, y_1 заданы точно, $|f_n| \leqslant M$, $\quad h \ll 1$. Какую вычислительную погрешность можно ожидать при вычислении y_n для больших значений n? Улучшится ли ситуация, если вычисления вести по формулам $\dfrac{z_{n+1} - z_n}{h} = f_n\,,\ \dfrac{y_n - y_{n-1}}{h} = z_n\,$?

$\lhd$ Формулы, приведенные в условии, являются численными алгоритмами решения задачи Коши для уравнения $y'' = f(x)$. Рассмотрим модельную задачу $y'' = M\,, y(0) = y'(0) = 0$, имеющую точное решение $y(x) = x^2 \dfrac{M}{2}$. Введем сетку с шагом h: $x_n = n\,h$ и будем искать приближенное решение по формуле

$$y_{n+1} = 2y_n - y_{n-1} + h^2 M\,, \quad n = 1, 2, \ldots\,; \quad y_0 = 0\,, y_1 = h^2 \frac{M}{2}\,.$$

При отсутствии ошибок округлений получим $y_n = (n\,h)^2 \dfrac{M}{2}$, т. е. проекцию точного решения на сетку.

Вычисления приводят к соотношениям

$$y_0^* = 0\,, y_1^* = h^2 \frac{M}{2} + \eta_1,$$

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + h^2 M + \eta_{n+1}\,, \quad n = 1, 2, \ldots\,.$$

Отсюда для погрешности $r_n = y_n^* - y_n$ получим

$$r_{n+1} = 2r_n - r_{n-1} + \eta_{n+1}\,, \quad n = 1, 2, \ldots\,; \quad r_0 = 0\,, r_1 = \eta_1\,.$$

Для простоты вычислений предположим, что все η_n постоянны и равны η, тогда для погрешности справедлива формула $r_n = \eta\frac{n^2+n}{2}$. Сопоставляя точное решение y_n и погрешность, приходим к относительной погрешности порядка $h^{-2}\frac{\eta}{M}$. Требование малости этой величины накладывает ограничение на шаг интегрирования h снизу, так как обычно $\eta \sim p^{1-t}$.

Аналогичные рассуждения для второго способа расчетов приводят к относительной погрешности порядка $h^{-1}\frac{\eta}{M}$, что, в свою очередь, приводит к более слабым ограничениям на h при одном и том же η. Другими словами, используя формулы

$$\frac{z_{n+1} - z_n}{h} = f_n\,, \quad \frac{y_n - y_{n-1}}{h} = z_n\,,$$

как правило, получаем меньшую вычислительную погрешность. $\triangleright$

1.2. Погрешность функции

Пусть искомая величина y является функцией параметров x_j, $j = 1, 2, \ldots, n$: $y = y(x_1, x_2, \ldots, x_n)$. Область G допустимого изменения параметров x_j известна, требуется получить приближение к y и оценить его погрешность. Если y^* — приближенное значение величины y, то *предельной абсолютной погрешностью* называют величину

$$A(y^*) = \sup_{(x_1, x_2, \ldots, x_n) \in G} |y(x_1, x_2, \ldots, x_n) - y^*|\,;$$

при этом *предельной относительной погрешностью* называют величину $R(y^*) = \dfrac{A(y^*)}{|y^*|}$.

1.18. Доказать, что предельная абсолютная погрешность $A(y^*)$ минимальна при

$$y^* = \frac{y_1 + y_2}{2},$$

где $y_1 = \inf\limits_G y(x_1, x_2, \ldots, x_n)$, $y_2 = \sup\limits_G y(x_1, x_2, \ldots, x_n)$.

$\triangleleft$ Используя определения величин y_1 и y_2, выражение для $A(y^*)$ перепишем в виде

$$A(y^*) = \sup_{y(x_1, x_2, \ldots, x_n) \in [y_1, y_2]} |y(x_1, x_2, \ldots, x_n) - y^*|\,,$$

при этом $A(y_1) = A(y_2) = y_2 - y_1$. Обозначим $A = y_2 - y_1$. Так как нас интересует минимальное значение величины $A(y^*)$, то достаточно проанализировать только $y^* \in [y_1, y_2]$. Это следует из того, что для $y^* \notin [y_1, y_2]$

справедливо неравенство $A(y^*) > A$. Введем для y^* параметризацию $y^* = \alpha\, y_1 + (1-\alpha)\, y_2$ с $\alpha \in [0,1]$ и рассмотрим предельную абсолютную погрешность

$$A(y^*) = \sup_{y\in[y_1,y_2]} |y - [\alpha\, y_1 + (1-\alpha)\, y_2]| = \\ = \max\{\alpha\, A(y_1),\ (1-\alpha)\, A(y_2)\} = A\, \max\{\alpha, 1-\alpha\}\,.$$

Минимум величины $\max\{\alpha, 1-\alpha\}$ равен $\frac{1}{2}$ и достигается при $\alpha = \frac{1}{2}$, т. е. минимум $A(y^*)$ имеет место при $y^* = \frac{y_1+y_2}{2}$. $\triangleright$

1.19. Показать, что предельная абсолютная погрешность суммы или разности чисел равна сумме их предельных абсолютных погрешностей.

$\triangleleft$ Если известны оценки $|x_j - x_j^*| \leqslant \Delta(x_j^*),\ j = 1,2$, то можно определить область G:

$$G = \left\{(x_1, x_2)\ :\ x_j^* - \Delta(x_j^*) \leqslant x_j \leqslant x_j^* + \Delta(x_j^*)\,,\ j = 1,2\right\}\,.$$

Рассмотрим в этой области функции $y_\pm = x_1 \pm x_2$ и их предельные абсолютные погрешности. Имеем

$$A(y^*) = \sup_{(x_1,x_2)\in G} |y_\pm - y_\pm^*| = \sup_{(x_1,x_2)\in G} |(x_1 \pm x_2) - (x_1^* \pm x_2^*)| \leqslant \\ \leqslant \sum_{j=1}^{2} \sup_{x_j} |x_j - x_j^*| = \Delta(x_1^*) + \Delta(x_2^*)\,. \qquad \triangleright$$

1.20. Показать, что предельная относительная погрешность произведения или частного с точностью до членов второго порядка малости равна сумме предельных относительных погрешностей.

$\triangleleft$ Если известны оценки $\dfrac{|x_j - x_j^*|}{|x_j^*|} \leqslant \delta(x_j^*),\ j = 1,2$, то можно определить область G:

$$G = \left\{(x_1, x_2)\,:\ x_j^* - \Delta(x_j^*) \leqslant x_j \leqslant x_j^* + \Delta(x_j^*)\,,\ j = 1,2\right\},$$

где $\Delta(x_j^*) = |x_j^*|\, \delta(x_j^*)$. Рассмотрим в этой области функцию $y = x_1 x_2$ и ее предельную относительную погрешность

$$R(y^*) = \frac{A(y^*)}{|y^*|} = \frac{1}{|x_1^* x_2^*|} \sup_{(x_1,x_2)\in G} |x_1 x_2 - x_1^* x_2^*| \leqslant \\ \leqslant \frac{1}{|x_1^* x_2^*|}\left(\Delta(x_1^*)x_2^* + \Delta(x_2^*)x_1^* + \Delta(x_1^*)\Delta(x_2^*)\right)\,.$$

Отбрасывая члены второго порядка малости, получим

$$R(y^*) \leqslant \frac{\Delta(x_1^*)}{|x_1^*|} + \frac{\Delta(x_2^*)}{|x_2^*|} = \delta(x_1^*) + \delta(x_2^*)\,.$$

Аналогично рассматривается случай функции $y = \dfrac{x_1}{x_2}$. $\triangleright$

1.21. Пусть $y = y(x_1, x_2, \ldots, x_n)$ — непрерывно дифференцируемая функция. Положим

$$A_{\sup}(y^*) = \sum_{j=1}^{n} B_j \Delta(x_j^*)\,, \quad \text{где } B_j = \sup_G \left| \frac{\partial y(x_1, x_2, \ldots, x_n)}{\partial x_j} \right| ;$$

$$A_{\mathrm{lin}}(y^*) = \sum_{j=1}^{n} b_j \Delta(x_j^*)\,, \quad \text{где } b_j = \left| \frac{\partial y(x_1, x_2, \ldots, x_n)}{\partial x_j} \right|_{\mathbf{x}=(x_1^*, x_2^*, \ldots, x_n^*)}$$

Доказать, что $A(y^*) \leqslant A_{\sup}(y^*)$, и если величина $\rho = \left(\sum\limits_{j=1}^{n} \Delta^2(x_j^*) \right)^{1/2}$ мала, то справедливо равенство $A_{\sup}(y^*) = A_{\mathrm{lin}}(y^*) + o(\rho)$.

$\lhd$ Используя формулу конечных приращений Лагранжа, получим

$$y(x_1, x_2, \ldots, x_n) - y^* = \sum_{j=1}^{n} b_j(\theta)(x_j - x_j^*)\,,$$

где

$$b_j(\theta) = \left. \frac{\partial y(x_1, x_2, \ldots, x_n)}{\partial x_j} \right|_{\mathbf{x}=\mathbf{x}(\theta)}\,,$$

$$\mathbf{x}(\theta) = (x_1^* + \theta_1(x_1 - x_1^*), \ldots, x_n^* + \theta_n(x_n - x_n^*))\,,\ \theta_j \in [0, 1]\,.$$

Отсюда следует $A(y^*) \leqslant A_{\sup}(y^*)$, так как $|b_j(\theta)| \leqslant B_j$.

В силу непрерывности производных $\frac{\partial y}{\partial x_j}$ справедливо представление $B_j = |b_j(0)| + o(1)$ при $\rho \to 0$. Поэтому величину $A_{\sup}(y^*)$ можно записать в виде $A_{\sup}(y^*) = A_{\mathrm{lin}}(y^*) + o(\rho)$, так как $b_j = |b_j(0)|$.

На практике часто используют, вообще говоря, неверную «оценку» $|y(x_1, x_2, \ldots, x_n) - y^*| \leqslant A_{\mathrm{lin}}(y^*)$, называемую *линейной оценкой погрешности*. Величина $A_{\mathrm{lin}}(y^*)$ вычисляется значительно проще, чем $A_{\sup}(y^*)$ или $A(y^*)$, но не следует забывать о требуемой малости ρ. $\rhd$

1.22. Пусть $y = x^{10}$, $x^* = 1$ и задано: 1) $\Delta(x^*) = 0{,}001$; 2) $\Delta(x^*) = 0{,}1$. Вычислить величины $A_{\sup}(y^*)$, $A_{\mathrm{lin}}(y^*)$, $A(y^*)$.

$\lhd$ 1) Здесь $y^* = 1$, $\frac{\partial y}{\partial x} = 10 \cdot x^9$, $b(0) = 10$. Пусть $\Delta(x^*) = 0{,}001$, тогда

$$B = \sup_{|x-1| \leqslant 0{,}001} |10 \cdot x^9| = 10{,}09\ldots\,,$$

$$A_{\sup}(y^*) = B\,\Delta(x^*) = 0{,}01009\ldots\,,$$

$$A_{\mathrm{lin}}(y^*) = |b(0)|\,\Delta(x^*) = 0{,}01\,,$$

$$A(y^*) = \sup_{|x-1| \leqslant 0{,}001} |x^{10} - 1| = 1{,}001^{10} - 1 = 0{,}010045\ldots\,.$$

В этом случае верхняя оценка, предельно точная оценка и линейная оценка отличаются несущественно.

2) Здесь

$$B = \sup_{|x-1|\leqslant 0,1} |10\cdot x^9| = 10\cdot(1,1)^{10} = 23,\dots,$$

$$A_{\sup}(y^*) = B\,\Delta(x^*) = 2,3\dots,$$

$$A_{\mathrm{lin}}(y^*) = |b(0)|\,\Delta(x^*) = 1,$$

$$A(y^*) = \sup_{|x-1|\leqslant 0,1} |x^{10}-1| = (1,1)^{10}-1 = 1,5\dots.$$

Различие между рассматриваемыми величинами в этом случае более заметно. $\triangleright$

1.23. Получить линейную оценку погрешности функции, заданной неявно уравнением $F(y, x_1, \dots, x_n) = 0$.

$\triangleleft$ Дифференцируя по x_j, имеем $\dfrac{\partial F}{\partial y}\,\dfrac{\partial y}{\partial x_j} + \dfrac{\partial F}{\partial x_j} = 0$, откуда $\dfrac{\partial y}{\partial x_j} = -\dfrac{\partial F}{\partial x_j}\left(\dfrac{\partial F}{\partial y}\right)^{-1}$. При фиксированных $x_1^*, \dots, x_n^*$ можно найти y^* как решение нелинейного уравнения $F(y, x_1^*, \dots, x_n^*) = 0$ с одним неизвестным y. Далее вычисляем значения $b_j = -\dfrac{\partial F}{\partial x_j}\left(\dfrac{\partial F}{\partial y}\right)^{-1}\Big|_{(y^*, x_1^*, \dots, x_n^*)}$, приводящие к искомой величине $A_{\mathrm{lin}}(y^*) = \sum\limits_{j=1}^{n} |b_j|\Delta(x_j^*)$. $\triangleright$

1.24. Пусть y^* — простой (не кратный!) корень уравнения $y^2 + b\,y + c = 0$, вычисленный при заданных приближенных значениях коэффициентов b^*, c^*, и известны погрешности $\Delta(b^*), \Delta(c^*)$. Доказать, что

$$A_{\mathrm{lin}}(y^*) = \frac{|y^*|\Delta(b^*) + \Delta(c^*)}{|2y^* + b^*|}\ .$$

У к а з а н и е. Воспользоваться решением 1.23, где $F(y, b, c) \equiv y^2 + b\,y + c = 0$ — неявная функция, и вычислить следующие величины:

$$b_1 = -\ \frac{\partial F}{\partial b}\left(\frac{\partial F}{\partial y}\right)^{-1}\Big|_{(y^*, b^*, c^*)} = -\ \frac{y^*}{2\,y^* + b^*}\ , \quad b_2 = -\ \frac{1}{2\,y^* + b^*}\ .$$

1.25. Показать, что если уравнение из 1.24 имеет кратный корень, то погрешность приближенного значения корня имеет порядок $O(\sqrt{\rho})$, где $\rho = \left(\Delta^2(b^*) + \Delta^2(c^*)\right)^{1/2} \ll 1$.

$\triangleleft$ Пусть уравнение $F(y, b, c) \equiv y^2 + b\,y + c = 0$ имеет y^* — двухкратный корень при $b = b^*$, $c = c^*$. Разложим F в ряд Тейлора в окрестности точки (y^*, b^*, c^*):

$$F(y, b, c) = F(y^*, b^*, c^*) + F_y(y^*, b^*, c^*)(y - y^*) + F_b(y^*, b^*, c^*)(b - b^*) +$$

$$+F_c(y^*, b^*, c^*)(c - c^*) + \frac{1}{2}\,F_{yy}(y^*, b^*, c^*)(y - y^*)^2 + o(\rho) = 0\,.$$

Из условия имеем

$$F(y^*, b^*, c^*) = F_y(y^*, b^*, c^*) = 0\,, \quad \frac{1}{2}\,F_{yy}(y^*, b^*, c^*) = 1\,,$$

что приводит к неравенству

$$(y-y^*)^2 \leqslant |F_b(y^*,b^*,c^*)|\,|b-b^*| + |F_c(y^*,b^*,c^*)|\,|c-c^*| + o(\rho)\,,$$

т. е. $|y-y^*| = O(\sqrt{\rho})$. ▷

1.26. Показать, что в случае, когда алгебраическое уравнение $\sum\limits_{i=0}^{N} a_i y^i = 0$ имеет корень кратности n, погрешность значения корня, вычисленного при заданных приближенных значениях коэффициентов a_i^* с известными погрешностями $\Delta(a_i^*)$, имеет порядок $O(\rho^{1/n})$, где $\rho = \left(\sum\limits_{i=0}^{N} \Delta^2(a_i^*)\right)^{1/2}$.

У к а з а н и е. Воспользоваться решением 1.25.

1.27. Имеется приближение y^* к простому корню уравнения $f(y) = 0$. Вывести приближенное равенство $y - y^* \approx -\dfrac{f(y^*)}{f'(y^*)}$.

◁ Рассмотрим более общее уравнение $f(y) = a$ и вычислим $a^* = f(y^*)$. При малых $|y^* - y|$ из равенства $f(y) - f(y^*) = a - a^*$ следует, что $f'(y^*)(y - y^*) \approx a - a^*$, откуда получаем

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

Заметим, что $f'(y^*) \neq 0$ в силу того, что y^* — простой корень. Полагая $a = 0$ (по условию), приходим к искомой формуле. ▷

1.28. С каким минимальным числом верных знаков надо взять $\lg 2$ для того, чтобы вычислить корни уравнения $y^2 - 2\,y + \lg 2 = 0$ с четырьмя верными знаками?

◁ Уточним условие. Если $\lg 2 = 0,30102999566\ldots$, то корни принимают значения $y_1 = 1,83604425979\ldots$ и $y_2 = 0,16395574020\ldots$. Требуется найти приближение к числу $\lg 2$, обеспечивающее значения корней $y_1^* = 1,836$ и $y_2^* = 0,164$. Теперь воспользуемся решением 1.24 при $b = -2$, $\Delta(b^*) = 0$ и $c = \lg 2$. После подстановки в $A_{\lim}(y^*) = \dfrac{|y^*|\Delta(b^*) + \Delta(c^*)}{|2y^* + b^*|}$ имеем

$$A_{\lim}(y_{1,2}^*) = \frac{\Delta(c^*)}{2\sqrt{1-c^*}} = \Delta(c^*)\cdot 0,5980544\ldots\,.$$

Из этой формулы следует: если требуется в решении получить n верных знаков, то достаточно в c^* взять также n верных знаков, так как постоянная, связывающая величины погрешностей, не превосходит единицы. Таким образом, требуется взять $\lg 2$ с четырьмя верными знаками, т. е. $\lg 2 \approx 0,301$.

Если вычисления провести аккуратно, то при $\lg 2 \approx 0,301$ получим $y_1^* = 1,83606\cdots \approx 1,836$ и $y_2 = 0,16393\cdots \approx 0,164$. Меньшее количество верных знаков брать нельзя: при $\lg 2 \approx 0,30$ имеем $y_1^* = 1,83666\cdots \approx 1,837$ и $y_2 = 0,16333\cdots \approx 0,163$. ▷

1.29. Пусть ограниченные по модулю величиной M коэффициенты уравнения $a\,y^2 + b\,y + c = 0$ заданы с одинаковой относительной погрешностью δ. Найти максимальную абсолютную (относительную) погрешность, с которой могут вычисляться их корни.

Указание. Воспользоваться решениями 1.24 и 1.25.

1.30. Найти приближенное значение интеграла $I_{100} = \int\limits_0^{2\pi} \sin^{100} x dx$ с относительной погрешностью не более 10%.

Указание. Вывести по индукции с помощью интегрирования по частям формулу для точного значения интеграла $I_{100} = \frac{100!\,2\pi}{2^{100}(50!)^2}$, затем применить формулу Стирлинга (см. указание к 5.69) с учетом 1.20.

Ответ: с заданной погрешностью $I_{100} \approx \frac{1}{2}$.

Глава 2

Разностные уравнения

Пусть неизвестная функция y и заданная функция f являются функциями одного целочисленного аргумента. Тогда линейное уравнение

$$a_0 y(k) + a_1 y(k+1) + \cdots + a_n y(k+n) = f(k), \quad k = 0, 1, 2, \ldots ,$$

где a_i, $i = 0, 1, \ldots, n$ — постоянные коэффициенты и $a_0 \neq 0$, $a_n \neq 0$, называют *линейным разностным уравнением n-го порядка с постоянными коэффициентами*. Если в этом уравнении положить $y(k+i) = y_{k+i}$ и $f(k) = f_k$, то оно принимает вид

$$a_0 y_k + a_1 y_{k+1} + \cdots + a_n y_{k+n} = f_k, \quad k = 0, 1, 2, \ldots .$$

Для однозначного определения решения требуется задать n условий, например,

$$y_i = b_i , \quad i = 0, 1, \ldots , n-1 .$$

Как в постановках задач, так и в методах решения, имеется глубокая аналогия между рассмотренным разностным уравнением и обыкновенным дифференциальным уравнением с постоянными коэффициентами

$$\tilde{a}_0 y(x) + \tilde{a}_1 y'(x) + \cdots + \tilde{a}_n y^{(n)}(x) = \tilde{f}(x) .$$

2.1. Однородные разностные уравнения

Если в разностном уравнении правая часть f_k равна нулю, то уравнение называют *однородным*. Напомним, как ищется общее решение однородного дифференциального уравнения с постоянными коэффициентами. Положим $y(x) = \exp(\lambda x)$. Подставляя это выражение в дифференциальное уравнение и сокращая на $\exp(\lambda x)$, получим характеристическое уравнение

$$\tilde{p}(\lambda) \equiv \sum_{j=0}^{n} \tilde{a}_j \lambda^j = 0.$$

Если $\lambda_1, \ldots, \lambda_r$ — различные корни этого уравнения кратности σ_1, …, σ_r соответственно, то общее решение можно записать в виде

$$y(x) = c_{11} \mathrm{e}^{\lambda_1 x} + c_{12} x \mathrm{e}^{\lambda_1 x} + \cdots + c_{1\sigma_1} x^{\sigma_1 - 1} \mathrm{e}^{\lambda_1 x} + \cdots$$
$$\cdots + c_{r1} \mathrm{e}^{\lambda_r x} + c_{r2} x \mathrm{e}^{\lambda_r x} + \cdots + c_{r\sigma_r} x^{\sigma_r - 1} e^{\lambda_r x} ,$$

где c_{ij} — произвольные постоянные.

Аналогично ищется решение разностного уравнения. Положим $y_k = \mu^k$. Подставляя это выражение в разностное уравнение и сокращая на μ^k, получим характеристическое уравнение

$$p(\mu) \equiv \sum_{j=0}^{n} a_j \mu^j = 0 .$$

Пусть $\mu_1, \ldots, \mu_r$ — его различные корни, $\sigma_1, \ldots, \sigma_r$ — их кратности. Тогда общее решение однородного разностного уравнения имеет вид

$$y_k = c_{11}\mu_1^k + c_{12}k\mu_1^k + \cdots\cdots + c_{1\sigma_1}k^{\sigma_1-1}\mu_1^k + \cdots$$
$$+c_{r1}\mu_r^k + c_{r2}k\mu_r^k + \cdots + c_{r\sigma_r}k^{\sigma_r-1}\mu_r^k\,,$$

где c_{ij} — произвольные постоянные. Таким образом, каждому корню μ кратности σ соответствует набор частных решений вида $\mu^k, k\mu^k, \ldots, k^{\sigma-1}\mu^k$.

2.1. Найти общее решение уравнения $by_{k+1} - cy_k + ay_{k-1} = 0$.

$\lhd$ Найдем корни характеристического уравнения $b\mu^2 - c\mu + a = 0$. Имеем

$$\mu_{1,2} = \frac{c \pm \sqrt{D}}{2b}\ , \quad D = c^2 - 4ab\,.$$

Рассмотрим следующие три случая:

а) $D > 0$, $\mu_1 \neq \mu_2$ — вещественные:

$$y_k = C_1\mu_1^k + C_2\mu_2^k\,.$$

б) $D < 0$, $\mu_{1,2} = \rho e^{\pm i\varphi}$ — комплексно-сопряженные.

Здесь $\rho = \sqrt{\frac{a}{b}}$, $\varphi = \begin{cases} \operatorname{arctg}\left(\frac{\sqrt{|D|}}{c}\right), & \frac{c}{b} > 0, \\ \pi - \operatorname{arctg}\left(\frac{\sqrt{|D|}}{c}\right), & \frac{c}{b} < 0, \\ \frac{\pi}{2}, & c = 0. \end{cases}$

При этом $y_k = \rho^k(C_1\cos k\varphi + C_2\sin k\varphi)$. Так записывают общее действительное решение; для комплексного решения можно использовать формулу из п. а).

в) $D = 0$, $\mu_1 = \mu_2 = \mu$ — кратные. Имеем

$$y_k = C_1\mu^k + C_2k\mu^k.$$

В предыдущих формулах C_1, C_2 — произвольные постоянные. $\rhd$

2.2. Найти общее действительное решение уравнения $y_{k+1} - y_k + 2y_{k-1} = 0$.

Ответ: $y_k = (\sqrt{2})^k(C_1\sin k\varphi + C_2\cos k\varphi)$, $\varphi = \operatorname{arctg}\sqrt{7}$.

2.3. Верно ли, что любое решение разностного уравнения

$$y_{k+1} - 5y_k + 6y_{k-1} = 0$$

удовлетворяет уравнению

$$y_{k+1} - 9y_k + 27y_{k-1} - 23y_{k-2} - 24y_{k-3} + 36y_{k-4} = 0\,?$$

Ответ: да, так как характеристический многочлен второго уравнения делится на характеристический многочлен первого без остатка.

2.4. Пусть φ_k и z_k — частные решения уравнения

$$a_1 y_{k+1} + a_0 y_k + a_{-1} y_{k-1} = 0, \quad a_1, a_{-1} \neq 0\,.$$

Доказать, что определитель матрицы

$$A_k = \begin{pmatrix} \varphi_k & \varphi_{k+1} \\ z_k & z_{k+1} \end{pmatrix}$$

либо равен нулю, либо отличен от нуля для всех k одновременно.

У к а з а н и е. Для определителя $I_k = \det A_k$ справедливо разностное уравнение $I_k = \frac{a_{-1}}{a_1} I_{k-1}$.

Соответствующее утверждение можно обобщить на случай разностных уравнений более высокого порядка. Равенство нулю определителя означает линейную зависимость соответствующих частных решений.

2.5. Найти решение разностной задачи

$$y_{k+4} + 2y_{k+3} + 3y_{k+2} + 2y_{k+1} + y_k = 0, \quad y_0 = y_1 = y_3 = 0,\ y_2 = -1\,.$$

У к а з а н и е. Характеристическое уравнение имеет следующий вид $(\mu^2 + \mu + 1)^2 = 0$. Отсюда получим

$$y_k = \frac{2(k-1)}{\sqrt{3}} \sin\frac{2\pi k}{3}\ .$$

2.6. Показать, что для чисел Фибоначчи

$$f_{k+1} = f_k + f_{k-1}, \quad f_0 = 0, \quad f_1 = 1$$

справедливо равенство

$$f_k f_{k+2} - f_{k+1}^2 = (-1)^{k+1}, \quad k = 0, 1, 2, \ldots\ .$$

У к а з а н и е. Формула для чисел Фибоначчи имеет вид

$$f_k = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^k - \left(\frac{1-\sqrt{5}}{2} \right)^k \right].$$

2.7. Вычислить определитель порядка k:

$$\Delta_k = \det \begin{pmatrix} b & c & 0 & . & . & . & 0 & 0 \\ a & b & c & 0 & . & . & . & 0 \\ 0 & a & b & c & 0 & . & . & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & . & . & . & 0 & a & b & c \\ 0 & . & . & . & . & 0 & a & b \end{pmatrix}.$$

$\lhd$ Разлагая определитель Δ_k по первой строке, получим следующую разностную задачу:

$$\Delta_k = b\Delta_{k-1} - ac\Delta_{k-2}\,, \quad \Delta_1 = b\,, \Delta_2 = b^2 - ac\,.$$

Отсюда формально находим $\Delta_0 = 1$, что упрощает последующие выкладки.

Найдем корни характеристического уравнения

$$\mu_{1,2} = \frac{b \pm \sqrt{b^2 - 4ac}}{2}\,.$$

Рассмотрим следующие два случая.

а) $D = \sqrt{b^2 - 4ac} \neq 0$, тогда

$$\Delta_k = C_1 \left(\frac{b-D}{2}\right)^k + C_2 \left(\frac{b+D}{2}\right)^k .$$

Из начальных условий $\Delta_0 = 1$, $\Delta_1 = b$ получаем линейную систему

$$C_1 + C_2 = 1, \quad \frac{C_1}{2}\,(b-D) + \frac{C_2}{2}\,(b+D) = b\,,$$

решение которой имеет вид $C_2 = \frac{1}{2}\left(1 + \frac{b}{D}\right)$, $C_1 = \frac{1}{2}\left(1 - \frac{b}{D}\right)$. Для случая ненулевого дискриминанта

$$\Delta_k = \frac{\left(b + \sqrt{b^2 - 4ac}\right)^{k+1} - \left(b - \sqrt{b^2 - 4ac}\right)^{k+1}}{2^{k+1}\sqrt{b^2 - 4ac}}\,.$$

б) $D = \sqrt{b^2 - 4ac} = 0$, тогда

$$\Delta_k = C_1 \left(\frac{b}{2}\right)^k + C_2 k \left(\frac{b}{2}\right)^k .$$

Из начальных условий получаем линейную систему

$$C_1 = 1, \quad C_1\,\frac{b}{2} + C_2\,\frac{b}{2} = b\,,$$

решение которой $C_1 = C_2 = 1$. Для случая нулевого дискриминанта

$$\Delta_k = \left(\frac{b}{2}\right)^k (1+k)\,.$$

Данное решение можно получить из вида Δ_k для $D \neq 0$ предельным переходом при $4ac \to b^2$. $\triangleright$

2.8. Используя разностное уравнение, записать формулу для вычисления интеграла

$$I_k(\alpha) = \frac{1}{\pi} \int\limits_0^\pi \frac{\cos(kx) - \cos(k\alpha)}{\cos x - \cos\alpha}\; dx\,,$$

где α — параметр из отрезка $[0, \pi]$.

У к а з а н и е. Можно показать, что

$$I_{k-1} + I_{k+1} = 2I_k \cos\alpha\,, \quad I_0 = 0, \quad I_1 = 1,$$

откуда для $0 < \alpha < \pi$ следует формула $I_k(\alpha) = \frac{\sin k\alpha}{\sin\alpha}$. Для оставшихся значений корни характеристического уравнения кратные, поэтому формула имеет другой вид.

2.9. Найти решение разностного уравнения

$$y_{k+2} - y_{k+1} + 2\,y_k - y_{k-1} + y_{k-2} = 0\,, \quad 2 \leqslant k \leqslant N-2\,,$$

удовлетворяющее следующим *краевым* условиям:

$$\begin{gathered} 2\,y_2 - y_1 + y_0 = 2\,, \\ y_3 - y_2 + y_1 - y_0 = 0\,, \\ y_{N-3} - y_{N-2} + y_{N-1} - y_N = 0\,, \\ 2\,y_{N-2} - y_{N-1} + y_N = 0\,. \end{gathered}$$

$\lhd$ Характеристическое уравнение имеет вид

$$\mu^4 - \mu^3 + 2\,\mu^2 - \mu + 1 = (\mu^2 - \mu + 1)\,(\mu^2 + 1)\,.$$

Следовательно, общее решение можно записать так:

$$y_k = C_1 \cos\tfrac{\pi}{3}\,k + C_2 \sin\tfrac{\pi}{3}\,k + C_3 \cos\tfrac{\pi}{2}\,k + C_4 \sin\tfrac{\pi}{2}\,k\,.$$

Для определения постоянных воспользуемся краевыми условиями

$$D \begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

с матрицей D следующего вида:

$$\begin{pmatrix} \cos\left(\frac{2\,\pi}{3}\right) & \sin\left(\frac{2\,\pi}{3}\right) & -1 & -1 \\ 1 & 0 & 0 & 0 \\ \cos\left(\frac{N\,\pi}{3}\right) & \sin\left(\frac{N\,\pi}{3}\right) & 0 & 0 \\ \cos\frac{(N-2)\,\pi}{3} & \sin\frac{(N-2)\,\pi}{3} & -\left(\cos\frac{N\,\pi}{2} + \sin\frac{N\,\pi}{2}\right) & \cos\frac{N\,\pi}{2} - \sin\frac{N\,\pi}{2} \end{pmatrix}$$

Определитель этой системы равен $-2\sin\left(\frac{N\,\pi}{3}\right)\cos\left(\frac{N\,\pi}{2}\right)$ и отличен от нуля, если N четное, но не кратное 3. В этом случае $C_1 = C_2 = 0\,, \quad C_3 = C_4 = -1\,.$ $\rhd$

О т в е т: если N четное, но не кратное 3, то решение имеет вид

$$y_k = -\left(\cos\tfrac{\pi}{2}\,k + \sin\tfrac{\pi}{2}\,k\right)\,, \quad 0 \leqslant k \leqslant N\,.$$

В противном случае решение либо не существует, либо оно не единственное.

2.10. Предложить удобную форму записи решения уравнения

$$y_{k+1} - 2\,p\,y_k + y_{k-1} = 0\,, \quad k = 1, 2, \ldots\,; \quad p > 0\,.$$

О т в е т: при $p < 1$ положим $p = \cos\alpha$ $(\alpha \neq 0)$, тогда $y_k = C_1 \cos k\alpha + C_2 \sin k\alpha$. При $p > 1$ положим $p = \cosh\alpha$, тогда $y_k = C_1 \cosh k\alpha + C_2 \sinh k\alpha$. При $p = 1$ имеем $y_k = C_1 + C_2 k$.

2.11. Показать, что если $-1 < \lambda < 1$, то любое решение разностного уравнения

$$y_{k+1} - 2\lambda y_k + y_{k-1} = 0$$

ограничено при $k \to \infty$. Если λ — любое комплексное число, не принадлежащее интервалу действительной оси $-1 < \lambda < 1$, то среди решений этого разностного уравнения имеются неограниченные при $k \to \infty$.

$\lhd$ Если z — корень характеристического уравнения $z^2 - 2\lambda z + 1 = 0$, то $\frac{1}{z}$ — другой его корень. Ограниченность решений разностного уравнения равносильна следующему условию: корни характеристического уравнения различны и лежат на единичной окружности. Поэтому (см. 2.10) решение ограничено, если только $z_{1,2} = \cos\alpha \pm \mathrm{i}\sin\alpha$, $\alpha \neq 0, \pi$. В этом случае $\lambda = \cos\alpha$. $\rhd$

2.12. Найти общее решение уравнения второго порядка: 1) $y_{k+2} - y_{k+1} - 2\,y_k = 0$; 2) $y_{k+2} - 5\,y_{k+1} + 4\,y_k = 0$; 3) $y_{k+2} - 4\,y_{k+1} + 5\,y_k = 0$.

2.13. Найти общее решение уравнения третьего порядка: 1) $y_{k+2} + y_{k+1} + 5\,y_k + 3\,y_{k-1} = 0$; 2) $y_{k+2} - 5\,y_{k+1} + 8\,y_k - 4\,y_{k-1} = 0$.

2.14. Найти общее решение уравнения четвертого порядка: 1) $y_{k+2} + 2\,y_k + y_{k-2} = 0$; 2) $y_{k+4} + y_k = 0$.

2.15. Доказать, что любое решение разностного уравнения

$$y_{k+1} - 12\,y_{k-1} + 2\,y_{k-2} + 27\,y_{k-3} - 18\,y_{k-4} = 0$$

однозначно представимо в виде суммы решений уравнений

$$y_{k+1} - 3\,y_{k-1} + 2\,y_{k-2} = 0 \quad \text{и} \quad y_{k+1} - 9\,y_{k-1} = 0\,.$$

2.16. Найти решение краевой задачи

$$y_{k+1} - y_k + y_{k-1} = 0\,, \quad 1 \leqslant k \leqslant N-1\,,$$
$$y_0 = 1\,, \quad y_N = 0\,.$$

Ответ: если N не кратно 3, то $y_k = \dfrac{\sin\left((N-k)\,\pi\right.}{3)}\sin(Nk\pi/3)/$. В противном случае решения не существует.

2.17. Найти решение системы $a_{k+1} = \frac{a_k + b_k}{2}$, $b_{k+1} = \frac{a_{k+1} + b_k}{2}$, если a_1, b_1 заданы.

Ответ: $a_k = \frac{a_1 + 2b_1}{3} + \frac{2(a_1 - b_1)}{3 \cdot 4^{k-1}}\,, \quad b_k = \frac{a_1 + 2b_1}{3} - \frac{a_1 - b_1}{3 \cdot 4^{k-1}}\,.$

2.18. Найти общее действительное решение уравнения

$$20y_{k-1} - 8y_k + y_{k+1} = 0\,.$$

Ответ: $y_k = \left(\sqrt{20}\right)^k \left(C_1 \sin k\varphi + C_2 \cos k\varphi\right), \quad \varphi = \operatorname{arctg}\left(\frac{1}{2}\right).$

2.19. Найти общее действительное решение уравнения

$$2y_{k-1} - 2y_k + y_{k+1} = 0\,.$$

Ответ: $y_k = \left(\sqrt{2}\right)^k \left(C_1 \sin\left(\frac{k\pi}{4}\right) + C_2 \cos\left(\frac{k\pi}{4}\right)\right).$

2.20. Найти общее действительное решение уравнения

$$26y_{k-1} + 10y_k + y_{k+1} = 0\,.$$

Ответ: $y_k = \left(\sqrt{26}\right)^k \left(C_1 \sin k\varphi + C_2 \cos k\varphi\right), \quad \varphi = \pi + \operatorname{arctg}\left(\frac{1}{5}\right).$

2.21. Найти общее действительное решение уравнения

$$13y_{k-1} + 4y_k + y_{k+1} = 0\,.$$

Ответ: $y_k = \left(\sqrt{13}\right)^k \left(C_1 \sin k\varphi + C_2 \cos k\varphi\right), \quad \varphi = \pi + \operatorname{arctg}\left(\frac{3}{2}\right).$

2.22. Найти решение разностной задачи

$$y_{k+2} + 4y_{k+1} + 4y_k = 0, \quad y_0 = 1,\, y_1 = 4\,.$$

Ответ: $y_k = (-2)^k(1 - 3k)\,.$

2.23. Найти решение разностной задачи

$$y_{k+2} + 3y_{k+1} + 2y_k = 0, \quad y_0 = 2,\, y_1 = 1\,.$$

Ответ: $y_k = (-1)^k(5 - 3 \cdot 2^k)\,.$

2.24. Найти решение разностной задачи

$$y_{k+2} + y_k = 0, \quad y_0 = 2,\, y_1 = 1\,.$$

Ответ: $y_k = 2\cos\left(\frac{\pi k}{2}\right) + \sin\left(\frac{\pi k}{2}\right).$

2.25. Найти решение разностной задачи

$$y_{k+1} - 4y_k + y_{k-1} + 6y_{k-2} = 0, \quad y_0 = 6,\, y_1 = 12,\, y_4 = 276\,.$$

Ответ: $y_k = (-1)^k + 2^{k+1} + 3^{k+1}\,.$

2.26. Найти общее решение уравнения

$$y_{k+4} - 2\,y_{k+3} + 3\,y_{k+2} + 2\,y_{k+1} - 4\,y_k = 0\,.$$

Ответ: $y_k = C_1 + C_2(-1)^k + 2^k \left(C_3 \cos\left(\frac{\pi k}{3}\right) + C_4 \sin\left(\frac{\pi k}{3}\right)\right).$

2.27. Найти общее решение уравнения

$$y_{k+4} - 7\,y_{k+3} + 18\,y_{k+2} - 20\,y_{k+1} + 8\,y_k = 0\,.$$

Ответ: $y_k = C_1 + 2^k \left(C_2 + C_3 k + C_4 k^2\right).$

2.28. Найти общее решение уравнения $y_{k+4} + 8\,y_{k+2} + 16\,y_k = 0\,.$

Ответ: $y_k = 2^k \left[(C_1 + C_2 k)\cos\left(\frac{\pi k}{2}\right) + (C_3 + C_4 k)\sin\left(\frac{\pi k}{2}\right)\right]\,.$

2.29. Вывести и решить разностное уравнение для коэффициентов ряда Тейлора функции $\dfrac{1}{t^2+t+1}$.

Указание. Полагая

$$\frac{1}{t^2+t+1} = f_0 + f_1 t + f_2 t^2 + \cdots + f_m t^m + \ldots\,,$$

найдем

$$1 = (t^2+t+1)\,\left(f_0 + f_1 t + f_2 t^2 + \cdots + f_m t^m + \ldots\right)\,,$$

откуда $f_0 = 1\,, f_0 + f_1 = 0\,, f_{k+2} + f_{k+1} + f_k = 0\,, k \geqslant 0\,.$

2.30. Пусть задана последовательность интегралов

$$I_k = \int\limits_0^\infty x^k \mathrm{e}^{-x} \sin x\, dx, \quad k \geqslant 0\,.$$

Показать, что для целых неотрицательных n справедливо равенство $I_{4n+3} = 0$.

$\lhd$ Заметим, что $I_k = \operatorname{Im}\left[\int\limits_0^\infty x^k \mathrm{e}^{-x+\mathrm{i}x}\, dx\right]$. Обозначим через K_k вещественную часть этого выражения:

$$K_k = \int\limits_0^\infty x^k \mathrm{e}^{-x} \cos x\, dx, \quad k \geqslant 0\,.$$

Интегрируя по частям, имеем систему разностных уравнений

$$I_k = \frac{k}{2}\,(I_{k-1} + K_{k-1})\,, \qquad K_k = \frac{k}{2}\,(-I_{k-1} + K_{k-1})$$

с начальными условиями $I_0 = K_0 = \frac{1}{2}$. Если положить

$$I_k = \frac{k!}{2^k}\; j_k, \quad K_k = \frac{k!}{2^k}\; l_k\,,$$

то исходная система с переменными коэффициентами переходит в систему с постоянными коэффициентами

$$j_k = j_{k-1} + l_{k-1},\; j_0 = \frac{1}{2}\,, \quad l_k = -j_{k-1} + l_{k-1},\; l_0 = \frac{1}{2}\,.$$

Исключая l_k, получим разностное уравнение второго порядка относительно j_k:

$$j_{k+1} - 2j_k + 2j_{k-1} = 0, \quad j_0 = \frac{1}{2}, \quad j_1 = 1.$$

Его решение имеет вид

$$j_k = \frac{1}{2}\left[(1+\mathrm{i})^{k-1} + (1-\mathrm{i})^{k-1}\right], \quad \mathrm{i} = \sqrt{-1}.$$

Отсюда находим

$$I_k = \frac{k!}{2^{k+1}}\left[(1+\mathrm{i})^{k-1} + (1-\mathrm{i})^{k-1}\right].$$

Заметим, что $(1+\mathrm{i})^4 = -4 = (1-\mathrm{i})^4$, следовательно,

$$j_{4n+3} = (-4)^n j_3 = \frac{(-4)^n}{2}\left[(1+2\mathrm{i}+\mathrm{i}^2) + (1-2\mathrm{i}+\mathrm{i}^2)\right] = 0\,,$$

откуда $I_{4n+3} = 0$. $\triangleright$

2.31. Для целых положительных чисел $a_0 > a_1$ наибольший общий делитель находится последовательным делением a_0 на a_1, затем a_1 — на первый остаток и т. д. Указать оценку сверху для числа делений (длину алгоритма Евклида).

$\triangleleft$ Обозначим частное от деления a_i на a_{i+1} через d_i и запишем систему равенств

$$\begin{aligned}
a_0 &= a_1 d_1 + a_2\,,\\
a_1 &= a_2 d_2 + a_3\,,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
a_{m-2} &= a_{m-1} d_{m-1} + a_m\,,\\
a_{m-1} &= a_m d_m\,.
\end{aligned}$$

Наибольшее количество операций деления m имеет место в том случае, когда все $d_1, d_2, \ldots, d_m$ равны единице (доказать почему!). Поэтому введем числа $y_0, y_1, \ldots, y_m$ при условиях $y_0 = 0,\ y_1 = 1,\ \ldots,\ y_{i+1} = y_{i-1} + y_i$, для которых справедливы неравенства

$$a_{m+1} = y_0,\ a_m \geqslant y_1,\ \ldots,\ a_2 \geqslant y_{m-1},\ a_1 \geqslant y_m.$$

Последнее из них можно использовать для определения m, если известно выражение $y_m = f(m)$. Но y_m — числа Фибоначчи, поэтому

$$y_m = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^m - \left(\frac{1-\sqrt{5}}{2}\right)^m\right],$$

т. е. при всех m справедливо неравенство

$$y_m > \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^m - 1$$

или

$$a_1 + 1 > \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^m.$$

Отсюда после логарифмирования имеем

$$m < \frac{\lg(1+a_1) + \lg\sqrt{5}}{\lg\left(\dfrac{1+\sqrt{5}}{2}\right)}\,.$$

Обозначим через p число цифр в a_1. Тогда числитель $\lg\left((1+a_1)\sqrt{5}\right) \approx p$. Поскольку $\lg\left(\dfrac{1+\sqrt{5}}{2}\right) < \dfrac{1}{5}$, получаем $m < 5p$. Это неравенство называют теоремой Ламе. $\triangleright$

2.32. Пусть задано k чисел: $f_0, f_1, \ldots, f_{k-1}$ и построена последовательность

$$f_k = \frac{1}{k}\sum_{j=0}^{k-1} f_j\,, \quad f_{k+1} = \frac{1}{k}\sum_{j=1}^{k} f_j\,, \quad f_{k+2} = \frac{1}{k}\sum_{j=2}^{k+1} f_j\,, \quad \ldots\,.$$

Найти $\lim\limits_{m\to\infty} f_m$.

$\lhd$ Функция f_m удовлетворяет разностному уравнению

$$f_{m+k} = \frac{1}{k}\sum_{j=0}^{k-1} f_{m+j}\,, \tag{2.1}$$

характеристическое уравнение которого имеет вид:

$$\mu^k - \frac{1}{k}\left(\mu^{k-1} + \mu^{k-2} + \cdots + 1\right) = 0\,.$$

Это уравнение имеет один корень, равный единице: $\mu_1 = 1$, остальные корни различны и по модулю меньше единицы: $|\mu_i| < 1\,,\ i = 2, 3, \ldots, k$. Поэтому общее решение f_m уравнения (2.1) имеет вид

$$f_m = C_1 + C_2\mu_2^m + C_3\mu_3^m + \cdots + C_k\mu_k^m\,.$$

Постоянные $C_1, C_2, \ldots, C_k$ находятся из начальных условий:

$$\begin{array}{l} C_1 \;+\; C_2 \qquad\quad +\;\ldots\;+\; C_k \qquad\quad = f_0\,,\\ C_1 \;+\; C_2\mu_2 \quad +\;\ldots\;+\; C_k\mu_k \quad = f_1\,,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ C_1 \;+\; C_2\mu_2^{k-1} \;+\;\ldots\;+\; C_k\mu_k^{k-1} \;= f_{k-1}\,. \end{array}$$

Эта система имеет единственное решение, так как все корни $\mu_i\,,\ i = 2, 3, \ldots, k,$ — простые.

Так как все $|\mu_i| < 1$, то $\lim\limits_{m\to\infty} f_m = C_1$. Для определения C_1, чтобы избежать решения системы относительно C_1, C_2, …, C_k, воспользуемся искусственным приемом. Коэффициент C_1 линейно зависит от начальных данных (доказать почему), т. е. $C_1 = \sum\limits_{j=0}^{k-1} \alpha_j f_j$, где α_j выражаются только через корни характеристического уравнения μ_i, следовательно, от начальных данных f_j не зависят. Напомним, что разностное уравнение k-го порядка однозначно определяется k подряд идущими значениями $f_0, f_1, \ldots, f_{k-1}$, или $f_1, f_2, \ldots, f_k$, или вообще $f_j, f_{j+1}, \ldots, f_{j+k-1}$ при любом $j \geqslant 0$. При этом для рассматриваемого уравнения всегда будем получать одно и то же решение f_m с одними и теми же постоянными $C_1, C_2, \ldots, C_k$. Поэтому можно написать равенство

$$C_1 = \sum_{j=0}^{k-1} \alpha_j f_j = \sum_{j=0}^{k-1} \alpha_j f_{j+1} = \sum_{j=0}^{k-1} \alpha_j f_{j+l}$$

с произвольным фиксированным l. Воспользуемся первым равенством сумм

$$\alpha_0 f_0 + \alpha_1 f_1 + \cdots + \alpha_{k-1} f_{k-1} = \alpha_0 f_1 + \alpha_1 f_2 + \cdots + \alpha_{k-1}\,\frac{f_0 + f_1 + \cdots + f_{k-1}}{k}\,.$$

Учитывая произвольность начальных данных $f_0, f_1, \ldots, f_{k-1}$, приравняем коэффициенты при одинаковых f_i из обеих частей последнего уравнения:

$$\begin{gathered}\alpha_0 = \tfrac{1}{k}\,\alpha_{k-1}\,,\\ \alpha_1 = \alpha_0 + \tfrac{1}{k}\,\alpha_{k-1} = 2\,\alpha_0\,,\\ \alpha_2 = \alpha_1 + \tfrac{1}{k}\,\alpha_{k-1} = 3\,\alpha_0\,,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \alpha_{k-1} = \alpha_{k-2} + \tfrac{1}{k}\,\alpha_{k-1} = k\,\alpha_0\,.\end{gathered}$$

Следовательно,

$$C_1 = \alpha_0 \sum_{j=0}^{k-1} (j+1) f_j\,,$$

и остается определить α_0. Для этого положим $f_0 = f_1 = \cdots = f_{k-1} = 1$; тогда $f_m \equiv 1$, и $\lim\limits_{m\to\infty} f_m = C_1 = 1$. Отсюда имеем $\alpha_0 = \dfrac{2}{k(k+1)}$, и окончательно получаем

$$\lim_{m\to\infty} f_m = C_1 = 2\,\frac{f_0 + 2\,f_1 + 3\,f_2 + \cdots + k\,f_{k-1}}{k(k+1)}\,.$$

Пусть $k = 2$, $f_0 = 1$, $f_1 = 2$, тогда последовательность f_m будет следующая: $1, 2, \frac{3}{2}, \frac{7}{4}, \frac{13}{8}, \ldots$, а ее предел равен $2\,\frac{1+4}{2(2+1)} = \frac{5}{3}$. ▷

2.2. Вспомогательные формулы

Пусть φ_i — функция целочисленного аргумента i. Введем обозначения для разностей первого порядка:

$$\Delta\varphi_i = \varphi_{i+1} - \varphi_i, \quad \nabla\varphi_i = \varphi_i - \varphi_{i-1}.$$

Разности более высокого порядка определяются рекуррентно:

$$\Delta^m\varphi_i = \Delta(\Delta^{m-1}\varphi_i), \quad m \geqslant 1.$$

Рассмотрим ряд полезных соотношений.

Формулы «разностного дифференцирования» произведения. Известна формула дифференцирования произведения функций

$$\frac{d}{dx}\,[u(x)v(x)] = u(x)\,\frac{dv(x)}{dx} + v(x)\,\frac{du(x)}{dx}\,.$$

Аналогичные формулы для разностей имеют вид

$$\begin{aligned}\Delta\,(uv)_i &= u_i\Delta v_i + v_{i+1}\Delta u_i = u_{i+1}\Delta v_i + v_i\Delta u_i,\\ \nabla\,(uv)_i &= u_{i-1}\nabla v_i + v_i\nabla u_i = u_i\nabla v_i + v_{i-1}\nabla u_i.\end{aligned}$$

При их проверке достаточно учесть, что $\Delta\varphi_{i-1} = \nabla\varphi_i$.

2.33. Получить выражения для $\Delta^k\varphi_i$ и $\nabla^k\varphi_i$ в виде линейной комбинации значений φ_j.

Ответ: $\Delta^k\varphi_i = \sum\limits_{j=0}^{k}(-1)^{k-j}C_k^j\varphi_{i+j}$, где C_k^j — биномиальные коэффициенты.

2.34. Найти общее решение уравнения $\Delta^3\varphi_k - 3\Delta\varphi_k + 2\varphi_k = 0$.

2.35. Решить уравнение $\Delta^3\varphi_k = 0$ при начальных условиях $\varphi_0 = \varphi_1 = 0$, $\varphi_2 = 1$.

Разностные аналоги интегрирования по частям. Рассмотрим выражение

$$\int\limits_a^b u'(x)v(x)dx = u(x)v(x)\Big|_a^b - \int\limits_a^b u(x)v'(x)dx$$

и введем суммы

$$(\varphi,\psi) = \sum_{i=1}^{N-1}\varphi_i\psi_i, \quad (\varphi,\psi] = \sum_{i=1}^{N}\varphi_i\psi_i, \quad [\varphi,\psi) = \sum_{i=0}^{N-1}\varphi_i\psi_i$$

— аналоги интеграла $\int\limits_a^b \varphi(x)\psi(x)dx$. С помощью формулы Абеля

$$\sum_{i=0}^{N-1}(a_{i+1}-a_i)b_i = -\sum_{i=0}^{N-1}(b_{i+1}-b_i)a_{i+1} + a_Nb_N - a_0b_0$$

можно показать справедливость формулы суммирования по частям:

$$(\varphi,\Delta\psi) = -(\nabla\varphi,\psi] + \varphi_N\psi_N - \varphi_0\psi_1.$$

2.36. Вычислить сумму $S_N = \sum\limits_{i=1}^{N} i2^i$.

$\lhd$ Положим $u_i = i, \Delta v_i = 2^i$. Имеем

$$v_{i+1} = v_i + 2^i = \sum_{k=0}^{i}2^k + v_0 = 2^{i+1} - 1 + v_0.$$

Чтобы выполнялось условие $v_{N+1} = 0$, достаточно положить $v_0 = 1-2^{N+1}$. Далее применим формулу суммирования по частям

$$\sum_{i=1}^{N} i2^i = \sum_{i=1}^{N} u_i\Delta v_i = -\sum_{i=1}^{N+1} v_i\Delta u_i + u_{N+1}v_{N+1} - u_0v_1 =$$

$$= -\sum_{i=1}^{N+1}\left(2^i - 2^{N+1}\right) = -2\left(2^{N+1}-1\right) + 2^{N+1}(N+1) = (N-1)2^{N+1} + 2. \quad \rhd$$

Ответ: $S_N = (N-1)2^{N+1} + 2$.

2.37. Вычислить сумму $S_N = \sum\limits_{i=1}^{N} ia^i,\ a \neq 1.$

Указание. Положить $u_i = i,\ \Delta v_i = a^i,\ v_i = \dfrac{a^i - a^N}{a-1},\ v_{N+1} = 0.$

Ответ: $S_N = \dfrac{a^{N+1}\,(N(a-1)-1)+a}{(a-1)^2}.$

2.38. Вычислить сумму $S_N = \sum\limits_{i=1}^{N} i(i-1).$

Ответ: $S_N = \dfrac{N^3 - N}{3}.$

Разностные формулы Грина. Формулы

$$\sum_a^b u(x)Lv(x)dx = -\sum_a^b k(x)u'(x)v'(x)dx - \sum_a^b p(x)u(x)v(x)dx + k(x)u(x)v'(x)\big|_a^b,$$

$$\sum_a^b \left[u(x)Lv(x) - Lu(x)v(x)\right]dx = k(x)\left[u(x)v'(x) - u'(x)v(x)\right]\big|_a^b,$$

где $Lv(x) = \left(k(x)v'(x)\right)' - p(x)v(x)$, называют соответственно *первой* и *второй формулами Грина* для оператора L.

2.39. Доказать справедливость соотношений

$$(\varphi, \Delta\nabla\psi) = -(\nabla\varphi, \nabla\psi) + \varphi_{N-1}\nabla\psi_N - \varphi_0\nabla\psi_1,$$
$$(\varphi, \Delta\nabla\psi) - (\Delta\nabla\varphi, \psi) = \varphi_{N-1}\psi_N - \varphi_N\psi_{N-1} + \varphi_1\psi_0 - \varphi_0\psi_1.$$

Указание. Воспользоваться формулой суммирования по частям.

2.40. Вывести формулы Грина для разностного оператора

$$\Lambda\varphi_i = \Delta\left(a_i\nabla\varphi_i\right) - d_i\varphi_i \equiv a_{i+1}\left(\varphi_{i+1} - \varphi_i\right) - a_i\left(\varphi_i - \varphi_{i-1}\right) - d_i\varphi_i.$$

Ответ: $(\psi, \Lambda\varphi) = -(a\nabla\varphi, \nabla\psi] - (d\varphi, \psi) + (a\psi\nabla\varphi)_N - \psi_0(a\nabla\varphi)_1,$
$(\varphi, \Lambda\psi) - (\psi, \Lambda\varphi) = a_N(\psi\nabla\varphi - \varphi\nabla\psi)_N - a_1(\psi_0\nabla\varphi_1 - \varphi_0\nabla\psi_1).$

2.3. Неоднородные разностные уравнения

Пусть y_k^0 — общее решение однородного, y_k^1 — частное решение неоднородного уравнения. Тогда общее решение линейного неоднородного уравнения с постоянными коэффициентами можно представить в виде их суммы

$$y_k = y_k^0 + y_k^1.$$

Если правая часть имеет специальный вид, то частное решение можно найти методом неопределенных коэффициентов. Пусть

$$f_k = \alpha^k\left(P_m(k)\,\cos\beta k + Q_n(k)\,\sin\beta k\right),$$

где $P_m(k)$, $Q_n(k)$ — многочлены степени m и n соответственно. Тогда частное решение ищут в виде

$$y_k^1 = k^s\,\alpha^k\,(R_l(k)\,\cos\beta k + T_l(k)\,\sin\beta k), \tag{2.2}$$

где $s = 0$, если $\alpha e^{\pm i\beta}$ не являются корнями характеристического уравнения, и s равно кратности корня в противном случае; $l = \max(m, n)$ — степень многочленов $R_l(k)$ и $T_l(k)$. Чтобы найти коэффициенты этих многочленов, надо подставить выражение (2.2) в неоднородное уравнение и приравнять коэффициенты при подобных членах.

2.41. Найти частное решение уравнения $2y_k - y_{k+1} = 1 + 2k - k^2$.

$\lhd$ Корень характеристического уравнения $\mu = 2$, поэтому частное решение ищем в виде $y_k^1 = bk^2 + ck + d$. Подставим его в уравнение

$$2bk^2 + 2ck + 2d - [b(k+1)^2 + c(k+1) + d] = 1 + 2k - k^2 \quad \forall\, k.$$

Совпадение коэффициентов при линейно независимых функциях приводит к следующим равенствам:

при k^2 $\quad 2b - b = -1$,
при k^1 $\quad 2c - (2b + c) = 2$,
при $k^0 = 1$ $\quad 2d - (b + c + d) = 1$.

Отсюда имеем: $b = -1$, $c = 0$, $d = 0$, следовательно, $y_k^1 = -k^2$. $\rhd$

2.42. Найти частное решение уравнения $2y_k - y_{k+1} = k\,2^k$.

$\lhd$ Корень характеристического уравнения $\mu = 2$, поэтому частное решение ищем в виде $y_k^1 = 2^k k(bk + c)$. Подставим его в уравнение

$$2^{k+1}(bk^2 + ck) - 2^{k+1}(b(k+1)^2 + c(k+1)) = k\,2^k \quad \forall\, k\,.$$

Совпадение коэффициентов при линейно независимых функциях приводит к следующим равенствам:

при $2^k\,k^2$ $\quad 2b - 2b = 0$,
при $2^k\,k^1$ $\quad 2c - (4b + 2c) = 1$,
при $2^k\,k^0$ $\quad -(2b + 2c) = 0$.

Отсюда имеем: $b = -\frac{1}{4}$, $c = \frac{1}{4}$, следовательно, $y_k^1 = 2^{k-2}(k - k^2)$. $\rhd$

2.43. Найти частное решение уравнения $2y_k - y_{k+1} = \sin k$.

$\lhd$ Корень характеристического уравнения $\mu = 2$, поэтому частное решение ищем в виде $y_k^1 = c\,\sin k + d\,\cos k$. Подставим его в уравнение

$$2(c\,\sin k + d\,\cos k) - (c\,\sin(k+1) + d\,\cos(k+1)) = \sin k \quad \forall\, k\,.$$

Так как $\sin(k+1) = \sin k\,\cos 1 + \cos k\,\sin 1$ и $\cos(k+1) = \cos k\,\cos 1 - \sin k\,\sin 1$, то совпадение коэффициентов при линейно независимых функциях приводит к следующим равенствам:

при $\sin k$ $\quad (2 - \cos 1)\,c + d\,\sin 1 = 1$,
при $\cos k$ $\quad (2 - \cos 1)\,d - c\,\sin 1 = 0$,

следовательно,

$$c = \frac{2-\cos 1}{5-4\cos 1}\,, \quad d = \frac{\sin 1}{5-4\cos 1}$$

и

$$y_k^1 = \frac{2-\cos 1}{5-4\ \cos 1}\sin k + \frac{\sin 1}{5-4\cos 1}\cos k\,. \qquad \triangleright$$

2.44. Найти решение разностной задачи $y_{k+1} - b\,y_k = a^k,\ y_0 = 1\ (a, b \neq 0)$.

О т в е т: корень характеристического уравнения $\mu = b$, поэтому возможны два случая:

при $b \neq a$ имеем $y_k = \frac{a-b-1}{a-b}\ b^k + \frac{1}{a-b}\ a^k$,

при $b = a$ имеем $y_k = a^{k-1}(a+k)$.

2.45. Найти решение разностной задачи $y_{k+1} - y_{k-1} = \frac{1}{k^2-1}$, $y_1 = 0$, $y_2 = 0$.

$\triangleleft$ Преобразуем уравнение к виду

$$y_{k+1} - y_{k-1} = -\frac{1}{2}\left(\frac{1}{k+1} - \frac{1}{k-1}\right).$$

Отсюда находим частное решение $y_k^1 = -\frac{1}{2k}$. Окончательно имеем $y_k = \frac{1}{8}\ (3-(-1)^k) - \frac{1}{2k}$. $\triangleright$

2.46. Найти решение разностной задачи с переменными коэффициентами

$$y_{k+1} - k y_k = 2^k\,k!\,, \quad k \geqslant 0\,.$$

$\triangleleft$ При $k = 0$ из уравнения получим $y_1 = 1$. Запишем исходное уравнение в следующем виде:

$$y_{k+1} = (2^k(k-1)! + y_k)k.$$

Воспользовавшись заменой $y_k = z_k\,(k-1)!$, приходим к разностной задаче для z_k

$$z_{k+1} - z_k = 2^k, \quad z_1 = 1\,.$$

Найдем ее решение: $z_k = 2^k - 1$, следовательно, $y_k = (k-1)!\,(2^k - 1)$. $\triangleright$

2.47. Найти решение нелинейной разностной задачи $y_{k+1} = \frac{y_k}{1+y_k}$, $y_0 = 1$.

$\triangleleft$ Исходное уравнение эквивалентно следующему:

$$y_{k+1} = \frac{1}{1/y_k + 1}\,.$$

Заменяя $y_k = \frac{1}{z_k}$, получаем $z_k = k+1$, откуда $y_k = \frac{1}{k+1}$. $\triangleright$

2.48. Найти решение нелинейной разностной задачи $y_{k+1} = \frac{a\,y_k + b}{c\,y_k + d}$, $y_0 = 1$, при условии $(a-d)^2 + 4\,b\,c > 0$.

$\lhd$ Положим $y_k = \frac{u_k}{v_k}$, тогда

$$\frac{u_{k+1}}{v_{k+1}} = \frac{a\,u_k + b\,v_k}{c\,u_k + d\,v_k}\,.$$

Рассмотрим систему

$$u_{k+1} = a\,u_k + b\,v_k\,, \quad v_{k+1} = c\,u_k + d\,v_k\,,$$

из которой следует уравнение второго порядка

$$v_{k+2} = (a+d)v_{k+1} - (ad - bc)v_k\,.$$

Его характеристическое уравнение имеет вид

$$\mu^2 - \mu(a+d) + ad - bc = 0\,,$$

а корни соответственно равны

$$\mu_{1,2} = \frac{a+d}{2} \pm \sqrt{\frac{(a+d)^2}{4} - (ad - bc)}\,.$$

Из условия на коэффициенты следует, что дискриминант больше нуля, значит, вещественные корни различны $\mu_1 \neq \mu_2$, следовательно, $v_k = A\,\mu_1^k + B\,\mu_2^k$. Из второго уравнения системы получаем:

$$u_k = \frac{v_{k+1} - d\,v_k}{c} = \frac{1}{c}\left[A\,\mu_1^k(\mu_1 - d) + B\,\mu_2^k(\mu_2 - d)\right]\,.$$

Подставим полученные выражения в y_k и разделим числитель и знаменатель на $A\,\mu_1^k$:

$$y_k = \frac{u_k}{v_k} = \frac{\mu_1 - d + K\left(\frac{\mu_2}{\mu_1}\right)^k(\mu_2 - d)}{c\left[1 + K\left(\frac{\mu_2}{\mu_1}\right)^k\right]}\,.$$

Здесь через K обозначена пока неизвестная постоянная ($K = \frac{B}{A}$). Определим ее из начального условия $y_0 = 1$:

$$1 = \frac{\mu_1 - d + K(\mu_2 - d)}{c(1+K)}\,,$$

отсюда $K = -\,\frac{\mu_1 - (c+d)}{\mu_2 - (c+d)}$. $\rhd$

2.49. Найти частное решение уравнения $y_{k+2} - y_{k+1} - 6\,y_k = 4 \cdot 2^k$.

Ответ: $y_k = -\,2^k$.

2.50. Найти решение разностной задачи

1) $y_{k+2} - 4\,y_k = 5 \cdot 3^k\,$, $\quad y_0 = 0\,, y_1 = 1\,,$
2) $y_{k+2} - 4\,y_k = 2^k\,$, $\quad y_0 = 0\,, y_1 = 1\,,$
3) $y_{k+2} - 4\,y_{k+1} + 4\,y_k = 2^k\,$, $\quad y_0 = 0\,, y_1 = 1\,.$

2.51. Найти решение разностной задачи

$$y_{k+4} - \frac{5}{2}\, y_{k+3} + \frac{5}{2}\, y_{k+1} - y_k = 1\,,$$

$$y_0 = 0\,,\ y_1 = 11\,,\ y_2 = -8\,,\ y_3 = 6\,.$$

Ответ: $\mu_1 = 1$ и $\mu_2 = -1$, $y_k = 8 \cdot (-1)^{k-1} + 8 \cdot 2^{-k} - k$.

2.52. Вычислить сумму $S_k = \sum\limits_{n=0}^{k} a_n$, $a_n = (1 + n + n^2)\cos \beta n\,.$

Указание. Решение удовлетворяет разностному уравнению $S_{k+1} - S_k = a_{k+1}$ и начальному условию $S_0 = a_0$.

2.53. Найти решение нелинейного уравнения $y_{k+1} = \dfrac{1}{2 - y_k}$.

◁ Преобразуем исходное уравнение к виду

$$y_{k+1}(1 - y_k) = 1 - y_{k+1}$$

и запишем его в более удобной форме

$$\frac{y_{k+1}}{1 - y_{k+1}} = \frac{1}{1 - y_k}\,.$$

Заменяя $y_k = 1 - \dfrac{1}{z_k}$, получаем разностную задачу для z_k: $z_{k+1} - z_k = 1$. Отсюда

$$y_k = \frac{y_0 + k(1 - y_0)}{1 + k(1 - y_0)}\ . \qquad \triangleright$$

2.54. Найти решение нелинейной разностной задачи $y_{k+1} = 2 - \dfrac{1}{y_k}$, $y_0 = 2\,.$

◁ Преобразуем исходное уравнение к виду

$$y_{k+1} - 1 = \frac{y_k - 1}{y_k};$$

сделав замену $y_k = 1 + \dfrac{1}{z_k}$, получим $z_{k+1} = z_k + 1$, откуда $y_k = \dfrac{k+2}{k+1}$. $\triangleright$

2.55. Найти решение нелинейного уравнения $y_{k+1}^2 - y_k^2 = 1\,,\ k \geqslant 0\,.$

Ответ: $y_k = \sqrt{k + C}$, $C \geqslant 0$.

2.56. Найти решение нелинейного уравнения $y_{k+1}^2 = 2\, y_k\,.$

◁ Прологарифмируем обе части уравнения и выполним замену $z_k = \log y_k$. Получаем уравнение

$$2\, z_{k+1} - z_k = \log 2\,,$$

общее решение которого

$$z_k = C_1 \left(\frac{1}{2}\right)^k + \log 2,$$

следовательно, $y_k = 2\, C^{(1/2)^k}$. $\triangleright$

2.57. Найти решение нелинейной разностной задачи

$$y_k y_{k+2}^3 = y_{k+1}^3 y_{k+3}\,, \quad y_0 = 1,\ y_1 = \mathrm{e}^{-1/2},\ y_2 = \mathrm{e}^{-2}\,.$$

Ответ: $y_k = \mathrm{e}^{-k^2/2}$ (см. решение 2.56).

2.58. Найти решение нелинейной разностной задачи $y_{k+1} = \dfrac{a\,y_k + b}{c\,y_k + d}$, $y_0 = 1$, при условии $(a-d)^2 + 4\,b\,c = 0$.

2.59. Найти частное решение уравнения $\frac{1}{8}\,y_{k-1} - \frac{3}{4}\,y_k + y_{k+1} = \left(\frac{1}{2}\right)^k$.

Ответ: $y_k = 4\,k\,2^{-k}$.

2.60. Найти частное решение уравнения $y_{k+1} - y_k - 12y_{k-1} = 4^k$.

Ответ: $y_k = \frac{k}{7}\ 4^k$.

2.61. Найти частное решение уравнения $3y_{k+1} + 17y_k - 6y_{k-1} = \left(\frac{1}{3}\right)^k$.

Ответ: $y_k = \frac{k}{19}\ 3^{-k}$.

2.62. Найти частное решение уравнения $y_{k+1} - 5y_k + 6y_{k-1} = 2^k$.

Ответ: $y_k = -k\,2^k$.

2.63. Найти общее решение уравнения $y_{k+1} - \frac{5}{2}\,y_k + y_{k-1} = \cos k$.

Ответ: $y_k = C_1\,2^k + C_2\,2^{-k} + \frac{2\cos k}{4\cos 1 - 5}$.

2.64. Найти общее решение уравнения $y_{k+2} - 2y_{k+1} - 3y_k + 4y_{k-1} = k$.

Ответ: $y_k = C_1 + C_2\left(\frac{1+\sqrt{17}}{2}\right)^k + C_3\left(\frac{1-\sqrt{17}}{2}\right)^k - \frac{3}{16}\,k - \frac{1}{8}\,k^2$.

2.65. Найти общее решение уравнения $y_{k+1} + y_k - 5y_{k-1} + 3y_{k-2} = 1$.

Ответ: $y_k = C_1 + C_2\,k + C_3\,(-3)^k + \frac{1}{8}\,k^2$.

2.66. Найти общее решение уравнения $y_{k+1} - 2y_k - 8y_{k-1} = \sin k$.

Ответ: $y_k = C_1\,(-2)^k + C_2\,4^k - \frac{7\cos 1 + 2}{D}\sin k - \frac{9\sin 1}{D}\cos k$, где $D = (2+7\cos 1)^2 + (9\sin 1)^2$.

Отыскание частного решения методом вариации постоянных. Пусть требуется найти частное решение уравнения

$$y_{k+2} + a_k y_{k+1} + b_k y_k = f_k\,, \quad b_k \neq 0\,, \quad k = 0, \pm 1, \pm 2, \ldots\,, \tag{2.3}$$

общее решение которого при $f_k \equiv 0$ имеет вид

$$y_k^0 = C^{(1)} y_k^{(1)} + C^{(2)} y_k^{(2)}\,,$$

где $y_k^{(1)}$ и $y_k^{(2)}$ — линейно независимые функции.

Будем искать частное решение y_k^1 в виде

$$Y_k = C_k^{(1)} y_k^{(1)} + C_k^{(2)} y_k^{(2)}\,, \tag{2.4}$$

считая $C_k^{(1)}$ и $C_k^{(2)}$ не постоянными, а переменными функциями аргумента k (при $f_k \neq 0$).

Из формулы (2.4) имеем

$$\begin{gathered} Y_{k+1} = C_{k+1}^{(1)} y_{k+1}^{(1)} + C_{k+1}^{(2)} y_{k+1}^{(2)} = \\ = C_k^{(1)} y_{k+1}^{(1)} + C_k^{(2)} y_{k+1}^{(2)} + y_{k+1}^{(1)} \Delta C_k^{(1)} + y_{k+1}^{(2)} \Delta C_k^{(2)}\,, \end{gathered}$$

где $\Delta C_k^{(j)} = C_{k+1}^{(j)} - C_k^{(j)}$, $j = 1, 2$. Потребуем, чтобы для всех k выполнялось равенство

$$y_{k+1}^{(1)} \Delta C_k^{(1)} + y_{k+1}^{(2)} \Delta C_k^{(2)} = 0\,, \tag{2.5}$$

тогда

$$Y_{k+1} = C_k^{(1)} y_{k+1}^{(1)} + C_k^{(2)} y_{k+1}^{(2)}\,; \tag{2.6}$$

увеличивая индекс k на единицу, получим

$$\begin{gathered} Y_{k+2} = C_{k+2}^{(1)} y_{k+2}^{(1)} + C_{k+2}^{(2)} y_{k+2}^{(2)} = \\ = C_{k+1}^{(1)} y_{k+2}^{(1)} + C_{k+1}^{(2)} y_{k+2}^{(2)} + \left[y_{k+2}^{(1)} \Delta C_{k+1}^{(1)} + y_{k+2}^{(2)} \Delta C_{k+1}^{(2)} \right]. \end{gathered}$$

В силу выполнения равенства (2.5) при замене k на $k+1$ выражение в квадратных скобках равно нулю, откуда

$$\begin{gathered} Y_{k+2} = C_{k+1}^{(1)} y_{k+2}^{(1)} + C_{k+1}^{(2)} y_{k+2}^{(2)} = \\ = C_k^{(1)} y_{k+2}^{(1)} + C_k^{(2)} y_{k+2}^{(2)} + y_{k+2}^{(1)} \Delta C_k^{(1)} + y_{k+2}^{(2)} \Delta C_k^{(2)}\,. \end{gathered} \tag{2.7}$$

Подставим выражения для Y_{k+1} и Y_{k+2} (формулы (2.6) и (2.7)) в исходное уравнение (2.3). Так как $y_k^{(1)}$ и $y_k^{(2)}$ — частные решения однородного уравнения, получим

$$\begin{aligned} f_k &= Y_{k+2} + a_k Y_{k+1} + b_k Y_k = C_k^{(1)} y_{k+2}^{(1)} + C_k^{(2)} y_{k+2}^{(2)} + y_{k+2}^{(1)} \Delta C_k^{(1)} + \\ &\quad + y_{k+2}^{(2)} \Delta C_k^{(2)} + a_k \left[C_k^{(1)} y_{k+1}^{(1)} + C_k^{(2)} y_{k+1}^{(2)} \right] + b_k \left[C_k^{(1)} y_k^{(1)} + C_k^{(2)} y_k^{(2)} \right] = \\ &= C_k^{(1)} \left[y_{k+2}^{(1)} + a_k y_{k+1}^{(1)} + b_k y_k^{(1)} \right] + C_k^{(2)} \left[y_{k+2}^{(2)} + a_k y_{k+1}^{(2)} + b_k y_k^{(2)} \right] + \\ &\quad + y_{k+2}^{(1)} \Delta C_k^{(1)} + y_{k+2}^{(2)} \Delta C_k^{(2)} = y_{k+2}^{(1)} \Delta C_k^{(1)} + y_{k+2}^{(2)} \Delta C_k^{(2)}\,. \end{aligned}$$

Таким образом, $\Delta C_k^{(1)}$ и $\Delta C_k^{(2)}$ должны при всех k удовлетворять системе уравнений

$$\begin{aligned} y_{k+1}^{(1)} \Delta C_k^{(1)} + y_{k+1}^{(2)} \Delta C_k^{(2)} &= 0, \\ y_{k+2}^{(1)} \Delta C_k^{(1)} + y_{k+2}^{(2)} \Delta C_k^{(2)} &= f_k. \end{aligned} \tag{2.8}$$

Напомним, что первое уравнение системы — это уравнение (2.5). Определитель системы (2.8), обозначим его через

$$\det_{k+1,k+2} = \begin{vmatrix} y_{k+1}^{(1)} & y_{k+1}^{(2)} \\ y_{k+2}^{(1)} & y_{k+2}^{(2)} \end{vmatrix},$$

отличен от нуля при всех k, так как $y_k^{(1)}$ и $y_k^{(2)}$ — линейно независимые решения. Поэтому можно записать

$$\Delta C_k^{(1)} = -\frac{y_{k+1}^{(2)}}{\det_{k+1,k+2}}\, f_k \equiv F_k^{(1)}\,, \qquad \Delta C_k^{(2)} = \frac{y_{k+1}^{(1)}}{\det_{k+1,k+2}}\, f_k \equiv F_k^{(2)}\,.$$

Из этих соотношений находим $C_k^{(1)}$ и $C_k^{(2)}$:

$$C_k^{(l)} = \sum_{j=1}^{k} F_{j-1}^{(l)} + C_0^{(l)}\,, \quad l = 1, 2\,.$$

Так как мы ищем частное решение уравнения (2.3), то можно положить $C_0^{(1)} = C_0^{(2)} = 0$. Окончательно получим

$$Y_k = \sum_{j=1}^{k} \frac{y_j^{(1)} y_k^{(2)} - y_k^{(1)} y_j^{(2)}}{\det_{j,j+1}}\, f_{j-1}\,.$$

2.67. Найти частное решение уравнения

$$y_{k+2} - 5\, y_{k+1} + 6\, y_k = 6^{k+1}\,.$$

$\lhd$ Линейно независимые решения однородного уравнения имеют вид

$$y_k^{(1)} = 2^k\,, \quad y_k^{(2)} = 3^k\,, \quad \det_{j,j+1} = \begin{vmatrix} 2^j & 3^j \\ 2^{j+1} & 3^{j+1} \end{vmatrix} = 6^j\,.$$

Воспользуемся формулой для частного решения

$$Y_k = \sum_{j=1}^{k} \frac{y_j^{(1)} y_k^{(2)} - y_k^{(1)} y_j^{(2)}}{\det_{j,j+1}}\, f_{j-1} = \sum_{j=1}^{k} \frac{2^j 3^k - 2^k 3^j}{6^j}\, 6^j =$$

$$= 3^k \sum_{j=1}^{k} 2^j - 2^k \sum_{j=1}^{k} 3^j = \frac{1}{2} \cdot 6^k + \frac{3}{2} \cdot 2^k - 2 \cdot 3^k\,. \qquad \rhd$$

2.68. Найти методом вариации постоянных формулу для решения разностного уравнения

$$y_{k+1} + a_k y_k = f_k\,, \quad a_k \neq 0\,, k = 0, \pm 1, \pm 2, \ldots\,.$$

2.69. Найти решение разностной задачи

$$y_{k+1} - a\, y_k = f_k\,, \quad k \geqslant 0\,,\ y_0 = c\,.$$

Ответ: $y_k = c\, a^k + \sum\limits_{j=0}^{k-1} a^j f_{k-j-1}$.

2.70. Пусть для элементов последовательности y_k справедливо

$$y_{k+1} \leqslant a\, y_k + f_k\,, \quad k \geqslant 0\,,\ y_0 = c,\ a > 0\,.$$

Найти оценку для y_k в зависимости от $a, c, f_i,\ i = 0, \ldots, k-1$.

$\lhd$ Из (2.69) следует, что решение уравнения

$$v_{k+1} = a\, v_k + f_k\,, v_0 = y_0$$

имеет вид $v_k = c\, a^k + \sum\limits_{j=0}^{k-1} a^j f_{k-j-1}$. Теперь покажем, что $y_k \leqslant v_k$. Вычтем уравнение из неравенства

$$y_{k+1} - v_{k+1} \leqslant a\,(y_k - v_k) \leqslant \cdots \leqslant a^{k+1}\,(y_0 - v_0) = 0\,.$$

Отсюда получаем $y_k \leqslant c\, a^k + \sum\limits_{j=0}^{k-1} a^j f_{k-j-1}\,, k \geqslant 0$. $\rhd$

2.71. Найти общее решение уравнения

$$y_{k+1} - \exp(2k)\, y_k = 6k^2 \exp(k^2 + k)\,.$$

$\lhd$ Найдем сначала решение однородного уравнения

$$y_{k+1} = \exp(2k)\, y_k = \exp(2k) \exp(2(k-1))\, y_{k-1}\,,$$

$$y_{k+1} = \cdots = \exp\left(2 \sum_{j=1}^{k} j\right) y_1 = \exp(k(k+1))\, y_1\,.$$

Отсюда имеем

$$y_k^0 = C\, \exp(k(k-1))\,.$$

Далее методом вариации постоянных найдем частное решение неоднородного уравнения

$$y_k^1 = 6 \exp(k(k-1)) \sum_{j=1}^{k-1} j^2 = k(k-1)(2k-1) \exp(k(k-1))\,.$$ $\rhd$

О т в е т: $y_k = [C + k(k-1)(2k-1)] \exp(k^2 - k)$.

2.72. Найти общее решение уравнения

$$a_k y_{k+2} + b_k y_{k+1} + c_k y_k = f_k\,,$$

где $a_k = k^2 - k + 1,\ b_k = -2(k^2+1),\ c_k = k^2 + k + 1,\ f_k = 2^k(k^2 - 3k + 1)$.

$\lhd$ Заметим, что $c_k = a_{k+1}$ и $b_k = -(a_k + c_k)$, поэтому данное уравнение можно переписать в виде

$$a_k(y_{k+2} - y_{k+1}) - a_{k+1}(y_{k+1} - y_k) = f_k\,. \tag{2.9}$$

Частные решения однородного уравнения $v_k^{(1)}$ и $v_k^{(2)}$ выделим условиями

$$v_0^{(1)} = v_1^{(1)} = 1\,,\ \ v_0^{(2)} = 0,\ v_1^{(2)} = 3\,.$$

Эти решения линейно независимы, так как определитель отличен от нуля:

$$\det \begin{vmatrix} v_0^{(1)} & v_1^{(1)} \\ v_0^{(2)} & v_1^{(2)} \end{vmatrix} = 3\,.$$

Решение $v_k^{(1)}$ находится легко: $v_k^{(1)} = 1$, а для определения $v_k^{(2)}$ преобразуем (2.9) при $f_k \equiv 0$. Имеем

$$y_{k+2} - y_{k+1} = \frac{a_{k+1}}{a_k}\,(y_{k+1} - y_k) = \frac{a_{k+1}}{a_{k-1}}\,(y_k - y_{k-1}) = \cdots = \frac{a_{k+1}}{a_0}\,(y_1 - y_0)\,.$$

Учитывая начальные значения для $v_k^{(2)}$, получим

$$v_{k+1}^{(2)} - v_k^{(2)} = 3a_k = 3(k^2 - k + 1)\,.$$

Окончательно имеем

$$v_k^{(2)} = k(k^2 - 3k + 5)\,.$$

Общее решение исходного однородного уравнения имеет вид

$$y_k^{(0)} = C_1 + C_2 k(k^2 - 3k + 5)\,.$$

Построим теперь частное решение неоднородного уравнения методом вариации постоянных

$$y_k^{(1)} = \sum_{j=0}^{k-2} \frac{v_k^{(2)} - v_{j+1}^{(2)}}{v_{j+2}^{(2)} - v_{j+1}^{(2)}}\,\frac{f_j}{a_j} = \sum_{j=0}^{k-2} \frac{v_k^{(2)} - v_{j+1}^{(2)}}{3a_{j+1}a_j}\left[2^{j+1}a_j - 2^j a_{j+1}\right] =$$

$$= \frac{1}{3}\sum_{j=0}^{k-2}\left[v_k^{(2)} - v_{j+1}^{(2)}\right]\left[\frac{2^{j+1}}{a_{j+1}} - \frac{2^j}{a_j}\right]\,.$$

Введем следующие обозначения: $x_j = v_k^{(2)} - v_{j+1}^{(2)}$, $z_j = \dfrac{2^j}{a_j}$, и перепишем частное решение в виде

$$y_k^{(1)} = \frac{1}{3}\sum_{j=0}^{k-2}\left[z_{j+1} - z_j\right] x_j\,.$$

Применяя формулу суммирования по частям, получаем

$$y_k^{(1)} = -\,\frac{1}{3}\sum_{j=0}^{k-1}\left[x_j - x_{j-1}\right] z_j + \frac{1}{3}\left[z_{k-1}x_{k-1} - z_0 x_{-1}\right]\,.$$

По определению x_k и z_k имеем

$$x_j - x_{j-1} = v_k^{(2)} - v_{j+1}^{(2)} = -3a_j\,,$$

$$x_{k-1} = v_k^{(2)} - v_k^{(2)} = 0\,,$$

$$x_{-1} = v_k^{(2)} - v_0^{(2)} = v_k^{(2)}\,.$$

Отсюда

$$y_k^{(1)} = \sum_{j=0}^{k-1} 2^j - \frac{1}{3}\,v_k^{(2)} = 2^k - 1 - \frac{1}{3}\,k(k^2 - 3k + 5)\,. \qquad \triangleright$$

Ответ: $y_k = 2^k - 1 - \frac{1}{3}\,(k^2 - 3k + 5) + C_1 + C_2 k(k^2 - 3k + 5)$.

2.4. Фундаментальное решение и функция Грина

Фундаментальным решением G_k называют решение разностного уравнения

$$a_0 y_k + a_1 y_{k+1} + \cdots + a_n y_{k+n} = f_k$$

с правой частью специального вида $f_k = \delta_k^0$, где

$$\delta_k^j = \begin{cases} 0 & \text{при} \quad k \neq j, \\ 1 & \text{при} \quad k = j. \end{cases}$$

2.73. Найти ограниченное фундаментальное решение уравнения

$$a\, y_k + b\, y_{k+1} = \delta_k^0 .$$

$\lhd$ Обозначим искомое фундаментальное решение через G_k. Для определения G_k имеем три группы уравнений:

$$\begin{cases} a\, G_k + b\, G_{k+1} = 0 & \text{при} \quad k \leqslant -1, \\ a\, G_0 + b\, G_1 = 1 & \text{при} \quad k = 0, \\ a\, G_k + b\, G_{k+1} = 0 & \text{при} \quad k \geqslant 1. \end{cases}$$

При $k \leqslant 0$ возьмем $G_k = 0$. Тогда все уравнения первой группы выполнены, из второго уравнения следует, что $G_1 = \frac{1}{b}$, а общее решение третьей группы уравнений имеет вид $G_k = C\,\mu^k$, где $\mu = -\frac{a}{b}$. Определяя константу C из G_1, получаем частное решение неоднородного уравнения

$$G_k^1 = \begin{cases} 0 & \text{при} \quad k \leqslant 0, \\ -\frac{1}{a}\left(-\frac{a}{b}\right)^k & \text{при} \quad k \geqslant 1. \end{cases}$$

Сложим полученное частное решение с общим решением $A\left(-\frac{a}{b}\right)^k$ однородного уравнения. В результате имеем

$$G_k = \begin{cases} A\left(-\frac{a}{b}\right)^k & \text{при} \quad k \leqslant 0, \\ \left(A - \frac{1}{a}\right)\left(-\frac{a}{b}\right)^k & \text{при} \quad k \geqslant 1. \end{cases}$$

Условие ограниченности выражается в виде зависимости постоянной A от величины $\left|\frac{a}{b}\right|$:

$$A = 0 \qquad \text{при} \quad \left|\frac{a}{b}\right| < 1,$$

$$\forall\, A \qquad \text{при} \quad \left|\frac{a}{b}\right| = 1,$$

$$A = 1/a \quad \text{при} \quad \left|\frac{a}{b}\right| > 1. \qquad \rhd$$

2.74. Пусть $\left|\frac{a}{b}\right| \neq 1$, $|f_k| \leqslant F$, а G_k — ограниченное фундаментальное решение уравнения

$$a\,y_k + b\,y_{k+1} = f_k.$$

Показать, что частным решением этого уравнения является абсолютно сходящийся ряд

$$y_k^1 = \sum_{n=-\infty}^{\infty} G_{k-n} f_n.$$

$\lhd$ Рассмотрим случай $\left|\frac{a}{b}\right| > 1$. Из 2.73 следует, что

$$G_{k-n} = \begin{cases} \frac{1}{a}\left(-\frac{a}{b}\right)^{k-n} & \text{при} \quad k \leqslant n, \\ 0 & \text{при} \quad k \geqslant n+1. \end{cases}$$

Каждый член ряда может быть оценен сверху членом сходящейся геометрической прогрессии

$$\left|\frac{1}{a}\left(-\frac{a}{b}\right)^{k-n} f_n\right| < \frac{F}{|a|}\left|\frac{b}{a}\right|^{n-k},$$

поэтому ряд сходится абсолютно. Кроме того, ряд является частным решением заданного уравнения

$$a\,y_k + b\,y_{k+1} = a\sum_{n=-\infty}^{\infty} G_{k-n} f_n + b\sum_{n=-\infty}^{\infty} G_{k+1-n} f_n =$$

$$= \sum_{n=-\infty}^{\infty} (a\,G_{k-n} + b\,G_{k+1-n})\, f_n = \sum_{n=-\infty}^{\infty} \delta_k^n\, f_n = f_k.$$

Для этого решения верна оценка

$$|y_k^1| \leqslant \frac{F}{|a|}\sum_{n=k}^{\infty}\left|\frac{b}{a}\right|^{n-k} = \frac{F}{|a|-|b|},$$

т. е. полученное частное решение является ограниченным.

Случай $\left|\frac{a}{b}\right| < 1$ рассматривается аналогично. $\rhd$

2.75. Найти ограниченное фундаментальное решение уравнения

$$y_{k-1} - 2\,y_k + y_{k+1} = \delta_k^0.$$

$\lhd$ Для определения G_k имеем три группы уравнений:

$$\begin{cases} G_{k-1} - 2\,G_k + G_{k+1} = 0 & \text{при} \quad k \leqslant -1, \\ G_{-1} - 2\,G_0 + G_1 = 1 & \text{при} \quad k = 0, \\ G_{k-1} - 2\,G_k + G_{k+1} = 0 & \text{при} \quad k \geqslant 1. \end{cases}$$

Общие решения первой и третьей групп имеют одинаковый вид, отличающийся только постоянными

$$G_k = \begin{cases} C_1^- + C_2^- k & \text{при} \quad k \leqslant 0, \\ C_1^+ + C_2^+ k & \text{при} \quad k \geqslant 0. \end{cases}$$

Так как G_0 входит во все три группы уравнений, то из полученных соотношений имеем $G_0 = C_1^- = C_1^+ = A$. Теперь воспользуемся уравнением при $k = 0$ для установления связи между C_2^- и C_2^+

$$(A - C_2^-) - 2A + (A + C_2^+) = 1.$$

Отсюда $C_2^- = B, C_2^+ = 1 + B$. Окончательное выражение для фундаментального решения имеет вид

$$G_k = \begin{cases} A + Bk & \text{при} \quad k \leqslant 0, \\ A + (B+1)k & \text{при} \quad k \geqslant 0. \end{cases}$$

Ограниченное решение не существует, поскольку B не может одновременно быть равным 0 и -1. $\triangleright$

2.76. Найти ограниченное фундаментальное решение уравнения

$$y_{k-1} - y_k + y_{k+1} = \delta_k^0.$$

Ответ: $G_k = \begin{cases} A\cos\frac{k\pi}{3} + \left[B + \frac{2\sqrt{3}}{3}(1-2A)\right]\sin\frac{k\pi}{3} & \text{при} \quad k \geqslant 0, \\ A\cos\frac{k\pi}{3} + B\sin\frac{k\pi}{3} & \text{при} \quad k \leqslant 0. \end{cases}$

2.77. Найти ограниченное фундаментальное решение уравнения

$$y_{k-1} - \frac{5}{2}\, y_k + y_{k+1} = \delta_k^0.$$

Ответ: $G_k = \begin{cases} A\,2^{-k} & \text{при} \quad k \geqslant 0, \\ A\,2^{k} & \text{при} \quad k \leqslant 0, \end{cases} \quad A = -\frac{2}{3}.$

2.78. Найти ограниченное фундаментальное решение уравнения

$$y_{k+1} - 5y_k + 6y_{k-1} = \delta_k^0.$$

Ответ: $G_k = \begin{cases} 2^k - 3^k & \text{при} \quad k \leqslant 0, \\ 0 & \text{при} \quad k \geqslant 0. \end{cases}$

2.79. Найти ограниченное фундаментальное решение уравнения

$$3y_{k-1} - \frac{13}{2}\, y_k + y_{k+1} = \delta_k^0.$$

Ответ: $G_k = \begin{cases} C\,6^k & \text{при} \quad k \leqslant 0, \\ C\,2^{-k} & \text{при} \quad k \geqslant 0, \end{cases} \quad C = -\frac{2}{11}.$

2.80. Найти ограниченное фундаментальное решение уравнения

$$\frac{1}{8}\, y_{k-1} - \frac{3}{4}\, y_k + y_{k+1} = \delta_k^0.$$

Ответ: $G_k = \begin{cases} 0 & \text{при} \quad k \leqslant 0, \\ 4\left(2^{-k} - 4^{-k}\right) & \text{при} \quad k \geqslant 0. \end{cases}$

2.81. Найти ограниченное фундаментальное решение уравнения

$$y_{k+1} - y_k - 12y_{k-1} = \delta_k^0.$$

Ответ: $G_k = \begin{cases} 0 & \text{при} \quad k \geqslant 0, \\ C\left((-3)^k - 4^k\right) & \text{при} \quad k \leqslant 0, \end{cases} \quad C = \frac{1}{7}.$

Разностная функция Грина для уравнения второго порядка. *Функцией Грина* $G_{k,i}$ разностной краевой задачи называют фундаментальное решение, удовлетворяющее однородным краевым условиям. Например, для задачи

$$\begin{gathered} \Lambda y_k \equiv \Delta\left(a_k \nabla y_k\right) - d_k y_k = \varphi_k,\ a_k > 0,\ d_k \geqslant 0,\ 1 \leqslant k \leqslant N-1, \\ y_0 = c, \quad y_N - y_{N-1} = \varphi_N, \end{gathered} \tag{2.10}$$

под функцией Грина понимают функцию $G_{k,i}$, определенную при фиксированном i для $0 \leqslant k \leqslant N$, которая удовлетворяет краевым условиям

$$G_{0,i} = 0, \quad G_{N,i} - G_{N-1,i} = 0,$$

и уравнению по переменной k

$$\Lambda G_{k,i} = \delta_k^i.$$

«Польза» от функции $G_{k,i}$ в первую очередь состоит в представлении частного решения неоднородного уравнения в виде

$$y_k = \sum_{i=1}^{N} G_{k,i}\varphi_i.$$

2.82. Построить функцию Грина для уравнения (2.10) при краевых условиях $y_0 = 0, y_N = 0$.

$\lhd$ Пусть u_k и v_k — решения задач Коши:

$$\begin{aligned} &\Lambda u_k = 0, \qquad u_0 = 0, \qquad a_1(u_1 - u_0) = 1, \\ &\Lambda v_k = 0, \qquad v_N = 0, \qquad -a_N(v_N - v_{N-1}) = 1. \end{aligned}$$

Покажем, что они обладают следующими свойствами:

1) u_k — монотонно возрастающая, а v_k — монотонно убывающая положительные функции, т. е. $u_k > 0$, $v_k > 0$, $u_k > u_{k-1}$, $v_k < v_{k-1}$;

2) $u_N = v_0$;

3) u_k и v_k — линейно независимые функции.

Докажем эти свойства.

1) Из уравнения $\Lambda u_k = 0$ и условия $a_1 u_1 = 1$ следует, что

$$a_k(u_k - u_{k-1}) = 1 + \sum_{i=1}^{k-1} d_i u_i.$$

Так как правая часть равенства больше нуля и $a_k > 0$, то последовательность u_k монотонна, т. е. $u_k - u_{k-1} > 0$, и положительна в силу $u_1 > 0$. Аналогично показывается, что $0 < v_k < v_{k-1}$.

2) Рассмотрим вторую формулу Грина

$$0 = \sum_{k=1}^{N-1} (u_k \Lambda v_k - v_k \Lambda u_k) = a_N(u_N v_{N-1} - v_N u_{N-1}) - a_1(u_1 v_0 - v_1 u_0).$$

Начальные условия u_0 и v_N дают следующие соотношения: $a_1 u_1 = 1$ и $a_N v_{N-1} = 1$, т. е.

$$0 = a_N u_N v_{N-1} - a_1 u_1 v_0 = u_N - v_0.$$

3) Применим вторую формулу Грина от $k = 1$ до $k = k_0$:

$$0 = \sum_{k=1}^{k_0-1} (u_k \Lambda v_k - v_k \Lambda u_k) = a_{k_0}(u_{k_0} v_{k_0-1} - v_{k_0} u_{k_0-1}) - v_0.$$

Введем обозначение: $\det_k = \begin{vmatrix} u_k & u_{k-1} \\ v_k & v_{k-1} \end{vmatrix}$. Тогда можно записать $0 = a_{k_0} \det_{k_0} - v_0$. Так как k_0 произвольно, то $a_{k_0} \det_{k_0} = v_0 > 0$, откуда следует, что $\det_{k_0} > 0$, т. е. линейная независимость u_k и v_k.

Будем теперь искать функцию Грина $G_{k,i}$ в виде

$$G_{k,i} = \begin{cases} A_i u_k & \text{при} \quad i \geqslant k, \\ B_i v_k & \text{при} \quad i \leqslant k. \end{cases}$$

При $i = k$ имеем $G_{k,k} = A_k u_k = B_k v_k$ и $\Lambda G_{k,k} = 1$. Перепишем последнее соотношение в виде

$$B_k \left[a_{k+1} v_{k+1} - (a_{k+1} + a_k + d_k) v_k\right] + a_k A_k u_{k-1} = 1,$$

или

$$B_k \left[\Delta (a_k \nabla v_k) - d_k v_k\right] + a_k \left[A_k u_{k-1} - B_k v_{k-1}\right] = 1.$$

Так как $\Lambda v_k = 0$, то из последнего уравнения имеем

$$A_k u_{k-1} - B_k v_{k-1} = \frac{1}{a_k}.$$

Таким образом, для определения A_k и B_k получена система уравнений

$$A_k u_k - B_k v_k = 0, A_k u_{k-1} - B_k v_{k-1} = \frac{1}{a_k},$$

определитель которой отличен от нуля в силу свойства 3). Решая систему, учтем равенство $a_k \det_k = v_0 = u_N$. Окончательно получаем

$$G_{k,i} = \begin{cases} v_i u_k & \text{при} \quad i \geqslant k, \\ u_i v_k & \text{при} \quad i \leqslant k. \end{cases} \qquad \triangleright$$

2.83. Построить функцию Грина для следующих краевых задач:

1) $\Delta^2 y_{k-1} = \varphi_k, y_0 = y_N = 0;$
2) $y_{k+1} - 2\cos\alpha y_k + y_{k-1} = \varphi_k, y_0 = y_N = 0;$
3) $y_{k+1} + 2y_k + y_{k-1} = \varphi_k, y_0 = y_1, \;\; y_N = 0;$
4) $y_{k+1} - 2\cos\alpha y_k + y_{k-1} = \varphi_k, \Delta y_0 = y_N = 0;$
5) $y_{k+1} - 3y_k + 2y_{k-1} = \varphi_k, y_0 = y_N = 0;$
6) $y_{k+1} - 2Ay_k - y_{k-1} = \varphi_k, \Delta y_0 = y_N = 0.$

2.84. Доказать, что решение разностной задачи

$$y_{k-1} - 2y_k + y_{k+1} = f_k, y_0 = \alpha, y_N = \beta$$

удовлетворяет неравенству

$$\max_k |y_k| \leqslant \max(|\alpha|, |\beta|) + \max_{1\leqslant k\leqslant N-1} |f_k| \, \frac{N^2}{8} \, .$$

У к а з а н и е. Решение удобно записать в виде

$$y_k = \alpha + k\,\frac{\beta-\alpha}{N} + \sum_{i=1}^{N-1} G_{k,i} f_i,$$

где функция Грина представима формулой

$$G_{k,i} = \begin{cases} \frac{i}{N}(k-N) & \text{при} \quad k \geqslant i, \\ \frac{k}{N}(i-N) & \text{при} \quad k \leqslant i, \end{cases} \quad i = 1, \ldots, N-1, \;\; k = 0, \ldots, N.$$

2.5. Задачи на собственные значения

2.85. Найти все решения задачи на собственные значения

$$\frac{y_{k+1} - y_{k-1}}{2h} = -\lambda\, y_k, \; 0 < k < N, \quad y_0 = y_N = 0, \;\; h = \frac{1}{N} \, .$$

$\lhd$ Перепишем разностное уравнение в виде

$$y_{k+1} + 2h\lambda\, y_k - y_{k-1} = 0.$$

Его характеристическое уравнение

$$\mu^2 + 2h\lambda\,\mu - 1 = 0$$

имеет корни $\mu_1 = -h\lambda + \sqrt{1 + h^2\lambda^2}$ и $\mu_2 = -h\lambda - \sqrt{1 + h^2\lambda^2}$. Можно показать (сделайте это самостоятельно), что при $\mu_1 = \mu_2$ существует только тривиальное решение $y_k \equiv 0$, поэтому общее решение разностного уравнения имеет вид

$$y_k = C_1\mu_1^k + C_2\mu_2^k.$$

Константы C_1 и C_2 определяются из системы

$$C_1 + C_2 = 0, \; C_1\mu_1^N + C_2\mu_2^N = 0,$$

откуда получаем, что $C_2 = -C_1$ и $C_1(\mu_1^N - \mu_2^N) = 0$, т. е. нетривиальное решение разностной задачи существует тогда и только тогда, когда $\mu_1^N = \mu_2^N$. Следовательно,

$$\frac{\mu_1}{\mu_2} = \exp\left(\mathrm{i}\frac{2\pi m}{N}\right), \quad m = 0, 1, \ldots, N-1.$$

Так как $\mu_1\mu_2 = -1$, то $\mu_1^2 = -\exp\left(\mathrm{i}\frac{2\pi m}{N}\right)$, откуда

$$\mu_1 = \mathrm{i}\exp\left(\mathrm{i}\frac{\pi m}{N}\right), \quad \mu_2 = \mathrm{i}\exp\left(-\mathrm{i}\frac{\pi m}{N}\right).$$

Поскольку

$$\mu_1 + \mu_2 = -2h\lambda = \mathrm{i}\left(\exp\left(\mathrm{i}\frac{\pi m}{N}\right) + \exp\left(-\mathrm{i}\frac{\pi m}{N}\right)\right) = 2\mathrm{i}\cos\frac{\pi m}{N},$$

имеем

$$\lambda^{(m)} = -\frac{\mathrm{i}}{h}\cos\frac{\pi m}{N}, \quad m = 0, 1, \ldots, N-1.$$

Соответствующие решения исходной задачи таковы:

$$y_k^{(m)} = C_1\left(\mu_1^k - \mu_2^k\right) = C_1\mathrm{i}^k\left(\exp\left(\mathrm{i}\frac{\pi km}{N}\right) - \exp\left(-\mathrm{i}\frac{\pi km}{N}\right)\right) =$$

$$= C_1\mathrm{i}^k 2\mathrm{i}\sin\frac{\pi km}{N} = C\,\mathrm{i}^k\sin\frac{\pi km}{N},\ m = 0, 1, 2 \ldots N-1.$$

При $m = 0$ имеем $y_k^{(0)} \equiv 0$, поэтому решение $(\lambda^{(0)}, y_k^{(0)})$ следует отбросить. Отметим, что количество нетривиальных решений равно $N-1$, что совпадает с размерностью задачи. $\triangleright$

2.86. Найти все решения задачи на собственные значения

$$\frac{y_{k+1} - 2\,y_k + y_{k-1}}{h^2} = -\lambda\,y_k,\ 0 < k < N, \quad y_0 = y_N = 0, \quad h = \frac{1}{N}.$$

$\triangleleft$ Характеристическое уравнение разностной задачи имеет вид

$$\mu^2 - (2 - h^2\lambda)\mu + 1 = 0.$$

Если корни характеристического уравнения вещественные, то разностная задача имеет только тривиальное решение. Действительно, пусть $\mu_1 \neq \mu_2$ — вещественные корни. Тогда общее решение имеет вид $y_k = C_1\mu_1^k + C_2\mu_2^k$ и для определения C_1 и C_2 из краевых условий имеем систему

$$C_1 + C_2 = 0,\ C_1\mu_1^N + C_2\mu_2^N = 0,$$

из которой следует $C_1\mu_1^N - C_1\mu_2^N = 0$. Так как $\mu_1 \neq \mu_2$, то $C_1 = C_2 = 0$, т. е. общее решение является нулевым. Аналогично рассматривается случай равных вещественных корней.

Поэтому следует рассмотреть случай комплексно-сопряженных корней $\mu_{1,2} = \cos\varphi \pm \mathrm{i}\sin\varphi$. В этом случае общее решение разностной задачи представляется в виде $y_k = C_1\cos k\varphi + C_2\sin k\varphi$. Из краевых условий получаем $C_1 = 0$ и $\sin N\varphi = 0$. Отсюда

$$\varphi = \frac{\pi m}{N}, \quad m = 0, \pm1, \pm2, \ldots.$$

Так как $\mu_1 + \mu_2 = 2 - h^2\lambda$, то $\cos\varphi = 1 - h^2\frac{\lambda}{2}$. Следовательно,

$$\lambda^{(m)} = \frac{2}{h^2}\left(1 - \cos\frac{\pi m}{N}\right) = \frac{4}{h^2}\sin^2\frac{\pi m}{2N}\,, \quad m = 0, 1, 2, \ldots, N-1.$$

Все собственные значения различны. Из представления общего решения разностной задачи следует, что собственные функции имеют вид

$$y_k^{(m)} = C_2 \sin\left(\frac{\pi k m}{N}\right), \quad m = 0, 1, 2, \ldots, N-1.$$

При $m = 0$ имеем $y_k^{(0)} \equiv 0$, поэтому решение $(\lambda^{(0)}, y_k^{(0)})$ следует отбросить.

Полезно провести аналогию с дифференциальной задачей:

$$y'' = -\lambda\, y,\ y(0) = y(1) = 0,$$
$$y_{(m)}(x) = C\sin(\pi m x),\ \lambda_{(m)} = (\pi m)^2,\ m = 1, 2, \ldots\,. \qquad \triangleright$$

2.87. Найти все решения задачи на собственные значения

$$\frac{y_{k+1} - 2\,y_k + y_{k-1}}{h^2} = -\lambda\, y_k\,, \quad h = \frac{1}{N}\,, \quad 1 \leqslant k \leqslant N-1\,,$$

$$\frac{2}{h^2}\,(y_1 - y_0) = -\lambda\, y_0\,, \quad -\frac{2}{h^2}\,(y_N - y_{N-1}) = -\lambda\, y_N\,.$$

$\triangleleft$ Введем обозначение $p = 1 - h^2\frac{\lambda}{2}$ и перепишем исходную задачу в виде

$$y_{k+1} - 2\,p\,y_k + y_{k-1} = 0\,, \quad 1 \leqslant k \leqslant N-1\,,$$
$$y_1 - p\,y_0 = 0\,, \quad y_{N-1} - p\,y_N = 0\,.$$

Корни характеристического уравнения $\mu^2 - 2\,p\,\mu + 1 = 0$ имеют вид $\mu_{1,2} = p \pm \sqrt{p^2 - 1}$. Отметим полезные соотношения:

$$\mu_1\mu_2 = 1\,, \quad p = \frac{\mu_1 + \mu_2}{2}\,.$$

Рассмотрим случай различных (не обязательно вещественных) корней: $\mu_1 \neq \mu_2$. Общее решение разностного уравнения имеет вид $y_k = C_1\mu_1^k + C_2\mu_2^k$. Воспользуемся этим решением для записи левого краевого условия (при $k = 0$). Имеем

$$C_1\mu_1 + C_2\mu_2 - p\,(C_1 + C_2) = 0\,.$$

Учитывая, что p — полусумма корней характеристического уравнения, отсюда получаем $C_1 = C_2$ $(\neq 0)$. Теперь оставшееся краевое условие можно записать в удобной форме

$$\mu_1^{N-1} + \mu_2^{N-1} - \frac{\mu_1 + \mu_2}{2}\left(\mu_1^N + \mu_2^N\right) = 0\,.$$

Используем равенство $\mu_1\mu_2 = 1$, имеем

$$\mu_1^{N-1} + \mu_2^{N-1} - \mu_1^{N+1} - \mu_2^{N+1} = 0\,,$$

или

$$\mu_1^{N-1}\,(1 - \mu_1^2) + \mu_2^{N-1}\,(1 - \mu_2^2) = 0\,.$$

Отсюда получаем

$$\mu_1^N\,(\mu_2-\mu_1)+\mu_2^N\,(\mu_1-\mu_2)=0\,.$$

В силу предположения о неравенстве корней имеем $\dfrac{\mu_1^N}{\mu_2^N}=1$, что дает $\mu_1^{2N}=1$ или

$$\mu_{1,2}=\cos\left(\frac{\pi\,m}{N}\right)\pm\mathrm{i}\sin\left(\frac{\pi\,m}{N}\right)\,,\quad m=1,2,\ldots,N-1\,.$$

Отсюда получаем, что $p=\cos\left(\dfrac{\pi\,m}{N}\right)$, и формулу для собственных значений

$$\lambda^{(m)}=\frac{4}{h^2}\sin^2\frac{\pi\,m}{2\,N}\,,\quad m=1,2,\ldots,N-1\,.$$

Приведем формулу для собственных функций

$$y_k^{(m)}=\tilde{C}\,\left[\exp\left(\mathrm{i}\,\frac{\pi\,m\,k}{N}\right)+\exp\left(-\,\mathrm{i}\,\frac{\pi\,m\,k}{N}\right)\right]=C\,\cos\frac{\pi\,m\,k}{N}\,.$$

Осталось рассмотреть случай кратных корней: $\mu_1=\mu_2=p$. Возможны два случая: $p=\pm1$, так как $\mu_1\mu_2=1$. При этом соответствующие собственные значения равны $\lambda=2\,\dfrac{1-p}{h^2}$. Их удобно включить в полученную ранее общую формулу, расширив границы индекса m от нуля до N, т. е. $\lambda^{(0)}=0$ $(p=1)$, $\lambda^{(N)}=\dfrac{4}{h^2}$ $(p=-1)$.

Аналогично поступим и с соответствующими собственными функциями. $\triangleright$

О т в е т: $\lambda^{(m)}=\dfrac{4}{h^2}\sin^2\dfrac{\pi\,m}{2\,N}\,,\quad m=0,1,\ldots,N\,,$

$$y_k^{(m)}=C\,\cos\frac{\pi\,m\,k}{N}\,,\quad k=0,1,\ldots,N\,.$$

2.88. Найти все решения задачи на собственные значения

$$\frac{y_{k+1}-2\,y_k+y_{k-1}}{h^2}=-\lambda\,y_k\,,\quad h=\frac{1}{N}\,,\quad 1\leqslant k\leqslant N-1\,,$$

$$y_0=0\,,\quad -\,\frac{2}{h^2}\,(y_N-y_{N-1})=-\lambda\,y_N\,.$$

О т в е т: $\lambda^{(m)}=\dfrac{4}{h^2}\sin^2\dfrac{\pi(2m-1)}{4\,N}\,,\quad m=1,2,\ldots,N\,,$

$$y_k^{(m)}=C\,\sin\frac{\pi(2m-1)k}{2N}\,,\quad k=0,1,\ldots,N\,.$$

2.89. Найти все решения задачи на собственные значения

$$\frac{y_{k+1}-2\,y_k+y_{k-1}}{h^2}=-\lambda\,y_k\,,\quad h=\frac{1}{N}\,,\quad 1\leqslant k\leqslant N-1\,,$$

$$\frac{2}{h^2}\,(y_1-y_0)=-\lambda\,y_0\,,\quad y_N=0\,.$$

Ответ: $\lambda^{(m)} = \frac{4}{h^2}\sin^2\frac{\pi(2m-1)}{4N}\,, \quad m = 1, 2, \ldots, N\,,$

$$y_k^{(m)} = C\,\cos\frac{\pi(2m-1)k}{2N}\,, \qquad k = 0, 1, \ldots, N\,.$$

2.90. Найти все решения задачи на собственные значения

$$\frac{y_{k+1} - 2\,y_k + y_{k-1}}{h^2} = -\lambda\,y_k, \quad y_k = y_{k+N}, \quad h = \frac{1}{N}, \quad k = 0, \pm1, \pm2, \ldots$$

Ответ: $\lambda^{(m)} = \frac{4}{h^2}\sin^2\frac{\pi\,m}{N}\,, \quad m = 0, 1, \ldots, N-1,$

$$y_k^{(m)} = C_1^{(m)}\cos\frac{2\,\pi\,m\,k}{N} + C_2^{(m)}\sin\frac{2\,\pi\,m\,k}{N}\,, \quad k - \text{целое}\,.$$

2.91. Найти все решения задачи на собственные значения

$$\frac{y_{k+1} - 2\,y_k + y_{k-1}}{h^2} = -\lambda\,y_k, \quad h = \frac{1}{N-1}\,, \quad 1 \leqslant k \leqslant N-1,$$

$$y_0 = y_1,\ y_N = y_{N-1}\,.$$

Ответ: $\lambda^{(m)} = \frac{4}{h^2}\sin^2\frac{\pi(m-1)}{2(N-1)}\,, \quad m = 1, \ldots, N-1,$

$$y_k^{(m)} = C\cos\frac{\pi(m-1)\,(2k-1)}{2(N-1)}\,, \quad k = 0, 1, \ldots, N\,.$$

2.92. Найти все решения задачи на собственные значения

$$\frac{y_{k+1} - 2\,y_k + y_{k-1}}{h^2} = -\lambda\,y_k, \quad h = \frac{1}{N-1}\,, \quad 1 \leqslant k \leqslant N-1,$$

$$y_0 = -y_1,\ y_N = -y_{N-1}\,.$$

Ответ: $\lambda^{(m)} = \frac{4}{h^2}\sin^2\frac{\pi m}{2(N-1)}\,, \qquad m = 1, \ldots, N-1,$

$$y_k^{(m)} = C\sin\frac{\pi m\,(2k-1)}{2(N-1)}\,, \qquad k = 0, 1, \ldots, N\,.$$

2.93. Найти все решения задачи на собственные значения

$$\frac{y_{k+1} - 2\,y_k + y_{k-1}}{h^2} = -\lambda\,y_k, \quad h = \frac{2}{2N-1}\,, \quad 1 \leqslant k \leqslant N-1,$$

$$y_0 = y_1,\ y_N = 0\,.$$

Ответ: $\lambda^{(m)} = \frac{4}{h^2}\sin^2\frac{\pi(2m-1)}{2(2N-1)}\,, \quad m = 1, \ldots, N-1,$

$$y_k^{(m)} = C\sin\frac{\pi(2m-1)(N-k)}{2N-1}\,, \quad k = 0, 1, \ldots, N\,.$$

2.94. Найти все решения задачи на собственные значения

$$\frac{y_{k+1}-2\,y_k+y_{k-1}}{h^2}=-\lambda\,y_k,\quad h=\frac{2}{2N-1}\,,\quad 1\leqslant k\leqslant N-1,$$

$$y_0=0,\ y_N=y_{N-1}\,.$$

Ответ: $\lambda^{(m)}=\frac{4}{h^2}\sin^2\frac{\pi(2m-1)}{2(2N-1)}\,,\quad m=1,\dots,N-1,$

$$y_k^{(m)}=C\sin\frac{\pi(2m-1)k}{2N-1}\,,\quad k=0,1,\dots,N\,.$$

2.95. Найти все решения задачи на собственные значения

$$\frac{y_{k+1}-y_{k-1}}{2h}=-\lambda\,y_k,\quad h=\frac{1}{N}\,,\quad 1\leqslant k\leqslant N-1,$$

$$y_0=y_1,\ y_N=y_{N-1}\,.$$

Ответ: $\lambda^{(m)}=-\frac{\mathrm{i}}{h}\cos\frac{\pi m}{N-1}\,,\quad m=1,\dots,N-2,$

$$y_k^{(m)}=C\mathrm{i}^k\left[\sin\frac{\pi mk}{N-1}-\mathrm{i}\sin\frac{\pi m(k-1)}{N-1}\right],\quad k=0,1,\dots,N\,.$$

$$\lambda^{(0)}=0,\quad y_k^{(0)}=C\neq 0\,.$$

Неравенства для сеточных функций. Учитывая определения из п. 2.2, введем следующие обозначения:

$$\|y\|=\sqrt{h}\,[y,y]^{1/2}\,,\qquad \|y\|_C=\max_{0\leqslant i\leqslant N}|y_i|\,,$$

$$(y_{\bar{x}})_i=\frac{\nabla y_i}{h}=\frac{y_i-y_{i-1}}{h}\,,\quad \|y_{\bar{x}}]\|=\sqrt{h}\,\left(y_{\bar{x}},y_{\bar{x}}\right]^{1/2}\,,$$

где h — постоянный шаг сетки $x_i=ih$.

2.96. Пусть $y_0=y_N=0$ и $N\,h=1$. Доказать неравенство

$$\|y\|_C\leqslant\frac{1}{2}\,\|y_{\bar{x}}]\|\,.$$

$\lhd$ Запишем на сетке $x_i=ih,\ 0\leqslant i\leqslant N$ для функции y_i тождество

$$y_i^2\equiv(1-x_i)y_i^2+x_iy_i^2\,.$$

Принимая во внимание условия $y_0=y_N=0$, имеем

$$y_i^2=\left(\sum_{k=1}^{i}(y_{\bar{x}})_k\,h\right)^2,\quad y_i^2=\left(\sum_{k=i+1}^{N}(y_{\bar{x}})_k\,h\right)^2.$$

Подставим полученные равенства в тождество и оценим его правую часть, используя неравенство Коши—Буняковского. Получаем

$$y_i^2\leqslant(1-x_i)\sum_{k=1}^{i}h\sum_{k=1}^{i}(y_{\bar{x}})_k^2\,h+x_i\sum_{k=i+1}^{N}h\sum_{k=i+1}^{N}(y_{\bar{x}})_k^2\,h=$$

$$=x_i(1-x_i)\sum_{k=1}^{N}(y_{\bar{x}})_k^2\,h\equiv x_i(1-x_i)\|y_{\bar{x}}]\|^2\,.$$

Максимум выражения $x(1-x)$ на отрезке $[0,1]$ равен $\frac{1}{4}$ и достигается при $x=\frac{1}{2}$, поэтому $y_i^2 \leqslant \frac{1}{4}\,\|y_{\bar{x}}]|^2$ и $\|y\|_C \leqslant \frac{1}{2}\,\|y_{\bar{x}}]|$. $\triangleright$

2.97. Пусть $y_0 = y_N = 0$ и $N\,h = l$. Доказать неравенство

$$\|y\|_C \leqslant \frac{\sqrt{l}}{2}\,\|y_{\bar{x}}]|\,.$$

Указание. Сделать замену $x = l\,x'$ и использовать решение 2.96.

2.98. Пусть $y_0 = 0$ и $N\,h = l$. Доказать неравенство

$$\|y\|_C \leqslant \sqrt{l}\|y_{\bar{x}}]|\,.$$

2.99. Для произвольной сеточной функции y_i, $0 \leqslant i \leqslant N$, $Nh = l$ доказать неравенства

$$\|y\|_C^2 \leqslant 2\,(l\,\|y_{\bar{x}}]|^2 + y_0^2) \quad \text{и} \quad \|y\|_C^2 \leqslant 2\,(l\,\|y_{\bar{x}}]|^2 + y_N^2)\,.$$

2.100. Пусть $y_0 = y_N = 0$ и $N\,h = l$. Доказать неравенство

$$\|y\| \leqslant \sqrt{l}\|y\|_C\,.$$

2.101. Пусть $y_0 = y_N = 0$ и $N\,h = l$. Доказать неравенство

$$\frac{h}{2}\,\|y_{\bar{x}}]| \leqslant \|y\| \leqslant \frac{l}{2\sqrt{2}}\,\|y_{\bar{x}}]|\,.$$

Указание. Так как

$$(-\Lambda y, y) = \|y_{\bar{x}}]|^2,$$

где

$$\Lambda y_i = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}\,, \quad 1 \leqslant i \leqslant N-1, \quad y_0 = y_N = 0,$$

то постоянные в сеточных неравенствах можно получить, определив экстремумы собственных значений оператора Λ. Из решения задачи на собственные значения $\Lambda y = -\lambda y$ следует

$$\lambda_{\min} = \frac{4}{h^2}\sin^2\frac{\pi h}{2l}\,, \quad \lambda_{\max} = \frac{4}{h^2}\cos^2\frac{\pi h}{2l}\,.$$

Постоянные в искомом неравенстве получаются из оценок снизу для $\lambda_{\min}$ и сверху для $\lambda_{\max}$.

2.102. Доказать тождество Лагранжа для сеточных функций

$$\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} x_i y_i\right)^2 = \frac{1}{2}\sum_{i,j=1}^{N}(x_i y_j - x_j y_i)^2\,.$$

2.103. Доказать неравенство для неотрицательных сеточных функций

$$\left(\prod_{i=1}^{N} x_i\right)^{1/N} + \left(\prod_{i=1}^{N} y_i\right)^{1/N} \leqslant \left(\prod_{i=1}^{N}(x_i + y_i)\right)^{1/N}\,.$$

2.104. Доказать при $0 < \theta < 1$ неравенство Гельдера для неотрицательных сеточных функций

$$\sum_{i=1}^{N} x_i y_i \leqslant \left(\sum_{i=1}^{N} x_i^{1/\theta} \right)^{\theta} \left(\sum_{i=1}^{N} y_i^{1/(1-\theta)} \right)^{1-\theta}.$$

2.105. Доказать при $0 < \theta < 1$ неравенство Минковского для неотрицательных сеточных функций

$$\left(\sum_{i=1}^{M} \left(\sum_{j=1}^{N} x_{ij} \right)^{1/\theta} \right)^{\theta} \leqslant \sum_{j=1}^{N} \left(\sum_{i=1}^{M} x_{ij}^{1/\theta} \right)^{\theta}.$$

Глава 3

Приближение функций и производных

Задачи приближения функции можно условно разделить на два множества. Задачи первого множества сводятся к приближенному восстановлению достаточно гладкой функции по ее заданным значениям в некоторых фиксированных точках. В задачах второго множества речь идет о наилучшем (в некоторой метрике) приближении — замене сложной с точки зрения вычислений функции ее более простым аналогом. Типичным при таком подходе является поиск приближения в виде линейной комбинации «удобных» функций, например ортогональных алгебраических или тригонометрических многочленов. Многообразие математических постановок приводит к большому количеству применяемых методов, каждый из которых может оказаться оптимальным в своем классе. В этой главе рассмотрены наиболее известные в теории приближений подходы для функций одного переменного.

3.1. Полиномиальная интерполяция

Пусть $a = x_1 < x_2 < \ldots < x_n = b$ — набор различных точек (узлов) на отрезке $[a, b]$, в которых заданы значения функции $f(x)$ так, что $f_i = f(x_i)$, $i = 1, \ldots, n$. Требуется построить многочлен наименьшей степени, принимающий в точках x_i значения f_i, и оценить погрешность приближения достаточно гладкой функции $f(x)$ этим многочленом на всем отрезке $[a, b]$.

Приведем в явном виде вспомогательные многочлены $\Phi_i(x)$ степени $n-1$, удовлетворяющие условиям $\Phi_i(x_i) = 1, \Phi_i(x_j) = 0$ при $j \neq i$. Имеем $\Phi_i(x) = \prod\limits_{\substack{j=1 \\ j\neq i}}^{n} \dfrac{x - x_j}{x_i - x_j}$. Запишем с их помощью формулу для искомого *многочлена Лагранжа* $L_n(x) = \sum\limits_{i=1}^{n} f_i \Phi_i(x)$. Так как существует единственный многочлен степени $n-1$, принимающий в n различных точках заданные значения, то многочлен $L_n(x)$ есть решение поставленной задачи.

Теорема. *Пусть n-я производная функции $f(x)$ непрерывна на отрезке $[a, b]$. Тогда для любой точки $x \in [a, b]$ существует точка $\xi \in [a, b]$ такая, что справедливо равенство*

$$f(x) - L_n(x) = \frac{f^{(n)}(\xi)}{n!}\ \omega_n(x), \quad \textit{где}\ \ \omega_n(x) = \prod_{i=1}^{n}(x - x_i).$$

Следствием этого представления является оценка погрешности в равномерной норме

$$\|f(x) - L_n(x)\| \leqslant \frac{\|f^{(n)}(x)\|}{n!}\ \|\omega_n(x)\|, \quad \text{где} \quad \|f(x)\| = \max_{x\in[a,b]} |f(x)|.$$

Величина $\lambda_n = \max\limits_{x\in[a,b]} \sum\limits_{i=1}^{n} |\Phi_i(x)|$ называется *константой Лебега интерполяционного процесса*. Скорость ее роста в зависимости от величины n существенно влияет как на сходимость $L_n(x)$ к $f(x)$, так и на оценку вычислительной погрешности интерполяции. Для равномерных сеток λ_n растет экспоненциально. Это приводит к тому, что построенный на равномерной сетке интерполяционный полином $L_n(x)$ при большом числе узлов может сильно отличаться от приближаемой функции. Так, например, для функции Рунге $f(x) = \dfrac{1}{25x^2+1}$ на отрезке $[-1,1]$ известно, что $\max\limits_{x\in[-1,1]} |L_n(x) - f(x)| \to \infty$ при $n \to \infty$. Для *чебышёвских узлов* соответствующий интерполяционный полином сходится к указанной функции; это верно и для произвольной непрерывно дифференцируемой функции: если $f(x)$ удовлетворяет неравенству $\max\limits_{[-1,1]} |f^{(m)}(x)| < \infty$, то для интерполяционного многочлена, построенного по чебышёвским узлам, справедливо соотношение $\max\limits_{[-1,1]} |f(x) - L_n(x)| = O(n^{-m} \ln n)$ при $n \to \infty$.

Если приближаемая функция не обладает достаточной гладкостью, то никакая *таблица узлов интерполяции* не может гарантировать сходимость интерполяционного процесса. Под таблицей узлов интерполяции на отрезке $[a,b]$ понимают любой треугольный массив

$$\begin{array}{cccc} x_1^1 & & & \\ x_1^2 & x_2^2 & & \\ x_1^3 & x_2^3 & x_3^3 & \\ \dots & \dots & \dots & \dots \end{array}$$

с тем свойством, что все $x_i^j \in [a,b]$ и элементы каждой строки различны.

Теорема Фабера. *Для любой заданной таблицы узлов интерполяции на отрезке $[a,b]$, существует непрерывная на этом отрезке функция $f(x)$ такая, что погрешность $||L_n(x) - f(x)||$ в равномерной норме не стремится к нулю при $n \to \infty$.*

3.1. Построить многочлен Лагранжа при $n = 3$ для следующих случаев:

1) $x_1 = -1, x_2 = 0, x_3 = 1,$ $f_1 = 3, f_2 = 2, f_3 = 5;$

2) $x_1 = 1, x_2 = 2, x_3 = 4,$ $f_1 = 3, f_2 = 4, f_3 = 6.$

Ответ: 1) $L_3(x) = 2x^2 + x + 2$; 2) $L_3(x) = x + 2$.

3.2. Построение многочлена Лагранжа $L_n(x)$ эквивалентно задаче нахождения коэффициентов c_i из системы уравнений $\sum\limits_{i=0}^{n-1} c_i x_j^i = f_j$ при $j = 1, \dots, n$. Показать, что эта система при больших n может быть близка к вырожденной.

У к а з а н и е. Определителем данной системы уравнений является определитель Вандермонда, следовательно, задача вычисления коэффициентов искомого многочлена имеет единственное решение. Пусть узлы интерполяции принадлежат отрезку $[0,1]$. Функции x^{n-2}, x^{n-1} при больших n на этом отрезке почти неразличимы, поэтому столбцы $(x_1^{n-2},\dots,x_n^{n-2})^T$ и $(x_1^{n-1},\dots,x_n^{n-1})^T$ матрицы получатся близкими.

3.3. Найти $\sum\limits_{i=1}^{n} x_i^p \Phi_i(x)$ при $p=0,\dots,n$.

О т в е т: x^p при $p=0,\dots,n-1$, и $x^n-\omega_n(x)$ при $p=n$.

3.4. Пусть на отрезке $[a,b]$ заданы *равноотстоящие* узлы: $x_i = a+\frac{b-a}{n-1}(i-1)$, $i=1,\dots,n$. Вычислить $\|\omega_n(x)\|$ при $n=2,3,4$.

$\lhd$ Пусть $n=3$. Выполним в формуле

$$\omega_3(x)=(x-a)\left(x-\frac{a+b}{2}\right)(x-b)$$

стандартную замену переменных

$$x=\frac{a+b}{2}+\frac{b-a}{2}\,y,\ \text{где}\ \ y\in[-1,1]\,.$$

В результате получим

$$\omega_3(y)=\left(\frac{b-a}{2}\right)^3(y^3-y).$$

Точки экстремума кубического многочлена y^3-y на $[-1,1]$ равны соответственно $y_{1,2}=\pm\frac{1}{\sqrt{3}}$. Следовательно,

$$\|\omega_3(x)\|=|\omega_3(y_{1,2})|=\frac{(b-a)^3}{12\sqrt{3}}\ .$$

Рассуждая аналогично для $n=2$ и $n=4$, получаем

$$\|\omega_2(x)\|=\frac{(b-a)^2}{4}\,,\quad \|\omega_4(x)\|=\frac{(b-a)^4}{81}\,. \qquad \rhd$$

3.5. Для многочлена $\omega_n(x)$ с равноотстоящими корнями на отрезке $[a,b]$ получить оценку $\|\omega_n(x)\|\leqslant\frac{(b-a)^n(n-1)!}{4(n-1)^n}$ при $n\geqslant 2$.

$\lhd$ Выполним в формуле

$$\omega_n(x)=\prod_{j=1}^{n}(x-x_j),$$

где $x_j=a+\frac{b-a}{n-1}(j-1)$, $j=1,\dots,n$, $n\geqslant 2$, замену переменных

$$x=\frac{na-b}{n-1}+\frac{b-a}{n-1}\,y,\ \text{где}\ \ y\in[1,n]\,.$$

В результате получим

$$\omega_n(x(y)) \equiv \omega_n(y) = \left(\frac{b-a}{n-1}\right)^n \prod_{j=1}^{n}(y-j).$$

Покажем, что справедливо неравенство

$$\max_{y\in[1,n]} \prod_{j=1}^{n} |y-j| \;\leqslant\; \frac{(n-1)!}{4}$$

с помощью специальной параметризации аргумента y. Пусть $y = k+t$, где k — целое. При $2 \leqslant k \leqslant n-1$ будем предполагать, что $|t| \leqslant \frac{1}{2}$; при $k=1$ параметр t принимает значение из отрезка $\left[0, \frac{1}{2}\right]$, а при $k=n$ — из отрезка $\left[-\frac{1}{2}, 0\right]$. Отметим равенство

$$\prod_{j=1}^{n} |y-j| = |t|\,(t+1)\dots(t+k-1)(1-t)\dots(n-k-t).$$

При $t>0$ справедливы неравенства

$$(t+1)\dots(t+k-1) < k! \quad \text{и} \quad |t|(1-t)\dots(n-k-t) < \tfrac{1}{4}\,(n-k)!,$$

а при $t<0$ — неравенства

$$|t|(t+1)\dots(t+k-1) < \tfrac{1}{4}\,(k-1)! \quad \text{и} \quad (1-t)\dots(n-k-t) < (n-k+1)!.$$

В обоих случаях использование соотношения

$$k!(n-k)! \leqslant (n-1)!\,, \;\; 1 \leqslant k < n$$

приводит к искомому неравенству.

Окончательно имеем

$$\|\omega_n(x)\| = \max_{x\in[a,b]} \left|\prod_{j=1}^{n}(x-x_j)\right| = \left(\frac{b-a}{n-1}\right)^n \max_{y\in[1,n]} \left|\prod_{j=1}^{n}(y-j)\right| \leqslant \frac{(b-a)^n(n-1)!}{4(n-1)^n}\,. \;\triangleright$$

3.6. Функция $f(x)$ приближается на $[a,b]$ по n равноотстоящим узлам $x_i = a + \frac{b-a}{n-1}\,(i-1), i = 1,\dots,n$. Найти наибольшее целое p в оценке погрешности $\|f(x) - L_n(x)\| \leqslant 10^{-p}$ в равномерной норме для следующих случаев: 1) $[0,0,1], f(x) = \sin 2x, n=2$; 2) $[-1,0], f(x) = \mathrm{e}^x, n=3$.

Ответ: 1) $p=3$; 2) $p=2$.

3.7. Приближение к числу $\ln 15,2$ вычислено следующим образом. Найдены точные значения $\ln 15$ и $\ln 16$ и построена линейная интерполяция между этими числами. Показать, что если x и y — соответственно точное и интерполированное значения $\ln 15,2$, то справедлива оценка $0 < x - y < 4 \cdot 10^{-4}$.

Указание. Использовать выпуклость функции $\ln x$ и представление погрешности (но не оценку погрешности!).

3.8. Функция $f(x) = \dfrac{1}{A^2 - x}$ приближается на $[-4, -1]$ многочленом Лагранжа по узлам $-4, -3, -2, -1$. При каких значениях A оценка погрешности в равномерной норме не превосходит 10^{-5}?

$\lhd$ Поскольку $f^{(4)}(x) = \dfrac{4!}{(A^2 - x)^5}$ и $\|\omega_4(x)\| = 1$, для оценки погрешности имеем

$$\|f(x) - L_4(x)\| \leqslant \left\| \frac{1}{(A^2 - x)^5} \right\| = \frac{1}{(A^2 + 1)^5} \leqslant 10^{-5}.$$

Следовательно, $|A| \geqslant 3$. $\rhd$

3.9. Доказать, что если узлы интерполяции расположены симметрично относительно некоторой точки c, а значения интерполируемой функции в симметричных узлах равны, то интерполяционный многочлен Лагранжа — функция, четная относительно точки c.

$\lhd$ Покажем сначала справедливость следующего представления: $\Phi_i(x) = \dfrac{\omega_n(x)}{(x - x_i)\omega_n'(x_i)}$. Действительно, так как $\omega_n'(x) = \sum\limits_{k=1}^{n} \prod\limits_{\substack{j=1 \\ j \neq k}}^{n} (x - x_j)$, и при $x = x_i, k \neq i$ каждое из произведений под знаком суммирования обращается в нуль, то $\omega_n'(x_i) = \prod\limits_{\substack{j=1 \\ j \neq i}}^{n} (x_i - x_j)$.

Без ограничения общности можно считать $c = 0$, т. е. $x_i = -x_{n+1-i}$, $i = 1, \ldots, n$. Рассмотрим теперь два слагаемых из общей формулы многочлена Лагранжа, соответствующих равным значениям функции f_k и f_{n+1-k} для некоторого k. Вынося одинаковый числовой множитель за скобку, получим

$$f_k \left[\frac{\omega_n(x)}{(x - x_k)\omega_n'(x_k)} + \frac{\omega_n(x)}{(x - x_{n+1-k})\omega_n'(x_{n+1-k})} \right] = \\ = f_k \left[\frac{\omega_n(x)}{(x - x_k)\omega_n'(x_k)} + \frac{\omega_n(x)}{(x + x_k)\omega_n'(-x_k)} \right].$$

Для четного n функция $\omega_n(x)$ — четная, а ее производная $\omega_n'(x)$ — нечетная. Поэтому выражение в квадратных скобках принимает вид $\dfrac{\omega_n(x)}{x^2 - x_k^2} \cdot \dfrac{2x_k}{\omega_n'(x_k)}$, являясь, очевидно, четной функцией.

Аналогично для нечетного n функция $\omega_n(x)$ — нечетная, а ее производная $\omega_n'(x)$ — четная, и выражение в квадратных скобках также является четной функцией. В данном случае $x = 0$ является узлом интерполяции с номером $k = \dfrac{n+1}{2}$, и у этого слагаемого нет пары. Но само слагаемое — четное, что и завершает доказательство.

Доказательство также может быть получено методом от противного из единственности многочлена Лагранжа для заданного набора узлов и значений, так как отражение относительно середины отрезка не меняет входных данных задачи. $\rhd$

3.10. Показать, что многочлен Лагранжа может быть построен рекуррентным способом:

$$L_1(x) = f(x_1),\ L_n(x) = L_{n-1}(x) + [f(x_n) - L_{n-1}(x_n)]\,\frac{\omega_{n-1}(x)}{\omega_{n-1}(x_n)}\,,\ n \geqslant 2,$$

где

$$\omega_1(x) = x - x_1,\ \omega_n(x) = \omega_{n-1}(x)\,(x - x_n).$$

3.11. Построить многочлен Лагранжа $L_n(x)$ степени $n-1$, удовлетворяющий условиям $L_n(x_k) = y_k$:

1) $n = 4;\ x_1 = 0,\ x_2 = 1,\ x_3 = 2,\ x_4 = 4;\ y_1 = 2,\ y_2 = 3,\ y_3 = 4,\ y_4 = 6;$

2) $n = 3;\ x_k = 2k - 1,\ y_k = 8\sin\frac{\pi}{6}\,(2k-1),\ k = 1, 2, 3.$

3.12. Построить интерполяционный многочлен для функции $f(x) = |x|$ по узлам $-1, 0, 1$.

3.13. Построить интерполяционный многочлен для функции $f(x) = x^2$ по узлам $0, 1, 2, 3$.

3.14. Построить многочлен Лагранжа $L_4(x)$ третьей степени, удовлетворяющий условиям $L_4(x_k) = y_k$: $x_k = k-5$, $y_k = 3k^3+2k^2+k+1$, $k = 1, 2, 3, 4$.

3.15. Функция $f(x)$ приближается на $[a, b]$ по n равноотстоящим узлам $x_i = a + \frac{b-a}{n-1}\,(i-1)$, $i = 1, \ldots, n$. Найти наибольшее целое p в оценке погрешности $\|f(x) - L_n(x)\| \leqslant 10^{-p}$ в равномерной норме для следующих случаев: 1) $f(x) = \frac{1}{\pi}\int\limits_0^{\pi} \cos(x\sin t)dt$, $[0, 1]$, $n = 3$; 2) $f(x) = \ln x$, $[1, 2]$, $n = 4$.

3.16. Оценить погрешность приближения функции e^x интерполяционным многочленом Лагранжа $L_2(x)$, построенным по узлам $x_0 = 0,0,\ x_1 = 0,1$, $x_2 = 0,2$, в точке: 1) $x = 0{,}05$; 2) $x = 0{,}15$.

3.17. Функция $\sin x$ приближается на отрезке $\left[0, \frac{\pi}{4}\right]$ интерполяционным многочленом по значениям в точках $0, \frac{\pi}{8}, \frac{\pi}{4}$. Оценить погрешность интерполяции на этом отрезке.

3.18. Функция $\ln x$ приближается на отрезке $[1, 2]$ интерполяционным многочленом третьей степени по четырем узлам 1, $\frac{4}{3}$, $\frac{5}{3}$, 2. Доказать, что погрешность интерполяции в равномерной норме не превосходит $\frac{1}{300}$.

3.19. Функция $f(x) = \exp(2x)$ приближается на отрезке $\left[-\frac{1}{2}, \frac{1}{2}\right]$ интерполяционным многочленом второй степени по трем узлам: $-\frac{1}{2}$, 0, $\frac{1}{2}$. Доказать, что погрешность интерполяции в равномерной норме не превосходит $\frac{\sqrt{3}}{9}$.

3.20. Оценить погрешность интерполяции функции $f(x) = \operatorname{arctg} x$ на отрезке $[0, 1]$ многочленом Лагранжа пятой степени, построенным по равноотстоящим узлам.

3.21. Оценить число равноотстоящих узлов интерполяции на отрезке $\left[0, \frac{\pi}{4}\right]$, обеспечивающее точность $\varepsilon \leqslant 10^{-2}$ приближения функции $f(x) = \sin x$.

3.22. Определить степень многочлена Лагранжа на равномерной сетке, обеспечивающую точность приближения функции e^x на отрезке $[0, 1]$ не хуже 10^{-3}.

3.23. Пусть функция $f(x) = \sin x$ задана на отрезке $[0, b]$. При каком b многочлен Лагранжа $L_3(x)$, построенный на равномерной сетке, приближает эту функцию с погрешностью $\varepsilon \leqslant 10^{-3}$?

3.24. Привести пример непрерывной на отрезке $[-1, 1]$ функции, для которой интерполяционный процесс Лагранжа на равномерной сетке расходится.

Ответ: например, функция Рунге или $|x|$.

3.25. Пусть функция $f(x)$ задана на $[a, b]$ и $\max\limits_{x\in[a,b]} |f''(x)| \leqslant 1$. Оценить погрешность приближения $f(x)$ кусочно-линейным интерполянтом, построенным на равномерной сетке с шагом h.

3.26. С каким шагом следует составлять таблицу функции $\sin x$ на $\left[0, \frac{\pi}{2}\right]$, чтобы погрешность линейной интерполяции не превосходила $0,5 \cdot 10^{-6}$?

3.27. Пусть $f \in C^{(1)}[a, b]$ и $p(x)$ — полином, аппроксимирующий $f'(x)$ с точностью ε в норме $C[a, b]$. Доказать, что полином $q(x) = f(a) + \int\limits_a^x p(t)dt$ аппроксимирует $f(x)$ с точностью $\varepsilon(b - a)$ в норме $C[a, b]$.

3.28. Построить многочлен $P_3(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, удовлетворяющий условиям: $P_3(-1) = 0$, $P_3(1) = 1$, $P_3(2) = 2$, $a_3 = 1$.

3.29. Построить многочлен $P_3(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, удовлетворяющий условиям: $P_3(0) = P_3(-1) = P_3(1) = 0$, $a_2 = 1$.

3.30. Построить многочлен $P_3(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, удовлетворяющий условиям: $P_3(-1) = 0$, $P_3(1) = 1$, $P_3(2) = 2$, $a_1 = 1$.

3.31. Построить многочлен $P_3(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, удовлетворяющий условиям: $P_3(-1) = P_3(-2) = P_3(1) = 0$, $a_0 = 1$.

3.32. Построить многочлен $P_4(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$, удовлетворяющий условиям: $\sum\limits_{i=0}^{4} a_i = 0, P(0) = 0, P(-1) = 1, P(2) = 2, P(3) = 3$.

3.33. Построить многочлен $P_4(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$, удовлетворяющий условиям: $P_4(1) = P_4(-1) = P_4'(0) = P_4''(0) = 0, P_4(0) = 1$.

3.34. Построить многочлен $P_4(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$, удовлетворяющий условиям: $P_4(0) = 0, P_4(1) = 1, P_4(2) = 2, P_4(3) = 3, \sum\limits_{i=1}^{4} a_i = 0$.

3.35. Доказать при целых t формулу:

$$L_n(x_0 + th) = \sum_{k=0}^{n-1} C_t^k \Delta^k f_0, \;\; \Delta^1 f_i = f_{i+1} - f_i, \; \Delta^0 f_i = f_i, \; x_{i+1} = x_i + h.$$

3.36. Доказать при целых t формулу:

$$L_n(x_0 - th) = \sum_{k=0}^{n-1} (-1)^k C_t^k \nabla^k f_0, \;\; \nabla^1 f_i = f_i - f_{i-1}, \; \nabla^0 f_i = f_i, \; x_{i+1} = x_i + h.$$

3.37. Доказать при целых t формулу:

$$L_n(x_0 + th) = \sum_{k=0}^{n-1} C_t^k \delta^k f_{k/2}, \;\; \delta^1 f_i = f_{i+1/2} - f_{i-1/2}, \; \delta^0 f_i = f_i, \; x_{i+1} = x_i + h.$$

3.38. Доказать, что если многочлен $P_s(x)$ степени $s-1$ удовлетворяет условиям

$$\begin{gathered}
P_s(x_1) = f(x_1), \quad \ldots, \quad P_s^{(M_1-1)}(x_1) = f^{(M_1-1)}(x_1), \\
P_s(x_2) = f(x_2), \quad \ldots, \quad P_s^{(M_2-1)}(x_2) = f^{(M_2-1)}(x_2), \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
P_s(x_n) = f(x_n), \quad \ldots, \quad P_s^{(M_n-1)}(x_n) = f^{(M_n-1)}(x_n), \\
M_1 + M_2 + \cdots + M_n = s,
\end{gathered}$$

то справедливо равенство

$$f(x) - P_s(x) = \frac{f^{(s)}(\xi)}{s!}\,\omega(x), \;\; \omega(x) = \prod_{i=1}^{n} (x - x_i)^{M_i}.$$

3.39. Пусть $a \leqslant x \leqslant b$ и $-1 \leqslant y \leqslant 1$ и узлы интерполяции x_i и y_i, $i = 1, \ldots, n$ связаны линейным соотношением $x_i = x(y_i) = \frac{a+b}{2} + \frac{b-a}{2}\, y_i$. Доказать, что константы Лебега $\lambda_n^{[a,b]}$ и $\lambda_n^{[-1,1]}$, соответствующие этим отрезкам, совпадают.

$\lhd$ По определению, вспомогательные многочлены $(n-1)$-й степени $\Phi_i(y)$, $i = 1, \ldots, n$ обладают свойством $\Phi_i(y_k) = \delta_i^k$. Положим в формуле для $\Phi_i(x)$, обладающей теми же свойствами, $x = x(y)$. Линейное преобразование не меняет степени многочлена. Кроме того, $\Phi_i(x_k) = \Phi_i(x(y_k)) = \Phi_i(y_k) = \delta_i^k$, т. е. два многочлена $(n-1)$-й степени, совпадают в n точках. Отсюда следует их тождественное совпадение, следовательно, равенство констант Лебега $\lambda_n^{[a,b]}$ и $\lambda_n^{[-1,1]}$.

Таким образом, величина λ_n не зависит от длины и расположения отрезка интерполяции $[a, b]$, а определяется только взаимным расположением узлов. $\rhd$

3.40. Показать, что для системы равноотстоящих узлов $\{x_i = i,\ i = 1, \dots, n\}$ при $n \geqslant 2$ справедлива оценка снизу для константы Лебега $\lambda_n \geqslant K\,\frac{2^n}{n^{3/2}}$ с постоянной K, не зависящей от n.

$\lhd$ По определению λ_n на отрезке $[1, n]$ имеем

$$\lambda_n = \max_{x\in[1,n]} \sum_{i=1}^{n} \left| \prod_{\substack{j=1\\ j\neq i}}^{n} \frac{x-j}{i-j} \right| .$$

Справедливы следующие соотношения:

$$\prod_{\substack{j=1\\ j\neq i}}^{n} |i-j| = (i-1)!\,(n-i)!\,, \qquad \prod_{j=1}^{n} \left(j - \tfrac{1}{2}\right) \geqslant \tfrac{n!}{2\sqrt{n}}\,,\ \ n \geqslant 1\,,$$

первое из которых очевидно, а второе доказывается по индукции. Проведем с их помощью оценку снизу для λ_n:

$$\lambda_n = \max_{x\in[1,n]} \sum_{i=1}^{n} \frac{1}{(i-1)!(n-i)!} \prod_{\substack{j=1\\ j\neq i}}^{n} |x-j| \geqslant \sum_{i=1}^{n} \frac{1}{(i-1)!(n-i)!} \prod_{\substack{j=1\\ j\neq i}}^{n} \left|\tfrac{3}{2} - j\right|$$

(использовано неравенство $\max\limits_{x\in[1,n]} |f(x)| \geqslant |f(3/2)|$). Для оценки произведения в правой части выполним преобразования:

$$\prod_{\substack{j=1\\ j\neq i}}^{n} \left|\tfrac{3}{2} - j\right| = \frac{1}{|i-\frac{3}{2}\,|} \prod_{j=1}^{n} \left|\tfrac{3}{2} - j\right| = \frac{1}{2|i-\frac{3}{2}\,|} \prod_{j=1}^{n-1} \left|\tfrac{1}{2} - j\right| \geqslant \frac{(n-1)!}{4(n-\frac{3}{2})\sqrt{n-1}}\ .$$

Наконец, получим искомое неравенство ($K = 1/8$):

$$\lambda_n \geqslant \frac{1}{4(n-\frac{3}{2})\sqrt{n-1}} \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!\,(n-i)!} \geqslant \frac{1}{4n^{3/2}} \sum_{i=1}^{n} C_{n-1}^{i-1} = \tfrac{1}{8}\ \frac{2^n}{n^{3/2}}\ . \qquad \rhd$$

3.41. Показать, что для системы равноотстоящих узлов $\{x_i = i,\ i = 1, \dots, n\}$ при $n \geqslant 2$ справедлива оценка сверху для константы Лебега $\lambda_n \leqslant K\,2^n$ с постоянной K, не зависящей от n.

$\lhd$ Покажем (ср. 3.5), что справедливо неравенство

$$\max_{x\in[1,n]} \prod_{\substack{j=1\\ j\neq i}}^{n} |x-j| \leqslant (n-1)!$$

с помощью специальной параметризации аргумента x. Пусть $x = k + t$, где k — целое. При $2 \leqslant k \leqslant n-1$ будем предполагать, что $|t| \leqslant \frac{1}{2}$; при $k = 1$ параметр t принимает значение из отрезка $\left[0, \frac{1}{2}\right]$, а при $k = n$ — из отрезка $\left[-\frac{1}{2}, 0\right]$. Отметим равенство

$$\prod_{\substack{j=1\\ j\neq i}}^{n} |x-j| = \left|\frac{t}{k-i+t}\right| (t+1)\dots(t+k-1)(1-t)\dots(n-k-t).$$

При $t > 0$ справедливы неравенства

$$(t+1)\dots(t+k-1) < k! \quad \text{и} \quad (1-t)\dots(n-k-t) < (n-k)!,$$

а при $t < 0$ — неравенства

$$(t+1)\dots(t+k-1) < (k-1)! \quad \text{и} \quad (1-t)\dots(n-k-t) < (n-k+1)!.$$

В обоих случаях использование соотношений

$$\left|\frac{t}{k-i+t}\right| \leqslant 1, \quad k!(n-k)! \leqslant (n-1)!\,, \quad 1 \leqslant k < n$$

приводит к искомому неравенству.

Тогда из решения 3.40 имеем

$$\lambda_n = \max_{x\in[1,n]} \sum_{i=1}^{n} \frac{1}{(i-1)!\,(n-i)!} \prod_{\substack{j=1\\ j\neq i}}^{n} |x-j| \leqslant \sum_{i=1}^{n} C_{n-1}^{i-1} = K\,2^n,\ K = \frac{1}{2}\,.$$

Оценка доказана. ▷

3.42. Определить узлы интерполяции, при которых константа Лебега λ_3 минимальна.

Ответ: константа Лебега не зависит от отрезка, поэтому будем считать, что $x \in [-1,1]$, тогда $x_1 = -\xi$, $x_2 = 0$, $x_3 = \xi$, где ξ — произвольное число из отрезка $\left[\frac{\sqrt{8}}{3}, 1\right]$; $\lambda_3 = \frac{5}{4}$.

3.43. Показать, что если $x_1, \dots, x_{2n}$ — вещественные, то функция $T(x) = \prod_{k=1}^{2n} \sin\frac{x-x_k}{2}$ является тригонометрическим полиномом вида $T(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx)$ с вещественными коэффициентами a_k, b_k.

3.44. Доказать, что интерполяционный тригонометрический полином $T(x)$, удовлетворяющий условиям $T(x_j) = y_j$, $j = 0, 1, \dots, 2n$, где $0 \leqslant x_0 < x_1 < \dots < x_{2n} < 2\pi$, может быть записан в виде

$$T(x) = \sum_{k=0}^{2n} y_k t_k(x)\,, \quad \text{где} \quad t_k(x) = \prod_{\substack{s=0\\ s\neq k}}^{2n} \sin\frac{x-x_s}{2} \Big/ \sin\frac{x_k-x_s}{2}\,.$$

3.45. Доказать, что для любых $x_0, x_1, \dots, x_{2n}$, удовлетворяющих условиям $0 \leqslant x_0 < x_1 < \dots < x_{2n} < 2\pi$, и для любых $y_0, y_1, \dots, y_{2n}$ существует единственный тригонометрический полином $T(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \sin kx)$, удовлетворяющий условиям $T(x_j) = y_j$, $j = 0, 1, 2, \dots, 2n$. Если при этом $y_0, y_1, \dots, y_{2n}$ — вещественные, то и коэффициенты a_k, b_k являются вещественными.

3.46. Доказать, что для любых $x_0, x_1, \dots, x_n$, удовлетворяющих условиям $0 \leqslant x_0 < x_1 < \dots < x_n < \pi$, и для любых $y_0, y_1, \dots, y_n$ существует единственный тригонометрический полином $C(x) = \sum_{k=0}^{n} a_k \cos kx$, удовлетворяющий условиям $C(x_j) = y_j$, $j = 0, 1, 2, \dots, n$.

3.47. Построить тригонометрический полином на отрезке $[0,1]$ по заданным значениям $f(0),\ f(h),\ f(2h),\ f(3h),\ h=\frac{1}{3}$.

3.48. 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 следующим условиям: $T_2(0)=0,\ T_2\left(\frac{\pi}{4}\right)=1,\ T_2\left(\frac{\pi}{2}\right)=1,\ T_2\left(\frac{3\pi}{4}\right)=1,$ $T_2(\pi)=1$.

3.49. Построить интерполяционный тригонометрический полином минимальной степени по заданным значениям $f(-\pi)=0,\ f\left(-\frac{\pi}{2}\right)=0,\ f\left(\frac{\pi}{2}\right)=1$.

3.50. Доказать, что тригонометрический полином $T_n(z)$ степени n имеет в любой полосе $\mathrm{Re}(z)\in[a,a+2\pi]$ ровно $2n$ корней.

3.51. Пусть $T_n(x)$ — тригонометрический интерполяционный многочлен степени n, построенный по равноотстоящим узлам на $[0,2\pi]$ для функции $f(x)\in C^{(\alpha)},\ \alpha>0$. Доказать, что в равномерной норме

$$\lim_{n\to\infty}||T_n-f||=0\,.$$

3.52. Вычислить для 2π-периодической функции

$$H(x)=\begin{cases}1 & \text{при} \quad x\in[0,\pi],\\ 0 & \text{при} \quad x\in(\pi,2\pi)\end{cases}$$

частичную сумму ряда Фурье $H_{2n}(x)$ и проанализировать их близость.

$\lhd$ При вычислении суммы первых $2n$ членов коэффициенты при косинусах равны нулю, поэтому

$$H_{2n}(x)=\frac{1}{2}+\frac{2}{\pi}\sum_{k=1}^{n}\frac{\sin(2k-1)x}{2k-1}\,.$$

Преобразуем полученное выражение

$$H_{2n}(x)=\frac{1}{2}+\frac{2}{\pi}\sum_{k=1}^{n}\int_0^x\cos(2k-1)t\,dt=$$

$$=\frac{1}{2}+\frac{2}{\pi}\int_0^x\sum_{k=1}^{n}\cos(2k-1)t\,dt=\frac{1}{2}+\frac{1}{\pi}\int_0^x\frac{\sin 2nt}{\sin t}\,dt,$$

из которого следует, что максимум и минимум для $0\leqslant x\leqslant\pi$ достигаются в точках

$$\frac{d}{dx}H_{2n}(x)=\frac{1}{\pi}\frac{\sin 2nx}{\sin x}=0,$$

т. е. при $x_m=\frac{m\pi}{2n},\ m=1,2,\ldots,2n-1$. При этом экстремумы чередуются. Непосредственные вычисления показывают, что $H_{2n}(0)=0{,}5$, $H_{2n}\left(\frac{\pi}{2n}\right)\to 1{,}08949\ldots$ с дальнейшим убыванием амплитуды колебаний по мере удаления от точки разрыва.

Отклонение разрывной функции от ее ряда Фурье часто называют *эффектом Гиббса.* $\rhd$

3.53. Функция двух переменных $f(x_1, x_2)$ аппроксимируется интерполяционным многочленом $P(x_1, x_2) = a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2$. При этом $f(0,0) = 1, f(1,0) = 2, f(0,1) = 4, f(1,1) = 3$. Найти $P\left(\frac{1}{2}, \frac{1}{2}\right)$.

3.54. Пусть $P(x_1, x_2)$ — многочлен от двух переменных степени не выше n по каждой переменной и $P\left(\frac{k}{n}, \frac{m}{n}\right) = 0,\ k, m = 0, 1, ..., n$. Доказать, что $P(x_1, x_2) \equiv 0$.

3.2. Многочлены Чебышёва

Имеется несколько способов определения последовательности многочленов Чебышёва первого рода. Рассмотрим некоторые из них.

а) *Рекуррентное соотношение*:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2x\,T_n(x) - T_{n-1}(x)\,.$$

б) *Тригонометрическая форма*. При любом η имеем

$$\cos\left((n+1)\eta\right) = 2\cos\eta\cos(n\eta) - \cos\left((n-1)\eta\right).$$

Полагая $\eta = \arccos x$, получаем

$$T_n(x) = \cos(n\arccos x).$$

Простое следствие: $|T_n(x)| \leqslant 1$ при $|x| \leqslant 1$.

в) *Разностное уравнение*. Рекуррентное соотношение является разностным уравнением по переменной n. Ему соответствует характеристическое уравнение $\mu^2 - 2x\mu + 1 = 0$. Следовательно, $\mu_{1,2} = x \pm \sqrt{x^2-1}$. При $x \neq \pm 1$ справедливо $T_n(x) = C_1\mu_1^n + C_2\mu_2^n$. Из начальных условий получаем $C_1 = C_2 = \frac{1}{2}$, что приводит к формуле

$$T_n(x) = \frac{1}{2}\left(\left(x + \sqrt{x^2-1}\right)^n + \left(x - \sqrt{x^2-1}\right)^n\right).$$

В силу непрерывности многочлена формула верна и при $x = \pm 1$.

Отметим, что все многочлены $T_{2n}(x)$ — четные, а $T_{2n+1}(x)$ — нечетные. При этом коэффициент при старшем члене равен 2^{n-1}.

3.55. Доказать следующие свойства многочленов Чебышёва:

1) $T_{2n}(x) = 2\,T_n^2(x) - 1$;

2) $I_{mn} = \int\limits_{-1}^{1} \frac{T_n(x)\,T_m(x)}{\sqrt{1-x^2}}\,dx = \begin{cases} 0 & \text{при } n \neq m, \\ \frac{\pi}{2} & \text{при } n = m \neq 0, \\ \pi & \text{при } n = m = 0; \end{cases}$

3) $\int\limits_{-1}^{x} T_n(y)dy = \frac{1}{2}\left(\frac{1}{n+1}\,T_{n+1}(x) - \frac{1}{n-1}\,T_{n-1}(x)\right) - \frac{(-1)^n}{n^2-1}\,,\ n \geqslant 2$;

4) $\left(1-x^2\right)T_n''(x) - x\,T_n'(x) + n^2\,T_n(x) = 0, \quad n \geqslant 0$.

$\lhd$ 1) Следствием тригонометрического тождества

$$\cos\left((n+m)\eta\right)+\cos\left((n-m)\eta\right)=2\cos(n\eta)\cos(m\eta)$$

является полиномиальное тождество

$$2T_n(x)\,T_m(x)=T_{n+m}(x)+T_{n-m}(x),\quad n\geqslant m\geqslant 0\,,$$

из которого при $n=m$ следует искомое выражение.

2) Положим $x=\cos\eta$, тогда $dx=-\sin\eta d\eta$ и

$$I_{mn}=\int\limits_0^{\pi}\cos(n\eta)\cos(m\eta)d\eta=\frac{\pi}{2}\left(\delta^0_{n-m}+\delta^0_{n+m}\right).$$

3) Так как $\dfrac{T_n'(x)}{n}=\dfrac{-\sin(n\arccos x)}{-\sqrt{1-x^2}}$, то, полагая $x=\cos\eta$, имеем

$$\frac{1}{2}\left(\frac{1}{n+1}\,T'_{n+1}(x)-\frac{1}{n-1}\,T'_{n-1}(x)\right)=$$

$$=\frac{\sin((n+1)\eta)-\sin((n-1)\eta)}{2\sin\eta}=\frac{2\cos(n\eta)\sin\eta}{2\sin\eta}=T_n(x)\,;$$

теперь искомое равенство справедливо с точностью до постоянной, которую легко определить, поскольку $T_n(-1)=(-1)^n$.

4) Непосредственно дифференцированием вычисляется $T_n''(x)$; напомним, что $(\arccos x)'=-(1-x^2)^{-1/2}$. $\rhd$

3.56. Пусть $x^2+y^2=1$. Доказать, что $T_{2n}(y)=(-1)^nT_{2n}(x)$.

3.57. Найти все нули многочленов Чебышёва $T_n(x)$.

Ответ: $x_m=\cos\dfrac{\pi\,(2m-1)}{2n}$, где $m=1,\ldots,n$ (все нули лежат внутри отрезка $[-1,1]$, их ровно n).

3.58. Найти все экстремумы многочлена Чебышёва $T_n(x)$ на отрезке $[-1,1]$.

Ответ: $x_{(m)}=\cos\dfrac{\pi m}{n}$, $m=0,\ldots,n$ (на $[-1,1]$ имеется $n+1$ экстремум и $T_n(x_{(m)})=(-1)^m$).

3.59. Доказать, что *приведенный многочлен Чебышёва* $\overline{T}_n(x)=2^{1-n}T_n(x)$ наименее уклоняется от нуля среди всех многочленов $P_n(x)$ со старшим коэффициентом 1 на отрезке $[-1,1]$, т. е.

$$\|P_n(x)\|=\max_{[-1,1]}|P_n(x)|\geqslant\max_{[-1,1]}|\overline{T}_n(x)|=2^{1-n}\,.$$

$\lhd$ Пусть $\|P_n(x)\|<2^{1-n}$. Тогда в точках экстремума многочлена Чебышёва знак разности $\overline{T}_n(x)-P_n(x)$ определяется знаком $\overline{T}_n(x)$:

$$\operatorname{sign}\left(\overline{T}_n(x_{(m)})-P_n(x_{(m)})\right)=\operatorname{sign}\left((-1)^m2^{1-n}-P_n(x_{(m)})\right)=(-1)^m\,.$$

При этом указанная разность является отличным от нуля многочленом степени $n-1$, но имеет n нулей, поскольку $n+1$ раз меняет знак в точках экстремума, т. е. $P_n(x)\equiv T_n(x)$, что невозможно в силу $\|P_n\|<\|T_n\|$. Полученное противоречие завершает доказательство. $\rhd$

3.60. Доказать единственность многочлена, наименее уклоняющегося от нуля на отрезке $[-1,1]$ среди всех многочленов со старшим коэффициентом 1.

3.61. Найти многочлен, наименее уклоняющийся от нуля на отрезке $[a,b]$ среди всех многочленов со старшим коэффициентом 1.

$\lhd$ Выполним линейную замену переменных $x = \frac{a+b}{2} + \frac{b-a}{2}\,x'$ для отображения отрезка $[-1,1]$ в заданный отрезок $[a,b]$. Многочлен $\overline{T}_n(x')$ при этом преобразуется в многочлен $\overline{T}_n\left(\frac{2x-(b+a)}{b-a}\right)$ со старшим коэффициентом $\left(\frac{2}{b-a}\right)^n$. В результате перенормировки и использования схемы доказательства из 3.59 имеем

$$\overline{T}_n^{[a,b]}(x) = (b-a)^n 2^{1-2n} T_n\left(\frac{2x-(b+a)}{b-a}\right). \qquad \rhd$$

3.62. Пусть $\omega_n(x) = \prod\limits_{i=1}^{n}(x-x_i)$. Показать, что при любом выборе узлов $x_i \in [a,b]$ имеет место неравенство $\|\omega_n(x)\| \geqslant (b-a)^n\,2^{1-2n}$. Сравнить полученный результат с аналогичным для равномерного распределения узлов.

У к а з а н и е. Использовать решения 3.61 и 3.5.

3.63. Пусть $0 \leqslant a < b$. В классе многочленов $P_n(x)$ степени n, удовлетворяющих условию $P_n(0) = c \neq 0$, найти наименее уклоняющийся от нуля на отрезке $[a,b]$ и вычислить его равномерную норму.

О т в е т: $P_n^*(x) = c\,\dfrac{T_n\left(\frac{2x-(a+b)}{b-a}\right)}{T_n\left(-\frac{a+b}{b-a}\right)}$, $\|P_n^*(x)\| = \dfrac{2\,c}{q_1^n + q_1^{-n}}$, $q_1 = \dfrac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}$.

3.64. Пусть $k \leqslant n$, $0 \leqslant a < b$. В классе многочленов $P_n(x)$ степени n, удовлетворяющих условию $P_n^{(k)}(0) = c \neq 0$, найти наименее уклоняющийся от нуля на отрезке $[a,b]$.

О т в е т: $P_n^*(x) = c\left(\frac{b-a}{2}\right)^k \dfrac{T_n\left(\frac{2x-(a+b)}{b-a}\right)}{T_n^{(k)}\left(\frac{a+b}{a-b}\right)}$.

3.65. Среди всех многочленов $P_n(x) = x^n + \ldots$ степени $n \geqslant 2$, удовлетворяющих условиям $P_n(-1) = P_n(1) = 0$, найти наименее уклоняющийся от нуля на $[-1,1]$.

О т в е т: $P_n^*(x) = 2^{1-n}\left(\cos\frac{\pi}{2n}\right)^{-n} T_n\left(x\cos\frac{\pi}{2n}\right)$.

3.66. Пусть $P_n(x)$ — многочлен степени n и $\max\limits_{x\in[-1,1]} |P_n(x)| = M$. Доказать, что для всех x, удовлетворяющих условию $|x| \geqslant 1$, выполняется неравенство $|P_n(x)| \leqslant M\,|T_n(x)|$, где $T_n(x)$ — многочлен Чебышёва степени n.

У к а з а н и е. Предположив противное, т. е. допустив существование такого ξ, $|\xi| \geqslant 1$, что $|P_n(\xi)| > M\,|T_n(\xi)|$, получить противоречие, доказав, что у полинома $Q_n(x) = \dfrac{P_n(\xi)}{T_n(\xi)}\ T_n(x) - P_n(x)$, как минимум, $n+1$ нуль.

3.67. Для производных многочлена Чебышёва получить представления следующего вида:

$$\frac{T'_{2n}}{2n} = 2(T_{2n-1} + T_{2n-3} + \cdots + T_1),\quad \frac{T'_{2n+1}}{2n+1} = 2(T_{2n} + T_{2n-2} + \cdots + T_2) + 1\,.$$

У к а з а н и е. Воспользоваться третьим свойством из 3.55 в виде

$$\frac{T'_n}{n} = 2T_{n-1} + \frac{T'_{n-2}}{n-2}\,,\quad n > 2\,.$$

3.68. Пусть функция $f(x)$ представима при $|x| \leqslant 1$ в виде $f(x) = \sum\limits_{k=0}^{\infty} a_k T_k(x)$, где $\sum\limits_{k=0}^{\infty} |a_k| < \infty$, $T_k(x)$ — многочлены Чебышёва. Доказать, что для всех $x \in [-1,1]$ справедливо равенство

$$\int\limits_{-1}^{x} f(t)dt = \frac{a_0}{2}\,x + \sum_{k=1}^{\infty} \frac{1}{2k}\,(a_{k-1} - a_{k+1})T_k(x) + a_0 - \frac{a_1}{4} + \sum_{k=2}^{\infty} \frac{(-1)^{k+1}a_k}{k^2-1}\,.$$

3.69. Вычислить значение многочлена Чебышёва n-й степени в точке: 1) $x = \frac{1}{2}$; 2) $x = -\frac{1}{2}$.

О т в е т: 1) $T_{3k}\left(\frac{1}{2}\right) = (-1)^k$, $\quad T_{3k\pm1}\left(\frac{1}{2}\right) = \frac{(-1)^k}{2}$;

2) $T_{3k}\left(-\frac{1}{2}\right) = 1$, $\qquad T_{3k\pm1}\left(-\frac{1}{2}\right) = -\frac{1}{2}\,.$

3.70. Вычислить значение первой производной многочлена Чебышёва n-й степени в точке: 1)$x = 1$; 2)$x = -1$.

О т в е т: 1) $T'_n(1) = n^2$; $\quad$ 2)$T'_n(-1) = (-1)^{n+1}n^2\,.$

3.71. Функция $f(x) = \sin 2x$ приближается многочленом Лагранжа на отрезке $[0,2]$ по n чебышёвским узлам: $x_i = 1 + \cos\frac{2i-1}{2n}\,\pi$, $i = 1,\ldots,n$. Найти наибольшее целое p в оценке погрешности в равномерной норме вида $\varepsilon \leqslant \frac{1}{3}\ 10^{-p}$, если $n = 6$.

О т в е т: $p = 2$.

3.72. Функция $f(x) = \cos x$ приближается многочленом Лагранжа на $[-1,1]$ по n чебышёвским узлам: $x_i = \cos\frac{2i-1}{2n}\,\pi$, $i = 1,\ldots,n$. Найти наибольшее целое p в оценке погрешности в равномерной норме вида $\varepsilon \leqslant 10^{-p}$, если $n = 5$.

О т в е т: $p = 3$.

3.73. Функция $f(x) = \mathrm{e}^x$ приближается на $[0,1]$ интерполяционным многочленом степени 3 с чебышёвским набором узлов интерполяции: $x_k = \frac{1}{2} + \frac{1}{2}\cos\frac{(2k-1)\pi}{8}$, $k = 1,2,3,4$. Доказать, что погрешность интерполяции в равномерной норме не превосходит величины $\mathrm{e}\cdot 10^{-3}$.

3.74. Среди всех многочленов вида $a_3x^3 + 2\,x^2 + a_1x + a_0$ найти наименее уклоняющийся от нуля на отрезке $[3,5]$.

Ответ: $P(x) = 4\,\dfrac{T_3(x-4)}{T_3^{(2)}(-4)} = -\dfrac{x^3}{6} + 2x^2 - \dfrac{63x}{8} + \dfrac{61}{6}$.

3.75. Среди всех многочленов вида $a_2x^2 + x + a_0$ найти наименее уклоняющийся от нуля на отрезке $[-1,1]$.

Ответ: $a_2 = -a_0$ при любом $|a_0| \leqslant \frac{1}{2}$.

3.76. Среди всех многочленов вида $5x^3 + a_2x^2 + a_1x + a_0$ найти наименее уклоняющийся от нуля на отрезке $[1,2]$.

Ответ: $P(x) = \frac{5}{32}\,T_3(2x-3) = \frac{5}{32}\,(32x^3 - 144x^2 + 210x - 99)$.

3.77. Среди всех многочленов вида $a_3x^3 + a_2x^2 + a_1x + 4$ найти наименее уклоняющийся от нуля на отрезке $[1,3]$.

Ответ: $P(x) = 4\,\dfrac{T_3(x-2)}{T_3(-2)} = -\left(\dfrac{8x^3}{13} - \dfrac{48x^2}{13} + \dfrac{90x}{13} - 4\right)$.

3.78. Среди всех многочленов вида $a_3x^3 + a_2x^2 + 3\,x + a_0$ найти наименее уклоняющийся от нуля на отрезке $[2,4]$.

Ответ: $P(x) = 3\,\dfrac{T_3(x-3)}{T_3'(-3)} = \dfrac{4x^3}{35} - \dfrac{36x^2}{35} + 3x - \dfrac{99}{35}$.

3.79. Доказать следующие представления многочленов Чебышёва:

1) $T_n(x) = \dfrac{(-1)^n 2^n n!}{(2n)!}\,\sqrt{1-x^2}\,\dfrac{d^n}{dx^n}\left((1-x^2)^{n-1/2}\right)$, $n \geqslant 0$;

2) $T_n(x) = \dfrac{1}{n!}\,\dfrac{d^n}{dt^n}\left(\dfrac{1-tx}{1-2tx+t^2}\right)\Big|_{t=0}$, $n \geqslant 0$;

3) $T_n(x) = \dfrac{1}{2}\,\dfrac{1}{n!}\,\dfrac{d^n}{dt^n}\left(\dfrac{1-t^2}{1-2tx+t^2}\right)\Big|_{t=0}$, $n \geqslant 1$;

4) $T_n(x) = -\,\dfrac{1}{2}\,\dfrac{1}{(n-1)!}\,\dfrac{d^n}{dt^n}\left(\ln(1-2tx+t^2)\right)\Big|_{t=0}$, $n \geqslant 1$;

5) $T_n(x) = \dfrac{n}{2}\sum\limits_{k=0}^{[n/2]}(-1)^k\,\dfrac{(n-k-1)!}{k!(n-2k)!}\,(2x)^{n-2k}$, $n \geqslant 1$.

3.80. Показать, что для системы узлов интерполяции $x_i = \cos\dfrac{2i-1}{2n}\,\pi$, $i = 1,\ldots,n$ (нули многочлена Чебышёва $T_n(x)$), справедлива асимптотическая оценка сверху для константы Лебега $\lambda_n \leqslant K\,\ln n$ с постоянной K, не зависящей от n.

$\lhd$ Рассмотрим функцию $\Lambda_n(x) = \sum\limits_{i=1}^{n} \left| \dfrac{\omega_n(x)}{(x-x_i)\omega_n'(x_i)} \right|$. По определению λ_n имеем $\lambda_n = \max\limits_{x\in[-1,1]} \Lambda_n(x)$. Учитывая выбор узлов интерполяции, получим

$$\Lambda_n(x) = \sum_{i=1}^{n} \frac{|\cos(n\arccos x)| \sin\frac{2i-1}{2n}\,\pi}{n|x-\cos\frac{2i-1}{2n}\,\pi|} = \sum_{i=1}^{n} \frac{|\cos(\pi n\varphi)| \sin\frac{2i-1}{2n}\,\pi}{n|\cos(\pi\varphi)-\cos\frac{2i-1}{2n}\,\pi|},$$

где сделана замена $x = \cos(\pi\varphi)$, а φ меняется на отрезке $[0,1]$. Обозначим эту сумму через $\theta(\varphi)$ и заметим, что, в силу симметрии узлов, $\Lambda_n(x)$ — четная функция, поэтому при оценке сверху для $\theta(\varphi)$ достаточно рассматривать только отрезок $\left[0, \frac{1}{2}\right]$.

Так как имеют место неравенства

$$\sin|\alpha| \leqslant |\alpha|,\ \sin|\beta| \geqslant \frac{2\sqrt{2}}{3\pi}\,|\beta| \quad \text{при } |\beta| \leqslant \frac{3}{4}\,\pi,$$

$$\sin|\beta| \geqslant \frac{2}{\pi}\,|\beta| \ \text{ при } \ |\beta| \leqslant \frac{\pi}{2},$$

то при $0 < \beta \leqslant \frac{\pi}{2}$, $0 < \alpha \leqslant \pi$, имеем

$$\frac{|\sin\alpha|}{|\cos\beta-\cos\alpha|} = \frac{|\sin\alpha|}{2\sin\left|\frac{\alpha+\beta}{2}\right|\sin\left|\frac{\alpha-\beta}{2}\right|} \leqslant \frac{3\pi^2}{2\sqrt{2}}\,\frac{\alpha}{|\alpha+\beta|\,|\alpha-\beta|},$$

откуда, если положить

$$\alpha = \frac{2i-1}{2n}\,\pi, \quad \beta = \pi\varphi = \frac{\pi}{2}\,\frac{2m-1-2t}{n},\ 1 \leqslant m \leqslant \frac{n+2}{2},\ 0 \leqslant t \leqslant \frac{1}{2},$$

следует, что

$$\frac{\sin\frac{2i-1}{2n}\,\pi}{|\cos\pi\varphi-\cos\frac{2i-1}{2n}\,\pi|} \leqslant \frac{3\pi n}{4\sqrt{2}}\,\frac{2i-1}{|m+i-1-t||m-i-t|}.$$

Параметризация $\varphi = \dfrac{2m-1-2t}{2n}$, $1 \leqslant m \leqslant 1+\frac{n}{2}$, корректна, так как, полагая m в указанных пределах и изменяя t на $\left[0, \frac{1}{2}\right]$, можно получить любое значение φ (либо $1-\varphi$) из отрезка $\left[0, \frac{1}{2}\right]$. Далее имеем

$$|\cos\pi n\varphi| = \left|\cos\frac{\pi}{2}\,(2m-1-2t)\right| = \sin\pi t \leqslant \pi t.$$

Используя два последних неравенства, оценим $\theta(\varphi)$:

$$\theta(\varphi) \leqslant C\sum_{i=1}^{n} \frac{(2i-1)t}{|m+i-1-t||m-i-t|}, \quad C = \frac{3\pi^2}{4\sqrt{2}}.$$

Отсюда следует, что при $m = 1$

$$\theta(\varphi) \leqslant C\left[\frac{1}{1-t} + t\sum_{i=2}^{n}\left(\frac{1}{i-1+t} + \frac{1}{i-t}\right)\right] \leqslant$$

$$\leqslant C\left(2 + \sum_{i=1}^{n-1}\frac{1}{i}\right) \leqslant C\left(3 + \int_1^n \frac{dt}{t}\right) = C\,(3+\ln n).$$

При $2 \leqslant m \leqslant 1 + \frac{n}{2}$ получаем

$$\theta(\varphi) \leqslant C \left[\frac{2m-1}{2m-1-t} + \frac{1}{2} \sum_{i=1}^{m-1} \left(\frac{1}{m-i-t} - \frac{1}{m+i-1-t} \right) + \right.$$

$$\left. + \frac{1}{2} \sum_{i=m+1}^{n} \left(\frac{1}{m+i-1-t} + \frac{1}{i-m+t} \right) \right] \leqslant C \left(4 + \sum_{i=2}^{n} \frac{1}{i} \right) \leqslant C \left(4 + \ln n \right).$$

Окончательно имеем

$$\lambda_n = \max_{\varphi \in [0,1]} \theta(\varphi) \leqslant C \left(4 + \ln n \right) \leqslant K \ln n \,. \qquad \triangleright$$

3.81. Доказать, что если узлы интерполяции на отрезке совпадают с нулями многочлена Чебышёва соответствующей степени, то справедливо неравенство $\lambda_n = \max\limits_{x} \sum\limits_{i=1}^{n} |\Phi_i(x)| \geqslant K \ln n$ с постоянной K, не зависящей от n.

3.82. Определить константу Лебега λ_3 для узлов интерполяции — нулей многочлена Чебышёва $T_3(x)$.

Ответ: $\lambda_3 = \frac{5}{3}$.

В приложениях встречаются также многочлены Чебышёва второго рода $U_n(x)$. Они удовлетворяют рекуррентному соотношению и начальным условиям: $U_{n+1}(x) = 2x\,U_n(x) - U_{n-1}(x)$, $U_0(x) = 1$, $U_1(x) = 2\,x$.

3.83. Показать, что для $U_n(x)$, $x \in \mathbf{R}$, справедливо представление

$$U_n(x) = \begin{cases} \dfrac{\sin((n+1)\arccos x)}{\sin(\arccos x)} & \text{при } |x| \leqslant 1\,, \\ \dfrac{\left(x + \sqrt{x^2-1}\right)^{n+1} - \left(x - \sqrt{x^2-1}\right)^{n+1}}{2\sqrt{x^2-1}} & \text{при } |x| \geqslant 1\,. \end{cases}$$

3.84. Показать, что общее решение разностного уравнения $y_{n+1}(x) - 2x\,y_n(x) + y_{n-1}(x) = 0$ представимо в виде: $y_n = C_1(x)T_n(x) + C_2(x)U_{n-1}(x)$.

Указание. Вычислить определитель

$$\begin{vmatrix} T_0(x) & T_1(x) \\ U_{-1}(x) & U_0(x) \end{vmatrix} = \begin{vmatrix} 1 & x \\ 0 & 1 \end{vmatrix} = 1\,;$$

откуда следует, что $T_n(x)$ и $U_{n-1}(x)$ — линейно независимы.

3.85. Проверить соотношения для $T_n(x)$ и $U_n(x)$:

1) $T_{n-1}(x) - xT_n(x) = (1-x^2)U_{n-1}(x)$;
2) $U_{n-1}(x) - xU_n(x) = -T_{n+1}(x)$;
3) $U_{n+i}(x) + U_{n-i}(x) = 2T_i(x)U_n(x)$;
4) $U_{in-1}(x) = 2U_{i-1}\left(T_n(x)\right)$.

3.86. Показать, что $\max\limits_{|x|\leqslant 1} |U_n(x)| = U_n(1) = n+1$.

3.87. Вычислить $I_{mn} = \int\limits_{-1}^{1} \sqrt{1-x^2}\, U_n(x)\, U_m(x)\, dx$.

3.3. Численное дифференцирование

Пусть известны значения функции $f(x)$ в точках $x_1, x_2, \ldots, x_n$ и требуется приближенно определить производную $f^{(k)}(x_0)$ для некоторого $0 \leqslant k \leqslant n-1$. Построим интерполяционный многочлен $L_n(x)$ и положим $f^{(k)}(x) \approx L_n^{(k)}(x)$; при этом для погрешности справедливо представление

$$f^{(k)}(x) - L_n^{(k)}(x) = \sum_{j=0}^{k} \frac{k!}{(k-j)!(n+j)!}\, f^{(n+j)}(\xi_j)\omega_n^{(k-j)}(x).$$

Для системы равноотстоящих узлов ($x_{i+1} - x_i = h$) часто используют другой подход, основанный на получении приближений для старших производных через приближения для младших, аналогично последовательному дифференцированию. Базовыми являются следующие выражения:

$$\partial f(x) = \frac{f(x+h) - f(x)}{h}\,,\ \bar{\partial} f(x) = \frac{f(x) - f(x-h)}{h}\,,\ \tilde{\partial} f(x) = \frac{\partial f(x) + \bar{\partial} f(x)}{2}\,,$$

которые являются простейшими аналогами первой производной функции $f(x)$. Их называют *разностями вперед, назад* и *центральной* соответственно. Для вывода оценок погрешностей при данном подходе удобно использовать разложения Тейлора.

Для получения формул численного дифференцирования на практике также используют *метод неопределенных коэффициентов*. Он заключается в следующем: искомую формулу записывают в виде $f^{(k)}(x_0) = \sum\limits_{i=1}^{n} c_i f(x_i) + R(f)$, и коэффициенты c_i определяют из системы линейных уравнений при $R(f) = 0$, для получения которой последовательно полагают $f(x)$ равной $1, x, x^2, \ldots, x^{n-1}$.

Будем далее использовать обозначение $f(x) \in C^{(r)}$, если функция $f(x)$ имеет на интересующем нас отрезке все непрерывные производные до порядка r включительно.

3.88. Показать, что в точке $x = x_i$ (одном из узлов интерполяции) справедлива оценка погрешности

$$|f'(x_i) - L_n'(x_i)| \leqslant \frac{1}{n!} \max_{x\in[x_1,x_n]} \left|f^{(n)}(x)\right| \prod_{\substack{j=1\\ j\neq i}}^{n} |x_i - x_j|.$$

У к а з а н и е. Использовать явное представление погрешности для производной многочлена Лагранжа.

3.89. Доказать равенства:

1) если $f \in C^{(2)}$, то $\partial f(x) - f'(x) = \frac{h}{2}\ f''(\xi)$, $x < \xi < x + h$;

2) если $f \in C^{(3)}$, то $\tilde{\partial} f(x) - f'(x) = \frac{h^2}{6}\ f'''(\xi)$, $x - h < \xi < x + h$.

Указание. Использовать разложение в ряд Тейлора.

3.90. Получить явные формулы для разностных аналогов старших производных: $f''(x) \approx \bar{\partial}\partial f(x)$, $f'''(x) \approx \tilde{\partial}\,\bar{\partial}\partial f(x)$, $f^{(4)}(x) \approx \bar{\partial}^2\partial^2 f(x)$.

Ответ: $\bar{\partial}\partial f(x) = \dfrac{f(x+h) - 2\,f(x) + f(x-h)}{h^2}$,

$$\tilde{\partial}\,\bar{\partial}\partial f(x) = \frac{f(x+2h) - 2\,f(x+h) + 2\,f(x-h) - f(x-2h)}{2\,h^3},$$

$$\bar{\partial}^2\partial^2 f(x) = \frac{f(x+2h) - 4\,f(x+h) + 6\,f(x) - 4\,f(x-h) + f(x-2h)}{h^4}.$$

3.91. Найти величину $K_i = K_i(h)$ в следующих равенствах:

1) если $f \in C^{(4)}$, то $\bar{\partial}\partial f(x) - f''(x) = K_2 f^{(4)}(\xi)$, $x - h < \xi < x + h$;

2) если $f \in C^{(5)}$, то $\tilde{\partial}\,\bar{\partial}\partial f(x) - f'''(x) = K_3 f^{(5)}(\xi)$, $x - 2h < \xi < x + 2h$;

3) если $f \in C^{(6)}$, то $\bar{\partial}^2\partial^2 f(x) - f^{(4)}(x) = K_4 f^{(6)}(\xi)$, $x - 2h < \xi < x + 2h$.

Ответ: 1) $K_2 = \frac{h^2}{12}$; 2) $K_3 = \frac{h^2}{4}$; 3) $K_4 = \frac{h^2}{6}$.

3.92. Считая, что значения функции в формулах численного дифференцирования для аналогов второй и четвертой производных из 3.91 заданы с абсолютной погрешностью ε, получить оценки полной погрешности этих формул как сумму погрешности метода и вычислительной погрешности. Найти оптимальный шаг h_0, при котором минимизируется величина оценки полной погрешности.

Указание. Решение провести по аналогии со следующим примером для разности вперед (см. также 1.6). Полная погрешность для разности вперед $\partial f(x)$ имеет вид

$$R_1(h, \varepsilon) = \left| \frac{f^*(x+h) - f^*(x)}{h} - f'(x) \right|,$$

где $f^*(x+h)$ и $f^*(x)$ — приближенные значения функции $f(x)$ в соответствующих точках. Добавляя в числитель дроби $\pm f(x+h)$ и $\pm f(x)$, после перегруппировки слагаемых получим

$$\left| \frac{f^*(x+h) - f(x+h)}{h} - \frac{f^*(x) - f(x)}{h} + \left(\frac{f(x+h) - f(x)}{h} - f'(x) \right) \right|.$$

Оценка вычислительной погрешности для каждого из двух первых слагаемых имеет вид $\frac{\varepsilon}{h}$, а погрешность метода в предположении ограниченности второй производной $|f''(\xi)| \leqslant M_2$ равна $\frac{hM_2}{2}$. Окончательно имеем $R_1(h, \varepsilon) \leqslant \frac{2\,\varepsilon}{h} + \frac{h\,M_2}{2}$. Для определения значения h_0, при котором минимизируется полная погрешность, необходимо правую часть полученного

выражения продифференцировать по h и приравнять к нулю. Решая уравнение $-2\varepsilon h^{-2} + \frac{M_2}{2} = 0$, находим $h_0 = 2\sqrt{\frac{\varepsilon}{M_2}}$ и $R_1(h_0, \varepsilon) = 2\sqrt{\varepsilon\, M_2}$.

Ответ: 1) $h_0 = 2\left(\frac{3\varepsilon}{M_4}\right)^{1/4}$ для $R_2(h,\varepsilon) \leqslant \frac{4\varepsilon}{h^2} + \frac{h^2 M_4}{12}$; 2) $h_0 = 2\left(\frac{3\varepsilon}{M_6}\right)^{1/6}$ для $R_4(h,\varepsilon) \leqslant \frac{16\varepsilon}{h^4} + \frac{h^2 M_6}{6}$.

3.93. Методом неопределенных коэффициентов построить формулы численного дифференцирования наиболее высокого порядка точности по h:

1) $f'(0) \approx \dfrac{a\,f(-2h) + b\,f(0) + c\,f(h)}{h}$;

2) $f''(0) \approx \dfrac{a\,f(-h) + b\,f(h) + c\,f(2h) + d\,f(3h)}{h^2}$.

Ответ: 1) $a = -\frac{1}{6},\ b = -\frac{1}{2},\ c = \frac{2}{3}$, 2) $a = \frac{1}{2},\ b = -2, c = 2,\ d = -\frac{1}{2}$.

3.94. Доказать, что $\tilde{\partial} f(0) - f'(0) = \frac{1}{4h} \int\limits_{-h}^{h} (h - |x|)^2 f'''(x)dx$.

Указание. Разбить интеграл на два, раскрывая модуль, и интегрировать по частям.

3.95. Используя формулу Тейлора с остаточным членом в интегральной форме

$$f(b) = f(a) + (b-a)f'(a) + \cdots + \frac{(b-a)^n}{n!}\, f^{(n)}(a) + \frac{1}{n!}\int\limits_a^b (b-\xi)^n f^{(n+1)}(\xi)\, d\xi,$$

получить оценки погрешности формул численного дифференцирования (постоянные C_1, C_2 не зависят от f и h)

$$|\bar{\partial} f(x) - f'(x)| \leqslant C_1 \int\limits_{x-h}^{x} |f''(\xi)|\, d\xi,$$

$$|\bar{\partial}\partial f(x) - f''(x)| \leqslant C_2\, h \int\limits_{x-h}^{x+h} |f^{(4)}(\xi)|\, d\xi.$$

3.96. Доказать справедливость следующих равенств:

$$\partial(f\,g) = f\,\partial g + g\,\partial f + h\,\partial f\,\partial g, \quad \bar{\partial}\left(\frac{f}{g}\right) = \frac{g\,\bar{\partial} f - f\,\bar{\partial} g}{g\,(g - h\,\bar{\partial} g)}\,.$$

3.97. Пусть вычислены точное и приближенное значения $f''(x_0)$ при заданных узлах интерполяции $x_{-l}, \ldots, x_0, \ldots, x_l,\ x_i - x_{i-1} = h$. Показать, что справедливо представление

$$f''(x_0) - L_n''(x_0) = \frac{2(-1)^l (l!)^2}{(2l+2)!}\, f^{(2l+2)}(\xi) h^{2l}.$$

3.98. Используя формулу Тейлора с остаточным членом в интегральной форме, получить оценки погрешности следующих формул численного дифференцирования (постоянные C_i не зависят от f и h):

1) $|\partial f(x) - f'(x)| \leqslant C_1 \int\limits_{x}^{x+h} |f''(\xi)|\, d\xi$;

2) $|\tilde{\partial} f(x) - f'(x)| \leqslant C_2 h \int\limits_{x-h}^{x+h} |f'''(\xi)|\, d\xi$;

3) $|2\partial f(x) - \tilde{\partial} f(x+h) - f'(x)| \leqslant C_3 h \int\limits_{x}^{x+2h} |f'''(\xi)|\, d\xi$;

4) $|2\bar{\partial} f(x) - \tilde{\partial} f(x-h) - f'(x)| \leqslant C_4 h \int\limits_{x-2h}^{x} |f'''(\xi)|\, d\xi$;

5) $|\bar{\partial}^2 \partial^2 f(x) - f^{(4)}(x)| \leqslant C_5 h \int\limits_{x-2h}^{x+2h} |f^{(6)}(\xi)|\, d\xi$.

3.99. Доказать справедливость следующих равенств:

1) $\bar{\partial}(f\,g) = f\,\bar{\partial} g + g\,\bar{\partial} f - h\,\bar{\partial} f\,\bar{\partial} g$;

2) $\tilde{\partial}(f\,g) = f\,\tilde{\partial} g + g\,\tilde{\partial} f + \dfrac{h^2}{2}\,(\bar{\partial}\partial f\,\tilde{\partial} g + \bar{\partial}\partial g\,\tilde{\partial} f)$;

3) $\partial\left(\dfrac{f}{g}\right) = \dfrac{g\,\partial f - f\,\partial g}{g\,(g + h\,\partial g)}$.

3.100. Получить формулу численного дифференцирования наиболее высокого порядка точности по h следующего вида:

1) $f'(0) \approx h^{-1}[a\,f(0) + b\,f(h) + c\,f(2h)]$;

2) $f'(0) \approx h^{-1}[a\,f(0) + b\,f(-h) + c\,f(2h)]$;

3) $f'(0) \approx h^{-1}[a\,f(0) + b\,f(-h) + c\,f(-2h)]$;

4) $f'(0) \approx h^{-1}[a\,f(0) + b\,f(2h) + c\,f(3h)]$

и найти h, при котором достигается минимум оценки полной погрешности, если $\max\limits_x |f^{(k)}(x)| \leqslant A_k$, и абсолютная вычислительная погрешность функции не превосходит ε, т. е. $\max\limits_x |f(x) - f^*(x)| \leqslant \varepsilon$.

Ответ: 1) $a = -\frac{3}{2}$, $b = 2$, $c = -\frac{1}{2}$; $h_0 = \left(\frac{6\varepsilon}{A_3}\right)^{1/3}$; 2) $a = \frac{1}{2}$, $b = -\frac{2}{3}$, $c = \frac{1}{6}$; $h_0 = \left(\frac{2\varepsilon}{A_3}\right)^{1/3}$; 3) $a = \frac{3}{2}$, $b = -2$, $c = \frac{1}{2}$; $h_0 = \left(\frac{6\varepsilon}{A_3}\right)^{1/3}$; 4) $a = -\frac{5}{6}$, $b = \frac{3}{2}$, $c = -\frac{2}{3}$; $h_0 = \left(\frac{3\varepsilon}{2A_3}\right)^{1/3}$.

3.101. Пусть $f \in C^{3,\lambda}$, $0 \leqslant \lambda \leqslant 1$, т. е. $f \in C^{(3)}$, $|f'''(x) - f'''(y)| \leqslant k|x-y|^{\lambda}\ \forall x, y$. Доказать, что $\bar{\partial}\partial f(x) - f''(x) = O(h^{1+\lambda})$.

3.102. Пусть числа α_j, не зависящие от h, порождают формулу численного дифференцирования максимального порядка точности среди формул вида $f^{(k)}(x) \approx h^{-k} \sum\limits_{j=-n}^{n} \alpha_j f(x+jh)$. Доказать, что:

1) $\alpha_j = \alpha_{-j}$, если k четное, $\alpha_j = -\alpha_{-j}$, если k нечетное;
2) формула с дополнительным слагаемым

$$f^{(k)}(x) \approx h^{-k} \sum_{j=-n}^{n+1} \beta_j f(x+jh)$$

не может иметь больший порядок точности; причем она имеет тот же порядок точности тогда и только тогда, когда $\beta_{n+1} = 0$, $\beta_j = \alpha_j$, $j = -n, -n+1, \ldots, n-1, n$.

3.103. Доказать, что если все точки x_i различны и удалены от точки x_0 на расстояние $O(h)$, где h — малая величина, то при гладкой $f(x)$ приближенная формула численного дифференцирования $f^{(k)}(x_0) \approx \sum\limits_{i=1}^{n} c_i f(x_i)$ имеет порядок погрешности $O(h^m)$. Здесь $m \geqslant j+1-k$, j — максимальная степень многочленов, для которых эта формула точна.

3.104. Найти аппроксимацию $f''(x)$ по равноотстоящим ($x_{i+1} - x_i = h$) узлам x_i, $x_{i\pm1}$, $x_{i\pm2}$ с максимально возможным порядком точности по h.

3.105. Найти коэффициенты формул численного дифференцирования максимальной степени точности:

1) $f'(x) \approx \dfrac{af(x) + bf(x+h) + cf(x-h)}{h}$;
2) $f'(x) \approx \dfrac{af(x) + bf(x+h) + cf(x-2h)}{h}$;
3) $f''(x) \approx \dfrac{af(x) + bf(x+h) + cf(x+2h)}{h^2}$;
4) $f''(x) \approx \dfrac{af(x) + bf(x+h) + cf(x-h)}{h^2}$;
5) $f''(x) \approx \dfrac{af(x) + bf(x-h) + cf(x-2h)}{h^2}$.

3.4. Многочлен наилучшего равномерного приближения

Пусть R — пространство ограниченных вещественных функций, определенных на отрезке $[a,b]$ вещественной оси, с нормой $\|f(x)\| = \sup\limits_{x\in[a,b]} |f(x)|$. Для элемента $f \in R$ отыскивается наилучшее приближение вида $Q_n(x) = \sum\limits_{j=0}^{n} a_j x^j$. Многочлен $Q_n^0(x)$ называется *многочленом наилучшего равномерного приближения* для функции $f(x)$, если для любого многочлена $Q_n(x)$ степени n справедливо неравенство $\|f - Q_n^0\| \leqslant \|f - Q_n\|$.

Такой многочлен существует всегда, а для непрерывной функции он определяется единственным образом.

Теорема Чебышёва. *Чтобы многочлен $Q_n(x)$ был многочленом наилучшего равномерного приближения непрерывной функции $f(x)$, необходимо и достаточно существования на $[a,b]$ по крайней мере $n+2$ точек $x_0 < \ldots < x_{n+1}$ таких, что*

$$f(x_i) - Q_n(x_i) = \alpha(-1)^i \|f - Q_n\|\,,$$

где $i = 0, \ldots, n+1$ и $\alpha = 1$ (или $\alpha = -1$) одновременно для всех i.

Точки $x_0, \ldots, x_{n+1}$, удовлетворяющие условию теоремы, называются *точками чебышёвского альтернанса.*

3.106. Построить многочлен наилучшего равномерного приближения степени $n = 50$ для $f(x) = \sin 100x$ на отрезке $[0, \pi]$.
Ответ: $Q_{50}(x) = 0$.

3.107. Пусть $f(x)$ — выпуклая непрерывная функция на $[a,b]$ и $Q_1^0(x)$ — ее многочлен наилучшего равномерного приближения первой степени. Доказать, что концы отрезка a и b входят в альтернанс.

$\lhd$ Выпуклая функция удовлетворяет неравенству

$$f\left(\frac{x_1+x_2}{2}\right) \leqslant \frac{f(x_1)+f(x_2)}{2}$$

для произвольных x_1, x_2 из отрезка $[a,b]$. Рассмотрим непрерывную выпуклую функцию $g(x) = f(x) - Q_1^0(x)$ (добавление к $f(x)$ линейной функции $Q_1^0(x)$ эти свойства сохраняет) и обозначим через $\{\xi_i\}$ множество точек альтернанса. Доказательство проведем от противного.

Пусть, например, $a \notin \{\xi_i\}$. Тогда для $\theta = \inf\limits_i \{\xi_i\}$ имеем $g(\theta) = M$ и $\theta > a$. Следовательно, в силу выпуклости $g(x)$ справедлива следующая цепочка неравенств для достаточно малого ε:

$$M = g(\theta) \leqslant \frac{g(\theta+\varepsilon)+g(\theta-\varepsilon)}{2} \leqslant \frac{M+g(\theta-\varepsilon)}{2} < \frac{M+M}{2} = M\,.$$

Полученное противоречие означает, что $a \in \{\xi_i\}$. Аналогично доказывается принадлежность множеству точек альтернанса другого конца отрезка. $\rhd$

3.108. Построить многочлен наилучшего равномерного приближения степени $n = 1$ для $f(x) = x^3$ на отрезке $[1,2]$.

$\lhd$ Введем обозначения: $L = \|f(x) - Q_1(x)\|$, $Q_1(x) = a_0 + a_1 x$, $[1,2] \equiv [a,b]$ и, воспользовавшись выпуклостью $f(x)$, запишем соотношения из теоремы Чебышёва:

$$\begin{aligned} f(a) - (a_0 + a_1\, a) &= \alpha\, L\,,\\ f(d) - (a_0 + a_1\, d) &= -\alpha\, L\,,\\ f(b) - (a_0 + a_1\, b) &= \alpha\, L\,. \end{aligned}$$

Кроме того, поскольку d — внутренняя точка альтернанса и $f(x)$ — дифференцируема, отсюда получаем недостающее уравнение

$$\left(f(x) - (a_0 + a_1\,x)\right)'\big|_{x=d} = 0. \qquad \triangleright$$

Ответ: $Q_1(x) = 7x - 3 - \frac{7}{3}\,\sqrt{\frac{7}{3}}$.

Задача также имеет наглядное геометрическое решение: строим по точкам $(a, f(a))$ и $(b, f(b))$ прямую $y_1(x) = a_1x + a_2$; находим такое $d \in [a, b]$, что $f'(d) = a_1$; проводим прямую $y_2(x)$, параллельную $y_1(x)$ и проходящую через точку $(d, f(d))$; функция $Q_1(x) = a_1x + a_0$, параллельная $y_1(x)$ и $y_2(x)$ и равноотстоящая от них, будет искомым многочленом наилучшего равномерного приближения согласно теореме Чебышёва.

3.109. Построить многочлен наилучшего равномерного приближения степени $n = 1$ для $f(x) = |x|$ на отрезке $[-1, 5]$.

Ответ: $Q_1(x) = \frac{2}{3}\,x + \frac{5}{6}$.

3.110. Построить многочлен наилучшего равномерного приближения $Q_n^0(x)$ степени n для $P_{n+1}(x) = a_{n+1}x^{n+1} + \cdots$ на отрезке $[a, b]$.

Указание. По определению многочлена наилучшего равномерного приближения, разность $P_{n+1}(x)$ и $Q_n^0(x)$ представляет собой наименее уклоняющийся на отрезке $[a, b]$ многочлен степени $(n + 1)$ со старшим коэффициентом a_{n+1}. Следовательно, $P_{n+1}(x) - Q_n^0(x) = a_{n+1}\bar{T}_{n+1}^{[a,b]}(x)$, где $\bar{T}_{n+1}^{[a,b]}(x)$ — приведенный многочлен Чебышёва. Отсюда имеем $Q_n^0(x) = P_{n+1}(x) - a_{n+1}\bar{T}_{n+1}^{[a,b]}(x)$. На отрезке $[a, b]$ точки альтернанса определяются экстремумами многочлена $\bar{T}_{n+1}^{[a,b]}(x)$.

3.111. Пусть $f^{(n+1)}(x)$ непрерывна, не меняет знак на $[a, b]$ и $Q_n(x)$ — многочлен наилучшего равномерного приближения степени n для $f(x)$. Оценить величины C_1 и C_2 в неравенстве $C_1 \leqslant \|f(x) - Q_n(x)\| \leqslant C_2$.

$\triangleleft$ По определению многочлена наилучшего равномерного приближения, $L = \|f(x) - Q_n(x)\|$ не превосходит нормы погрешности приближения $f(x)$ интерполяционным многочленом $L_{n+1}(x)$ по узлам, являющимся нулями многочлена Чебышёва, т. е.

$$L \leqslant \max_{[a,b]} \left|f^{(n+1)}(x)\right| \ \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}\,.$$

С другой стороны, по теореме Чебышёва разность $f(x) - Q_n(x)$ обращается в нуль в $(n + 1)$-й точке, которые можно рассматривать как узлы интерполяции $y_1, \ldots, y_{n+1}$. Поэтому верно представление погрешности следующего вида:

$$f(x) - Q_n(x) = f^{(n+1)}(\xi)\ \frac{\omega_{n+1}(x)}{(n+1)!}\ ,$$

где $\omega_{n+1}(x) = (x - y_1)\cdots(x - y_{n+1})$ и $\xi = \xi(x) \in [a, b]$. Пусть точка x_0 такова, что $|\omega_{n+1}(x_0)| = \|\omega_{n+1}(x)\|$. Тогда

$$L \geqslant |f(x_0) - Q_n(x_0)| = \left|f^{(n+1)}(\xi(x_0))\right| \ \frac{|\omega_{n+1}(x_0)|}{(n+1)!} \ .$$

Поскольку $\|\omega_{n+1}(x)\| \geqslant (b-a)^{n+1}/2^{2n+1}$, окончательно имеем

$$L \geqslant \min_{[a,b]} \left|f^{(n+1)}(x)\right| \ \frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!} \ .$$

Таким образом, если $f^{(n+1)}(x)$ сохраняет знак и меняется не очень сильно, то разница между погрешностями приближения функции $f(x)$ многочленом наилучшего равномерного приближения и интерполяционным многочленом по нулям многочленов Чебышёва несущественна. $\triangleright$

3.112. Пусть $f(x)$ — непрерывная нечетная функция на отрезке $[-1, 1]$. Показать, что многочлен наилучшего равномерного приближения произвольной степени n — также нечетная функция.

$\triangleleft$ Пусть $Q_n(x)$ — многочлен наилучшего равномерного приближения $f(x)$ на $[-1, 1]$. Тогда $|f(x) - Q_n(x)| \leqslant L = \|f(x) - Q_n(x)\|$. Заменяя x на $-x$ и умножая выражение под знаком модуля на -1, получим $|-f(-x) - (-Q_n(-x))| \leqslant L$ или $|f(x) - (-Q_n(-x))| \leqslant L$. Следовательно, $-Q_n(-x)$ также многочлен наилучшего равномерного приближения $f(x)$ на $[-1, 1]$. По теореме единственности имеем $Q_n(x) = -Q_n(-x)$, что и требовалось показать.

Аналогично рассматривается случай четной $f(x)$. $\triangleright$

3.113. Получить оценку вида $C_n \leqslant \|\sin x - Q_n(x)\| \leqslant 2\,C_n$ для многочлена наилучшего равномерного приближения степени n на $\left[-\frac{\pi}{3}, \frac{\pi}{3}\right]$.

Ответ: $\|\sin x - Q_{2n-1}(x)\| = \|\sin x - Q_{2n}(x)\|$;
$C_{2n-1} = C_{2n} = \left(\frac{\pi}{6}\right)^{2n+1} \frac{1}{(2n+1)!}$.

3.114. Построить функцию $f(x)$ и ее многочлен наилучшего равномерного приближения $Q_n(x)$, не удовлетворяющие теоремам Чебышёва и единственности.

Ответ: $f(x) = \operatorname{sign} x$ на $[-1, 1]$, $Q_1(x) = \alpha\, x$, $\alpha \in [0, 2]$.

3.115. Построить многочлен наилучшего равномерного приближения степени n для функции $f(x)$ на отрезке $[a, b]$:

1) $n = 1$, $f(x) = x^3$, $[-1, 1]$;
2) $n = 3$, $f(x) = \exp(x^2)$, $[-1, 1]$;
3) $n = 3$, $f(x) = 3\sin^2 10x + |x^2 - 7x + 10|$, $[3, 4]$.

Ответ: 1) $Q_1(x) = \frac{3}{4}\, x$; 2) $Q_3(x) = (\mathrm{e} - 1)x^2 + \frac{\mathrm{e}}{2} - \frac{1}{2}(\mathrm{e} - 1)\ln(\mathrm{e} - 1)$;
3) $Q_3(x) = -x^2 + 7x - \frac{17}{2}$.

3.116. Построить многочлен наилучшего равномерного приближения первой степени для функции $f(x) = \sqrt{x^2+1}$ на отрезке $[0,1]$.

3.117. Построить многочлен наилучшего равномерного приближения четвертой степени для функции $f(x) = \sin(6\pi x)$ на отрезке $[0,\pi]$.

3.118. Построить многочлен наилучшего равномерного приближения степени n для функции $f(x)$ на отрезке $[a,b]$:

1) $n=2,\ f(x)=x^3,\ a=0,\ b=1$;
2) $n=2,\ f(x)=x^4,\ a=-1,\ b=1$;
3) $n=1,\ f(x)=\sin x,\ a=-\pi,\ b=\pi$;
4) $n=3,\ f(x)=|x^2-7x+10|,\ a=3,\ b=4$;
5) $n=30,\ f(x)=2x^2+3x+\cos 50x,\ a=0,\ b=\pi$;
6) $n=1,\ f(x)=1+x^p, p>0,\ a=0,\ b=1$;
7) $n=2,\ f(x)=2x^2+3x+5,\ a=1,\ b=7$.

3.119. Получить оценку вида

$$\frac{C}{2} \leqslant \|\cos x - Q_4^0(x)\|_{C[\frac{\pi}{6},\frac{\pi}{2}]} \leqslant C$$

с явным выражением для C, где $Q_4^0(x)$ — многочлен наилучшего равномерного приближения четвертой степени.

3.120. Доказать, что $\|\exp(x) - Q_4^0(x)\|_{C[0,1]} \geqslant \frac{1}{64\,000}$, где $Q_4^0(x)$ — многочлен наилучшего равномерного приближения четвертой степени.

3.121. Рассматривается задача наилучшего равномерного приближения функции $\exp(x)$ на $[-1,1]$. Показать, что $10^{-6} \leqslant \|\exp(x) - Q_6^0(x)\|_{C[-1,1]} \leqslant 10^{-5}$, где $Q_6^0(x)$ — многочлен наилучшего равномерного приближения шестой степени.

3.122. Показать, что чебышёвский альтернанс для функции $\exp(x)$ всегда содержит крайние точки отрезка, на котором решается задача наилучшего равномерного приближения.

3.123. Привести пример функции и соответствующего ей многочлена наилучшего равномерного приближения, для которых среди точек чебышёвского альтернанса нет граничных точек отрезка, на котором решается задача приближения.

3.124. Привести пример функции и соответствующего ей многочлена наилучшего равномерного приближения 6-й степени, для которых имеется 99 точек чебышёвского альтернанса.

У к а з а н и е. На отрезке $[a,b]$ сначала зафиксировать многочлен 6-й степени $Q_6(x)$ и построить вокруг него «коридор» из двух многочленов: $Q_+(x) = Q_6(x)+\varepsilon$ и $Q_-(x) = Q_6(x)-\varepsilon$. Затем внутри «коридора» провести колеблющуюся относительно многочлена $Q_6(x)$ непрерывную функцию $f(x)$, имеющую наперед заданное (например, 99) количество точек касания (альтернанса) с обеими границами $Q_+(x)$ и $Q_-(x)$.

3.125. Пусть $\sum\limits_{k=0}^{\infty} a_k T_k(x)$ — некоторый ряд по системе многочленов Чебышёва $T_k(x)$. Доказать, что каждая частичная сумма ряда $S_n(x) = \sum\limits_{k=0}^{n} a_k T_k(x)$ — многочлен наилучшего равномерного приближения степени n на $[-1, 1]$ для $S_{n+1}(x)$.

3.126. Функция $f(x) = \frac{1}{x+9}$ приближается на $[-1, 1]$ многочленом первой степени следующими способами:

1) наилучшее равномерное приближение;

2) отрезок ряда Тейлора в точке $x = 0$;

3) интерполяция с оптимальными узлами $x_{1,2} = \pm\frac{1}{\sqrt{2}}$.

Построить эти многочлены и вычислить нормы погрешностей в $C[-1, 1]$.

3.127. Функция $f(x) = \exp(-x)$ приближается на $[-1, 1]$ многочленом первой степени следующими способами:

1) наилучшее равномерное приближение;

2) наилучшее приближение в $L_2(-1, 1)$;

3) отрезок ряда Тейлора в точке $x = 0$, т. е. интерполяция с узлами $x_1 = x_2 = 0$;

4) интерполяция с узлами $x_1 = -1, x_2 = 1$;

5) интерполяция с оптимальными узлами $x_{1,2} = \pm\frac{1}{\sqrt{2}}$.

Построить эти многочлены и вычислить нормы погрешностей в $C[-1, 1]$.

3.128. Найти многочлен наилучшего равномерного приближения степени $n = 1$ для функции $f(x) = 1 + \sqrt{x}$ на отрезке $[0, 1]$.

Ответ: $Q_1(x) = x + \frac{9}{8}$.

3.129. Найти многочлен наилучшего равномерного приближения степени $n = 3$ для функции $f(x) = \sin x^2$ на отрезке $[-\sqrt{\pi}, \sqrt{\pi}]$.

Ответ: $Q_3(x) = \frac{1}{2}$.

3.130. Найти многочлен наилучшего равномерного приближения степени $n = 1$ для функции $f(x) = |x|$ на отрезке $[-1, 2]$.

Ответ: $Q_1(x) = \frac{1}{3}(x + 2)$.

3.131. Найти многочлен наилучшего равномерного приближения степени $n = 2$ для функции $f(x) = x^3$ на отрезке $[-1, 1]$.

Ответ: $Q_1(x) = \frac{3}{4}\, x$.

3.132. Найти для функции $\exp(x)$ наилучшее приближение многочленом нулевой степени в норме $L_1(0,1)$, где $\|f\|_{L_1(0,1)} = \int\limits_0^1 |f(x)|dx$.

3.133. Пусть P_2 — пространство алгебраических многочленов второй степени с нормой $||p|| = |p(-1)|+|p(0)|+|p(1)|$. Найти наилучшее приближение функции $p(x) = x^2 \in P_2$ константой.

3.134. Пусть $n \geqslant 1$ и заданы (x_k, y_k), $k = 0, 1, ..., n$. Найти линейную функцию $p(x) = ax + b$, минимизирующую функционал

$$\sum_{k=0}^{n} (y_k - ax_k - b)^2 .$$

3.135. Пусть A и $\mathbf{x}$ — вещественные симметричная матрица размерности $n \times n$ и n-мерный вектор, $f(t) = ||A\mathbf{x} - t\mathbf{x}||_2 = \sqrt{(A\mathbf{x} - t\mathbf{x}, A\mathbf{x} - t\mathbf{x})}$. Доказать, что $f(t)$ достигает минимума при $t = \dfrac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}$.

3.136. Найти для функции $f(x)$ наилучшее приближение в норме $L_2(a, b)$, $\|f\|^2_{L_2(0,1)} = \int\limits_0^1 |f(x)|^2 dx$, алгебраическими многочленами $P_n(x)$ степени n:

1) $a = -1,\ b = 1,\ f(x) = |x|,\ n = 1$;
2) $a = -1,\ b = 1,\ f(x) = x^2,\ n = 1$;
3) $a = -1,\ b = 1,\ f(x) = x^3,\ n = 1$;
4) $a = -1,\ b = 1,\ f(x) = x^3,\ n = 2$;
5) $a = 0,\ b = \pi,\ f(x) = \sin x,\ n = 2$;
6) $a = 0,\ b = 2,\ f(x) = x^3,\ n = 3$.

3.137. Для заданной функции $f(x)$ найти алгебраический многочлен $P_n(x)$ степени n, минимизирующий весовой функционал

$$\int\limits_{-1}^{1} \frac{(f(x) - P_n(x))^2}{\sqrt{1 - x^2}}\, dx,$$

где 1) $f(x) = \frac{1}{4}(x^2 + 2x + 1)$, $n = 1$; 2) $f(x) = x^2$, $n = 1$; 3) $f(x) = x^3$, $n = 2$.

3.138. Показать, что построение коэффициентов многочлена наилучшего приближения для функции $f(x)$ в пространстве $L_2(0,1)$ приводит к системе уравнений с *матрицей Гильберта*: $h_{ij} = \dfrac{1}{i + j - 1}$, $1 \leqslant i, j \leqslant n$.

$\lhd$ Наилучшее приближение ищется в виде $\sum\limits_{j=1}^{n} a_j x^{j-1}$ с неизвестными коэффициентами a_j, которые определяются из условия минимума функционала $\int\limits_0^1 \left(f(x) - \sum\limits_{j=1}^{n} a_j x^{j-1} \right)^2 dx$. Дифференцируя функционал по a_i

и приравнивания производные к нулю, получим уравнения

$$\int_0^1 \left(f(x) - \sum_{j=1}^{n} a_j x^{j-1} \right) x^{i-1} dx = 0\,, \quad i = 1, 2, \ldots, n$$

или

$$\sum_{j=1}^{n} \frac{a_j}{i+j-1} = \int_0^1 f(x)\, x^{i-1} dx\,, \quad i = 1, 2, \ldots, n\,. \qquad \triangleright$$

3.5. Приближение сплайнами

Пусть на отрезке $[a, b]$ вещественной оси задана сетка: $a = x_0 < x_1 < \ldots$ $\ldots < x_n = b$, $P_m(x)$ — множество многочленов степени не выше m $(m \geqslant 1)$, $C^{(r)}[a, b]$ — множество функций, имеющих на $[a, b]$ непрерывные производные до r-го порядка включительно $(r \geqslant 0)$.

Функцию $S_m(x) = S_{m,k}(x)$ называют *полиномиальным сплайном степени m дефекта k $(1 \leqslant k \leqslant m)$ с узлами $\{x_i\}$, $i = 0, 1, \ldots, n$, для функции $f(x) \in C[a, b]$,* если выполнены следующие условия:

1) *на каждом из отрезков $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, n-1$, она является многочленом, т. е. $S_m(x) \in P_m(x)$;*

2) *на всем отрезке $[a, b]$ обладает непрерывностью производных, т. е. $S_m(x) \in C^{(m-k)}[a, b]$.*

Ниже термин «дефекта k» будем опускать, так как далее рассматривается только случай $k = 1$.

Сплайн называется *интерполяционным*, если в узлах $\{x_i\}$ справедливы равенства $S_m(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$.

3.139. Построить линейный интерполяционный сплайн по значениям $f(0), f(1)$.

Ответ: $S_1(x) = f(0)(1 - x) + f(1)x$.

3.140. Получить оценку погрешности приближения функции $f(x)$ линейным интерполяционным сплайном на равномерной сетке с шагом h, если $f(x) \in C^{(2)}[0, 1]$.

$\triangleleft$ Пусть $x_i = i\,h$, $h = \dfrac{1}{n}$, $i = 0, 1, \ldots, n$; тогда линейный интерполяционный сплайн на отрезке $[x_{i-1}, x_i]$ имеет вид

$$S_1(x) = f(x_{i-1})\, \frac{x_i - x}{h} + f(x_i)\, \frac{x - x_{i-1}}{h}\,.$$

Если $f(x) \in C^{(2)}[0, 1]$, то из оценки погрешности для интерполяционного многочлена Лагранжа следует, что

$$\max_{[x_{i-1}, x_i]} |f(x) - S_1(x)| \leqslant \max_{[x_{i-1}, x_i]} |f''(x)|\, \frac{h^2}{8}\,.$$

Это неравенство справедливо на любом отрезке $[x_{i-1}, x_i]$, значит, на $[0, 1]$ в целом. $\triangleright$

3.141. Обозначим через M_i значения второй производной $S_3''(x)$ кубического интерполяционного сплайна в узлах $\{x_i\}$, $i = 0, 1, \ldots, n$. Показать, что они удовлетворяют системе линейных уравнений $C\,M = d$, где

$$c_{ij} = \begin{cases} \dfrac{h_i}{6} & \text{при} \quad j = i-1, \\ \dfrac{h_i + h_{i+1}}{3} & \text{при} \quad j = i, \\ \dfrac{h_{i+1}}{6} & \text{при} \quad j = i+1, \\ 0 & \text{при} \quad |j-i| > 1; \end{cases} \qquad d_i = \frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i}\,,$$

$$i = 1, 2, \ldots, n-1, \quad h_i = x_i - x_{i-1}.$$

$\lhd$ По определению, $S_3''(x)$ — линейная на каждом отрезке $[x_{i-1}, x_i]$ функция. В силу ее непрерывности в концах отрезков имеем представление:

$$S_3''(x) = M_{i-1}\,\frac{x_i - x}{h_i} + M_i\,\frac{x - x_{i-1}}{h_i}\,.$$

Двухкратно интегрируя и учитывая условия $S_3(x_i) = f_i\,, S_3(x_{i-1}) = f_{i-1}$, получим аналитическое представление кубического интерполяционного сплайна на отрезке $[x_{i-1}, x_i]$:

$$S_3(x) = M_{i-1}\,\frac{(x_i - x)^3}{6\,h_i} + M_i\,\frac{(x - x_{i-1})^3}{6\,h_i} +$$

$$+ \left(f_{i-1} - \frac{M_{i-1}h_i^2}{6}\right)\frac{x_i - x}{h_i} + \left(f_i - \frac{M_i h_i^2}{6}\right)\frac{x - x_{i-1}}{h_i}\,.$$

Вычислим производную сплайна $S_3'(x)$ слева в точке x_i, воспользовавшись представлением на $[x_{i-1}, x_i]$:

$$S_3'(x_i - 0) = M_{i-1}\,\frac{h_i}{6} + M_i\,\frac{h_i}{3} + \frac{f_i - f_{i-1}}{h_i}\,,$$

и аналогично найдем производную сплайна $S_3'(x)$ справа в точке x_i, воспользовавшись представлением на $[x_i, x_{i+1}]$:

$$S_3'(x_i + 0) = -M_i\,\frac{h_{i+1}}{3} - M_{i+1}\,\frac{h_{i+1}}{6} + \frac{f_{i+1} - f_i}{h_{i+1}}\,.$$

Непрерывность $S_3'(x)$ в точках $x_i\,, i = 1, \ldots, n-1$, т. е. $S_3'(x_i - 0) = S_3'(x_i + 0)$, порождает искомую систему из $(n-1)$ уравнения относительно $(n+1)$-го неизвестного. $\rhd$

3.142. Построить кубический интерполяционный сплайн по значениям $f(0), f(1), f(2)$.

$\lhd$ Из решения 3.141 следует, что здесь неизвестными являются величины M_0, M_1, M_2, удовлетворяющие уравнению

$$\tfrac{1}{6}\,M_0 + \tfrac{2}{3}\,M_1 + \tfrac{1}{6}\,M_2 = f(2) - 2\,f(1) + f(0).$$

При этом искомый сплайн имеет следующий вид:

на отрезке $[0,1]$

$$S_3(x) = M_0\,\frac{(1-x)^3}{6} + M_1\,\frac{x^3}{6} + \left(f(0) - \frac{M_0}{6}\right)(1-x) + \left(f(1) - \frac{M_1}{6}\right)x\,;$$

на отрезке $[1,2]$

$$S_3(x) = M_1\,\frac{(2-x)^3}{6} + M_2\,\frac{(x-1)^3}{6} +$$
$$+ \left(f(1) - \frac{M_1}{6}\right)(2-x) + \left(f(2) - \frac{M_2}{6}\right)(x-1).$$

У построенного сплайна две степени свободы, которые фиксируются заданием M_0 и M_2 или уравнениями для них. *Естественному сплайну* соответствуют значения $M_0 = M_2 = 0$. $\triangleright$

3.143. Пусть в 3.141 $M_0 = M_n = 0$.

Показать, что в этом случае решение системы $C\,M = d$ удовлетворяет неравенству

$$\max_{1\leqslant i\leqslant n-1} |M_i| \leqslant 3\,\frac{\max\limits_{1\leqslant i\leqslant n-1} |d_i|}{\min\limits_{1\leqslant i\leqslant n-1} h_i}\,.$$

$\triangleleft$ Пусть $\max\limits_i |M_i| = |M_j|\,, 1 \leqslant j \leqslant n-1$. Рассмотрим j-е уравнение системы

$$d_j = M_{j-1}\,\frac{h_j}{6} + M_j\,\frac{h_j + h_{j+1}}{3} + M_{j+1}\,\frac{h_{j+1}}{6}\,,$$

из которого следует неравенство:

$$|d_j| \geqslant |M_j|\,\frac{h_j + h_{j+1}}{3} - \left(|M_{j-1}|\,\frac{h_j}{6} + |M_{j+1}|\,\frac{h_{j+1}}{6}\right) \geqslant |M_j|\,\frac{h_j + h_{j+1}}{6}\,,$$

так как $|M_{j\pm1}| \leqslant |M_j|$. Оценивая левую часть неравенства сверху через $\max\limits_i |d_i|$ и множитель в его правой части снизу, как

$$\min_i \frac{h_i + h_{i+1}}{6} \geqslant \frac{1}{3}\,\min_i h_i,$$

приходим к искомому неравенству. $\triangleright$

3.144. Пусть $f(x) \in C^{(4)}[a,b]$, $\max\limits_{[a,b]} |f^{(4)}(x)| \leqslant A_4$, задана сетка с постоянным шагом $h_i = h$, и дополнительные условия для определения кубического интерполяционного сплайна имеют следующий вид:

$$S_3'(x_0 + 0) = f'(x_0), \quad S_3'(x_n - 0) = f'(x_n).$$

Показать, что справедлива оценка погрешности

$$|S_3^{(l)}(x) - f^{(l)}(x)| \leqslant C_l A_4 h^{4-l}, \quad l = 0,1,2,3.$$

$\lhd$ *Поточечное неравенство для второй производной.* Воспользуемся решением 3.141. Разделив обе части i-го уравнения системы $C\,M = d$ на $\frac{h}{6}$, приведем его к виду

$$\frac{1}{2}\,M_{i-1} + 2\,M_i + \frac{1}{2}\,M_{i+1} = 3\,\frac{f_{i+1} - 2\,f_i + f_{i-1}}{h^2}\,, \quad i = 1, 2, \ldots, n-1.$$

Вычисляя производную сплайна $S_3'(x)$ справа в точке x_0, воспользовавшись представлением на $[x_0, x_1]$

$$S_3'(x_0 + 0) = -\frac{h}{6}\,(2\,M_0 + M_1) + \frac{f_1 - f_0}{h} = f'(x_0),$$

получим первое (для $i = 0$) уравнение системы

$$2\,M_0 + M_1 = \frac{6}{h}\,\left(\frac{f_1 - f_0}{h} - f'(x_0)\right).$$

Последнее уравнение (для $i = n$) строим аналогично

$$M_{n-1} + 2\,M_n = \frac{6}{h}\,\left(f'(x_n) - \frac{f_n - f_{n-1}}{h}\right).$$

Положим $\varphi_i = f''(x_i)$ и вычтем из обеих частей уравнения $C\,M = d$ выражение $C\,\varphi$. Имеем $C(M - \varphi) = d - C\,\varphi$. Для полученной системы, используя решение 3.143, можно получить оценку $\max\limits_i |M_i - \varphi_i| \leqslant \max\limits_i |d_i - (C\varphi)_i|$. Представляет интерес величина $d_i - (C\varphi)_i$ в правой части неравенства. Рассмотрим ее для $i = 0$. Получаем

$$\frac{6}{h}\,\left(\frac{f_1 - f_0}{h} - f'(x_0)\right) - (2\,f''(x_0) + f''(x_0 + h)) =$$

$$= \frac{6}{h}\,\left(\frac{f_0 + hf'(x_0) + \frac{h^2}{2}\,f''(x_0) + \frac{h^3}{6}\,f^{(3)}(x_0) + \frac{h^4}{24}\,f^{(4)}(\xi_1) - f_0}{h} - \right.$$

$$\left. - f'(x_0)\right) - \left(2\,f''(x_0) + f''(x_0) + hf^{(3)}(x_0) + \frac{h^2}{2}\,f^{(4)}(\eta_1)\right) =$$

$$= \frac{h^2}{4}\,\left(f^{(4)}(\xi_1) - 2f^{(4)}(\eta_1)\right).$$

Мы пришли к неравенству $|d_0 - (C\varphi)_0| \leqslant c_0 h^2 A_4$ с постоянной $c_0 = \frac{3}{4}$. Аналогичная оценка справедлива для $i = n$, в которой также $c_n = \frac{3}{4}$.

Далее потребуются два следствия формулы Тейлора:

$$\frac{f_{i+1} - 2\,f_i + f_{i-1}}{h^2} = f''(x_i) + \frac{h^2}{12}\,f^{(4)}(\xi_i),$$

$$f_{i+1} + 4\,f_i + f_{i-1} = 6\,f_i + h^2 f''(\tilde{\eta}_i).$$

Применим их для получения оценок при $1 \leqslant i \leqslant n-1$:

$$3\,\frac{f_{i+1} - 2\,f_i + f_{i-1}}{h^2} - \left(\frac{1}{2}\,f''(x_{i+1}) + 2\,f''(x_i) + \frac{1}{2}\,f''(x_{i-1})\right) =$$

$$= 3\,f''(x_i) + \frac{h^2}{4}\,f^{(4)}(\xi_i) - \left(3\,f''(x_i) + \frac{h^2}{2}\,f^{(4)}(\eta_i)\right) =$$

$$= \frac{h^2}{4}\,\left(f^{(4)}(\xi_i) - 2f^{(4)}(\eta_i)\right),$$

т. е. $|d_i-(C\varphi)_i| \leqslant c_i h^2 A_4$. Таким образом, получено поточечное неравенство $\max\limits_{0\leqslant i\leqslant n} |M_i - f''(x_i)| \leqslant \tilde{C}_2 h^2 A_4, \quad \tilde{C}_2 = \frac{3}{4}$. Напомним, что $M_i = S_3''(x_i)$.

Оценка для второй производной. Рассмотрим на отрезке $[x_{i-1}, x_i]$ разность

$$f''(x) - S_3''(x) = f''(x) - \left(M_{i-1}\,\frac{x_i - x}{h} + M_i\,\frac{x - x_{i-1}}{h}\right) \pm$$

$$\pm f''(x_{i-1})\,\frac{x_i - x}{h} \pm f''(x_i)\,\frac{x - x_{i-1}}{h}\,.$$

Знак $\pm$ здесь и далее означает одновременное добавление и вычитание соответствующего слагаемого. Преобразуем эту разность к виду

$$\left[f''(x) - \left(f''(x_{i-1})\,\frac{x_i - x}{h} + f''(x_i)\,\frac{x - x_{i-1}}{h}\right)\right] +$$

$$+ \left(f''(x_{i-1}) - M_{i-1}\right)\frac{x_i - x}{h} + \left(f''(x_i) - M_i\right)\frac{x - x_{i-1}}{h}\,.$$

Первое слагаемое можно оценить как приближение функции $f''(x)$ ее линейным интерполянтом, а для оставшихся двух слагаемых можно применить полученное ранее поточечное неравенство. Учитывая, что величины $|x - x_{i-1}|, |x - x_i|$ не превосходят h, окончательно получим

$$\max_{x_{i-1}\leqslant x\leqslant x_i} |S_3''(x) - f''(x)| \leqslant \frac{A_4}{2}\,\frac{h^2}{4} + 2\,\tilde{C}_2 h^2 A_4 = C_2 h^2 A_4,\ C_2 = \frac{13}{8}\,.$$

Правая часть неравенства не зависит от конкретного отрезка $[x_{i-1}, x_i]$, поэтому полученная оценка справедлива для $x_0 \leqslant x \leqslant x_n$.

Оценка для третьей производной. Эта оценка является следствием поточечной оценки для второй производной и явного представления $S_3^{(3)}(x)$ на $[x_{i-1}, x_i]$:

$$S_3^{(3)}(x) = \frac{M_i - M_{i-1}}{h}\,.$$

Получаем

$$f^{(3)}(x) - S_3^{(3)}(x) + \frac{\pm f''(x_i) \pm f''(x_{i-1})}{h} =$$

$$= \left(f^{(3)}(x) - \frac{f''(x_i) - f''(x_{i-1})}{h}\right) + \frac{f''(x_i) - M_i}{h} - \frac{f''(x_{i-1}) - M_{i-1}}{h}\,.$$

Для оценки первого слагаемого разложим $f''(x_i)$ и $f''(x_{i-1})$ в точке x. Получаем

$$f''(x_i) = f''(x) + (x_i - x) f^{(3)}(x) + \frac{(x_i - x)^2}{2}\,f^{(4)}(\xi_+),$$

$$f''(x_{i-1}) = f''(x) + (x_{i-1} - x) f^{(3)}(x) + \frac{(x_{i-1} - x)^2}{2}\,f^{(4)}(\xi_-).$$

Теперь приходим к неравенству

$$\left|f^{(3)}(x) - \frac{1}{h}\left(h f^{(3)}(x) + \frac{(x_i - x)^2}{2}\,f^{(4)}(\xi_+) - \frac{(x_{i-1} - x)^2}{2}\,f^{(4)}(\xi_-)\right)\right| \leqslant$$

$$\leqslant 2\,\frac{h}{2}\,A_4 = h\,A_4.$$

Для оценок оставшихся слагаемых можно воспользоваться поточечной оценкой для второй производной разности, что приводит к окончательному результату

$$\max_x |f^{(3)}(x) - S_3^{(3)}(x)| \leqslant h\,A_4 + 2\,\tilde{C}_2 h\,A_4 = C_3 h\,A_4, \quad C_3 = \frac{5}{2}.$$

Оценка для первой производной. Эта оценка следует из непрерывной оценки для второй производной. Так как $f(x_i) = S_3(x_i)$, $f(x_{i-1}) = S_3(x_{i-1})$, то на отрезке $[x_{i-1}, x_i]$ существует точка ξ_i такая, что $f'(\xi_i) = S_3'(\xi_i)$. Отсюда, по формуле конечных приращений Лагранжа, получим

$$f'(x) - S_3'(x) = (f'(x) - S_3'(x)) - (f'(\xi_i) - S_3'(\xi_i)) = \\ = (x - \xi_i)(f''(\theta_i) - S_3''(\theta_i)),$$

где $x \leqslant \theta_i \leqslant \xi_i$. Поскольку $|x - \xi_i| \leqslant h$, сразу имеем оценку

$$\max_x |f'(x) - S_3'(x)| \leqslant C_1 h^3 A_4, \quad C_1 = C_2 = \frac{13}{8}.$$

Оценка для разности. Эту оценку получают из непрерывной оценки для второй производной тем же способом, что и при выводе оценки погрешности интерполяции многочленом Лагранжа. Рассмотрим функцию $g(z) = f(z) - S_3(z) - R(z - x_i)(z - x_{i-1})$ на отрезке $[x_{i-1}, x_i]$, где число R определяется из условия $g(x) = 0$, $x \neq x_i, x_{i-1}$. Таким образом, на этом отрезке существуют три точки: x, x_i, x_{i-1}, в которых $g(z)$ обращается в нуль. Поэтому в силу теоремы Ролля существуют две точки, в которых $g'(z)$ обращается в нуль, и, наконец, найдется точка ξ такая, что $g''(\xi) = 0$. Отсюда получаем $g''(\xi) = f''(\xi) - S_3''(\xi) - 2\,R = 0$, следовательно,

$$|R| = \tfrac{1}{2}\,|f''(\xi) - S_3''(\xi)| \leqslant \tfrac{1}{2}\,C_2 h^2 A_4.$$

Вспоминая, что точка x выбиралась из условия $g(x) = 0$, приходим к оценке

$$\max_x |f(x) - S_3(x)| \leqslant |R| \max_x |(x - x_i)(x - x_{i-1})| \leqslant \frac{1}{2}\,C_2 h^2 A_4\,\frac{h^2}{4} = C_0 h^4 A_4,$$

с константой $C_0 = \dfrac{C_2}{8} = \dfrac{13}{64}$. Все требуемые оценки получены. $\triangleright$

3.145. На сетке с постоянным шагом h построены естественные сплайны $S_3(x)$ и $S_3^*(x)$ при использовании точных f_i и приближенных f_i^* значений функции, так что $|f_i - f_i^*| \leqslant \varepsilon$. Показать справедливость оценки

$$\max_x |S_3(x) - S_3^*(x)| \leqslant K\,\varepsilon, \quad K = 10.$$

$\triangleleft$ Пусть $x \in [x_{i-1}, x_i]$; тогда, используя аналитическое представление сплайна из 3.141, получим

$$\max_{[x_{i-1}, x_i]} |S_3(x) - S_3^*(x)| \leqslant$$

$$\leqslant |M_{i-1} - M_{i-1}^*|\,\frac{h^2}{3} + |M_i - M_i^*|\,\frac{h^2}{3} + |f_{i-1} - f_{i-1}^*| + |f_i - f_i^*|.$$

Разность $M_i - M_i^*$ удовлетворяет уравнению

$$C(M_i - M_i^*) = d_i - d_i^* = \frac{1}{h}\left[f_{i+1} - 2\,f_i + f_{i-1} - (f_{i+1}^* - 2\,f_i^* + f_{i-1}^*)\right].$$

Отсюда на основании решения 3.143 для коэффициентов естественного сплайна имеем оценку $\max\limits_i |M_i - M_i^*| \leqslant \frac{12}{h^2}\,\varepsilon$. Поэтому справедливо неравенство

$$\max_{[x_{i-1},x_i]} |S_3(x) - S_3^*(x)| \leqslant \frac{12}{h^2}\,\varepsilon\,\frac{2\,h^2}{3} + 2\,\varepsilon = 10\,\varepsilon.$$

Правая часть неравенства не зависит от рассматриваемого отрезка $[x_{i-1}, x_i]$, значит, оно справедливо для $x_0 \leqslant x \leqslant x_n$.

Встречается термин *вычислительная устойчивость сплайна*. Это означает, что возмущение сплайна пропорционально возмущению исходных данных с некоторой абсолютной постоянной. В рассмотренном примере получена оценка с постоянной $K = 10$. $\triangleright$

Используют также *локальные (аппроксимационные) сплайны*, значения которых в узлах, как правило, не совпадают со значениями $f(x)$. Это обстоятельство не принципиально, так как сами значения $f(x)$ обычно известны приблизительно. Рассмотрим построение локального сплайна третьей степени на сетке с постоянным шагом $h = x_{i+1} - x_i$, $i = 0, 1, \ldots, n-1$, для отрезка $[0, 1]$. Возьмем *стандартный* сплайн $B(x)$, определяемый соотношениями

$$B(x) = \begin{cases} \frac{2}{3} - x^2 + |x|^3/2 & \text{при} \quad |x| \leqslant 1, \\ (2 - |x|)^3/6 & \text{при} \quad 1 \leqslant |x| \leqslant 2, \\ 0 & \text{при} \quad 2 \leqslant |x|. \end{cases}$$

Локальные сплайны третьей степени $B_2^{(1)}(x)$ и $B_2^{(2)}(x)$ записываются в виде

$$B_2^{(k)}(x) = \sum_{i=-1}^{n+1} \alpha_i^{(k)} B\left(\frac{x - ih}{h}\right), \quad k = 1, 2,$$

и отличаются выбором коэффициентов.

При $k = 1$ доопределяют значения f_{-1} и f_{n+1} линейной интерполяцией по значениям f_0, f_1 и f_n, f_{n-1} соответственно и полагают $\alpha_i = f_i$ ($f_i = f(x_i)$) при $-1 \leqslant i \leqslant n+1$. При $k = 2$ доопределяют значения f_{-2}, f_{-1} и f_{n+1}, f_{n+2} кубической интерполяцией по значениям f_0, f_1, f_2, f_3 и $f_n, f_{n-1}, f_{n-2}, f_{n-3}$ соответственно и полагают $\alpha_i = \frac{8f_i - f_{i+1} - f_{i-1}}{6}$.

Значения полученных сплайнов в узлах сетки равны некоторому среднему значений функции в ближайших узлах.

3.146. Показать, что при любых $\alpha_i^{(k)}$, $k = 1, 2$, функции $B_2^{(k)}(x)$ являются сплайнами третьей степени, причем они тождественно равны нулю вне отрезка $[-3h, 1 + 3h]$.

$\lhd$ Справедливость первого утверждения следует из свойств стандартного сплайна $B(x)$: он является кусочно-кубической функцией, имеющей в точках $\pm1, \pm2$ непрерывные производные до второго порядка включительно (проверяется непосредственно). Линейная комбинация таких функций удовлетворяет определению кубического сплайна.

Далее рассмотрим в формуле $B_2^{(k)}(x)$ множитель $B\left(\frac{x+h}{h}\right)$ при α_{-1}. Эта функция обращается в нуль при $\left|\frac{x+h}{h}\right| \geqslant 2$, т. е. при $x \leqslant -3\,h$ и $x \geqslant h$. Аналогично множитель $B\left(\frac{x-(n+1)h}{h}\right)$ при α_{n+1} обращается в нуль при $x \leqslant x_n - h = 1 - h$ и $x \geqslant 1 + 3\,h$. Эти слагаемые являются крайними в сумме (первым и последним), поэтому определяют область, где $B_2^{(k)}(x) \not\equiv 0$, а именно отрезок $[-3h, 1 + 3h]$. Областью определения приближаемой функции $f(x)$ является отрезок $[0, 1]$. $\rhd$

3.147. Записать значения f_{-1} и f_{n+1}, необходимые для определения локального сплайна $B_2^{(1)}(x)$.

$\lhd$ Построим многочлен Лагранжа первой степени для $f(x)$ по значениям f_0, f_1:

$$L_2(x) = f_0\,\frac{x_1 - x}{h} + f_1\,\frac{x - x_0}{h}\,, \quad x_i = x_0 + i\,h, \quad i = 0, 1,$$

и вычислим его значение в точке $x = x_0 - h$. Имеем $f_{-1} = L_2(x_0 - h) = 2\,f_0 - f_1$. Аналогично по значениям f_n, f_{n-1} строится величина $f_{n+1} = 2\,f_n - f_{n-1}$. $\rhd$

3.148. Записать значения f_{-2}, f_{-1} и f_{n+1}, f_{n+2}, необходимые для определения локального сплайна $B_2^{(2)}(x)$.

$\lhd$ Построим многочлен Лагранжа третьей степени для $f(x)$ по значениям f_n, f_{n-1}, f_{n-2}, f_{n-3}. Имеем

$$L_4(x) = \sum_{i=1}^{4} f_{n+1-i} \prod_{\substack{j=1\\ j\neq i}}^{4} \frac{x - x_{n+1-j}}{x_{n+1-i} - x_{n+1-j}}\,,$$

где $x_i = x_0 + i\,h$; $i = n, n-1, n-2, n-3$; $h = \frac{x_n - x_0}{n}$, и вычислим значения многочлена в точках $x_n + h$ и $x_n + 2\,h$. Получаем

$$f_{n+1} = L_4(x_n + h) = 4\,f_n - 6\,f_{n-1} + 4\,f_{n-2} - f_{n-3},$$
$$f_{n+2} = 10\,f_n - 20\,f_{n-1} + 15\,f_{n-2} - 4\,f_{n-3}.$$

Аналогично по значениям f_0, f_1, f_2, f_3 строят величины

$$f_{-2} = 10\,f_0 - 20\,f_1 + 15\,f_2 - 4\,f_3, \quad f_{-1} = 4\,f_0 - 6\,f_1 + 4\,f_2 - f_3. \quad \rhd$$

3.149. Показать, что величина $B_2^{(1)}(x)$ зависит только от значений f_i в четырех ближайших к x точках x_i, а величина $B_2^{(2)}(x)$ — в шести точках.

$\lhd$ Пусть $x \in [x_{k-1}, x_k]$, тогда $x = \theta\, h$, $k-1 \leqslant \theta \leqslant k$. Неравенство $\left|\frac{x-i\,h}{h}\right| \equiv |\theta - i| < 2$ выполняется только для значений $i = k-2, k-1, k, k+1$. Так как стандартный сплайн $B(x)$ равен нулю при $|x| \geqslant 2$, то $B_2^{(1)}(x)$ зависит только от значений f_i, $i = k-2, k-1, k, k+1$. Напомним, что для $B_2^{(1)}(x)$ коэффициенты определяются как $\alpha_i = f_i$.

Анализ для $B_2^{(2)}(x)$ проводится аналогично, только зависимость от значений в шести точках связана с другой формулой для коэффициентов: $\alpha_i = \frac{8f_i - f_{i+1} - f_{i-1}}{6}$. $\rhd$

3.150. Показать, что $B_2^{(i)}(x_0) = f_0$, $B_2^{(i)}(x_n) = f_n$, $i = 1, 2$; $B_2^{(2)}(x_1) = f_1$, $B_2^{(2)}(x_{n-1}) = f_{n-1}$.

$\lhd$ Покажем в качестве примера равенство $B_2^{(2)}(x_n) = f_n$. Так как $x_n = n\,h$, получим

$$B_2^{(2)}(x_n) = \sum_{i=-1}^{n+1} \alpha_i B\left(\frac{nh - ih}{h}\right) = \alpha_{n+1}B(-1) + \alpha_n B(0) + \alpha_{n-1}B(1) =$$

$$= \frac{\alpha_{n+1} + 4\,\alpha_n + \alpha_{n-1}}{6} = \frac{1}{6}\left(\frac{-f_{n+2} + 8f_{n+1} - f_n}{6} + 4\,\frac{-f_{n+1} + 8\,f_n - f_{n-1}}{6} + \right.$$

$$\left. + \frac{-f_n + 8f_{n-1} - f_{n-2}}{6}\right) = \frac{1}{36}\left(-f_{n+2} + 4f_{n+1} + 30f_n + 4f_{n-1} - f_{n-2}\right) = f_n.$$

Для получения последнего равенства использованы выражения для f_{n+2} и f_{n+1} из 3.148. $\rhd$

3.151. Пусть $|f^{(4)}(x)| \leqslant A_4$. Показать, что

$$\left|\left(B_2^{(2)}(x)\right)^{(l)} - f^{(l)}(x)\right| \leqslant C_l A_4 h^{4-l}, \quad l = 0, 1, 2, 3.$$

$\lhd$ Для сплайна $B_2^{(2)}(x)$ будем использовать обозначение $B_2(x)$, чтобы избежать недоразумений с символами производных.

Поточечное неравенство для второй производной. Рассмотрим выражение $B_2''(x)$ в одном из узлов $x_k = k\,h$. Имеем

$$B_2''(x_k) = \frac{1}{h^2}\left(\alpha_{k-1}B''(1) + \alpha_k B''(0) + \alpha_{k+1}B''(-1)\right) =$$

$$= \frac{\alpha_{k-1} - 2\,\alpha_k + \alpha_{k+1}}{h^2} = \frac{-f_{k-2} + 10\,f_{k-1} - 18\,f_k + 10\,f_{k+1} - f_{k+2}}{6\,h^2}\,.$$

Используя разложения в ряд Тейлора для величин $f_{k\pm1}$, $f_{k\pm2}$ в точке $x = x_k$, получим $B_2''(x_k) = f''(x_k) + \tilde{C}_2 h^2 f^{(4)}(\xi_k)$.

Оценка для второй производной. Эта оценка выводится из поточечного неравенства как для интерполяционного сплайна (см. решение 3.144), если в приведенных там выкладках $S_3(x)$ заменить на $B_2(x)$.

Оценка третьей производной. В этом случае необходимо отметить, что на отрезке $[x_{k-1}, x_k]$ справедливо равенство

$$B_2^{(3)}(x) = \frac{-\alpha_{k-2} + 3\,\alpha_{k-1} - 3\,\alpha_{k+1} + \alpha_{k+2}}{h^3} = \frac{B_2''(x_k) - B_2''(x_{k-1})}{h}\,.$$

Дальнейшие рассуждения такие же, как для интерполяционного сплайна (см. решение 3.144).

Оценка для первой производной. Рассмотрим на отрезке $[x_{k-1}, x_k]$ функцию $g(x)$, про которую известно следующее:

1)$g'(x)$ непрерывна; 2)$|g'(x)| \leqslant K$. Тогда

$$g(x) = \int\limits_{x_{k-1}}^{x} g'(\xi)d\xi + g(x_{k-1}) \quad \text{и} \quad |g(x)| \leqslant K\,h + |g(x_{k-1})|.$$

В рассматриваемом случае $g(x) = B_2'(x) - f'(x)$, и имеется оценка $|g'(x)| = |B_2''(x) - f''(x)| \leqslant K = C_2A_4h^2$. Для нахождения недостающей величины $|g(x_{k-1})|$ рассмотрим значение $B_2'(x)$ в узлах $x_k = k\,h$

$$B_2'(x_k) = \frac{1}{h}\,(\alpha_{k-1}B'(1) + \alpha_kB'(0) + \alpha_{k+1}B'(-1)) = \frac{\alpha_{k+1} - \alpha_{k-1}}{2\,h} =$$

$$= \frac{f_{k-2} - 8\,f_{k-1} + 8\,f_{k+1} - f_{k+2}}{12\,h} = f'(x_k) + \tilde{C}_1 f^{(4)}(\xi_k)\,h^3.$$

Откуда и следует искомая оценка:

$$\max_x |f'(x) - B_2'(x)| \leqslant C_1h^3A_4.$$

Оценка для разности. Эта оценка получается таким же способом: $g(x) = B_2(x) - f(x)$. В данном случае $K = C_1A_4h^3$,

$$B_2(x_k) = \alpha_{k-1}B(1) + \alpha_kB(0) + \alpha_{k+1}B(-1) =$$

$$= \frac{\alpha_{k+1} + 4\,\alpha_k + \alpha_{k-1}}{6} = f(x_k) + \tilde{C}_0 f^{(4)}(\xi_k)\,h^4,$$

что приводит к завершающей оценке для $l = 0$. $\triangleright$

3.152. Пусть $|f^{(2)}(x)| \leqslant A_2$. Показать, что

$$\left|\left(B_2^{(1)}(x)\right)^{(l)} - f^{(l)}(x)\right| \leqslant C_lA_2h^{2-l}, \quad l = 0, 1.$$

Глава 4

Численное интегрирование

Рассмотрим интеграл вида

$$I(f) = \int_a^b p(x)\, f(x)\, dx\,,$$

где $[a, b]$ — конечный или бесконечный промежуток числовой оси и $f(x)$ — произвольная функция из некоторого класса F. Если не оговорено противное, то считаем, что все $f(x)$ непрерывны на отрезке $[a, b]$. Заданную функцию $p(x)$ называют *весовой*. Будем предполагать, что на $[a, b]$ она измерима, тождественно не равна нулю (как правило, почти всюду положительна) и ее произведение на любую $f(x) \in F$ суммируемо.

Для приближенного вычисления интеграла $I(f)$ строят линейные квадратурные формулы (*квадратуры*) следующего вида:

$$S_n(f) = \sum_{i=1}^{n} c_i\, f(x_i)\,.$$

Постоянные c_i называются *коэффициентами* (*весами*) квадратуры, x_i — ее *узлами*.

Для каждой функции $f(x) \in F$ погрешность квадратурной формулы $S_n(f)$ определяется как $R_n(f) = I(f) - S_n(f)$. При этом оценкой погрешности на классе F называют величину

$$R_n(F) = \sup_{f\in F} |R_n(f)|\,, \quad \|R_n(F)\| = \sup_{f\in F, \|f\|_F \neq 0} \frac{|R_n(f)|}{\|f\|_F}\,.$$

На практике часто используют оценки сверху для $|R_n(f)|$, которые будем обозначать через R_n.

4.1. Интерполяционные квадратуры

Имеется большая группа квадратурных формул, построенных на основе замены $f(x)$ алгебраическим интерполяционным многочленом. Пусть на конечном промежутке $[a, b]$ по заданному набору различных узлов $\{x_i\}_{i=1}^n$ функция $f(x)$ приближается интерполяционным многочленом Лагранжа $L_n(x)$ степени $n - 1$

$$L_n(x) = \sum_{i=1}^{n} f(x_i) \prod_{\substack{j=1\\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}\,.$$

Положим

$$S_n(f) = \int_a^b p(x)\, L_n(x)\, dx\,.$$

Отсюда получаем явные формулы для набора коэффициентов $\{c_i\}_{i=1}^n$ и оценку погрешности R_n такую,что $|I(f) - S_n(f)| \leqslant R_n$,

$$c_i = \int_a^b p(x) \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}\, dx, \quad R_n = \frac{\left\|f^{(n)}(x)\right\|}{n!} \int_a^b |p(x)|\,|\,\omega_n(x)|\, dx\,,$$

где

$$\left\|f^{(n)}(x)\right\| = \max_{x\in[a,b]} \left|f^{(n)}(x)\right|\,, \quad \omega_n(x) = \prod_{i=1}^{n}(x - x_i)\,.$$

В оценках, приведенных ниже, также используется равномерная норма.

Квадратурные формулы интерполяционного типа, построенные в случае весовой функции $p(x) \equiv 1$ для системы равноотстоящих узлов $x_i = a+(i-1)\,\frac{b-a}{n-1}$, $i = 1, \ldots, n$, называют *формулами Ньютона—Котеса*.

4.1. Получить формулы Ньютона—Котеса и соответствующие оценки погрешностей при числе узлов $n = 1, 2, 3$.

У к а з а н и е. При вычислении интегралов использовать замену переменной $x = x(t) = \frac{b+a}{2} + \frac{b-a}{2}\, t$. В частности,

$$\int_a^b |\omega_n(x)|\, dx = \left(\frac{b-a}{2}\right)^{n+1} \int_{-1}^{1} |\omega_n^0(t)|\, dt\,,$$

где $\omega_n^0(t) = \prod_{i=1}^{n}(t - t_i)$, а t_i являются образами узлов x_i на отрезке $[-1, 1]$.

О т в е т: $n = 1$ — формула прямоугольников

$$S_1(f) = (b-a) f\left(\frac{a+b}{2}\right), \quad R_1 = \|f'(x)\|\, \frac{(b-a)^2}{4}\,;$$

$n = 2$ — формула трапеций

$$S_2(f) = \frac{b-a}{2}\,(f(a) + f(b)), \quad R_2 = \|f''(x)\|\, \frac{(b-a)^3}{12}\,;$$

$n = 3$ — формула парабол (Симпсона)

$$S_3(f) = \frac{b-a}{6}\left(f(a) + 4\, f\left(\frac{a+b}{2}\right) + f(b)\right), \quad R_3 = \|f^{(3)}(x)\|\, \frac{(b-a)^4}{192}\,.$$

4.2. Рассмотрим формулы прямоугольников и трапеций. Какая из них имеет лучшую точность?

$\lhd$ Можно сравнивать точность только для функций из одного класса, поэтому необходимо получить для формулы прямоугольников другую оценку погрешности. Воспользуемся в качестве приближения к функции $f(x)$ отрезком ряда Тейлора в точке $\frac{a+b}{2}$. Имеем

$$f(x) = f\left(\frac{a+b}{2}\right) + f'\left(\frac{a+b}{2}\right)\left(x - \frac{a+b}{2}\right) + \frac{f''(\xi)}{2}\left(x - \frac{a+b}{2}\right)^2\,.$$

Тогда для квадратурной формулы $\widetilde{S}_1(f)$, полученной с помощью интегрирования двух первых слагаемых, справедливо равенство

$$\widetilde{S}_1(f) = \int_a^b \left[f\left(\frac{a+b}{2}\right) + f'\left(\frac{a+b}{2}\right)\left(x - \frac{a+b}{2}\right)\right] dx = S_1(f)\,,$$

при этом оценка погрешности принимает вид

$$\widetilde{R}_1 = \frac{\|f''(x)\|}{2} \int_a^b \left(x - \frac{a+b}{2}\right)^2 dx = \|f''(x)\| \, \frac{(b-a)^3}{24}\ .$$

Следовательно, на классе функций с непрерывной второй производной формула прямоугольников имеет оценку погрешности в два раза меньшую, чем формула трапеций. $\triangleright$

В общем случае оценка погрешности для формул Ньютона—Котеса имеет следующий вид:

при нечетных n

$$\tilde{R}_n = \frac{\left\|f^{(n+1)}(x)\right\|}{(n+1)!} \left| \int_a^b x\,\omega_n(x)\,dx \right|,$$

при четных n

$$\tilde{R}_n = \frac{\left\|f^{(n)}(x)\right\|}{n!} \left| \int_a^b \omega_n(x)\,dx \right|.$$

Отсюда можно получить известную оценку погрешности для формулы Симпсона: $\tilde{R}_3 = \|f^{(4)}(x)\| \, \dfrac{(b-a)^5}{2880}$.

4.3. Пусть весовая функция $p(x)$ четная, узлы x_i расположены симметрично относительно нуля, т. е. $x_{n+1-i} = -x_i$, $i = 1, \ldots, n$. Доказать, что в интерполяционной квадратурной формуле для вычисления интеграла $I(f) = \int\limits_{-a}^{a} p(x)\, f(x)\, dx$ коэффициенты, соответствующие симметричным узлам, равны, т. е. $c_{n+1-i} = c_i$, $i = 1, \ldots, n$.

У к а з а н и е. В формуле для коэффициента квадратуры

$$c_{n+1-i} = a \int_{-1}^{1} p\,(at) \prod_{j \neq n+1-i} \frac{t - t_j}{t_i - t_j}\, dt$$

заменить узлы на симметричные $t_{n+1-i} = -t_i$, $t_j = -t_{n+1-j}$, формально поменять индекс в произведении и использовать свойство определенного интеграла $\int\limits_{-1}^{1} g(t)dt = \int\limits_{-1}^{1} g(-t)dt$.

4.4. Доказать, что для погрешности квадратурной формулы трапеций справедливо представление

$$R_2(f) = \int_a^b f(x)\,dx - \frac{b-a}{2}\,(f(a) + f(b)) = \frac{1}{2}\int_a^b (a-\xi)(b-\xi) f''(\xi)\,d\xi.$$

Указание. Проинтегрировать правую часть равенства по частям два раза или использовать формулу Тейлора с остаточным членом в интегральной форме.

Составные квадратурные формулы. Рассмотрим задачи на построение составных квадратурных формул и вывод оценок их погрешностей. Пусть $h = \frac{b-a}{N}$ и $x_k = a + kh$, $k = 0, 1, \ldots, N$. Введем следующие обозначения: $I^{(k)}(f) = \int\limits_{x_k}^{x_{k+1}} p(x) f(x)\, dx$, $S_n^{(k)}(f) = S_n(f)$ для отрезка $[x_k, x_{k+1}]$, $k = 0, \ldots, N-1$.

Исходный интеграл $I(f)$ равен $I(f) = \sum\limits_{k=0}^{N-1} I^{(k)}(f)$, поэтому соответствующая составная квадратурная формула принимает вид $S_n^N(f) = \sum\limits_{k=0}^{N-1} S_n^{(k)}(f)$, а для ее погрешности справедливо неравенство $|R_n^N(f)| \leqslant \sum\limits_{k=0}^{N-1} \left|R_n^{(k)}(f)\right|$. Например, в случае составной формулы прямоугольников

$$S_1^N(f) = \frac{b-a}{N} \sum_{k=0}^{N-1} f\left(x_k + \frac{h}{2}\right)$$

для погрешности на отрезке $[x_k, x_{k+1}]$ имеем неравенство

$$\left|R_1^{(k)}(f)\right| \leqslant \|f''(x)\| \frac{(x_{k+1} - x_k)^3}{24} = \|f''(x)\| \frac{h^3}{24} = \|f''(x)\| \frac{(b-a)^3}{24\, N^3}\ .$$

Следовательно, для всего отрезка $[a, b]$ оценка погрешности получается суммированием по всем $[x_k, x_{k+1}]$

$$R_1^N = \|f''(x)\| \frac{(b-a)^3}{24\, N^2}\ .$$

4.5. Для вычисления $\int\limits_0^1 f(x)\, dx$ применяется составная формула трапеций. Оценить минимальное число разбиений N, обеспечивающее точность $0,5 \cdot 10^{-3}$ на следующих классах функций: 1) $\|f''(x)\| \leqslant 1$; 2) $\int\limits_0^1 |f''(x)|\, dx \leqslant 1$.

Ответ: 1) $N = 13$; 2) $N = 16$ (для этого случая полезно использовать 4.4).

4.6. Для составной квадратурной формулы трапеций с шагом $h = \frac{b-a}{N}$, погрешность которой имеет вид

$$R_2^N(f) = \int_a^b f(x)\, dx - \frac{b-a}{N} \left(\frac{1}{2} f(x_0) + \frac{1}{2}\ f(x_N) + \sum_{k=1}^{N-1} f(x_k)\right)$$

получить оценки погрешности следующего вида:

$$1)\ R_2^N = \frac{h^2}{8}\int\limits_a^b |f''(x)|\,dx; \quad 2)\ R_2^N = \frac{h^2}{2}\sqrt{\frac{b-a}{30}}\left(\int\limits_a^b |f''(x)|^2 dx\right)^{1/2}.$$

Указание. Использовать 4.4. Для второго случая дополнительно ввести функцию

$$\varphi_k(x) = \begin{cases} (x_k - x)(x_{k+1} - x) & \text{на } [x_k, x_{k+1}], \\ 0 & \text{вне } [x_k, x_{k+1}]. \end{cases}$$

Тогда имеют место соотношения

$$\sum_{k=0}^{N-1} \int\limits_{x_k}^{x_{k+1}} (x_k - \xi)(x_{k+1} - \xi) f''(\xi)\,d\xi =$$

$$= \sum_{k=0}^{N-1} \int\limits_a^b \varphi_k(\xi) f''(\xi)\,d\xi = \int\limits_a^b f''(\xi) \sum_{k=0}^{N-1} \varphi_k(\xi)\,d\xi\,,$$

к последнему из которых следует применить неравенство Коши—Буняковского.

4.7. Вычислить интеграл $\int\limits_0^1 \exp(x^2)dx$ по формуле Ньютона—Котеса с узлами $x_1 = 0$, $x_2 = \frac{1}{4}$, $x_3 = \frac{1}{2}$, $x_4 = \frac{3}{4}$, $x_5 = 1$ и оценить погрешность.

4.8. Найти оценку погрешности вычисления интеграла $\int\limits_0^1 f(x)\,dx$ при $f(x) = \frac{1}{1+x^2}$ по составной квадратурной формуле

$$S(f) = \frac{f(0) + 4f(0,1) + 2f(0,2) + 4f(0,3) + \ldots + 4f(0,9) + f(1,0)}{30}.$$

Указание. Покажем, что $\|f^{(n)}(x)\| = n!$. Для этого введем функцию $y = \operatorname{arctg} x$. Тогда $y' = f(x)$. Используя обратную функцию $x = \operatorname{tg} y$, получим

$$y' = \cos^2 y, \quad y'' = -2y' \cos y \sin y, \quad \ldots\,.$$

Эти выражения можно преобразовать к виду

$$y' = \cos y \sin\left(y + \frac{\pi}{2}\right),$$

$$y'' = \cos^2 y \sin 2\left(y + \frac{\pi}{2}\right),$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$y^{(n)} = (n-1)! \cos^n y \sin n\left(y + \frac{\pi}{2}\right).$$

Отсюда следует $\|f^{(n)}(x)\| = \|y^{(n+1)}(x)\| = n!$.

Ответ: $\|f^{(4)}(x)\|\,\frac{1}{2880 \cdot 5^4} = \frac{1}{75000}$.

4.9. Найти оценку погрешности вычисления интеграла $\int\limits_0^1 f(x)\,dx$ при $f(x) = \dfrac{1}{1+x^2}$ по составной квадратурной формуле

$$S(f) = \frac{f(0) + 2f(0,1) + 2f(0,2) + \ldots + 2f(0,9) + f(1,0)}{20}\,.$$

Ответ: $\|f''(x)\| \frac{1}{12 \cdot 10^2} = \frac{1}{600}$ (см. указание к 4.8).

4.10. Оценить число разбиений отрезка N для вычисления интеграла $\int\limits_0^1 \sin(x^2)\,dx$ по составной квадратурной формуле трапеций, обеспечивающее точность 10^{-4}.

Ответ: $N \geqslant 45 > \left[\sqrt{\frac{\theta}{12}} \cdot 10^2\right] + 1, \quad \theta = \max(2, 4\sin 1 - 2\cos 1) < 2,29\,.$

4.11. Оценить число разбиений отрезка N для вычисления интеграла $\int\limits_0^1 \exp(x^2)\,dx$ по составной квадратурной формуле прямоугольников, обеспечивающее точность 10^{-4}.

Ответ: $N \geqslant [50\sqrt{\mathrm{e}}] + 1 = 83\,.$

4.12. Оценить число узлов составной квадратурной формулы трапеций для вычисления интеграла $\int\limits_0^1 \exp(x^2)dx$, обеспечивающее точность $\varepsilon \leqslant 10^{-4}$.

4.13. Оценить число узлов составной квадратурной формулы Симпсона для вычисления интеграла $\int\limits_0^2 f(x)dx$, обеспечивающее точность $\varepsilon \leqslant 0,5 \cdot 10^{-4}$ на классе функций, удовлетворяющих условию $\|f^{(4)}(x)\| \leqslant 1$.

4.14. Записать квадратурную формулу для вычисления с точностью 10^{-4} интегралов $I(f) = \int\limits_0^\infty \mathrm{e}^{-x} f(x)dx$, $I(f) = \int\limits_0^\infty x\mathrm{e}^{-x} f(x)dx$, если для некоторого фиксированного $k \geqslant 1$ выполнено неравенство $\|f^{(k)}(x)\| \leqslant 1$.

4.15. Доказать справедливость следующих представлений погрешностей квадратурных формул:

1) $\int\limits_a^b f(x)dx - \frac{b-a}{8}\left(f(a)+3f\left(\frac{2a+b}{3}\right)+3f\left(\frac{a+2b}{3}\right)+f(b)\right) =$

$= -\left(\frac{b-a}{3}\right)^5 \frac{3}{80}\, f^{(4)}(\xi), \quad a<\xi<b;$

2) $\int\limits_a^b f(x)dx - \frac{b-a}{90}\left(7f(a)+32f\left(\frac{3a+b}{4}\right)+12f\left(\frac{a+b}{2}\right)+32f\left(\frac{a+3b}{4}\right)+\right.$

$\left.+\,7f(b)\right) = -\left(\frac{b-a}{4}\right)^7 \frac{8}{945}\, f^{(6)}(\xi), \quad a<\xi<b;$

3) $\int\limits_a^b f(x)dx - \frac{b-a}{2}\left(f(a)+f(b)\right) - \frac{(b-a)^2}{12}\left(f'(a)-f'(b)\right) =$

$= \frac{(b-a)^5}{720}\, f^{(4)}(\xi), \quad a<\xi<b.$

4.16. Показать, что ни для какой системы узлов и коэффициентов погрешность квадратурной формулы $R_n(f)$ не стремится сильно к нулю на пространстве непрерывных функций ($f(x)\in C[a,b]$). Более того, всегда справедливо равенство

$$\|R_n(C)\| = \int\limits_a^b p(x)\,dx + \sum_{i=1}^n |c_i|.$$

$\lhd$ Рассмотрим непрерывную на отрезке $[a,b]$ функцию $f(x)$ такую, что

$$\max_{[a,b]}|f(x)| = 1, \quad \int\limits_a^b p(x)f(x)\,dx \geqslant \int\limits_a^b p(x)\,dx - \varepsilon$$

и $f(x_i) = -\operatorname{sign} c_i,\ i=1,2,\ldots,n$, где ε — произвольно малое число. Она строится конструктивно по заданным узлам x_i и весам c_i: $f(x)\equiv 1$ вне малых окрестностей точек x_i, внутри них — непрерывная функция $|f(x)|\leqslant 1$ и $f(x_i) = -\operatorname{sign} c_i$. Для такой функции справедливо неравенство

$$R_n(f) = I(f) - S_n(f) \geqslant \int\limits_a^b p(x)\,dx + \sum_{i=1}^n |c_i| - \varepsilon,$$

и так как $|R_n(f)|\leqslant \|R_n(C)\| \max\limits_{[a,b]}|f(x)| = \|R_n(C)\|$, то имеет место оценка

$$\|R_n(C)\| \geqslant \int\limits_a^b p(x)\,dx + \sum_{i=1}^n |c_i| - \varepsilon,$$

откуда в силу произвольности ε следует, что

$$\|R_n(C)\| \geqslant \int\limits_a^b p(x)\,dx + \sum_{i=1}^n |c_i|.$$

Неравенство противоположного знака устанавливается просто, поэтому искомое утверждение доказано.

Этот факт иллюстрирует «пессимистическую» точку зрения, согласно которой проблема численного интегрирования непрерывных функций, вообще говоря, неразрешима. Однако для практических приложений более важен факт существования системы узлов и весовых коэффициентов таких, что $R_n(f) \to 0$ слабо на $C[a,b]$ при $n \to \infty$. Это тем более важно, что в приложениях приходится иметь дело не со всем пространством $C[a,b]$, а с некоторым его компактным подмножеством, для элементов которого можно указать порядок стремления к нулю величины $R_n(f)$. $\triangleright$

4.17. Пусть $C_q = \int\limits_a^b |f^{(q)}(x)|dx < \infty$, $q = 1, 2$. Получить оценку погрешности формулы трапеций $|R_2^N(f)| \leqslant \tau_q C_q h^q$, где τ_q — абсолютная постоянная, h — шаг интегрирования.

4.18. Пусть $C_q = \int\limits_a^b |f^{(q)}(x)|dx < \infty$, $q = 1, 2, 3, 4$. Получить оценку погрешности формулы Симпсона $|R_3^N(f)| \leqslant \rho_q C_q h^q$, где ρ_q — абсолютная постоянная, h — шаг интегрирования.

В упражнениях 4.19–4.21 рассматривается приближенное вычисление интеграла $I(f_b) = \int\limits_0^1 f_b(x)dx$ от функции с параметром $|b| < 1$:

$$f_b(x) = \begin{cases} 0 & \text{при } x = 0, \\ x^b & \text{при } x \in (0,1]. \end{cases}$$

4.19. Интеграл $I(f_b)$ вычисляется по составной квадратурной формуле трапеций с постоянным шагом $\frac{1}{N}$. Доказать, что суммарная погрешность удовлетворяет следующему соотношению: $|R_2^N(f)| \leqslant \frac{D_1(b)}{N^{1+b}}$, $D_1(b) \neq 0$.

4.20. Интеграл $I(f_b)$ вычисляется по составной квадратурной формуле трапеций с распределением узлов $x_q = \varphi\left(\frac{q}{N}\right)$, $\varphi(t) = t^{3/(1+b)}$. Доказать, что суммарная погрешность удовлетворяет соотношению $|R_2^N(f)| \leqslant \leqslant \frac{D_2(b)}{N^2}$, $D_2(b) \neq 0$.

4.21. Интеграл $I(f_b)$ вычисляется по составной квадратурной формуле трапеций с распределением узлов $x_q = \varphi\left(\frac{q}{N}\right)$, $\varphi(t) = t^a$. Доказать, что при $a > \frac{2}{b+1}$ суммарная погрешность удовлетворяет соотношению $|R_2^N(f)| \leqslant \frac{D(a,b)}{N^2}$, $D(a,b) \neq 0$.

Проверить, что $D(a,b) > D_2(b)$, где $D_2(b)$ определено в 4.20.

4.2. Метод неопределенных коэффициентов

Если интегралы вида $\int\limits_a^b p(x)x^k\,dx$ вычисляются просто, то при заданном наборе различных узлов можно найти коэффициенты c_i из условия точности квадратурной формулы $S_n(f) = \sum\limits_{i=1}^{n} c_i\,f(x_i)$ для произвольного многочлена наиболее высокой степени, т. е. из равенств $I(x^k) = S_n(x^k)$, $k = 0, 1, \ldots, (n-1)$. Полученная система линейных уравнений относительно c_i имеет единственное решение.

Если квадратура точна для многочлена степени m (говорят, что она имеет алгебраический порядок точности, равный m), то справедливо равенство $R_n(f) = R_n(f - P_m)$. Взяв в качестве $P_m(x)$ интерполяционный многочлен для $f(x)$, построенный по нулям многочлена Чебышёва, можно получить оценку

$$|R_n(f)| \leqslant \frac{\|f^{(m+1)}\|}{(m+1)!}\,\frac{(b-a)^{m+1}}{2^{2m+1}}\left(\int\limits_a^b |p(x)|dx + \sum_{i=1}^{n}|c_i|\right).$$

Из условия точности квадратурной формулы для функций заданного вида можно выписать уравнения (в общем случае нелинейные) не только для определения коэффициентов, но и для узлов квадратуры.

Квадратурными формулами Чебышёва называют квадратуры с одинаковыми коэффициентами, т. е.

$$S_n(f) = c\sum_{i=1}^{n} f(x_i), \quad c = \frac{1}{n}\int\limits_a^b p(x)\,dx.$$

Их построение заключается в нахождении узлов x_i из условия точности для многочлена максимально высокой степени. Квадратуры Чебышёва (их удается построить при $n = 1, 2, 3, 4, 7, 10$) обычно применяют, если значения $f(x_i)$ известны с независимыми случайными погрешностями. В этом случае выбор равных коэффициентов обеспечивает минимальную дисперсию $S_n(f)$.

4.22. Получить формулу Симпсона методом неопределенных коэффициентов.

У к а з а н и е. Сначала построить формулу на отрезке $[-1, 1]$, а затем отобразить ее на $[a, b]$.

О т в е т: $S_3(f) = \frac{b-a}{6}\left(f(a) + 4\,f\left(\frac{a+b}{2}\right) + f(b)\right)$.

4.23. Для формул трапеций и Симпсона найти оценки погрешности, следующие из метода неопределенных коэффициентов.

О т в е т: для формулы трапеций $|R_2(f)| \leqslant \frac{(b-a)^3}{8}\,\|f''\|$, так как $m = 1$; для формулы Симпсона $|R_3(f)| \leqslant \frac{(b-a)^5}{1536}\,\|f^{(4)}\|$, так как $m = 3$.

4.24. Для вычисления интегралов $I(f)$:

$$1)\ \int_0^2 (x+1)f(x)dx; \quad 2)\ \int_{-1}^0 x^2 f(x)dx; \quad 3)\ \int_{-1}^1 x^2 f(x)dx$$

построить формулы вида $S(f) = c_1\, f(\tilde{x}) + c_2\, f(x_2)$ с одним фиксированным узлом $\tilde{x} = 0$, точные для многочленов максимально высокой степени.

Ответ: 1) $S(f) = \frac{11}{15}\, f(0) + \frac{49}{15}\, f\left(\frac{10}{7}\right)$; 2) $S(f) = \frac{1}{48}\, f(0) + \frac{5}{16}\, f\left(-\frac{4}{5}\right)$; 3) $S(f) = \frac{2}{3}\, f(0)$.

4.25. Рассмотрим многочлен

$$P_n(x) = (x - x_1)\, \ldots\, (x - x_n) = x^n + a_1 x^{n-1} + a_2 x^{n-2} + \ \ldots\ + a_n.$$

Доказать, что величины $B_j = \sum\limits_{k=1}^{n} x_k^j$, $j = 1, \ldots n$, удовлетворяют равенствам

$$B_1 = -a_1,$$
$$a_1 B_1 + B_2 = -2a_2,$$
$$a_2 B_1 + a_1 B_2 + B_3 = -3a_3,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$a_{n-1} B_1 + a_{n-2} B_2 + \ldots + a_1 B_{n-1} + B_n = -n a_n.$$

$\lhd$ Представим производную $P_n(x)$ в виде

$$P_n'(x) = P_n(x)\ \frac{d}{dx} \ln P_n(x) = \sum_{k=1}^{n} \frac{P_n(x)}{x - x_k}\,,$$

где

$$\frac{P_n(x)}{x - x_k} = x^{n-1} + (a_1 + x_k)x^{n-2} + (a_2 + a_1 x_k + x_k^2)x^{n-3} + \ \ldots$$
$$\ldots + (a_{n-1} + a_{n-2} x_k + \cdots + a_1 x_k^{n-2} + x_k^{n-1}).$$

Положим $a_0 = 1$. Тогда соотношение для производной можно записать в виде

$$\sum_{k=0}^{n-1} (n-k) a_k x^{n-k-1} = n x^{n-1} + (n a_1 + B_1)\, x^{n-2} +$$
$$+ (n a_2 + a_1 B_1 + B_2)\, x^{n-3} + \ \ldots\ + (n a_{n-1} + a_{n-2} B_1 + \ \ldots\ + a_1 B_{n-2} + B_{n-1}).$$

Из равенства коэффициентов при одинаковых степенях x и следуют соотношения для $a_1, \ldots, a_{n-1}$. Последнее соотношение (для a_n) получается в результате сложения равенств

$$P_n(x_k) = \sum_{j=0}^{n} a_j x_k^{n-j}, \quad k = 1, 2, \ \ldots\ , n,$$
$$a_{n-1} B_1 + a_{n-2} B_2 + \ \ldots\ + a_1 B_{n-1} + B_n = -a_n n. \qquad \rhd$$

4.26. Построить квадратурные формулы Чебышёва на отрезке $[-1,1]$ с весом $p(x)\equiv 1$ для $n=2,3,4$.

Указание. Для $f(x)=x^j$, $j=1,2,\dots,n$, имеем следующие соотношения:

$$I(x^j)=S_n(x^j),\quad \text{или}\quad \frac{1-(-1)^{j+1}}{j+1}=\frac{2}{n}\sum_{k=1}^{n}x_k^j=\frac{2}{n}\,B_j,$$

где B_j определены в 4.25. Решая эти системы, получаем

$$P_2(x)=x^2-\tfrac{1}{3}\,,\quad P_3(x)=x^3-\tfrac{1}{2}\,x,\quad P_4(x)=x^4-\tfrac{2}{3}\,x^2+\tfrac{1}{45}\,.$$

4.27. Показать, что квадратурная формула

$$S_n(f)=\frac{\pi}{n}\sum_{j=1}^{n}f\left(\cos\frac{2j-1}{2n}\,\pi\right)$$

для вычисления интегралов $I(f)=\int\limits_{-1}^{1}\dfrac{f(x)}{\sqrt{1-x^2}}\,dx$ точна для всех алгебраических многочленов степени $2n-1$.

$\lhd$ Представим произвольный многочлен $P_{2n-1}(x)$ степени $2n-1$ в виде суммы многочленов Чебышёва: $P_{2n-1}(x)=\sum\limits_{m=0}^{2n-1}a_mT_m(x)$, для которых $T_m(x)=\cos(m\arccos x)$, и проверим утверждение.

При $m=0$ имеем

$$I(T_0)=\int_{-1}^{1}\frac{1}{\sqrt{1-x^2}}\,dx=\pi,\quad S_n(T_0)=\pi.$$

При $m>0$ справедливо свойство ортогональности $I(T_m\,T_0)=0$. Для квадратурной формулы выполним преобразования

$$S_n(T_m)=\frac{\pi}{n}\sum_{j=1}^{n}\cos(m\arccos x_j)=\frac{\pi}{n}\sum_{j=1}^{n}\cos m\,\frac{(2j-1)\pi}{2n}=$$

$$=\frac{\pi}{2n}\sum_{j=1-n}^{n}\exp\left(\frac{m(2j-1)\pi\mathrm{i}}{2n}\right).$$

Далее используем формулу суммы членов геометрической прогрессии $\sum\limits_{j=1}^{n}aq^{j-1}=\dfrac{a(q^n-1)}{q-1}$, $q=\exp\left(\dfrac{m\pi\mathrm{i}}{n}\right)$, и окончательно для $m=1,\dots,2n-1$ получаем

$$S_n(T_m)=\frac{\pi}{2n}\,\frac{\exp\left(\frac{m(2n+1)\pi\mathrm{i}}{2n}\right)-\exp\left(\frac{m(1-2n)\pi\mathrm{i}}{2n}\right)}{\exp\left(\frac{m\pi\mathrm{i}}{n}\right)-1}=0.\qquad\rhd$$

4.28. Показать, что квадратурная формула

$$S_n(f) = \frac{\pi}{n+1} \sum_{j=1}^{n} \sin^2 \left(\frac{\pi j}{n+1}\right) f\left(\cos \frac{\pi j}{n+1}\right)$$

для вычисления интегралов $I(f) = \int\limits_{-1}^{1} f(x)\sqrt{1-x^2}\,dx$ точна для всех алгебраических многочленов степени $2n-1$.

4.29. Показать, что квадратурная формула

$$S_n(f) = \frac{\omega}{n} \sum_{j=0}^{n-1} f\left(\frac{j\omega}{n}\right)$$

для вычисления интегралов $I(f) = \int\limits_{0}^{\omega} f(x)\,dx$ точна для всех тригонометрических многочленов с периодом ω степени не выше $n-1$.

$\lhd$ Рассмотрим величины $I(f)$ и $S_n(f)$ для функций вида $f(x) = \exp\left(2\pi m\mathrm{i}\,\frac{x}{\omega}\right)$, $m = 0, 1, \ldots, n$. При этом для интегралов имеем

$$I(f) = \begin{cases} \omega & \text{при} \quad m = 0, \\ 0 & \text{при} \quad m \neq 0. \end{cases}$$

Используя квадратурную формулу, получаем

$$S_n(f) = \frac{\omega}{n} \sum_{j=0}^{n-1} \exp\left(2\pi m\mathrm{i}\,\frac{j}{n}\right) =$$

$$= \begin{cases} \dfrac{\omega}{n} \displaystyle\sum_{j=0}^{n-1} 1 = \omega & \text{при} \quad \dfrac{m}{n} \text{ — целом}, \\ \dfrac{\exp(2\pi m\mathrm{i}) - 1}{\exp\left(\dfrac{2\pi m\mathrm{i}}{n}\right) - 1} = 0 & \text{при} \quad \dfrac{m}{n} \text{ — не целоем}. \end{cases}$$

Приведенное выражение означает, что квадратурная формула точна для всех $\sin\left(\frac{2\pi m x}{\omega}\right)$ и $\cos\left(\frac{2\pi m x}{\omega}\right)$, если $m = 0$ или $\frac{m}{n}$ не целое, т. е. точна для всех тригонометрических многочленов степени не выше $n-1$. Из явного выражения для $S_n(f)$ следует, что эта формула также точна для функции $\sin\left(\frac{2\pi n x}{\omega}\right)$. $\rhd$

4.30. Пусть T — треугольник на плоскости, $S(T)$ — его площадь, A, B, C — середины сторон. Показать, что квадратурная формула

$$I(f) = \iint\limits_{T} f(x)\,dx \approx \tfrac{1}{3}\, S(T)(f(A) + f(B) + f(C)),$$

где $x = (x_1, x_2)$, $dx = dx_1 dx_2$, точна для всех многочленов второй степени вида

$$a_0 + a_1 x_1 + a_2 x_2 + a_{11} x_1^2 + a_{12} x_1 x_2 + a_{22} x_2^2.$$

Указание. Линейным невырожденным преобразованием, якобиан которого постоянен и не равен нулю, произвольный треугольник перевести в равнобедренный прямоугольный, далее проверка утверждения становится простой.

4.31. Пусть P — прямоугольник на плоскости, $S(P)$ — его площадь, A, B, C, D — середины сторон, E — точка пересечения диагоналей. Показать, что квадратурная формула

$$I(f) = \iint\limits_P f(x)\,dx \approx \tfrac{1}{6}\,S(P)\,(f(A) + f(B) + f(C) + f(D) + 2f(E))$$

точна для всех алгебраических многочленов от двух переменных третьей степени.

Указание. Линейным невырожденным преобразованием, якобиан которого постоянен и не равен нулю, произвольный прямоугольник перевести в квадрат, симметричный относительно нуля.

4.32. Для вычисления интегралов $I(f)$:

1) $\int\limits_0^2 f(x)\,dx$; 2) $\int\limits_0^1 f(x)\,dx$; 3) $\int\limits_{-1}^0 f(x)\,dx$; 4) $\int\limits_{-2}^0 f(x)\,dx$

построить квадратурную формулу Чебышёва с тремя узлами.

Ответ: 1) $P_3(x) = x^3 - 3x^2 + \frac{5}{2}\,x - \frac{1}{2}\,, x_1 = 1, x_{2,3} = 1 \pm \frac{1}{\sqrt{2}}\,,\ c = \frac{2}{3}$; 2) $P_3(x) = x^3 - \frac{3}{2}\,x^2 + \frac{5}{8}\,x - \frac{1}{16}\,,\ x_1 = \frac{1}{2}\,,\ x_{2,3} = \frac{1}{2} \pm \frac{1}{2\sqrt{2}}\,,\ c = \frac{1}{3}$; 3) $x_1 = -\frac{1}{2}\,,$ $x_{2,3} = -\frac{1}{2} \pm \frac{1}{2\sqrt{2}}\,,\ c = \frac{1}{3}$; 4) $x_1 = -1,\ x_{2,3} = -1 \pm \frac{1}{\sqrt{2}}\,,\ c = \frac{2}{3}$.

4.33. Построить квадратурную формулу вида $S(f) = c_1 f(0) + c_2 f(x_2)$, точную для многочленов максимально высокой степени для вычисления интегралов $I(f)$:

1) $\int\limits_{-2}^0 x^2 f(x)dx$; 2) $\int\limits_0^1 x f(x)dx$; 3) $\int\limits_0^{\pi/2} \cos(x) f(x)dx$; 4) $\int\limits_0^2 (x+2) f(x)dx$.

Ответ: 1) $S_2(f) = \frac{1}{6}\,f(0) + \frac{5}{2}\,f\left(-\frac{8}{5}\right)$; 2) $S_2(f) = \frac{1}{18}\,f(0) + \frac{4}{9}\,f\left(\frac{3}{4}\right)$; 3) $S_2(f) = \frac{4(\pi-3)}{\pi^2-8}\,f(0) + \frac{(\pi-2)^2}{\pi^2-8}\,f\left(\frac{\pi^2-8}{2(\pi-2)}\right)$; 4) $S_2(f) = \frac{26}{21}\,f(0) +$ $+ \frac{100}{21}\,f\left(\frac{7}{5}\right)$.

4.34. Определить параметры c_1, c_2, x_2 так, чтобы квадратурная формула $S(f) = c_1 f(a) + c_2 f(x_2)$ для вычисления интегралов $\int\limits_a^b f(x)dx$ была точной на многочленах максимально высокой степени.

4.35. Определить параметры c_1, c_2, c_3, x_2 так, чтобы квадратурная формула $S(f) = c_1 f(-1) + c_2 f(x_2) + c_3 f(1)$ для вычисления интегралов $I(f) = \int\limits_{-1}^{1} x^2 f(x)dx$ была точной на многочленах максимально высокой степени.

4.36. Для вычисления интегралов $I(f) = \int\limits_{0}^{1} f(x)dx$ построить квадратурную формулу $S_2(f) = c_1 f(0) + c_2 f\left(\frac{2}{3}\right)$, точную для многочленов максимально высокой степени.

4.37. Для вычисления интегралов $I(f) = \int\limits_{0}^{1} f(x)dx$ построить квадратурную формулу $S_2(f) = c_1 f\left(\frac{1}{2}\right) + c_2 f\left(\frac{2}{3}\right)$, точную для многочленов максимально высокой степени.

4.38. Для вычисления интегралов $I(f) = \int\limits_{0}^{2} f(x)dx$ построить квадратурную формулу $S_3(f) = c_1 f(0) + c_2 f\left(\frac{1}{2}\right) + c_3 f(2)$, точную для многочленов максимально высокой степени.

4.39. Для вычисления интегралов $I(f) = \int\limits_{a}^{b} \mathrm{e}^{\alpha x} f(x)dx$ построить квадратурную формулу $S_2(f) = c_1 f(a) + c_2 f(b)$, точную для многочленов максимально высокой степени.

Указание. Получить систему уравнений для коэффициентов квадратурной формулы

$$c_1 + c_2 = \frac{1}{\alpha}\left(\mathrm{e}^{\alpha b} - \mathrm{e}^{\alpha a}\right), \quad c_1 a + c_2 b = \frac{1}{\alpha}\left[\mathrm{e}^{\alpha b}\left(b - \frac{1}{\alpha}\right) - \mathrm{e}^{\alpha a}\left(a - \frac{1}{\alpha}\right)\right].$$

4.40. Для вычисления интегралов $I(f) = \int\limits_{a}^{b} f(x)dx$ построить квадратурную формулу

$$S_4(f) = c_1 f(a) + c_2 f\left(a + \frac{b-a}{3}\right) + c_3 f\left(a + 2\,\frac{b-a}{3}\right) + c_4 f(b),$$

точную для многочленов максимально высокой степени.

4.41. Пусть $f \in C^{(1)}[-1,1]$ и $P_5(x)$ — алгебраический многочлен пятой степени, удовлетворяющий условиям $P(x_k) = f(x_k)$, $P'(x_k) = f'(x_k)$, $k = 1,2,3$, где $x_1 = -1$, $x_2 = 0$, $x_3 = 1$. Рассмотрим квадратурную формулу

$$S_5(f) = \frac{1}{15}\left(7f(-1) + 16f(0) + 7f(1) + f'(-1) - f'(1)\right).$$

Проверить, что $S_5(f)$ точна на многочленах пятой степени $\int\limits_{-1}^{1} P_5(x)dx = S_5(P_5)$, но найдется многочлен степени 6, на котором она не точна.

4.3. Квадратурные формулы Гаусса

Рассмотрим следующую задачу: при заданном числе узлов n построить для вычисления интегралов вида $I(f) = \int\limits_a^b p(x)f(x)\,dx$ квадратурную формулу

$$S_n(f) = \sum_{i=1}^{n} c_i f(x_i), \tag{4.1}$$

точную для многочленов максимально высокой степени. Весовая функция $p(x)$ предполагается почти всюду положительной.

В этой постановке имеется $2n$ свободных параметров (узлы x_i и коэффициенты c_i неизвестны), поэтому можно попытаться построить квадратуру, точную для многочленов степени $2n-1$. Несложно убедиться в том, что не существует квадратуры с n узлами, точной для всех многочленов степени $2n$. Действительно, возьмем $P_{2n}(x) = (x-x_1)^2\cdots(x-x_n)^2$. Тогда $0 = S_n(P_{2n}) \neq I(P_{2n}) > 0$.

Важную роль при построении *квадратурных формул Гаусса* (4.1) играют ортогональные многочлены на отрезке $[a,b]$ с весом $p(x) > 0$ почти всюду. Они могут быть получены, например, в результате стандартной процедуры ортогонализации, примененной к системе $\{1, x, \dots, x^k, \dots\}$, при скалярном произведении

$$(f,g) = \int\limits_a^b p(x)f(x)g(x)\,dx\,.$$

Пусть на отрезке $[a,b]$ имеется система ортогональных многочленов с весом $p(x)$

$$1, \psi_1(x), \psi_2(x), \dots, \psi_k(x), \dots\,.$$

Тогда многочлен k-й степени $\psi_k(x)$ ортогонален произвольному многочлену $P_l(x)$ при $l = 0, \dots, k-1$. Действительно, многочлен $P_l(x)$ представим в виде $P_l(x) = \sum\limits_{j=0}^{l} c_j\psi_j(x)$, и при $k \neq l$ имеют место равенства

$$\int\limits_a^b p(x)\psi_k(x)\psi_l(x)\,dx = 0\,.$$

На практике наиболее употребительны следующие ортогональные многочлены:

Лежандра ($[-1,1],\ p(x) \equiv 1$),

Чебышёва первого рода $\left([-1,1],\ p(x)=\dfrac{1}{\sqrt{1-x^2}}\right)$,

Лагерра $([0,\infty),\ p(x)=\mathrm{e}^{-x})$,

Эрмита $\left((-\infty,\infty),\ p(x)=\mathrm{e}^{-x^2}\right)$.

Здесь в скобках указаны промежуток интегрирования и весовая функция.

При построении квадратурных формул Гаусса базовым является следующее утверждение:

Теорема. *Пусть $x_1,\dots,x_n$ — нули ортогонального на $[a,b]$ с весом $p(x)$ многочлена $\psi_n(x)$ степени n и (4.1) — квадратура, точная для многочленов степени $n-1$. Тогда квадратура (4.1) точна для многочленов степени $2n-1$.*

На основании этого утверждения процесс построения квадратуры может быть разбит на два последовательных этапа: — нахождение нулей ортогонального многочлена; — нахождение весов методом неопределенных коэффициентов.

Приведем оценку погрешности формул Гаусса

$$R_n=\left\|f^{(2n)}(x)\right\|\int\limits_a^b p(x)\,\frac{\psi_n^2(x)}{(2n)!}\,dx\,,$$

которая для отрезка $[-1,1]$ и веса $p(x)\equiv 1$ имеет вид

$$R_n=\left\|f^{(2n)}(x)\right\|\frac{2^{2n+1}(n!)^4}{((2n)!)^3(2n+1)}\ .$$

4.42. Методом ортогонализации построить несколько первых многочленов Лежандра со старшим коэффициентом 1, ортогональных на отрезке $[-1,1]$ с весом $p(x)\equiv 1$.

Ответ: $\psi_0=1,\quad \psi_1=x,\quad \psi_2=x^2-\frac{1}{3}\,,\quad \psi_3=x^3-\frac{3}{5}\,x,\ \dots$.

4.43. Доказать, что ортогональный многочлен степени n имеет ровно n различных корней на отрезке $[a,b]$.

$\lhd$ Если $\psi_n(x)$ имеет на $[a,b]$ только $r<n$ нулей нечетной кратности, то многочлен

$$Q_{n+r}(x)=\psi_n(x)\prod_{l=1}^{r}(x-x_l)$$

не меняет знака на этом отрезке. Следовательно, интеграл $\int\limits_a^b p(x)Q_{n+r}(x)\,dx$ отличен от нуля, что противоречит свойству ортогональности $\psi_n(x)$ любому многочлену низшей степени. $\rhd$

4.44. Построить квадратуру Гаусса с одним узлом для вычисления интеграла: 1) $I(f)=\int\limits_0^1 xf(x)\,dx$; 2) $I(f)=\int\limits_0^1 \mathrm{e}^x f(x)\,dx$.

Ответ: 1) $\frac{1}{2}\,f\left(\frac{2}{3}\right)$; 2) $(e-1)\,f\left(\frac{1}{e-1}\right)$.

4.45. Построить квадратуру Гаусса с двумя узлами для вычисления интеграла: 1) $I(f)=\int\limits_{-1}^1 x^2 f(x)\,dx$; 2) $I(f)=\int\limits_{-\pi/2}^{\pi/2} \cos(x) f(x)\,dx$.

Ответ: 1) $\frac{1}{3}\left(f\left(\sqrt{\frac{3}{5}}\right)+f\left(-\sqrt{\frac{3}{5}}\right)\right)$; 2) $f\left(\sqrt{\frac{\pi^2}{4}-2}\right)+f\left(-\sqrt{\frac{\pi^2}{4}-2}\right)$.

4.46. Построить квадратуру Гаусса с тремя узлами для вычисления интеграла $I(f)=\int\limits_{-1}^1 f(x)\,dx$.

Ответ: $\frac{5}{9}\,f\left(-\sqrt{\frac{3}{5}}\right)+\frac{8}{9}\,f(0)+\frac{5}{9}\,f\left(\sqrt{\frac{3}{5}}\right)$.

4.47. Доказать, что все коэффициенты квадратуры Гаусса положительны.

$\lhd$ Рассмотрим многочлен степени $k=2n-2$ вида $P_k(x)=\left(\prod\limits_{\substack{i=1\\ i\neq k}}^{n}(x-x_i)\right)^2$. Для интеграла от этого многочлена формула Гаусса дает точный результат

$$\int\limits_a^b p(x)P_k(x)\,dx=\sum_{j=1}^{n} c_j P_k(x_j)=\sum_{\substack{j=1\\ j\neq k}}^{n} c_j P_k(x_j)+c_k P_k(x_k)\,.$$

Так как справедливо $P_k(x_j)=0$ при $j\neq k$, то имеет место равенство

$$c_k=\frac{\int\limits_a^b p(x)P_k(x)\,dx}{P_k(x_k)}>0\,. \qquad \rhd$$

4.48. Пусть весовая функция $p(x)$ четная относительно середины отрезка интегрирования — точки $\frac{a+b}{2}$. Доказать, что узлы квадратуры Гаусса для вычисления интегралов $I(f)=\int\limits_a^b p(x)f(x)\,dx$ расположены симметрично относительно $\frac{a+b}{2}$, а соответствующие симметричным узлам коэффициенты квадратуры равны.

Ответ: симметрия узлов квадратуры следует из решения 4.70, а равенство коэффициентов — следствие симметрии узлов (см. 4.3).

4.49. Пусть $R_n(f)$ — погрешность для функции $f(x) = x^{2n}$ квадратурной формулы Гаусса с n узлами для отрезка $[-1, 1]$ и весовой функции $p(x) = \frac{1}{\sqrt{1-x^2}}$. Вычислить $R_n(f)$ и показать, что $\lim\limits_{n\to\infty} 2^{2n-1}|R_n(f)| = \pi$.

4.50. Пусть $R_n(f)$ — погрешность для функции $f(x) = x^{2n}$ квадратурной формулы Гаусса с n узлами для отрезка $[-1, 1]$ и весовой функции $p(x) = \sqrt{1-x^2}$. Вычислить $R_n(f)$ и показать, что $\lim\limits_{n\to\infty} 2^{2n}|R_n(f)| = \pi$.

4.51. Пусть $f(x)$ — непрерывная на отрезке $[a, b]$ функция. Доказать, что для формул Гаусса $|R_n(f)| \to 0$ при $n \to \infty$.

Указание. Квадратурная формула и вычисляемый по ней интеграл определяют линейные функционалы на пространстве непрерывных функций. Поэтому здесь применима теорема Банаха о сходимости последовательности линейных операторов (необходимым и достаточным условием сходимости является выполнение следующих двух требований: 1) сходимость на множестве элементов, всюду плотном в пространстве, где определены операторы; 2) ограниченность в совокупности норм операторов.)

Для квадратур Гаусса положительность коэффициентов гарантирует выполнение второго требования. Проверить, что оценка погрешности дает сходимость по n для произвольного алгебраического многочлена, откуда следует выполнение первого требования.

4.52. Построить составную квадратурную формулу Гаусса с двумя узлами на каждом отрезке разбиения для вычисления интеграла $I(f) = \int\limits_a^b \mathrm{e}^{\alpha x} f(x)dx$, где $\mathrm{e}^{\alpha x}$ — весовая функция. Оценить погрешность построенной формулы.

4.53. Доказать, что не существует квадратур $\int\limits_a^b f(x)dx \approx \sum\limits_{i=1}^n c_i f(x_i)$ с n узлами, точных для всех тригонометрических полиномов степени n с весовой функцией $p(x) \equiv 1$.

4.54. Построить квадратурную формулу Гаусса с одним узлом для вычисления интеграла $I(f)$: 1) $\int\limits_0^1 x^2 f(x)dx$; 2) $\int\limits_{-1}^1 |x| f(x)dx$.

4.55. Построить квадратурную формулу Гаусса с двумя узлами для вычисления интеграла $I(f)$: 1) $\int\limits_{-1}^1 |x| f(x)dx$; 2) $\int\limits_{-1}^1 x^4 f(x)dx$.

4.56. Построить квадратурную формулу Гаусса с двумя узлами для вычисления интеграла $I(f) = \int\limits_0^1 p(x) f(x)dx$, $p(x)$ — весовая функция:

1) $p(x) = x$; 2) $p(x) = \sin(\pi x)$; 3) $p(x) = \mathrm{e}^x$; 4) $p(x) = \cos\left(x - \frac{1}{2}\right)$; 5) $p(x) = 1 - x$; 6) $p(x) = \mathrm{e}^{-x}$.

4.57. Показать, что квадратурная формула

$$S_3(f) = \frac{\sqrt{\pi}}{6}\left(f\left(-\sqrt{\frac{3}{2}}\right) + 4f(0) + f\left(\sqrt{\frac{3}{2}}\right)\right)$$

для вычисления интегралов $I(f) = \int\limits_{-\infty}^{+\infty} \exp(-x^2) f(x)dx$ точна для всех алгебраических многочленов пятой степени.

4.58. Показать, что квадратурная формула

$$S_3(f) = \frac{\pi}{3}\left(f\left(-\frac{\sqrt{3}}{2}\right) + f(0) + f\left(\frac{\sqrt{3}}{2}\right)\right)$$

для вычисления интегралов $I(f) = \int\limits_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx$ точна для всех алгебраических многочленов пятой степени.

4.59. Построить квадратуру Гаусса с четырьмя узлами для вычисления интеграла $I(f) = \int\limits_{-1}^{1} f(x)\,dx$.

Ответ: $-x_{-1} = x_1 = \sqrt{\frac{15 - 2\sqrt{30}}{35}}$, $\quad c_{-1} = c_1 = \frac{18 + \sqrt{30}}{36}$,

$-x_{-2} = x_2 = \sqrt{\frac{15 + 2\sqrt{30}}{35}}$, $\quad c_{-2} = c_2 = \frac{18 - \sqrt{30}}{36}$.

4.60. На интервале $(-\infty, \infty)$ найти ортогональный многочлен $\psi_3(x) = x^3 + \ldots$ при заданной весовой функции $p(x) = \exp(-x^2)$.

Ответ: $\psi_3(x) = x^3 - \frac{3}{2}\,x$.

4.61. На отрезке $[-1, 1]$ найти ортогональный многочлен $\psi_3(x) = x^3 + \ldots$ при заданной весовой функции $p(x) = \frac{1}{\sqrt{1-x^2}}$.

Ответ: $\psi_3(x) = x^3 - \frac{3}{4}\,x$.

4.62. На отрезке $[-1, 1]$ найти ортогональный многочлен $\psi_3(x) = x^3 + \ldots$ при заданной весовой функции $p(x) = \sqrt{1-x^2}$.

Ответ: $\psi_3(x) = x^3 - \frac{1}{2}\,x$.

4.63. На полуинтервале $[0, \infty)$ найти ортогональный многочлен $\psi_3(x) = x^3 + \ldots$ при заданной весовой функции $p(x) = \exp(-x)$.

Ответ: $\psi_3(x) = x^3 - 9x^2 + 18x - 6$.

4.64. Построить квадратурную формулу Гаусса с двумя узлами для вычисления интегралов $I(f) = \int\limits_0^\pi \sin(x)\, f(x)\, dx$.

Ответ: $S_2(f) = f\left(\frac{\pi + \sqrt{\pi^2 - 8}}{2}\right) + f\left(\frac{\pi - \sqrt{\pi^2 - 8}}{2}\right)$.

4.65. Построить квадратурную формулу Гаусса с двумя узлами для вычисления интегралов $I(f) = \int\limits_0^\infty \exp(-x)\, f(x)\, dx$.

Ответ: $S_2(f) = \frac{2+\sqrt{2}}{4}\, f(2 - \sqrt{2}) + \frac{2-\sqrt{2}}{4}\, f(2 + \sqrt{2})$.

4.66. Построить квадратурную формулу Гаусса с двумя узлами для вычисления интегралов $I(f) = \int\limits_0^1 \left(x - \frac{1}{2}\right)^2 f(x)\, dx$.

Ответ: $S_2(f) = \frac{1}{24}\left[f\left(\frac{1}{2} + \sqrt{\frac{3}{20}}\right) + f\left(\frac{1}{2} - \sqrt{\frac{3}{20}}\right)\right]$.

4.67. Доказать, что ни с каким весом $p(x) > 0$ многочлены $\{x^m\}_{m=0}^\infty$ не могут быть ортогональны на $[0, 1]$.

4.68. Пусть на отрезке $[a, b]$ имеется система ортогональных многочленов $\{\psi_n(x)\}$ с весом $p(x)$ и старшим коэффициентом, равным единице. Доказать, что среди всех многочленов степени n вида $P_n(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_0$ минимальную норму $\|P_n\|^2 = \int\limits_a^b p(x) P_n^2(x)\, dx$ имеет ортогональный многочлен $\psi_n(x)$.

$\lhd$ Пусть $P_n(x)$ — произвольный многочлен степени n со старшим коэффициентом, равным единице. Тогда $P_n(x) = \psi_n(x) + r_{n-1}(x)$, и из ортогональности $\psi_n(x)$ любому многочлену низшей степени следует

$$\|P_n(x)\|^2 = \|\psi_n(x)\|^2 + \|r_{n-1}(x)\|^2 . \qquad \rhd$$

4.69. Для ортогональных многочленов вида $\psi_n(x) = x^n + \ldots$ показать справедливость рекуррентного соотношения

$$\psi_n(x) = (x - b_n)\psi_{n-1}(x) - c_n\psi_{n-2}(x)$$

с коэффициентами $b_n = \frac{(x\psi_{n-1}, \psi_{n-1})}{(\psi_{n-1}, \psi_{n-1})}$ и $c_n = \frac{(\psi_{n-1}, \psi_{n-1})}{(\psi_{n-2}, \psi_{n-2})} > 0$.

$\lhd$ Представим многочлен $x\psi_{n-1}$ в виде $\sum\limits_{k=0}^{n} \alpha_k \psi_k$, где коэффициенты α_j определяются из условий ортогональности

$$(x\psi_{n-1}, \psi_j) = \alpha_j(\psi_j, \psi_j),\ j = 0, \ldots, n .$$

При $j < n-2$ имеем

$$(x\psi_{n-1}, \psi_j) = (\psi_{n-1}, x\psi_j) = (\psi_{n-1}, Q_{j+1}(x)) = 0\,,$$

т. е. все $\alpha_j = 0$ при $j < n-2$ (здесь $Q_{j+1}(x) = x\psi_j$ — многочлен степени $j+1$). Таким образом,

$$x\psi_{n-1} = \alpha_n\psi_n + \alpha_{n-1}\psi_{n-1} + \alpha_{n-2}\psi_{n-2}\,,$$

при этом в силу равенства коэффициентов при старшей степени x, $\alpha_n = 1$. Отсюда следует, что

$$\psi_n(x) = (x - \alpha_{n-1})\psi_{n-1} - \alpha_{n-2}\psi_{n-2}, \quad b_n \equiv \alpha_{n-1}, \quad c_n \equiv \alpha_{n-2}\,.$$

Умножая скалярно равенство на ψ_{n-1}, получаем $b_n = \dfrac{(x\psi_{n-1}, \psi_{n-1})}{(\psi_{n-1}, \psi_{n-1})}$. Умножая скалярно равенство на ψ_{n-2}, с учетом $(x\psi_{n-1}, \psi_{n-2}) = (\psi_{n-1}, \psi_{n-1})$, находим $c_n = \dfrac{(\psi_{n-1}, \psi_{n-1})}{(\psi_{n-2}, \psi_{n-2})} > 0\,.$ $\triangleright$

4.70. Доказать, что ортогональные многочлены на симметричном относительно нуля отрезке с четным весом $p(x)$ обладают свойством $\psi_n(-x) = (-1)^n\psi_n(x)$.

Указание. $\psi_0(x) = 1$, $\psi_1(x) = x$. Продолжить решение по индукции, используя рекуррентное соотношение из 4.69 и доказав, что $b_n \equiv 0$.

4.71. Пусть задан отрезок $[a, b]$. Доказать, что при $b > a \geqslant 0$ все коэффициенты ортогонального многочлена отличны от нуля.

$\triangleleft$ Все корни x_k многочлена $\psi_n(x)$ положительны, а его коэффициенты выражаются через величины $B_j = \sum\limits_{k=1}^{n} x_k^j$ (см. 4.25). Доказательство также можно построить на основе теоремы Виета. $\triangleright$

4.72. Доказать, что нули ортогональных многочленов с фиксированным на отрезке $[a, b]$ весом $p(x) > 0$ перемежаются, т. е.

$$a < x_1^{(n)} < x_1^{(n-1)} < \ldots < x_{n-1}^{(n-1)} < x_n^{(n)} < b.$$

$\triangleleft$ Подставим $x = x_i^{(n)}$ в рекуррентное соотношение (см. 4.69)

$$\psi_{n+1} = (x - \alpha_n)\psi_n - \alpha_{n-1}\psi_{n-1}\,.$$

Учитывая, что здесь $\alpha_{n-1} > 0$, имеем

$$\psi_{n+1}\left(x_i^{(n)}\right) + \alpha_{n-1}\psi_{n-1}\left(x_i^{(n)}\right) = 0\,.$$

Пусть утверждение верно для некоторого n. Отсюда и из равенств

$$\operatorname{sign}\psi_{n-1}(b) = 1, \quad \operatorname{sign}\psi_{n-1}(a) = (-1)^{n-1}$$

следует, что

$$\psi_{n-1}\left(x_i^{(n)}\right) = (-1)^{n-i},$$

а знаки $\operatorname{sign}\psi_{n+1}\left(x_i^{(n)}\right) = -\operatorname{sign}\psi_{n-1}\left(x_i^{(n)}\right)$ противоположны. Так как

$$\operatorname{sign}\psi_{n+1}(b) = 1 \quad \text{и} \quad \operatorname{sign}\psi_{n+1}(a) = (-1)^{n+1},$$

то $\psi_{n+1}(x)$ имеет чередующиеся знаки в последовательно расположенных точках $a, x_1^{(n)}, \ldots, x_n^{(n)}, b$, что и завершает доказательство. $\triangleright$

4.73. Доказать, что для многочленов Лежандра

$$L_n(x) = \frac{1}{2^n n!}\,\frac{d^n}{dx^n}\left((x^2-1)^n\right)$$

справедливы следующие соотношения:

1) $L_n(x) = \dfrac{1}{n!}\,\dfrac{d^n}{dt^n}\left(\dfrac{1}{\sqrt{1-2tx+t^2}}\right)\Bigg|_{t=0}$, $n \geqslant 0$;

2) $(n+1)L_{n+1}(x) - (2n+1)xL_n(x) + nL_{n-1}(x) = 0$, $n \geqslant 1$;

3) $L'_{n+1}(x) - L'_{n-1}(x) = (2n+1)L_n(x)$, $n \geqslant 1$;

4) $L'_{n+1}(x) - xL'_n(x) = (n+1)L_n(x)$, $n \geqslant 0$;

5) $xL'_n(x) - L'_{n-1}(x) = nL_n(x)$, $n \geqslant 1$;

6) $(x^2-1)L'_n(x) = nxL_n(x) - nL_{n-1}(x)$, $n \geqslant 1$;

7) $(1-x^2)L''_n(x) - 2xL'_n(x) + n(n+1)L_n(x) = 0$, $n \geqslant 0$;

8) $\displaystyle\int_{-1}^{1} x^k L_n(x)dx = \begin{cases} 0, & \text{если } 0 \leqslant k \leqslant n-1, \\ \dfrac{2^{n+1}(n!)^2}{(2n+1)!}, & \text{если } k = n; \end{cases}$

9) $\displaystyle\int_{-1}^{1} L_k(x)L_m(x)dx = \begin{cases} 0, & \text{если } k \neq m, \\ \dfrac{2}{2k+1}, & \text{если } k = m; \end{cases}$

10) Если $L_n(x_k) = 0$, то $\displaystyle\int_{-1}^{1} \frac{L_n(x)}{x - x_k}\,dx = -\frac{2}{(n+1)L_{n+1}(x_k)}$, $n \geqslant 1$;

11) $\displaystyle L_n(x) = 2^{-n}\sum_{k=0}^{[n/2]} (-1)^k C_n^k C_{2n-2k}^n x^{n-2k}$, $n \geqslant 0$.

4.74. Пусть $x_1, x_2, \ldots, x_n$ — корни многочлена Лежандра $L_n(x)$ и $\gamma_k = \displaystyle\int_{-1}^{1} \prod_{\substack{j=1 \\ j\neq k}}^{n} \frac{x - x_j}{x_k - x_j}\,dx$. Доказать, что если $f(x), g(x)$ — алгебраические многочлены степени $n-1$, то $\displaystyle\int_{-1}^{1} f(x)g(x)dx = \sum_{k=1}^{n} \gamma_k f(x_k)g(x_k)$.

4.75. Доказать следующие свойства узлов и коэффициентов квадратурной формулы Гаусса $S_n(f)$ для вычисления интегралов $\displaystyle\int_{-1}^{1} f(x)dx$:

1) $L_n(x_k) = 0$, $k = 1, 2, \ldots, n$, где $L_n(x)$ — ортогональный многочлен Лежандра степени n;

2) $c_k = -\dfrac{2}{(n+1)L_{n+1}(x_k)L'_n(x_k)}$, $\quad k = 1, 2, \ldots, n$;

3) $c_k = \dfrac{2(1-x_k^2)}{n^2(L_{n-1}(x_k))^2}$, $\quad k = 1, 2, \ldots, n$;

4) $c_k = \dfrac{2}{nL_{n-1}(x_k)L'_n(x_k)}$, $\quad k = 1, 2, \ldots, n$.

4.76. Для вычисления интегралов $\int\limits_{-1}^{1} f(x)dx$ построить *квадратурную формулу Маркова—Радо*

$$S_n(f) = c_1 f(-1) + \sum_{i=2}^{n} c_i f(x_i),$$

точную для произвольного многочлена степени $2n-2$.

4.77. Для вычисления интегралов $\int\limits_{-1}^{1} f(x)dx$ построить *квадратурную формулу Маркова—Лобатто*

$$S_n(f) = c_1 f(-1) + \sum_{i=2}^{n-1} c_i f(x_i) + c_n f(1),$$

точную для произвольного многочлена степени $2n-3$.

Указание. Представить исходный интеграл в виде

$$I(f) = \int\limits_{-1}^{1} Q_1(x)dx + \int\limits_{-1}^{1} \frac{f(x) - Q_1(x)}{1-x^2}\, p(x)dx,$$

где

$$Q_1(x) = f(-1)\,\frac{1-x}{2} + f(1)\,\frac{1+x}{2},\ p(x) = 1 - x^2.$$

Построить квадратурную формулу Гаусса с $n-2$ узлами, соответствующую весовой функции $p(x)$:

$$\int\limits_{-1}^{1} q(x)p(x)dx \approx \sum_{j=2}^{n-1} d_j q(x_j).$$

Показать, что квадратурная формула

$$\int\limits_{-1}^{1} f(x)dx \approx \int\limits_{-1}^{1} Q_1(x)dx + \sum_{j=2}^{n-1} d_j\,\frac{f(x_j) - Q_1(x_j)}{1-x_j^2}$$

при $c_j = \dfrac{d_j}{1-x_j^2}$, $j = 2, \ldots, n-1$, является искомой.

4.4. Главный член погрешности

Будем считать промежуток $[a,b]$ конечным и предположим, что $f(x)$ имеет на $[a,b]$ непрерывные производные до порядка $m+s$. Для квадратурной формулы $S_n(f)=\sum\limits_{i=1}^{n} c_i f(x_i)$, имеющей алгебраический порядок точности $m-1$, справедливо равенство

$$I(f)=\int\limits_a^b p(x)\,f(x)\,dx = S_n(f)+R_n(f).$$

Используя формулу Тейлора для $f(a+(x-a))$ с остаточным членом в интегральной форме $\int\limits_a^x \frac{(x-t)^m}{m!}\,f^{(m+1)}(t)dt$, можно получить следующее представление погрешности $R_n(f)$:

$$R_n(f)=\int\limits_a^b f^{(m)}(t)K(t)dt.$$

Здесь ядро $K(t)$ имеет вид

$$K(t)=\int\limits_t^a p(x)\,\frac{(x-t)^{m-1}}{(m-1)!}\,dx-\sum_{i=1}^{n} c_i E(x_i-t)\,\frac{(x_i-t)^{m-1}}{(m-1)!}\,,$$

где «гасящая» функция $E(x)$ определяется формулой

$$E(x)=\begin{cases}1 & \text{при} \quad x>0,\\ \frac{1}{2} & \text{при} \quad x=0,\\ 0 & \text{при} \quad x<0.\end{cases}$$

Имеет место представление Эйлера для погрешности:

$$R_n(f)\equiv R_m(f)=A_0\left[f^{(m-1)}(b)-f^{(m-1)}(a)\right]+\ldots$$
$$\cdots+A_{s-1}\left[f^{(m+s-2)}(b)-f^{(m+s-2)}(a)\right]+R_{m+s}(f),$$
$$A_j=\frac{1}{b-a}\int\limits_a^b L_j(t)dt,\; L_{j+1}(t)=\int\limits_a^t\left[A_j-L_j(x)\right]dx, L_0(t)=K(t),$$
$$R_{m+s}(f)=\int\limits_a^b f^{(m+s)}(t)L_s(t)dt.$$

Главным членом погрешности обычно называют первое слагаемое в этом представлении. Формула Эйлера позволяет с точностью до $O\left(h^{m+2}\right)$ определить значение главного члена погрешности.

Правило Рунге. Пусть на отрезке длины h для вычисления интеграла $I(f)$ используется некоторая квадратурная формула $S_h(f)$, имеющая

алгебраический порядок точности $m-1$. Разлагая $f(x)$ в ряд Тейлора в середине отрезка (точке c), получим

$$I(f) - S_h(f) = \alpha f^{(m)}(c) h^{m+1} + O\left(h^{m+2}\right).$$

Обозначим через $S_{h/2}(f)$ составную формулу, полученную с помощью формулы $S_h(f)$ для двух половинок отрезка длины h. Тогда при том же α находим

$$I(f) - S_{h/2}(f) = \alpha f^{(m)}(c)\, \frac{h^{m+1}}{2^m} + O\left(h^{m+2}\right).$$

Следовательно, с точностью до членов $O\left(h^{m+2}\right)$ справедливо следующее *правило Рунге*:

$$I(f) - S_{h/2}(f) \approx \frac{S_{h/2}(f) - S_h(f)}{2^m - 1}.$$

4.78. Пусть интеграл $I(f) = \int\limits_a^b f(x)\,dx$, где $f(x)$ — гладкая функция, вычисляют по составной формуле трапеций $S_2^N(f)$ с постоянным шагом $h = \frac{b-a}{N}$.

1) Показать, что суммарная погрешность удовлетворяет соотношению

$$R_2^N = a_1 h^2 + a_2 h^4 + a_3 h^6 + \ldots .$$

2) Показать, что

$$R_2^N(f) = I(f) - S_2^N(f) = -\frac{h^2}{12} \int\limits_a^b f''(x)dx + Z(f),\ Z(f) = o(h^2).$$

3) Пусть $|f^{(3)}(x)| \leqslant M_3$ на отрезке $[a,b]$. Показать, что $|Z(f)| \leqslant c_3 M_3 (b-a) h^3$.

4) Пусть $|f^{(4)}(x)| \leqslant M_4$ на отрезке $[a,b]$. Показать, что $|Z(f)| \leqslant c_4 M_4 (b-a) h^4$.

Указание. Пусть $[x_i, x_{i+1}]$ — один из подотрезков длины h, на которые разбит отрезок $[a,b]$, и пусть $\bar{x} = \frac{x_i + x_{i+1}}{2}$. Используя тейлоровское разложение подынтегральной функции в точке $\bar{x}$, получить следующие представления:

$$\int\limits_{x_i}^{x_{i+1}} f(x)\,dx = hf(\bar{x}) + \frac{h^3}{24}\, f''(\bar{x}) + \frac{h^5}{1920}\, f^{(4)}(\bar{x}) + \ldots ,$$

$$\int\limits_{x_i}^{x_{i+1}} f(x)\,dx = \frac{f(x_i) + f(x_{i+1})}{2}\, h - \frac{h^3}{12}\, f''(\bar{x}) - \frac{h^5}{480}\, f^{(4)}(\bar{x}) - \ldots .$$

4.79. Пусть $I(f) = \int\limits_0^1 f(x)dx$ вычисляют по составной формуле трапеций с переменным шагом интегрирования: $x_i = \varphi(ih)$, $\varphi(t)$ — гладкая функция, $\varphi(0) = 0$, $\varphi(1) = 1$. Доказать, что главный член погрешности есть

$$-\frac{h^2}{12}\int\limits_0^1 f''(\varphi(t))(\varphi'(t))^3 dt.$$

Указание. Применить 4.78, учитывая справедливость равенства $x_{i+1} - x_i = h\,\varphi'(ih) + o(1)$.

4.80. Пусть интеграл $I(f) = \int\limits_a^b f(x)\,dx$, где $f(x)$ — гладкая функция, вычисляют по составной формуле Симпсона $S_3^N(f)$ с постоянным шагом $h = \frac{b-a}{N}$. Показать, что для составной формулы Симпсона суммарная погрешность удовлетворяет соотношению

$$R_3^N(f) = b_1 h^4 + b_2 h^6 + \ldots\,.$$

4.81. Пусть интеграл $I(f) = \int\limits_0^1 x^\lambda f(x)\,dx$, где $f(x)$ — гладкая функция и $f(0) \neq 0$, вычисляют по составной формуле трапеций с постоянным шагом $h = \frac{1}{N}$. Показать, что при $-1 < \lambda < 1$ суммарная погрешность удовлетворяет соотношению $R_2^N = a_1 h^{1+\lambda} + a_2 h^{2+\lambda} + \ldots$.

4.82. Используя значения S_h и $S_{h/2}$ квадратуры с главным членом погрешности ch^m, т. е. $I = S_h + ch^m$, построить квадратурную формулу более высокого порядка точности.

Ответ: $S_{h,h/2} = S_{h/2} + \frac{S_{h/2} - S_h}{2^m - 1}$.

4.83. Показать, что при применении правила Рунге к формуле трапеций получается формула Симпсона. Насколько при этом увеличится порядок главного члена погрешности?

Указание. В обозначениях 4.82 имеем $m = 2$, $S_{h,h/2} = S_{h/2} + \frac{1}{3}(S_{h/2} - S_h)$ при $S_h = \frac{b-a}{2}(f(a) + f(b))$, $b - a = h$. Порядок главного члена погрешности увеличится на 2.

4.84. Показать, что операция построения формулы

$$S_{h,h/2} = S_{h/2} + \frac{S_{h/2} - S_h}{2^m - 1}$$

является экстраполяционной, т. е. при $S_h \neq S_{h/2}$ величина $S_{h,h/2}$ всегда лежит вне отрезка с концами S_h и $S_{h/2}$.

◁ Действительно, если $S_{h/2} > S_h$, то $S_{h,h/2} > S_{h/2} > S_h$. Если $S_{h/2} < S_h$, то $S_{h,h/2} < S_{h/2} < S_h$. ▷

4.85. Пусть для вычисления интеграла I от некоторой функции используется квадратурная формула S_h, фактический порядок главного члена погрешности p которой неизвестен для данной функции. Предложить способ численной оценки значения порядка p.

$\lhd$ Возможен следующий способ (*процесс Эйткена*), являющийся обобщением правила Рунге. Пусть I — точное значение интеграла. Запишем его приближенные значения с шагами h, $\frac{h}{2}$ и $\frac{h}{4}$. Если учитывать только главный член погрешности, то получаем систему трех уравнений

$$I = S_h + ch^p, \quad I = S_{h/2} + \frac{1}{2^p}\, ch^p, \quad I = S_{h/4} + \frac{1}{4^p}\, ch^p,$$

в которой значения I, c и p неизвестны. Из первого и второго уравнений имеем $ch^p \left(1 - \frac{1}{2^p}\right) = S_{h/2} - S_h$. Из второго и третьего уравнений находим

$$\frac{1}{2^p}\, ch^p \left(1 - \frac{1}{2^p}\right) = S_{h/4} - S_{h/2}.$$

Из последних двух равенств получаем уравнение для определения p:

$$2^p = \frac{S_{h/2} - S_h}{S_{h/4} - S_{h/2}}\,.$$

Выражение для главного члена погрешности имеет вид

$$ch^p = \frac{\left(S_{h/2} - S_h\right)^2}{2S_{h/2} - S_h - S_{h/4}}\,. \qquad \rhd$$

Упражнения 4.86–4.88 иллюстрируют возможные обобщения правила Рунге.

4.86. Пусть задан некоторый метод вычисления интеграла с погрешностью $I(f) - S_h(f) = ch^m + O(h^{m+1})$ и вычислен интеграл с шагом h_1 и с шагом $h_2 = \frac{h_1}{\lambda}$. Показать, что

$$I(f) - S_{h_2}(f) \approx \frac{S_{h_2}(f) - S_{h_1}(f)}{\lambda^m - 1}\,.$$

Имеется в виду предельный переход при $h_2 \to 0, \lambda = \text{const} > 1$.

4.87. Пусть $I(f) - S_h(f) = ch^m + O(h^{m+1})$ и вычислен интеграл с шагом h_1 и с шагом $h_2 = \frac{h_1}{\lambda}$. Доказать, что

$$I(f) - S_{h_2}(f) \approx \frac{S_{h_2}(f) - S_{h_1}(f)}{\left(\frac{h_1}{h_2}\right)^m - 1}$$

при следующих условиях: $h_1 \to 0$, $\frac{\lambda - 1}{h_1} \to \infty$.

4.88. Пусть $I(f) - S_h(f) = ch^m + O(h^{m+2})$ и вычислен интеграл с шагом h_1 и с шагом $h_2 = \frac{h_1}{\lambda}$. Доказать, что

$$I(f) - S_{h_2}(f) \approx \frac{S_{h_2}(f) - S_{h_1}(f)}{\left(\frac{h_1}{h_2}\right)^m - 1}$$

при следующих условиях: $h_1 \to 0, h_1 > h_2$.

4.5. Функции с особенностями

Быстро осциллирующие функции. Пусть требуется вычислить интеграл $\int\limits_a^b \exp\{\mathrm{i}\omega x\} f(x)dx$, где $\omega(b-a) \gg 1$, $f(x)$ — гладкая функция. Функции $\mathrm{Re}\,(\exp\{\mathrm{i}\omega x\} f(x))$, $\mathrm{Im}\,(\exp\{\mathrm{i}\omega x\} f(x))$ имеют на рассматриваемом отрезке примерно $\omega\frac{b-a}{\pi}$ нулей. Многочлен степени n имеет не более n нулей на этом отрезке, поэтому такие функции могут быть хорошо приближены многочленами степени n лишь при $n \gg \omega\frac{b-a}{\pi}$. Следовательно, для непосредственного вычисления интегралов от таких функций потребуется применение квадратур, точных для многочленов очень высокой степени.

Более выгодным может оказаться использование $\exp\{\mathrm{i}\omega x\}$ в качестве весовой функции. Задавшись узлами интерполирования

$$x_j = \frac{b+a}{2} + \frac{b-a}{2} d_j\,, \quad j = 1, 2, \ldots, n\,,$$

построим многочлен Лагранжа $L_n(x)$ и рассмотрим квадратурную формулу

$$\begin{aligned} S_n^\omega(f) &= \int\limits_a^b \exp\{\mathrm{i}\omega x\} L_n(x)dx = \\ &= \frac{b-a}{2} \exp\left\{\mathrm{i}\omega\,\frac{a+b}{2}\right\} \sum_{j=1}^n D_j\left(\omega\,\frac{b-a}{2}\right) f(x_j), \end{aligned} \tag{4.2}$$

где

$$D_j(p) = \int\limits_{-1}^1 \left(\prod_{k\neq j} \frac{\xi - d_k}{d_j - d_k}\right) \exp\{\mathrm{i}p\xi\} d\xi,\ p = \omega\,\frac{b-a}{2}\,.$$

При этом оценка погрешности

$$R_n = D\,(d_1, \ldots, d_n) \max_{x\in[a,b]} \left|f^{(n)}(x)\right| \left(\frac{b-a}{2}\right)^{n+1}$$

не зависит от ω.

4.89. Для приближенного вычисления интегралов вида

$$I(f) = \int_0^1 \sin(100\pi x) f(x)\, dx$$

построить методом неопределенных коэффициентов квадратурную формулу с заданными узлами $S(f) = c_1\, f(0) + c_2\, f(1)$, точную для многочленов наиболее высокой степени.

Ответ: $c_1 = -c_2 = \frac{1}{100\pi}$.

4.90. Для приближенного вычисления интегралов от быстро осциллирующих функций вида $I(f) = \int_0^1 \cos(10^4\pi x)\, f(x)\, dx$ построить методом неопределенных коэффициентов квадратурную формулу с заданными узлами $S(f) = c_1\, f(0) + c_2\, f(1)$, точную для многочленов наиболее высокой степени.

Ответ: $c_1 = c_2 = 0$.

4.91. Построить формулу вида (4.2) для $n = 2,\ d_1 = -1,\ d_2 = 1$.

Ответ: $p = \omega\frac{b-a}{2}$,

$$D_1(p) = \int_{-1}^1 \frac{1-\xi}{2} \exp\{\mathrm{i}p\xi\} d\xi = \frac{\sin p}{p} + \frac{p\cos p - \sin p}{p^2}\,\mathrm{i},$$

$$D_2(p) = \int_{-1}^1 \frac{1+\xi}{2} \exp\{\mathrm{i}p\xi\} d\xi = \frac{\sin p}{p} - \frac{p\cos p - \sin p}{p^2}\,\mathrm{i}.$$

4.92. Показать, что при малых ω формулы, полученные в 4.91, могут иметь большую вычислительную погрешность.

Указание. При малых ω величина p мала. Функции $\cos p$ и $\sin p$ вычисляются с погрешностями $O(2^{-t})$ и $O(p2^{-t})$ соответственно, где t — длина мантиссы. Как следствие коэффициенты $D_1(p)$ и $D_2(p)$ из 4.91 приобретают погрешность $\frac{O(2^{-t})}{p}$.

4.93. Построить формулу вида (4.2) для $n = 3,\ d_1 = -1,\ d_2 = 0,\ d_3 = 1$ (формула Филона).

4.94. Построить формулу вида (4.2) для $n = 5,\ d_1 = -1,\ d_2 = -0.5,\ d_3 = 0,\ d_4 = 0.5,\ d_5 = 1$.

Вычисление интегралов от функций с особенностями. Значительная часть реально встречающихся подынтегральных функций — это функции с особенностями, причем особенность может содержаться либо в самой функции, либо в ее производных. Если нерегулярность функции не вызвана колебательным характером ее поведения, то для вычисления

больших серий интегралов такого типа используют специальные приемы: выделение особенности в весовую функцию, разбиение интеграла на части, аддитивное представление подынтегральной функции, замену переменных и т. д.

4.95. Пусть вычисляется интеграл $I = \int\limits_0^1 f(x)dx$, причем $f(x)$ может быть представлена в виде $f(x) = g(x)x^\alpha$, где $\alpha \in (0,1)$, $g(x)$ — гладкая функция, $g(0) \neq 0$. Построить квадратурную формулу $S(g) = \sum\limits_{j=0}^{M} D_j g(jh)$ с оценкой погрешности вида $\text{const} \cdot \max\limits_{x\in[0,1]} |g''(x)| \cdot M^{-2}$.

Указание. Выделить функцию x^α в качестве весовой, а $g(x)$ на каждом отрезке разбиения заменить многочленом Лагранжа первой степени.

4.96. Пусть вычисляется интеграл $\int\limits_0^1 \frac{f(x)\lambda}{\lambda^2 + x^2}\, dx$, $f \in C^{(2)}[0,1]$, $|\lambda| \ll 1$. Показать, что при использовании составной формулы трапеций с постоянным шагом $h = \frac{1}{M}$ суммарная погрешность оценивается величиной $\text{const} \cdot \min\left(\frac{h}{\lambda}\,,\, \frac{h^2}{\lambda^2}\right)$.

4.97. Для вычисления интеграла $\int\limits_0^1 \frac{f(x)\lambda}{\lambda^2 + x^2}\, dx$, $f \in C^{(1)}[0,1]$, $|\lambda| \ll 1$, используется следующая квадратурная формула с постоянным шагом $h = \frac{1}{M}$:

$$S(f) = \sum_{j=1}^{M} f(\xi_j) \left[\operatorname{arctg}\left(\frac{jh}{\lambda}\right) - \operatorname{arctg}\left(\frac{(j-1)h}{\lambda}\right)\right],$$

где $(j-1)h \leqslant \xi_j \leqslant jh$. Получить оценку погрешности вида $|R^M| \leqslant \text{const} \cdot \max\limits_{x\in[0,1]} |f'(x)| \cdot M^{-1}$.

4.98. Предложить способ вычисления интеграла $\int\limits_0^1 \frac{\ln x}{1+x^2}\, dx$ по составной квадратурной формуле с постоянным шагом h, чтобы погрешность имела порядок $O\left(h^2\right)$.

Указание. Представить подынтегральную функцию в виде $f(x) = G(x) + g(x)$, где $G(x) = \ln x$, $g(x) = -\frac{x^2 \ln x}{1+x^2}$, вычислить $\int\limits_0^1 G(x)dx$ в явном виде.

4.99. Предложить способ вычисления интеграла $\int\limits_0^1 \frac{\ln x}{1+x^2}\,dx$ по составной квадратурной формуле с постоянным шагом h, чтобы погрешность имела порядок $O\left(h^4\right)$.

Указание. См. указание к 4.98 при $G(x) = (1-x^2)\ln x$.

4.100. Пусть $f(x)$ — достаточно гладкая функция. Предложить квадратурную формулу для вычисления интеграла $\int\limits_0^1 f(x)x^{-\alpha}\sin(\omega x)dx$, где $\alpha > 1,\ \omega \gg 1,\ f(0) \neq 0$.

Указание. Разбить отрезок интегрирования на $[0,\varepsilon]$ и $[\varepsilon,1]$ с $\varepsilon \approx \frac{1}{\omega}$. На первом отрезке $\sin(\omega x)$ не является осциллирующей, поэтому в качестве весовой функции можно взять $x^{-\alpha}$, а на втором отрезке использовать неравномерные узлы и весовую функцию $\sin(\omega x)$.

4.101. Построить квадратурную формулу для вычисления с точностью $\varepsilon \leqslant 10^{-4}$ интеграла $\int\limits_1^\infty \frac{f(x)}{1+x^2}\,dx$, если для некоторого фиксированного $k \geqslant 1$ справедливо неравенство $|f^{(k)}(x)| \leqslant A_k$.

4.102. Построить квадратурную формулу для вычисления с точностью $\varepsilon \leqslant 10^{-4}$ интеграла $\int\limits_0^\infty f(x)\mathrm{e}^{-x}dx$, если для некоторого фиксированного $k \geqslant 1$ справедливо неравенство $|f^{(k)}(x)| \leqslant A_k$.

4.103. Построить квадратурную формулу (не проводя замену переменных) для вычисления с точностью $\varepsilon \leqslant 10^{-3}$ интеграла $\int\limits_0^1 \frac{f(x)}{\sqrt{x}}\,dx$, если для некоторого фиксированного $k \geqslant 1$ справедливо неравенство $|f^{(k)}(x)| \leqslant A_k$.

4.104. Построить квадратурную формулу для вычисления с точностью $\varepsilon \leqslant 10^{-4}$ интеграла $\int\limits_0^1 f(x)\,\frac{\sqrt{\sin x}}{x}\,dx$, если для некоторого фиксированного $k \geqslant 1$ справедливо неравенство $|f^{(k)}(x)| \leqslant A_k$.

Глава 5

Матричные вычисления

Значительная часть задач вычислительной математики может быть сформулирована в терминах матричного анализа, который необходим при исследовании вопросов корректности, устойчивости и сходимости различных методов. Алгоритмы решения систем линейных алгебраических уравнений — важная часть методов решения уравнений в частных производных.

Глава посвящена вопросам теории устойчивости для матричных задач, приведены наиболее известные прямые и итерационные алгоритмы решения систем линейных алгебраических уравнений, подробно разобраны различные способы их построения. Отдельно рассмотрена задача наименьших квадратов для переопределенных системы уравнений. Приведены алгоритмы для решения проблемы собственных значений.

5.1. Векторные и матричные нормы

Нормой вектора $\mathbf{x} = (x_1, \ldots, x_n)^T$ называется функционал, обозначаемый $||\mathbf{x}||$ и удовлетворяющий условиям:

$$||\mathbf{x}|| > 0,\ \mathbf{x} \neq 0,\ ||\mathbf{0}|| = 0;$$
$$||\alpha\mathbf{x}|| = |\alpha|\,||\mathbf{x}||;$$
$$||\mathbf{x} + \mathbf{y}|| \leqslant ||\mathbf{x}|| + ||\mathbf{y}||.$$

Наиболее употребительны следующие нормы:

$$||\mathbf{x}||_\infty = \max_{1\leqslant i\leqslant n} |x_i|, \quad ||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|, \quad ||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{(\mathbf{x}, \mathbf{x})}.$$

Нормы $||\cdot||_{\mathrm{I}}$ и $||\cdot||_{\mathrm{II}}$ называются *эквивалентными*, если для всех $\mathbf{x} \in \mathbf{R}^n$ справедливы неравенства с одними и теми же положительными постоянными c_1 и c_2:

$$c_1\,||\mathbf{x}||_{\mathrm{II}} \leqslant ||\mathbf{x}||_{\mathrm{I}} \leqslant c_2\,||\mathbf{x}||_{\mathrm{II}}.$$

Нормой матрицы A называется функционал, обозначаемый $||A||$ и удовлетворяющий условиям:

$$||A|| > 0,\ A \neq 0,\ ||0|| = 0;$$
$$||\alpha\,A|| = |\alpha|\,||A||;$$
$$||A + B|| \leqslant ||A|| + ||B||;$$
$$||A\,C|| \leqslant ||A||\,||C||.$$

Пусть задана некоторая векторная норма $||\cdot||_v$. Тогда матричную норму можно определить как операторную

$$||A||_v = \sup_{||\mathbf{x}||_v \neq 0} \frac{||A\,\mathbf{x}||_v}{||\mathbf{x}||_v} = \sup_{||\mathbf{x}||_v = 1} ||A\,\mathbf{x}||_v.$$

В этом случае матричная норма называется *подчиненной* соответствующей векторной норме $\|\cdot\|_v$.

Квадратную матрицу A размерности $n\times n$ с элементами $a_{ij}=\delta_i^j$ будем обозначать I и называть *единичной матрицей*. Если не оговаривается противное, в задачах и упражнениях имеются в виду матричные нормы, для которых справедливо $\|I\|=1$. В частности, указанное равенство выполнено для любой подчиненной нормы.

5.1. Является ли выражение $\min(|x_1|+2|x_2|, 2|x_1|+|x_2|)$ нормой вектора $\mathbf{x}$ в $\mathbf{R}^2$?

О т в е т: нет, поскольку неравенство треугольника не выполнено, например для векторов $(1,0)^T$ и $(0,1)^T$.

5.2. Является ли выражение $\max\limits_{t\in[0,1]}\left|\sum\limits_{k=1}^{n} x_k\, t^{k-1}\right|$ нормой вектора $\mathbf{x}$ в $\mathbf{R}^n$?

О т в е т: да.

5.3. Найти константы эквивалентности, связывающие нормы $||\mathbf{x}||_\infty$, $||\mathbf{x}||_1$, $||\mathbf{x}||_2$, а также векторы, на которых они достигаются.

$\lhd$ Из неравенств $\max\limits_{1\leqslant i\leqslant n}|x_i|\leqslant\sum\limits_{i=1}^{n}|x_i|\leqslant n\max\limits_{1\leqslant i\leqslant n}|x_i|$ следует

$$||\mathbf{x}||_\infty\leqslant||\mathbf{x}||_1\leqslant n\,||\mathbf{x}||_\infty.$$

Так как $\sum\limits_{i=1}^{n}x_i^2\leqslant\left(\sum\limits_{i=1}^{n}|x_i|\right)^2$, то $||\mathbf{x}||_2\leqslant||\mathbf{x}||_1$. Из неравенства Коши—Буняковского имеем

$$\sum_{i=1}^{n}|x_i|\leqslant\left(\sum_{i=1}^{n}1\right)^{1/2}\left(\sum_{i=1}^{n}x_i^2\right)^{1/2}=n^{1/2}\left(\sum_{i=1}^{n}x_i^2\right)^{1/2}.$$

Следовательно,

$$n^{-1/2}||\mathbf{x}||_1\leqslant||\mathbf{x}||_2\leqslant||\mathbf{x}||_1.$$

Из неравенств $\max\limits_{1\leqslant i\leqslant n}x_i^2\leqslant\sum\limits_{i=1}^{n}x_i^2\leqslant n\max\limits_{1\leqslant i\leqslant n}x_i^2$ имеем

$$||\mathbf{x}||_\infty\leqslant||\mathbf{x}||_2\leqslant n^{1/2}\,||\mathbf{x}||_\infty.$$

Легко заметить, что в полученных неравенствах константы эквивалентности достигаются на векторах либо с равными компонентами, либо с единственной ненулевой компонентой. $\rhd$

5.4. Пусть B — симметричная положительно определенная матрица.
1) Доказать, что величину $\sqrt{(B\mathbf{x},\mathbf{x})}$ можно принять за норму вектора $\mathbf{x}$;
2) найти константы эквивалентности, связывающие эту норму с нормой $||\mathbf{x}||_2$.

$\lhd$ 1) Для доказательства того, что соответствующее выражение определяет норму, достаточно проверить неравенство треугольника. 2) Найдем константы эквивалентности. Так как $B = B^T > 0$, то собственные векторы матрицы различны и ортогональны. Пусть $\mathbf{e}_1, \ldots, \mathbf{e}_n$ — ортонормированная система собственных векторов матрицы B (т. е. $(\mathbf{e}_i, \mathbf{e}_j) = \delta_i^j$), а $\lambda_1, \ldots, \lambda_n$ — соответствующие собственные значения. Любой вектор $\mathbf{x}$ представим в виде $\mathbf{x} = \sum\limits_{i=1}^{n} c_i \mathbf{e}_i$. Поэтому

$$(B\mathbf{x}, \mathbf{x}) = \left(\sum_{i=1}^{n} \lambda_i c_i \mathbf{e}_i, \sum_{i=1}^{n} c_i \mathbf{e}_i \right) = \sum_{i=1}^{n} \lambda_i c_i^2.$$

Отсюда для произвольного вектора $\mathbf{x}$ имеем

$$\min_i \lambda_i (\mathbf{x}, \mathbf{x}) \leqslant (B\mathbf{x}, \mathbf{x}) \leqslant \max_i \lambda_i (\mathbf{x}, \mathbf{x}), \quad (\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} c_i^2.$$

Так как все $\lambda_i(B) > 0$, то полученное неравенство означает эквивалентность евклидовой норме $||\mathbf{x}||_2$ с постоянными

$$\tilde{c}_1 = \sqrt{\min_i \lambda_i}, \quad \tilde{c}_2 = \sqrt{\max_i \lambda_i}. \qquad \rhd$$

5.5. Найти матричные нормы, подчиненные векторным нормам $||\cdot||_\infty$, $||\cdot||_1$ и $||\cdot||_2$.

$\lhd$ Получим оценку сверху для величины $||A\mathbf{x}||_\infty$:

$$||A\mathbf{x}||_\infty = \max_i \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leqslant \max_i \left(\sum_{j=1}^{n} |a_{ij}| \max_j |x_j| \right) \leqslant$$

$$\leqslant \max_i \left(\sum_{j=1}^{n} |a_{ij}| \right) ||\mathbf{x}||_\infty .$$

Покажем, что эта оценка достигается. Пусть максимум по i имеет место при $i = l$; тогда возьмем $\mathbf{x} = (\operatorname{sign}(a_{l1}), \operatorname{sign}(a_{l2}), \ldots, \operatorname{sign}(a_{ln}))^T$. Имеем $||\mathbf{x}||_\infty = 1$ и точные равенства во всей цепочке выше. Таким образом, $\|A\|_\infty = \max\limits_i \left(\sum\limits_{j=1}^{n} |a_{ij}| \right)$. Аналогично показывается, что $\|A\|_1 = \max\limits_j \left(\sum\limits_{i=1}^{n} |a_{ij}| \right)$.

По определению матричной нормы, подчиненной евклидовой векторной норме,

$$||A||_2 = \sup_{\mathbf{x} \neq 0} \frac{||A\mathbf{x}||_2}{||\mathbf{x}||_2} = \sup_{\mathbf{x} \neq 0} \sqrt{\frac{(A\mathbf{x}, A\mathbf{x})}{(\mathbf{x}, \mathbf{x})}} = \sup_{\mathbf{x} \neq 0} \sqrt{\frac{(A^T A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}}.$$

Имеем, что $(A^T A)^T = A^T (A^T)^T = A^T A$, т. е. матрица $B = A^T A$ — симметричная, и $(A^T A\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) \geqslant 0$, следовательно, все $\lambda(B) \geqslant 0$.

Рассуждая далее, как в 5.4, получим, что $\sup\limits_{x\neq 0}\dfrac{(B\mathbf{x},\mathbf{x})}{(\mathbf{x},\mathbf{x})}=\max\limits_i\lambda_i(B)$, а равенство достигается на соответствующем собственном векторе. Поэтому $||A||_2=\sqrt{\max\limits_i\lambda_i(A^TA)}$.

Отметим важный частный случай симметричной матрицы: $A=A^T$. Тогда $||A||_2=\max\limits_i|\lambda_i(A)|$. $\triangleright$

5.6. Доказать, что модуль любого собственного значения матрицы не больше любой ее нормы.

$\triangleleft$ Зафиксируем произвольный собственный вектор $\mathbf{x}$ матрицы A и построим квадратную матрицу X, столбцами которой являются векторы $\mathbf{x}$. Получим равенство $\lambda X=AX$. Отсюда следует: $|\lambda|\,\|X\|\leqslant\|A\|\,\|X\|$, т. е. $|\lambda|\leqslant\|A\|$. $\triangleright$

5.7. Пусть A — вещественная матрица размерности $m\times n$. Доказать следующие свойства спектральной нормы $||A||_2$:

1) $||A||_2=\sup\limits_{\substack{||\mathbf{x}||_2=1\\||\mathbf{y}||_2=1}}|\mathbf{y}^TA\mathbf{x}|$; 2) $||A^T||_2=||A||_2$; 3) $||A^TA||_2=||AA^T||_2=||A||_2^2$.

$\triangleleft$ Для доказательства свойства 1) надо показать, что существуют такие векторы $\mathbf{x}$ и $\mathbf{y}$ единичной длины, на которых максимум достигается. В силу неравенства Коши—Буняковского и с учетом того, что спектральная норма подчинена евклидовой векторной норме, получаем неравенство

$$|\mathbf{y}^TA\mathbf{x}|=|(\mathbf{y},A\mathbf{x})|\leqslant||\mathbf{y}||_2\,||A\mathbf{x}||_2\leqslant||\mathbf{y}||_2\,||\mathbf{x}||_2\,||A||_2=||A||_2.$$

Пусть вектор $\mathbf{x}$ такой, что $||A\mathbf{x}||_2=||A||_2$, т. е. на нем достигается максимум в определении подчиненной нормы, и возьмем $\mathbf{y}=\dfrac{A\mathbf{x}}{||A\mathbf{x}||_2}$. Тогда $||\mathbf{y}||_2=1$ и

$$|\mathbf{y}^TA\mathbf{x}|=\frac{(A\mathbf{x})^T}{||A\mathbf{x}||_2}\,A\mathbf{x}=\frac{(A\mathbf{x},A\mathbf{x})}{||A\mathbf{x}||_2}=||A\mathbf{x}||_2=||A||_2.$$

Следовательно, искомые векторы $\mathbf{x}$ и $\mathbf{y}$ построены и свойство 1) доказано.

Из свойства 1) и равенства

$$||A^T||_2=\sup_{\substack{||\mathbf{x}||_2=1\\||\mathbf{y}||_2=1}}|\mathbf{y}^TA^T\mathbf{x}|=\sup_{\substack{||\mathbf{x}||_2=1\\||\mathbf{y}||_2=1}}(\mathbf{y},A^T\mathbf{x})=$$

$$=\sup_{\substack{||\mathbf{x}||_2=1\\||\mathbf{y}||_2=1}}(A\mathbf{y},\mathbf{x})=\sup_{\substack{||\mathbf{x}||_2=1\\||\mathbf{y}||_2=1}}|\mathbf{x}^TA\mathbf{y}|=||A||_2$$

следует свойство 2).

Покажем справедливость свойства 3). Из свойства 2) следует неравенство

$$||A^TA||_2\leqslant||A^T||_2||A||_2=||A||_2^2.$$

Возьмем такой вектор $\mathbf{x}$, что $||\mathbf{x}||_2 = 1$ и $||A\,\mathbf{x}||_2 = ||A||_2$, и применим свойство 1) к матрице A^TA, положив $\mathbf{y} = \mathbf{x}$. Получаем неравенство

$$||A^TA||_2 \geqslant |\mathbf{x}^TA^TA\mathbf{x}| = (A\mathbf{x}, A\mathbf{x}) = ||A\,\mathbf{x}||_2^2 = ||A||_2^2.$$

Из этих двух неравенств следует $||A^TA||_2 = ||A||_2^2$. Аналогично показывается, что $||AA^T||_2 = ||A||_2^2$. Таким образом, свойство 3) доказано. $\triangleright$

5.8. Пусть A — вещественная прямоугольная матрица. Показать, что умножение ее справа или слева на ортогональную матрицу Q соответствующих размеров не меняет ее спектральную норму.

$\triangleleft$ Из свойства 3) спектральной нормы (см. 5.7) следует, что

$$||QA||_2^2 = ||(QA)^TQA||_2 = ||A^TQ^TQA||_2 = ||A^TA||_2 = ||A||_2^2.$$

Из свойства 2) и полученного равенства имеем

$$||AQ||_2 = ||(AQ)^T||_2 = ||Q^TA^T||_2 = ||A^T||_2 = ||A||_2.$$

В частности, из равенств

$$||Q\mathbf{x}||_2^2 = (Q\mathbf{x}, Q\mathbf{x}) = (Q\mathbf{x})^TQ\mathbf{x} = \mathbf{x}^TQ^TQ\mathbf{x} = \mathbf{x}^T\mathbf{x} = (\mathbf{x}, \mathbf{x}) = ||\mathbf{x}||_2^2$$

получаем, что умножение ортогональной матрицы на вектор $\mathbf{x}$ сохраняет его длину. $\triangleright$

5.9. Используя выражения для матричных норм из 5.5, показать справедливость неравенства $\|A\|_2^2 \leqslant \|A\|_1\|A\|_\infty$.

$\triangleleft$ Модуль любого собственного значения матрицы не больше любой ее нормы (см. 5.6), поэтому имеем

$$\|A\|_2^2 = \max\lambda(A^TA) \leqslant \|A^TA\|_1 \leqslant \|A\|_1\|A^T\|_1 = \|A\|_1\|A\|_\infty. \quad \triangleright$$

5.10. Рассмотрим функцию от элементов матрицы

$$\eta(A) = \max_{i,j}|a_{ij}|, \quad 1 \leqslant i, j \leqslant n.$$

Показать, что $\eta(A)$ не является нормой в пространстве матриц (хотя и является нормой вектора с компонентами a_{ij} в $\mathbf{R}^{n^2}$).

$\triangleleft$ Для любой матричной нормы справедливо неравенство $||AB|| \leqslant ||A||\,||B||$. Рассмотрим матрицы $A = B$, $a_{ij} = b_{ij} = 1\ \forall i, j$ для которых имеют место соотношения $\eta(AB) = n,\ \eta(A) = \eta(B) = 1$, противоречащие приведенному выше неравенству. $\triangleright$

5.11. Доказать, что выражение $M(A) = n\,\eta(A)$ (см. 5.10) является матричной нормой.

$\triangleleft$ Заметим, что требует проверки только четвертое условие из определения матричной нормы: $M(AB) \leqslant M(A)M(B)$.

$$M(AB) = n\max_{i,j}\left|\sum_{k=1}^n a_{ik}b_{kj}\right| \leqslant n\max_{i,j}\sum_{k=1}^n |a_{ik}b_{kj}| \leqslant$$

$$\leqslant n\sum_{k=1}^n \eta(A)\,\eta(B) = n\eta(A)\,n\eta(B) = M(A)M(B). \quad \triangleright$$

5.12. Доказать, что для векторов $\mathbf{x} = (x_1, x_2)^T$ и $h > 0$ выражение $||\mathbf{x}||_h = \max\left(|x_1|, \frac{|x_2 - x_1|}{h}\right)$ является нормой. Найти матричную норму, подчиненную этой векторной норме.

$\lhd$ Найдем матричную норму. Заметим, что $||\mathbf{x}||_h = ||\mathbf{y}||_\infty$, где

$$\mathbf{y} = S\,\mathbf{x}\,, \qquad S = \begin{pmatrix} 1 & 0 \\ -\frac{1}{h} & \frac{1}{h} \end{pmatrix}.$$

Поэтому

$$||A||_h = \sup_{\mathbf{x}\neq 0} \frac{||A\mathbf{x}||_h}{||\mathbf{x}||_h} = \sup_{\mathbf{x}\neq 0} \frac{||SA\mathbf{x}||_\infty}{||S\mathbf{x}||_\infty} = \sup_{\mathbf{y}\neq 0} \frac{||SAS^{-1}\mathbf{y}||_\infty}{||\mathbf{y}||_\infty} =$$

$$= ||SAS^{-1}||_\infty = \left|\left|\begin{pmatrix} a_{11} + a_{12} & a_{12}h \\ \frac{a_{21} + a_{22} - a_{11} - a_{12}}{h} & a_{22} - a_{12} \end{pmatrix}\right|\right|_\infty =$$

$$= \max\left(|a_{11} + a_{12}| + h\,|a_{12}|, |a_{22} - a_{12}| + \frac{1}{h}\,|a_{21} + a_{22} - a_{11} - a_{12}|\right). \quad \rhd$$

5.13. Доказать, что выражение $N(A) = \left(\sum\limits_{i,j=1}^{n} a_{ij}^2\right)^{1/2}$ является матричной нормой. Найти наилучшие константы эквивалентности, связывающие $N(A)$ и нормы $\|\cdot\|_1$, $\|\cdot\|_2$ $\|\cdot\|_\infty$.

Ответ: $\frac{1}{\sqrt{n}}\,N(A) \leqslant \|A\|_1 \leqslant \sqrt{n}\,N(A)$, $\frac{1}{\sqrt{n}}\,N(A) \leqslant \|A\|_2 \leqslant N(A)$, $\frac{1}{\sqrt{n}}\,N(A) \leqslant \|A\|_\infty \leqslant \sqrt{n}\,N(A)$. Матричную норму $N(A)$ называют *нормой Фробениуса (нормой Шура, евклидовой матричной нормой)* и обозначают $\|A\|_F$.

5.14. Пусть числа $d_k > 0$, $k = 1, \ldots, n$. Доказать, что $\max\limits_k\,(d_k|x_k|)$ — норма вектора $\mathbf{x}$. Найти норму матрицы, подчиненную этой векторной норме.

Ответ: $\|DAD^{-1}\|_\infty$, где $D = \mathrm{diag}(d_1, \ldots, d_n)$.

5.15. Пусть числа $d_k > 0$, $k = 1, \ldots, n$. Доказать, что $\sum\limits_{k=1}^{n} d_k|x_k|$ — норма вектора $\mathbf{x}$. Найти норму матрицы, подчиненную этой векторной норме.

Ответ: $\|DAD^{-1}\|_1$, где $D = \mathrm{diag}(d_1, \ldots, d_n)$.

5.16. Пусть числа $d_k > 0$, $k = 1, \ldots, n$. Доказать, что $\sqrt{\sum\limits_{k=1}^{n} d_k\,x_k^2}$ — норма вектора $\mathbf{x}$. Найти норму матрицы, подчиненную этой векторной норме.

Ответ: $\|DAD^{-1}\|_2$, где $D = \mathrm{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})$.

5.17. Доказать, что $\max\limits_{1\leqslant i\leqslant n}\left(\left|\sum\limits_{k=1}^{i} x_k\right|\right)$ — норма вектора $\mathbf{x}$. Найти норму матрицы, подчиненную этой векторной норме.

Ответ: $\|LAL^{-1}\|_\infty$, где $l_{ij} = 1$ при $i \leqslant j$ и $l_{ij} = 0$ при $i > j$.

5.18. Пусть $M(A) = n \cdot \max\limits_{1 \leqslant i,j \leqslant n} |a_{ij}|$. Найти наилучшие константы C_1, C_2 в неравенстве $C_1\, M(A) \leqslant \|A\|_2 \leqslant C_2\, M(A)$.

Ответ: $C_1 = \frac{1}{n}, C_2 = 1$.

5.19. Пусть $M(A) = n \cdot \max\limits_{1 \leqslant i,j \leqslant n} |a_{ij}|$. Найти наилучшие константы C_1, C_2 в неравенстве $C_1\, M(A) \leqslant \|A\|_1 \leqslant C_2\, M(A)$.

Ответ: $C_1 = \frac{1}{n}, C_2 = 1$.

5.20. Пусть $||\mathbf{x}||_p = \left(\sum\limits_{i=1}^{n} |x_i|^p\right)^{1/p}$, $p \geqslant 1$. Доказать *неравенство Йенсена*

$$||\mathbf{x}||_p \leqslant ||\mathbf{x}||_q, \quad 1 \leqslant q \leqslant p \leqslant \infty.$$

$\lhd$ Считаем, что $\mathbf{x} \neq 0$, так как иначе неравенство тривиально. Пусть для определенности $|x_1| = \max\limits_i |x_i|$. Тогда

$$||\mathbf{x}||_p = |x_1|\left(1 + \sum_{i=2}^{n} \alpha_i^p\right)^{1/p}, \quad ||\mathbf{x}||_q = |x_1|\left(1 + \sum_{i=2}^{n} \alpha_i^q\right)^{1/q},$$

$\alpha_i = |x_1|^{-1}|x_i| \leqslant 1$. Так как $q \leqslant p$, то $\alpha_i^p \leqslant \alpha_i^q$, а $t^{1/p} \leqslant t^{1/q}$ для $t \geqslant 1$. Таким образом,

$$\left(1 + \sum_{i=2}^{n} \alpha_i^p\right)^{1/p} \leqslant \left(1 + \sum_{i=2}^{n} \alpha_i^q\right)^{1/q}. \qquad \rhd$$

5.21. Доказать, что при $\mathbf{x} \in \mathbf{R}^n$ справедливо равенство $\lim\limits_{p \to \infty} ||\mathbf{x}||_p = ||\mathbf{x}||_\infty$.

$\lhd$ Из неравенств $||\mathbf{x}||_\infty \leqslant ||\mathbf{x}||_p \leqslant n^{1/p}||\mathbf{x}||_\infty$ в пределе при $p \to \infty$ получаем требуемый результат. $\rhd$

5.22. Доказать, что сходимость в любой норме в пространстве $\mathbf{R}^n$ эквивалентна покоординатной сходимости.

Указание. Воспользоваться эквивалентностью любых норм в конечномерном пространстве.

5.23. Пусть $||\cdot||$ — векторная норма в $\mathbf{R}^m$ и $A \in \mathbf{R}^{m \times n}$ — прямоугольная матрица, размерности $m \times n$. Показать, что если ранг матрицы $\operatorname{rank}(A) = n$, то $||A\mathbf{x}||$ — векторная норма в $\mathbf{R}^n$.

Указание. Убедиться в справедливости первого условия в определении нормы.

5.24. Проверить, что $||\mathbf{x}||_p = \left(\sum\limits_{i=1}^{n} |x_i|^p\right)^{1/p}$, $p \geqslant 1$, является нормой в пространстве $\mathbf{C}^n$ векторов с комплексными координатами. Показать, что при $\mathbf{x} \in \mathbf{C}^n$ справедливо неравенство $||\mathbf{x}||_p \leqslant ||\operatorname{Re}(\mathbf{x})||_p + ||\operatorname{Im}(\mathbf{x})||_p$. Найти такую максимальную постоянную $c_0 > 0$, что $c_0(||\operatorname{Re}(\mathbf{x})||_2 + ||\operatorname{Im}(\mathbf{x})||_2) \leqslant ||\mathbf{x}||_2$ для всех $\mathbf{x} \in \mathbf{C}^n$.

Указание. Обозначить $x_k = a_k + \mathrm{i}\, b_k$ и воспользоваться неравенством треугольника для векторов, координатами которых являются $|a_k|$ и $|b_k|$.

Ответ: $c_0 = \frac{1}{\sqrt{2}}$.

5.25. Пусть $||\cdot||$ — некоторая норма в $\mathbf{R}^n$. Доказать, что функционал

$$||\mathbf{x}||_* = \sup_{\mathbf{y}\neq 0} \frac{(\mathbf{x},\mathbf{y})}{||\mathbf{y}||}$$

также задает норму в $\mathbf{R}^n$, называемую *двойственной* к $||\cdot||$. Найти норму, двойственную к $||\cdot||_\infty$.
Ответ: $||\cdot||_1$.

5.26. Пусть $1 \leqslant p \leqslant \infty$ и B — любая подматрица квадратной матрицы A. Доказать, что $||B||_p \leqslant ||A||_p$.

$\lhd$ Пусть A — матрица, размерности $n \times n$, и B — некоторая ее подматрица размерности $n_1 \times n_2$, $n_i \leqslant n$. Используя при необходимости перестановки строк и столбцов (это не влияет на норму матрицы), представим A в виде

$$A = \begin{pmatrix} B & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

По определению, $||A||_p = \sup\limits_{||\mathbf{x}||_p=1} ||A\mathbf{x}||_p$. Пусть $\mathbf{x}^*, ||\mathbf{x}^*||_p = 1$ — такой вектор, что $||B||_p = ||B\mathbf{x}^*||_p$ и $\tilde{\mathbf{x}} = (x_1^*, \ldots, x_{n_2}^*, 0, \ldots, 0)^T$. В этом случае $||\tilde{\mathbf{x}}||_p = 1$ и

$$||A||_p \geqslant ||A\tilde{\mathbf{x}}||_p \geqslant ||B\mathbf{x}^*||_p = ||B||_p. \qquad \rhd$$

5.27. Доказать, что если матрица $D = \mathrm{diag}(d_1, d_2, ..., d_k) \in \mathbf{R}^{m\times n}$, где $k = \min\{m, n\}$, то $||D||_p = \max\limits_i |d_i|$.

5.28. Пусть B — невырожденная матрица, $||\cdot||$ — некоторая норма в пространстве векторов размерности n. Доказать, что $||\mathbf{x}||_* = ||B\mathbf{x}||$ также является нормой в пространстве векторов. Какая норма в пространстве матриц порождается нормой $||\mathbf{x}||_*$ в пространстве векторов?
Указание. Воспользоваться 5.12 и 5.23.

5.29. Показать, что если A — невырожденная матрица, то для нормы матрицы, подчиненной векторной норме, справедливо равенство

$$||A^{-1}||^{-1} = \inf_{\mathbf{x}\neq 0} \frac{||A\mathbf{x}||}{||\mathbf{x}||}.$$

Указание. По определению,

$$||A^{-1}|| = \sup_{\mathbf{y}\neq 0} \frac{||A^{-1}\mathbf{y}||}{||\mathbf{y}||} = \sup_{\mathbf{x}\neq 0} \frac{||\mathbf{x}||}{||A\mathbf{x}||}.$$

Используя далее определения inf и sup, доказать, что

$$\left(\sup_{\mathbf{x}\neq 0} \frac{||\mathbf{x}||}{||A\mathbf{x}||}\right)^{-1} = \inf_{\mathbf{x}\neq 0} \frac{||A\mathbf{x}||}{||\mathbf{x}||}.$$

5.30. Доказать неравенство $||A||_2 \leqslant ||A||^{1/2}||A^T||^{1/2}$ для любой нормы A, подчиненной какой-либо векторной норме.

У к а з а н и е. Воспользоваться решениями 5.6 и 5.9.

5.31. Доказать, что если $A = A^T$, то $||A||_2 = \sup\limits_{\mathbf{x}\neq 0} \dfrac{|(A\mathbf{x},\mathbf{x})|}{||\mathbf{x}||_2^2}$.

У к а з а н и е. Воспользоваться решением 5.5.

5.32. Пусть $A = A^T > 0$ и $||\mathbf{x}||_A = (A\mathbf{x},\mathbf{x})^{1/2}$. Доказать, что для произвольного многочлена $p_m(t)$ степени $m \geqslant 0$ верно равенство

$$||p_m(A)||_A = ||p_m(A)||_2.$$

$\lhd$ Известно, что симметричная и положительно определенная матрица A имеет квадратный корень $A^{1/2}$. Пусть Q — ортогональная матрица, i-й столбец которой является i-м собственным вектором из полной ортонормированной системы собственных векторов A, а D — диагональная матрица с i-м собственным числом A в i-й строке. Тогда $A = QDQ^T$ и $A^{1/2} = QD^{1/2}Q^T$. Матрица $A^{1/2}$ коммутирует с A и любой ее степенью, а также и с $p_m(A)$. Используя этот факт, а также определения нормы $||\cdot||_2$ и энергетической нормы $||\cdot||_A$, получаем требуемое утверждение из следующей цепочки равенств:

$$||p_m(A)||_A = \sup_{\mathbf{x}\neq 0} \frac{(Ap_m(A)\mathbf{x}, p_m(A)\mathbf{x})^{1/2}}{(A\mathbf{x},\mathbf{x})^{1/2}} =$$

$$= \sup_{\mathbf{x}\neq 0} \frac{(p_m(A)A^{1/2}\mathbf{x}, p_m(A)A^{1/2}\mathbf{x})^{1/2}}{||A^{1/2}\mathbf{x}||_2} = \sup_{\mathbf{y}\neq 0} \frac{||p_m(A)\mathbf{y}||_2}{||\mathbf{y}||_2} = ||p_m(A)||_2.$$

$\rhd$

5.33. Доказать, что если матрица A — вещественная и $(A\mathbf{x},\mathbf{x}) > 0$ для всех вещественных $\mathbf{x} \neq 0$, то существует такая постоянная $\delta > 0$, не зависящая от $\mathbf{x}$, что $(A\mathbf{x},\mathbf{x}) \geqslant \delta\,(\mathbf{x},\mathbf{x})$.

$\lhd$ Всякая вещественная матрица A представима в виде $A = S + K$, где $S = \dfrac{A + A^T}{2}$ — симметричная, а $K = \dfrac{A - A^T}{2}$ — кососимметричная матрицы. При этом для любого вещественного $\mathbf{x} \neq 0$ имеем $(A\mathbf{x},\mathbf{x}) = (S\mathbf{x},\mathbf{x}) \geqslant \delta\,(\mathbf{x},\mathbf{x})$, где $\delta \geqslant 0$ — минимальное собственное значение матрицы S. Из неравенства $(A\mathbf{x},\mathbf{x}) > 0$ следует, что $\delta > 0$. $\rhd$

5.34. Привести пример положительно определенной вещественной матрицы, спектр которой не является вещественным.

О т в е т: матрица

$$A = \begin{pmatrix} a & 1 & 0 & \cdots & 0 \\ -1 & a & 0 & \cdots & 0 \\ 0 & 0 & & & \\ & & & B & \\ 0 & 0 & & & \end{pmatrix}$$

с положительной константой a и симметричной положительно определенной подматрицей B положительно определена, но имеет пару комплексно-сопряженных собственных значений $\lambda_{1,2} = a \pm \mathrm{i}$.

5.35. Доказать, что матричные нормы, определенные равенствами $M(A) = n \cdot \max\limits_{1 \leqslant i,j \leqslant n} |a_{ij}|$ и $N(A) = (\sum\limits_{i,j=1}^{n} a_{ij}^2)^{1/2}$, не подчинены никаким векторным нормам.

Указание. Воспользоваться тем фактом, что для любой подчиненной нормы справедливо следующее равенство: $\|I\| = 1$, где $I = \mathrm{diag}(1, \ldots, 1)$.

5.36. Показать, что для любого собственного значения $\lambda(A)$ невырожденной матрицы A справедлива оценка $\dfrac{1}{||A^{-1}||} \leqslant |\lambda(A)|$.

Указание. Воспользоваться решением 5.6.

5.37. Доказать, что для любого собственного значения $\lambda(A)$ матрицы A справедливо неравенство $|\lambda(A)| \leqslant \inf\limits_{k} ||A^k||^{1/k}$, где k — натуральное число.

Указание. Воспользоваться решением 5.6.

5.38. Доказать, что если A — нормальная матрица ($AA^T = A^TA$), то $||A||_2 = \rho(A)$, где $\rho(A) = \max\limits_{i} |\lambda_i(A)|$ — *спектральный радиус* матрицы A.

Указание. Воспользоваться тем фактом, что нормальная матрица имеет полную ортонормированную систему собственных векторов.

5.39. Убедиться, что матрица A размерности $n \times n$ при $n \geqslant 2$ не определяется однозначно значениями квадратичной формы $(A\mathbf{x}, \mathbf{x})$ на произвольном векторе $\mathbf{x}$, т. е. найдутся две различные матрицы A и B, для которых $(A\mathbf{x}, \mathbf{x}) \equiv (B\mathbf{x}, \mathbf{x})$ для любых вещественных $\mathbf{x}$.

Указание. Воспользоваться решением 5.33.

5.40. Доказать, что всякая норма $\|\cdot\|_m$ матрицы согласована с какой-либо векторной нормой $\|\cdot\|_v$, т. е. верна оценка $\|A\,\mathbf{x}\|_v \leqslant \|A\|_m \|\mathbf{x}\|_v$.

$\lhd$ Пусть для матрицы A определена некоторая матричная норма $\|A\|_m$. Тогда определим функционал $\|\mathbf{x}\|_v$ следующим образом:

$$\|\mathbf{x}\|_v = \left\| \begin{pmatrix} 0 & 0 & \ldots & 0 & x_1 \\ 0 & 0 & \ldots & 0 & x_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & x_n \end{pmatrix} \right\|_m .$$

Непосредственно проверяется, что $\|\mathbf{x}\|_v$ удовлетворяет всем условиям векторной нормы и согласован с исходной матричной. $\rhd$

5.41. Пусть A — матрица размерности $n \times n$, $\rho(A)$ — ее спектральный радиус и задано число $\varepsilon > 0$. Доказать, что существует по крайней мере одна матричная норма, для которой имеют место оценки

$$\rho(A) \leqslant \|A\| \leqslant \rho(A) + \varepsilon.$$

$\lhd$ Из курса линейной алгебры (теорема Шура об унитарной триангуляции) известно, что найдутся такие унитарная матрица $U\,(U^* = U^{-1})$ и верхняя треугольная матрица R, что $A = U\,R\,U^*$. Положим $D_t = \operatorname{diag}(t, t^2, t^3, \ldots, t^n)$ и вычислим матрицу

$$D_t R D_t^{-1} = \begin{pmatrix} \lambda_1 & t^{-1}r_{12} & t^{-2}r_{13} & \ldots & t^{-n+1}r_{1n} \\ 0 & \lambda_2 & t^{-1}r_{23} & \ldots & t^{-n+2}r_{2n} \\ 0 & 0 & \lambda_3 & \ldots & t^{-n+3}r_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \lambda_n \end{pmatrix}.$$

При достаточно большом $t > 0$ сумма модулей наддиагональных элементов матрицы $D_t R D_t^{-1}$ не превосходит ε. В частности, это приводит к неравенству $\|D_t R D_t^{-1}\|_1 \leqslant \rho(A) + \varepsilon$. Теперь определим матричную норму с помощью формулы

$$\|A\| = \|D_t U^* A\, U\, D_t^{-1}\|_1 = \| \left(U\, D_t^{-1}\right)^{-1} A\, \left(U\, D_t^{-1}\right) \|_1.$$

Таким образом, выбор достаточно большого t в приведенной выше формуле приводит к оценке сверху, а оценка снизу следует из 5.6. $\rhd$

5.2. Элементы теории возмущений

Рассмотрим систему линейных алгебраических уравнений

$$A\,\mathbf{x} = \mathbf{b}$$

с квадратной невырожденной матрицей A. При ее решении в результате вычислений с конечной разрядностью вместо $\mathbf{x}$ получается *приближенное* решение $\tilde{\mathbf{x}}$, которое можно рассматривать как *точное* решение *возмущенной* системы

$$(A + \delta A)\,\tilde{\mathbf{x}} = \mathbf{b},$$

где матрица возмущений δA мала в каком-либо смысле.

Другой источник ошибок в $\tilde{\mathbf{x}}$ определяется возмущениями δA и $\delta\mathbf{b}$ в элементах матрицы A и в компонентах вектора правой части $\mathbf{b}$ (например, вследствие ошибок округлений, возникающих в процессе ввода вещественных чисел в память компьютера).

Чтобы оценить насколько приближенное решение $\tilde{\mathbf{x}}$ отличается от точного решения $\mathbf{x}$, используют нормы векторов и подчиненные нормы матриц.

Пусть в системе $A\mathbf{x} = \mathbf{b}$ возмущается только вектор $\mathbf{b}$, т. е. вместо исходной системы решается возмущенная система $A\tilde{\mathbf{x}} = \tilde{\mathbf{b}} \equiv \mathbf{b} + \delta\mathbf{b}$,

и пусть $\tilde{\mathbf{x}}$ — точное решение возмущенной системы. Тогда для относительной ошибки в $\tilde{\mathbf{x}}$ верна оценка

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||}{||\mathbf{x}||} \leqslant ||A||\,||A^{-1}||\,\frac{||\mathbf{b}-\tilde{\mathbf{b}}||}{||\mathbf{b}||} = ||A||\,||A^{-1}||\,\frac{||\mathbf{b}-A\tilde{\mathbf{x}}||}{||\mathbf{b}||}\ .$$

Величину $||A||\,||A^{-1}||$ называют *числом обусловленности* матрицы A и часто обозначают $\mathrm{cond}(A)$. Для вырожденных матриц $\mathrm{cond}(A) = \infty$. Конкретное значение $\mathrm{cond}(A)$ зависит от выбора матричной нормы, однако в силу их эквивалентности при практических оценках этим различием можно пренебречь.

Из приведенного выше неравенства следует, что даже если *вектор невязки* $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$ мал, относительные возмущения в решении могут быть большими, если $\mathrm{cond}(A)$ велико (такие матрицы называют *плохо обусловленными*).

5.42. Доказать неравенство

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||}{||\mathbf{x}||} \leqslant \mathrm{cond}(A)\,\frac{||\mathbf{b}-\tilde{\mathbf{b}}||}{||\mathbf{b}||} = \mathrm{cond}(A)\,\frac{||\mathbf{b}-A\tilde{\mathbf{x}}||}{||\mathbf{b}||}\ .$$

$\lhd$ Из равенства $A^{-1}\mathbf{r} = A^{-1}\mathbf{b} - A^{-1}A\tilde{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{x}}$ следует, что

$$||\mathbf{x}-\tilde{\mathbf{x}}|| \leqslant ||A^{-1}||\,||\mathbf{r}||. \tag{5.1}$$

Из равенства $\mathbf{b} = A\mathbf{x}$ имеем, что $||\mathbf{b}|| = ||A\mathbf{x}|| \leqslant ||A||\,||\mathbf{x}||$, т. е.

$$||\mathbf{x}|| \geqslant \frac{||\mathbf{b}||}{||A||}\,. \tag{5.2}$$

Разделив неравенство (5.1) на неравенство (5.2), получим

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||}{||\mathbf{x}||} \leqslant ||A||\,||A^{-1}||\,\frac{||\mathbf{r}||}{||\mathbf{b}||} = \mathrm{cond}(A)\,\frac{||\mathbf{r}||}{||\mathbf{b}||} = \mathrm{cond}(A)\,\frac{||\mathbf{b}-A\tilde{\mathbf{x}}||}{||\mathbf{b}||}\,.$$

Отсюда видно, что если матрица A плохо обусловлена, то даже очень маленькая невязка не может гарантировать малость относительной ошибки в $\tilde{\mathbf{x}}$. С другой стороны, может оказаться так, что достаточно точное решение имеет большую невязку. Рассмотрим пример:

$$A = \begin{pmatrix} 1{,}000 & 1{,}001 \\ 1{,}000 & 1{,}000 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 2{,}001 \\ 2{,}000 \end{pmatrix}$$

Точное решение системы $A\mathbf{x} = \mathbf{b}$ имеет вид $\mathbf{x} = (1,1)^T$. Однако вектор $\tilde{\mathbf{x}} = (2,0)^T$, который никак нельзя назвать близким к $\mathbf{x}$, дает маленькую невязку $\mathbf{r} = (10^{-3},0)^T$.

Возьмем теперь $\mathbf{b} = (1,0)^T$. Тогда вектор $\mathbf{x} = (-1000,1000)^T$ — точное решение системы. Вектор $\tilde{\mathbf{x}} = (-1001,1000)^T$ достаточно близок к $\mathbf{x}$ в смысле относительной погрешности, однако $\tilde{\mathbf{x}}$ дает большую невязку $\mathbf{r} = (1,1)^T$, близкую по норме к вектору $\mathbf{b}$. $\rhd$

5.43. Найти решения двух систем с близкими коэффициентами:

$$\begin{cases} x+3y &= 4\,, \\ x+3,00001y &= 4,00001\,; \end{cases} \qquad \begin{cases} x+3y &= 4\,, \\ x+2,99999y &= 4,00001 \end{cases}$$

и объяснить полученный результат.

Ответ: $(1,1)^T$ и $(7,-1)^T$. Обе матрицы получены малыми возмущениями одной вырожденной матрицы. В данном случае это приводит к большой разнице в решениях систем.

5.44. Показать, что $\mathrm{cond}(A) \geqslant 1$ для любой матрицы A и $\mathrm{cond}_2(Q) = 1$ для ортогональной матрицы Q.

$\lhd$ Так как $I = AA^{-1}$, то

$$1 = ||I|| = ||A\,A^{-1}|| \leqslant ||A||\,||A^{-1}|| = \mathrm{cond}(A)\,.$$

Умножение матрицы на ортогональную не меняет ее спектральную норму, поэтому

$$||Q||_2 = ||QI||_2 = ||I||_2 = 1 \quad \text{и} \quad ||Q^T||_2 = ||Q^T I||_2 = ||I||_2 = 1\,.$$

Таким образом, $\mathrm{cond}_2(Q) = ||Q||_2\,||Q^{-1}||_2 = ||Q||_2\,||Q^T||_2 = 1\,.$ $\rhd$

5.45. Можно ли утверждать, что если определитель матрицы мал, то матрица плохо обусловлена?

$\lhd$ Пусть дана диагональная матрица $D = \varepsilon I$, где $\varepsilon > 0$ — малое число и I — единичная матрица. Определитель $\det(D) = \varepsilon^n$ мал, тогда как матрица D хорошо обусловлена, поскольку

$$\mathrm{cond}(D) = ||D||\,||D^{-1}|| = \varepsilon||I||\,\varepsilon^{-1}||I^{-1}|| = 1\,.$$

Рассмотрим теперь матрицу

$$A = \begin{pmatrix} 1 & -1 & -1 & \dots & -1 \\ 0 & 1 & -1 & \dots & -1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix},$$

у которой определитель равен 1, и вычислим ее число обусловленности. Для этого возьмем произвольный вектор $\mathbf{b} \neq 0$ и, решая систему $A\mathbf{x} = \mathbf{b}$ с помощью обратной подстановки, построим элементы обратной матрицы A^{-1}:

$$\begin{aligned} x_n &= b_n, \\ x_{n-1} &= b_{n-1} + b_n, \\ x_{n-2} &= b_{n-2} + b_{n-1} + 2b_n, \\ x_{n-3} &= b_{n-3} + b_{n-2} + 2b_{n-1} + 2^2 b_n, \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ x_1 &= b_1 + b_2 + 2b_3 + \dots + 2^{n-3}b_{n-1} + 2^{n-2}b_n. \end{aligned}$$

Запишем полученную обратную матрицу:

$$A^{-1} = \begin{pmatrix} 1 & 1 & 2 & 4 & \dots & 2^{n-3} & 2^{n-2} \\ 0 & 1 & 1 & 2 & \dots & 2^{n-4} & 2^{n-3} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Следовательно, $\|A^{-1}\|_\infty = 1 + 1 + 2 + 2^2 + \dots + 2^{n-2} = 2^{n-1}$. Так как $\|A\|_\infty = n$, то $\mathrm{cond}_\infty(A) = n\,2^{n-1}$, т. е. матрица A плохо обусловлена, хотя $\det(A) = 1$.

Рассмотренные примеры показывают, что обусловленность матрицы зависит не только от величины определителя. $\triangleright$

5.46. Пусть дана матрица A размерности $n \times n$ с параметром $|a| \neq 1$

$$A = \begin{pmatrix} 1 & a & 0 & \dots & 0 & 0 \\ 0 & 1 & a & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & a \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Вычислить $\mathrm{cond}_\infty(A)$ и оценить возмущение в компоненте x_1 решения системы $A\mathbf{x} = \mathbf{b}$, если компонента b_n вектора $\mathbf{b}$ возмущена на ε.

$\triangleleft$ Как в 5.45, методом обратной подстановки получим обратную матрицу

$$A^{-1} = \begin{pmatrix} 1 & -a & (-a)^2 & \dots & (-a)^{n-2} & (-a)^{n-1} \\ 0 & 1 & -a & \dots & (-a)^{n-3} & (-a)^{n-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & -a \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Тогда

$$\|A\|_\infty = 1 + |a|,$$

$$\|A^{-1}\|_\infty = 1 + |a| + a^2 + \dots + |a|^{n-1} = \frac{|a|^n - 1}{|a| - 1},$$

$$\mathrm{cond}_\infty(A) = \frac{(|a| + 1)\,(|a|^n - 1)}{|a| - 1}.$$

Отсюда видно, что матрица A плохо обусловлена при $|a| > 1$ и хорошо обусловлена при $|a| < 1$. Например, при $n = 20$ и $a = 5$ имеем $\mathrm{cond}_\infty(A) \approx 10^{14}$.

Пусть компонента b_n задана с ошибкой ε. Тогда вычисленное значение $\tilde{x}_1$ компоненты x_1 имеет вид

$$\tilde{x}_1 = b_1 - ab_2 + \dots + (-a)^{n-2}b_{n-1} + (-a)^{n-1}(b_n + \varepsilon) = x_1 + (-a)^{n-1}\varepsilon\,.$$

Следовательно, при $|a| > 1$ возмущение в b_n увеличивается в компоненте x_1 в $|a|^{n-1}$ раз, а при $|a| < 1$ во столько же раз уменьшается. $\triangleright$

5.47. Пусть A — матрица размерности $n \times n$ с элементами $a_{ij} = \{p \text{ для } i = j, \quad q \text{ для } i = j-1, \quad 0 \text{ для остальных индексов}\}$. Вычислить матрицу A^{-1} и показать, что при $|q| < |p|$ матрица A хорошо обусловлена, а при $|q| > |p|$ и больших значениях n — плохо обусловлена.

У к а з а н и е. Воспользоваться решением 5.46.

5.48. Пусть матрица A определена как в 5.47. Выразить явно решение системы $A\mathbf{x} = \mathbf{b}$ через правую часть.

У к а з а н и е. Воспользоваться решением 5.46.

5.49. Решается система $A\mathbf{x} = \mathbf{b}$ с матрицей

$$A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & \varepsilon & \varepsilon \\ 1 & \varepsilon & \varepsilon \end{pmatrix}, \quad |\varepsilon| \ll 1.$$

В результате замены $x_1' = x_1$, $x_2' = \varepsilon x_2$, $x_3' = \varepsilon x_3$ для нахождения новых неизвестных $\mathbf{x}'$ имеем систему $A'\mathbf{x}' = \mathbf{b}'$ с матрицей

$$A' = \begin{pmatrix} \varepsilon & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

В каком случае число обусловленности меньше?

$\lhd$ Имеем

$$A^{-1} = \begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1-\varepsilon}{4\varepsilon} & \frac{1+\varepsilon}{4\varepsilon} \\ \frac{1}{2} & \frac{1+\varepsilon}{4\varepsilon} & \frac{1-\varepsilon}{4\varepsilon} \end{pmatrix}, \quad (A')^{-1} = \begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1-\varepsilon}{4} & \frac{1+\varepsilon}{4} \\ \frac{1}{2} & \frac{1+\varepsilon}{4} & \frac{1-\varepsilon}{4} \end{pmatrix},$$

поэтому число обусловленности исходной матрицы стремится к бесконечности при $\varepsilon \to 0$, а число обусловленности матрицы A' остается ограниченным. $\rhd$

5.50. Пусть $A = A^T > 0$, $\lambda(A) \in [m, M]$ и $A \neq \beta I$, где I — единичная матрица. Доказать, что $\operatorname{cond}_2(A + \alpha I)$ монотонно убывает по α при $\alpha > 0$.

У к а з а н и е. Вычислить $\operatorname{cond}_2(A + \alpha I) = \frac{M+\alpha}{m+\alpha} = 1 + \frac{M-m}{m+\alpha}$.

5.51. Существуют ли несимметричные матрицы, для которых справедливо $\operatorname{cond}^2(A) = \operatorname{cond}(A^2) > 1$?

О т в е т: примером такой матрицы является

$$A = \begin{pmatrix} 10^3 & 0 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 10^{-3} \end{pmatrix},$$

$$\lambda(A^T A) \in \left\{10^6, 10^{-6}, 4, 5 \pm \sqrt{4,25}\right\},$$

$$\operatorname{cond}(A^2) = \|A^2\|\|A^{-2}\| = 10^{12}; \quad \operatorname{cond}(A) = \|A\|\|A^{-1}\| = 10^6.$$

5.52. Доказать неравенство для квадратных невырожденных матриц размерности $n \times n$

$$\frac{1}{n} \leqslant \frac{\mathrm{cond}_1(A)}{\mathrm{cond}_2(A)} \leqslant n.$$

$\lhd$ Воспользовавшись неравенством для векторных норм

$$||\mathbf{x}||_2 \leqslant ||\mathbf{x}||_1 \leqslant \sqrt{n}\,||\mathbf{x}||_2,$$

получим $\frac{1}{\sqrt{n}}\,||A||_2 \leqslant ||A||_1 \leqslant \sqrt{n}\,||A||_2$, откуда и следует требуемый результат. $\rhd$

5.53. Оценить снизу и сверху $\mathrm{cond}_\infty(A)$ невырожденной матрицы A размерности $n \times n$, используя границы собственных чисел матрицы $A^T A$: $\lambda(A^T A) \in [\alpha, \beta]$.

$\lhd$ Из неравенства для чисел обусловленности в матричных нормах $\|\cdot\|_\infty$ и $\|\cdot\|_2$ и равенства $\|A\|_2^2 = \lambda_{\max}(A^T A)$ следует, что

$$\frac{1}{n}\sqrt{\frac{\beta}{\alpha}} \leqslant \mathrm{cond}_\infty(A) \leqslant n\sqrt{\frac{\beta}{\alpha}}. \qquad \rhd$$

5.54. Получить неравенство $\mathrm{cond}(A) \geqslant \left|\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}\right|$ для произвольной невырожденной матрицы A и любой матричной нормы, используемой при определении числа обусловленности.

У к а з а н и е. Воспользоваться решением 5.6.

5.55. Доказать, что $\mathrm{cond}(AB) \leqslant \mathrm{cond}(A)\,\mathrm{cond}(B)$ для любой заданной нормы в определении числа обусловленности и для любых невырожденных квадратных матриц.

5.56. Оценить $\mathrm{cond}_2(A)$ матрицы A размерности $n \times n$

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 \end{pmatrix}.$$

У к а з а н и е. Воспользоваться для собственных векторов $\mathbf{y}^{(j)}$, $j = 1, \dots, n$, матрицы A явной формулой (см. 2.86): $\mathbf{y}_k^{(j)} = \sin\frac{\pi\, j\, k}{n+1}$. Соответствующие собственные числа: $\lambda^{(j)} = 4\sin^2\frac{\pi\, j}{2(n+1)}$, так что

$$\mathrm{cond}_2(A) = \mathrm{ctg}^2\frac{\pi}{2(n+1)} \approx \frac{4(n+1)^2}{\pi^2}.$$

5.57. Оценить $\mathrm{cond}_2(A)$ матрицы A размерности $n \times n$

$$A = \frac{1}{6}\begin{pmatrix} 4 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 4 & 1 \\ 0 & 0 & 0 & 0 & \dots & 1 & 4 \end{pmatrix}.$$

У к а з а н и е. Для собственных векторов $\mathbf{y}^{(j)}, j = 1, \dots, n$ матрицы A найти явную формулу (см. 2.86). Получим $\mathbf{y}_k^{(j)} = \sin\frac{\pi\, j\, k}{n+1}$. Соответствующие собственные числа: $\lambda^{(j)} = \frac{2}{3} + \frac{1}{3}\cos\frac{\pi\, j}{n+1}$, так что $\mathrm{cond}_2(A) \leqslant 3$.

5.58. Пусть I — единичная матрица, δI — ее возмущение и $||\delta I|| < 1$. Показать, что матрица $I - \delta I$ невырожденная и выполнена оценка

$$||(I - \delta I)^{-1}|| \leqslant \frac{1}{1 - ||\delta I||}\,.$$

$\lhd$ Возьмем произвольный вектор $\mathbf{x} \neq 0$. Так как $1 - ||\delta I|| > 0$ и $||\mathbf{x}|| = ||(\mathbf{x} - \delta I\mathbf{x}) + \delta I\mathbf{x}|| \leqslant ||\mathbf{x} - \delta I\mathbf{x}|| + ||\delta I\mathbf{x}||$, то

$$||(I - \delta I)\,\mathbf{x}|| = ||\mathbf{x} - \delta I\mathbf{x}|| \geqslant ||\mathbf{x}|| - ||\delta I\mathbf{x}|| \geqslant$$
$$\geqslant ||\mathbf{x}|| - ||\delta I||\,||\mathbf{x}|| = (1 - ||\delta I||)\,||\mathbf{x}|| > 0\,.$$

Следовательно, если $\mathbf{x} \neq 0$, то $(I - \delta I)\,\mathbf{x} \neq 0$, т. е. матрица $I - \delta I$ невырожденная. Из тождества $(I - \delta I)\,(I - \delta I)^{-1} = I$ получаем $(I - \delta I)^{-1} = I + \delta I\,(I - \delta I)^{-1}$. Отсюда

$$||(I - \delta I)^{-1}|| \leqslant ||I|| + ||\delta I||\,||(I - \delta I)^{-1}|| = 1 + ||(I - \delta I)^{-1}||\,||\delta I||\,.$$

Из этого неравенства следует решение задачи (ее называют *задачей о возмущении единичной матрицы*). $\rhd$

5.59. Пусть I — единичная матрица, δI — ее возмущение и $||\delta I|| < 1$. Получить оценку отклонения матрицы I от матрицы $(I - \delta I)^{-1}$.

$\lhd$ Из $(I - \delta I)^{-1} = I + \delta I(I - \delta I)^{-1}$ (см. 5.58) получим $I - (I - \delta I)^{-1} = -\delta I(I - \delta I)^{-1}$. Отсюда в силу неравенства из 5.58

$$||I - (I - \delta I)^{-1}|| \leqslant ||\delta I||\,||(I - \delta I)^{-1}|| \leqslant \frac{||\delta I||}{1 - ||\delta I||}\,.$$ $\rhd$

5.60. Пусть A — невырожденная матрица, δA — ее возмущение и $||A^{-1}\delta A|| < 1$. Показать, что матрица $A + \delta A$ невырожденная и выполнена оценка

$$||(A + \delta A)^{-1}|| \leqslant \frac{||A^{-1}||}{1 - ||A^{-1}\delta A||}\,.$$

$\lhd$ Имеем $A+\delta A = A(I+A^{-1}\delta A)$. Поскольку $||A^{-1}\delta A|| < 1$, из 5.58 следует, что матрица $I + A^{-1}\delta A$ невырожденная. Это означает, что и матрица $A + \delta A$ также не вырождена.

Из равенства $(A + \delta A)^{-1} = (I + A^{-1}\delta A)^{-1}A^{-1}$, в силу неравенства из 5.58, следует, что

$$||(A + \delta A)^{-1}|| \leqslant ||(I + A^{-1}\delta A)^{-1}||\,||A^{-1}|| \leqslant \frac{||A^{-1}||}{1 - ||A^{-1}\delta A||}\,. \qquad \rhd$$

5.61. Пусть A — невырожденная матрица, δA — ее возмущение и $||A^{-1}\delta A|| < 1$. Получить оценку отклонения матрицы $(A + \delta A)^{-1}$ от A^{-1}.

$\lhd$ Из равенства $(A + \delta A)^{-1} = (I + A^{-1}\delta A)^{-1}A^{-1}$ следует, что $A^{-1} - (A + \delta A)^{-1} = (I - (I + A^{-1}\delta A)^{-1})A^{-1}$. Тогда в силу неравенства из 5.59

$$||A^{-1} - (A + \delta A)^{-1}|| \leqslant ||I - (I + A^{-1}\delta A)^{-1}||\,||A^{-1}|| \leqslant \frac{||A^{-1}\delta A||}{1 - ||A^{-1}\delta A||}\;||A^{-1}||.$$

Относительная ошибка в матрице $(A+\delta A)^{-1}$ оценивается неравенством

$$\frac{||A^{-1} - (A + \delta A)^{-1}||}{||A^{-1}||} \leqslant \frac{||A^{-1}||\,||\delta A||}{1 - ||A^{-1}\delta A||} = \frac{\mathrm{cond}(A)}{1 - ||A^{-1}\delta A||}\;\frac{||\delta A||}{||A||}\,. \qquad \rhd$$

5.62. Оценить снизу число обусловленности $\mathrm{cond}_2(A)$ матрицы:

$$1)\;\; A = \begin{pmatrix} 10 & 10 & 30 \\ 0,1 & 0,5 & 0,1 \\ 0,03 & 0,01 & 0,01 \end{pmatrix}; \quad 2)\;\; A = \begin{pmatrix} 1 & 20 & -400 \\ 0,2 & -2 & -20 \\ -0,04 & -0,2 & 1 \end{pmatrix}.$$

5.63. Система $A\mathbf{x} = \mathbf{b}$, где

$$A = \begin{pmatrix} 2 & -1 & 1 \\ -1 & 10^{-10} & 10^{-10} \\ 1 & 10^{-10} & 10^{-10} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 2(1 + 10^{-10}) \\ -10^{-10} \\ 10^{-10} \end{pmatrix},$$

имеет решение $\mathbf{x} = (10^{-10}, -1, 1)^T$. Доказать, что если $(A + \delta A)\mathbf{y} = \mathbf{b}$, $\|\delta A\| \leqslant 10^{-8}\|A\|$, то $\|\mathbf{x} - \mathbf{y}\| \leqslant 10^{-7}$. Это означает, что относительно малые изменения в элементах матрицы A не приводят к большим изменениям в решении, хотя $\mathrm{cond}_\infty(A)$ имеет порядок 10^{10}.

5.64. Пусть $A = \begin{pmatrix} 100 & 99 \\ 99 & 98 \end{pmatrix}$. Доказать, что данная матрица имеет наибольшее число обусловленности $\mathrm{cond}_2(A)$ из всех невырожденных матриц второго порядка, элементами которых являются положительные целые числа, меньшие или равные 100.

$\lhd$ Введем обозначения для элементов матрицы A

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

и найдем $\mathrm{cond}_2(A)$ в явном виде

$$\|A\|_2 = \sqrt{\max\lambda(A^TA)}, \quad \|A^{-1}\|_2 = \sqrt{\max\lambda((A^{-1})^TA^{-1})} = \\ = \sqrt{\max\lambda((A^TA)^{-1})} = \frac{1}{\sqrt{\min\lambda(A^TA)}}\,.$$

Имеем

$$\mathrm{cond}_2(A) = \sqrt{\frac{\max\lambda(A^TA)}{\min\lambda(A^TA)}}.$$

Введем матрицу

$$B = A^TA = \begin{pmatrix} a^2+c^2 & ab+cd \\ ab+cd & b^2+d^2 \end{pmatrix}$$

и запишем ее характеристический многочлен

$$p_B(\lambda) = \lambda^2 - \lambda\,\mathrm{tr}\,B + \det B, \;\; \mathrm{tr}B = a^2+b^2+c^2+d^2.$$

Его корни равны $\lambda_{1,2} = \frac{\mathrm{tr}\,B \pm \sqrt{\mathrm{tr}^2B - 4\det B}}{2}$. Так как $\mathrm{tr}\;B > 0$, то

$$\mathrm{cond}_2(A) = \sqrt{\frac{\mathrm{tr}\,B + \sqrt{\mathrm{tr}^2B - 4\det B}}{\mathrm{tr}\,B - \sqrt{\mathrm{tr}^2B - 4\det B}}} = \frac{\mathrm{tr}\,B + \sqrt{\mathrm{tr}^2B - 4\det B}}{\sqrt{4\det B}} = \\ = \frac{\mathrm{tr}\,B}{2\sqrt{\det B}} + \sqrt{\frac{\mathrm{tr}^2B}{4\det B} - 1}.$$

Таким образом, значение $\mathrm{cond}_2(A)$ максимально, если максимально $\frac{\mathrm{tr}^2\,(A^TA)}{\det\,(A^TA)}$. Имеем $\mathrm{tr}^2(A^TA) = \left(a^2+b^2+c^2+d^2\right)^2$,

$$\det\,(A^TA) = (a^2+c^2)\,(b^2+d^2) - (ab+cd)^2 = \\ = a^2d^2 + b^2c^2 - 2abcd = (ad-bc)^2 = \left|\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right|^2 \equiv \det{}^2A\,,$$

следовательно, значение $\frac{a^2+b^2+c^2+d^2}{\det^2A}$ должно быть максимальным. Отсюда получаем, что выражение $a^2+b^2+c^2+d^2$ максимально при условии $\det A = \pm 1$. Действительно, если модуль определителя больше 1, то $\mathrm{tr}B$ необходимо увеличить больше, чем в два раза. При ограничении $a_{ij} \leqslant 100$ это невозможно. Таким образом, при $n = 98$ можно воспользоваться любой из следующих матриц:

$$A_1 = \begin{pmatrix} n+2 & n+1 \\ n+1 & n \end{pmatrix}, \quad A_2 = \begin{pmatrix} n+1 & n+2 \\ n & n+1 \end{pmatrix}, \\ A_3 = \begin{pmatrix} n+1 & n \\ n+2 & n+1 \end{pmatrix}, \quad A_4 = \begin{pmatrix} n & n+1 \\ n+1 & n+2 \end{pmatrix}. \qquad \triangleright$$

5.65. Пусть при некотором $1 > q > 0$ для элементов каждой строки i невырожденной матрицы A выполнено неравенство $q\,|a_{ii}| \geqslant \sum\limits_{j\neq i} |a_{ij}|$. Оценить снизу и сверху $\operatorname{cond}_\infty(A)$, используя только диагональные элементы матрицы и параметр q.

$\lhd$ Отметим оценки

$$\max_i |a_{ii}| \leqslant \|A\|_\infty \leqslant (1+q)\max_i |a_{ii}|.$$

Введем обозначение $C = A^{-1}$ и заметим, что для $\forall i,j$ справедливо $|c_{ij}| \leqslant \|C\|_\infty$. При каждом i имеем $(A\,C = I)$

$$\sum_{k=1}^{n} a_{ik}c_{ki} = 1, \quad 1 \leqslant \sum_{k=1}^{n} |a_{ik}||c_{ki}| \leqslant |a_{ii}|\,(1+q)\,\|C\|_\infty.$$

Отсюда получаем оценку снизу для нормы матрицы A^{-1}

$$\|A^{-1}\|_\infty = \|C\|_\infty \geqslant \frac{1}{(1+q)\min\limits_i |a_{ii}|}\,,$$

следовательно,

$$\operatorname{cond}_\infty(A) = \|A\|_\infty\|A^{-1}\|_\infty \geqslant \frac{1}{1+q}\,\frac{\max\limits_i |a_{ii}|}{\min\limits_i |a_{ii}|}\,.$$

Если правая часть неравенства не превышает единицы, то полученная оценка малосодержательна.

В силу невырожденности матрицы A, все диагональные элементы a_{ii} отличны от нуля, поэтому можно построить матрицы

$$J = \operatorname{diag}(a_{11}^{-1}, a_{22}^{-1}, \ldots, a_{nn}^{-1}), \quad B = JA - I\,.$$

Отметим, что $\|B\|_\infty \leqslant q < 1$ в силу цепочки неравенств

$$\max_i |b_{i1}x_1 + \cdots + b_{in}x_n| \leqslant \max_i \sum_k |b_{ik}x_k| \leqslant \|\mathbf{x}\|_\infty \max_i \sum_k |b_{ik}| \leqslant q\,\|\mathbf{x}\|_\infty.$$

Отсюда следует справедливость представления

$$A^{-1} = (I+B)^{-1}J = (I - B + B^2 - B^3 + \cdots)\,J,$$

так как ряд является сходящимся.

Далее для произвольного вектора $\mathbf{x}$ получаем оценку

$$\begin{aligned}\|A^{-1}\mathbf{x}\|_\infty &= \|(I - B + B^2 - B^3 + \cdots)\,J\mathbf{x}\|_\infty \leqslant \\ &\leqslant \|J\mathbf{x}\|_\infty + q\,\|J\mathbf{x}\|_\infty + q^2\|J\mathbf{x}\|_\infty + \cdots = \frac{1}{1-q}\;\|J\mathbf{x}\|_\infty \leqslant \\ &\leqslant \frac{1}{1-q}\;\frac{1}{\min\limits_i |a_{ii}|}\;\|\mathbf{x}\|_\infty.\end{aligned}$$

Следовательно,

$$\operatorname{cond}_\infty(A) = \|A\|_\infty\|A^{-1}\|_\infty \leqslant \frac{1+q}{1-q}\;\frac{\max\limits_i |a_{ii}|}{\min\limits_i |a_{ii}|}\;. \qquad \rhd$$

Ответ: $\dfrac{1}{1+q}\;\dfrac{\max|a_{ii}|}{\min|a_{ii}|} \leqslant \operatorname{cond}_\infty(A) \leqslant \dfrac{1+q}{1-q}\;\dfrac{\max|a_{ii}|}{\min|a_{ii}|}\;.$

5.66. Пусть R — верхняя треугольная матрица размерности $n\times n$, у которой: 1) $|r_{ij}|\leqslant 1$ для всех i, j; 2) $r_{ii}=1$ для всех i. Найти максимально возможное значение числа обусловленности $\operatorname{cond}_\infty(R)$.

$\lhd$ Рассмотрим вспомогательные матрицы A_k размерности $(k+1)\times(k+1)$ с элементами $|a_{ij}|\leqslant 1$ следующей структуры:

$$a_{ij}=\begin{cases} a_{ii} & \text{при} \quad i=j,\\ 1 & \text{при} \quad i=j+1,\\ 0 & — \quad \text{иначе.}\end{cases}$$

Для определителя A_k из разложения по первому столбцу следует оценка

$$|\det(A_k)|\leqslant |a_{11}|\left|\det\left(A_{k-1}^{(1)}\right)\right|+\left|\det\left(A_{k-1}^{(2)}\right)\right|\leqslant 2|\det(A_{k-1})|\leqslant$$
$$\leqslant 4|\det(A_{k-2})|\leqslant\ldots\leqslant 2^k,$$

поскольку

$$|\det(A_1)|=\left|\det\begin{pmatrix} a_{11} & a_{12}\\ 1 & a_{22}\end{pmatrix}\right|\leqslant 2,\quad |\det(A_0)|=|a_{11}|\leqslant 1.$$

Выше было использовано обозначение $A_{k-1}^{(l)}$ $(l=1,2)$ для подматриц k-го порядка, получающихся из исходной матрицы A_k вычеркиванием первого столбца и l-й строки.

Рассмотрим теперь обратную к R матрицу R^{-1} с элементами

$$r_{ij}^{(-1)}=\begin{cases} 1 & \text{при} \quad i=j,\\ 0 & \text{при} \quad i>j,\\ q_{ij} & \text{при} \quad i<j.\end{cases}$$

Так как $\det(R)=1$, то q_{ij} имеет смысл алгебраического дополнения элемента r_{ji} в определителе матрицы R. При этом его значение равно (с точностью до знака) определителю матрицы, у которой диагональные элементы не превышают единицы, на нижней побочной диагонали ровно $k=j-i-1$ единиц, а остальные элементы равны нулю. Отсюда имеем

$$|q_{ij}|\leqslant|\det(A_{j-i-1})|\leqslant 2^{j-i-1}.$$

Рассмотрим случай максимально возможных значений q_{ij}:

$$R^{-1}=\begin{pmatrix} 1 & 1 & 2 & \ldots & 2^{n-3} & 2^{n-2}\\ 0 & 1 & 1 & \ldots & 2^{n-4} & 2^{n-3}\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & 0 & \ldots & 0 & 1\end{pmatrix}.$$

При этом исходная матрица R однозначно определяется как

$$R=\begin{pmatrix} 1 & -1 & -1 & \ldots & -1\\ 0 & 1 & -1 & \ldots & -1\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & 0 & \ldots & 1\end{pmatrix}.$$

Легко проверить, что

$$\|R^{-1}\|_\infty = 1 + 1 + 2 + \cdots + 2^{n-2} = 2^{n-1}, \quad \|R\|_\infty = n,$$

т. е. построена матрица, на которой одновременно достигаются максимально возможные значения как $\|R\|_\infty$, так и $\|R^{-1}\|_\infty$ среди всех матриц из заданного класса. $\triangleright$

О т в е т: $\max \operatorname{cond}_\infty(R) = n\,2^{n-1}$.

5.67. Показать, что определитель D_n *матрицы Коши* K_n с элементами $k_{ij} = \dfrac{1}{a_i + b_j}$, $1 \leqslant i, j \leqslant n$ равен

$$D_n = \prod_{1 \leqslant i < j \leqslant n} (a_i - a_j)(b_i - b_j) \left(\prod_{1 \leqslant i,j \leqslant n} (a_i + b_j) \right)^{-1}.$$

$\triangleleft$ Вычтем первый столбец определителя последовательно из второго, третьего, ..., n-го столбцов, а затем вынесем $b_1 - b_2$ за знак определителя из второго столбца, $b_1 - b_3$ — из третьего и т. д. Затем вынесем $(a_1 + b_1)^{-1}$ из первой строки, $(a_2 + b_1)^{-1}$ — из второй строки и т. д. Далее вычтем первую строку последовательно из второй, третьей, ..., n-й строки, а затем вынесем за знак определителя

$$(a_1 - a_2) \ldots (a_1 - a_n)(a_1 + b_2)^{-1} \ldots (a_1 + b_n)^{-1}.$$

В результате останется определитель матрицы Коши $(n-1)$-го порядка. Поэтому искомая формула получается по индукции. $\triangleright$

5.68. Пусть задана матрица Гильберта H_n с элементами $h_{ij} = \dfrac{1}{i + j - 1}$, $1 \leqslant i, j \leqslant n$. Показать, что элементами матрицы H_n^{-1} являются целые числа, которые можно вычислить по формуле

$$a_{ij} = (-1)^{i+j} \frac{(i + n - 1)!\,(j + n - 1)!}{[(i-1)!]^2 [(j-1)!]^2 (n-i)!(n-j)!(i+j-1)}.$$

$\triangleleft$ Рассмотрим матрицу Коши K_n с элементами $k_{ij} = \dfrac{1}{a_i + b_j}$, $1 \leqslant i, j \leqslant n$. Ее определитель вычислен в 5.67. Элементы матрицы K_n^{-1} являются отношениями алгебраических дополнений к определителю исходной матрицы. Миноры матрицы Коши снова являются матрицами Коши. Поэтому можно получить явные выражения для элементов K_n^{-1}:

$$b_{ij} = \prod_{k=1}^{n} (a_j + b_k)(a_k + b_i) \left[(a_j + b_i) \prod_{k \neq j} (a_j - a_k) \prod_{k \neq i} (b_i - b_k) \right]^{-1}.$$

Полагая $a_i = i$, $b_i = i - 1$, получим частный случай матрицы Коши — матрицу Гильберта и искомую формулу для элементов H_n^{-1}. $\triangleright$

5.69. Оценить рост числа обусловленности $\mathrm{cond}_\infty(H_n)$ матрицы Гильберта с элементами $h_{ij} = \frac{1}{i+j-1}$, $1 \leqslant i, j \leqslant n$ относительно параметра размерности n.

У к а з а н и е. Величина $\frac{(i+n-1)!}{((i-1)!)^2(n-i)!}$ принимает максимальное значение при $i = \left[\frac{n}{\sqrt{2}}\right]$, поэтому для элементов a_{ij} матрицы H_n^{-1} (см. 5.68) по формуле Стирлинга

$$n! = \sqrt{2\,\pi\,n}\,n^n \mathrm{e}^{-n}\mathrm{e}^{\theta(n)}\,, \quad |\theta(n)| \leqslant \frac{1}{12\,n}\,,$$

имеем асимптотику

$$\max_{i,j}|a_{ij}| = \frac{1}{4\sqrt{2}\pi^2 n}\left(\sqrt{2}+1\right)^{4n}\left(1+O\left(\frac{1}{n}\right)\right).$$

Отсюда следует равенство

$$\|H_n^{-1}\|_\infty = \frac{1}{(2\,\pi)^{3/2}2^{7/4}\sqrt{n}}\left(\sqrt{2}+1\right)^{4n}\left(1+O\left(\frac{1}{n}\right)\right).$$

Так как

$$\ln n \leqslant \|H_n\|_\infty = \sum_{j=1}^{n}\frac{1}{j} \leqslant 3\,\ln n \quad (\text{для}\, n \geqslant 2)\,,$$

то главный член асимптотики $\mathrm{cond}_\infty(H_n)$ имеет вид $\mathrm{const}\cdot 4^n \ln\frac{n}{\sqrt{n}}$.

5.70. Доказать *неравенство Адамара* для квадратных матриц вида $A = A^T > 0$:

$$\det(A) \leqslant \prod_{i=1}^{n} a_{ii}\,.$$

$\lhd$ Положим $d_i = \frac{1}{\sqrt{a_{ii}}}$ и пусть $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$. Неравенство $\det(A) \leqslant a_{11}a_{22}\ldots a_{nn}$ равносильно условию $\det(DAD) \leqslant 1$, и в дальнейшем достаточно рассматривать матрицу A, все диагональные элементы которой равны единице. Если $\lambda_1, \lambda_2, \ldots, \lambda_n$ — собственные значения матрицы A (обязательно положительные), то

$$\det(A) = \prod_{i=1}^{n}\lambda_i \leqslant \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right)^n = \left(\frac{1}{n}\ \mathrm{tr}A\right)^n = 1\,.$$

Здесь мы воспользовались неравенством между арифметическим и геометрическим средними неотрицательных чисел. Равенство средних имеет место тогда и только тогда, когда все $\lambda_i = 1$. В силу симметрии, матрица A диагонализуема. При единичных диагональных элементах и собственных значениях это равносильно тому, что A является единичной матрицей. Соответственно равенство в исходном неравенстве достигается тогда и только тогда, когда A — диагональная матрица. $\rhd$

5.71. Показать, что для произвольной квадратной матрицы C справедливы неравенства

$$|\det(C)| \leqslant \prod_{i=1}^{n} \left(\sum_{j=1}^{n} |c_{ij}|^2 \right)^{1/2}, \quad |\det(C)| \leqslant \prod_{j=1}^{n} \left(\sum_{i=1}^{n} |c_{ij}|^2 \right)^{1/2},$$

а равенства в них достигаются тогда и только тогда, когда строки (соответственно столбцы) матрицы C попарно ортогональны.

$\lhd$ Если матрица C вырождена, то доказывать нечего. В случае невырожденной матрицы C нужно применить неравенство из 5.70 к положительно определенной матрице $A = C\,C^T$ и извлечь квадратный корень из обеих частей неравенства. Правая часть доказываемого неравенства — квадратный корень из произведения диагональных элементов матрицы A, а левая часть — квадратный корень из определителя этой матрицы. Строки матрицы C попарно ортогональны тогда и только тогда, когда A — диагональная матрица, а это и есть случай равенства в 5.70. Второе искомое неравенство получается применением первого к матрице C^T. $\rhd$

5.3. Точные методы

К точным методам решения системы $A\mathbf{x} = \mathbf{b}$ линейных алгебраических уравнений относятся алгоритмы, которые при отсутствии ошибок округления, позволяют точно вычислить искомый вектор $\mathbf{x}$ за конечное число логических и арифметических операций. Если число ненулевых элементов матрицы имеет порядок n^2, то большинство алгоритмов такого рода позволяют найти решение за $O(n^3)$ арифметических действий. Данная оценка, а также необходимость хранения всех элементов матрицы в памяти компьютера накладывают существенное ограничение на область применимости точных методов. Однако для решения задач размерности n менее 10^4 разумно применять точные алгоритмы. При численном решении задач математической физики часто требуется обращать матрицы блочно-диагонального вида. В этом случае удается построить точные методы с меньшим по порядку числом арифметических действий. К таким алгоритмам относят методы прогонки, стрельбы, Фурье (базисных функций).

Наиболее известным из точных методов, применяемых для решения задач с матрицами общего вида, является *метод исключения Гаусса*. В предположении, что коэффициент $a_{11} \neq 0$, уравнения исходной системы заменяем следующими:

$$\begin{cases} x_1 + \sum\limits_{j=2}^{n} \dfrac{a_{1j}}{a_{11}}\, x_j = \dfrac{b_1}{a_{11}}\,, \\ \sum\limits_{j=2}^{n} \left(a_{ij} x_j - \dfrac{a_{1j}}{a_{11}}\, a_{i1} x_j \right) = b_i - \dfrac{b_1}{a_{11}}\, a_{i1}, \quad i = 2, \ldots, n, \end{cases}$$

т. е. первое уравнение делим на a_{11}, затем, умноженное на соответствующий коэффициент a_{i1}, вычитаем из последующих уравнений. В полученной системе $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$ неизвестное x_1 исключено из всех уравнений,

кроме первого. Далее при условии, что коэффициент $a_{22}^{(1)}$ матрицы $A^{(1)}$ отличен от нуля, исключаем x_2 из всех уравнений, кроме первого и второго, и т. д. В результате получаем систему $A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$ с верхней треугольной матрицей. Данную последовательность вычислений называют *прямым ходом метода Гаусса*. Из последнего уравнения приведенной системы определяем компоненту решения x_n. Далее подставляем x_n в $(n-1)$-е уравнение, находим x_{n-1} и т. д. Соответствующую последовательность вычислений называют *обратным ходом метода Гаусса*. Если на k-м шаге прямого хода коэффициент $a_{kk}^{(k-1)}$ равен нулю, то k-ю строку уравнения переставляют с произвольной l-й строкой, $l > k$ — с ненулевым коэффициентом $a_{lk}^{(k-1)}$ при x_k. Такая строка всегда найдется, если $\det(A) \neq 0$.

Если на k-м шаге прямого хода диагональный элемент $a_{kk}^{(k-1)}$ отличен от нуля, но его абсолютное значение мало, то коэффициенты очередной матрицы $A^{(k)}$ будут вычислены с большой абсолютной погрешностью. Полученное в результате решение может значительно отличаться от точного. Поэтому при практической реализации метода Гаусса требуют на каждом шаге прямого хода переставлять на k-е место строку с максимальным по модулю элементом $a_{lk}^{(k-1)}$ среди всех $l \geqslant k$. Такую модификацию называют *методом Гаусса с частичным выбором главного элемента*. Данный алгоритм позволяет гарантированно найти приближенное решение $\tilde{\mathbf{x}}$ с малой нормой невязки $\|\mathbf{b} - A\tilde{\mathbf{x}}\|$ но, возможно, с большой относительной ошибкой $\dfrac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|}$.

5.72. Показать, что реализации прямого и обратного хода метода Гаусса требуют по порядку $\frac{2}{3}n^3$ и n^2 арифметических действий соответственно.

Указание. Число умножений прямого хода приближенно равно $n^2 + (n-1)^2 + \ldots + 1 \approx \int\limits_0^n x^2 dx = \frac{n^3}{3}$, имеем столько же сложений.

5.73. Показать, что прямой ход метода Гаусса соответствует последовательному умножению исходной системы на некоторые диагональные матрицы C_k и нижние треугольные матрицы C'_k. Определить вид матриц C_k и C'_k.

Указание. Матрица C_k получается из матрицы I заменой диагонального элемента k-й строки на элемент $\left(a_{k,k}^{(k-1)}\right)^{-1}$. Матрица C'_k получается из матрицы I заменой k-го столбца на столбец $\left(0, \ldots, 1, -a_{k+1,k}^{(k-1)}, -a_{k+2,k}^{(k-1)}, \ldots, -a_{n,k}^{(k-1)}\right)^T$.

Таким образом, метод Гаусса соответствует неявному разложению исходной матрицы A на произведение нижней треугольной матрицы L и верхней треугольной матрицы R с $r_{kk} = 1$. Действительно, как следует из 5.73, $CA = R$, где R — верхняя треугольная матрица с единичной

диагональю, а $C = C_n C'_{n-1} C_{n-1} \dots C'_1 C_1$ — нижняя треугольная матрица. Поэтому $A = LR$, где $L = C^{-1}$. Аналогично можно построить разложение $A = LR$ с $l_{kk} = 1$.

5.74. Показать, что прямой ход метода Гаусса с частичным выбором главного элемента соответствует последовательному умножению исходной системы на некоторые диагональные матрицы C_k, нижние треугольные матрицы C'_k и матрицы перестановок P_k. Определить вид матриц C_k, C'_k и P_k.

Ответ: Матрицы C_k, C'_k совпадают с матрицами из 5.73, матрицы P_k получаются из единичной матрицы I некоторой перестановкой строк.

5.75. Пусть система $A\mathbf{x} = \mathbf{b}$ с матрицей $A = \begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix}$ решается методом LR-разложения: $A = LR$, $L\mathbf{y} = \mathbf{b}$, $R\mathbf{x} = \mathbf{y}$. Вычислить $\operatorname{cond}_\infty(L)$ и $\operatorname{cond}_\infty(R)$, если LR-разложение строится методом Гаусса: а) без выбора главного элемента; б) с выбором главного элемента.

$\lhd$ а) Применим схему без выбора главного элемента:

$$\begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{pmatrix}.$$

Отсюда для определения элементов матриц L и R получаем систему линейных алгебраических уравнений $l_{11}r_{11} = \varepsilon$, $l_{11}r_{12} = 1$, $l_{21}r_{11} = 1$, $l_{21}r_{12} + l_{22}r_{22} = 1$. Для определенности положим $l_{11} = l_{22} = 1$. Тогда

$$L = \begin{pmatrix} 1 & 0 \\ \frac{1}{\varepsilon} & 1 \end{pmatrix}, \qquad R = \begin{pmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{1}{\varepsilon} \end{pmatrix},$$

$$L^{-1} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{\varepsilon} & 1 \end{pmatrix}, \quad R^{-1} = \begin{pmatrix} \frac{1}{\varepsilon} & -\frac{1}{\varepsilon - 1} \\ 0 & \frac{\varepsilon}{\varepsilon - 1} \end{pmatrix}.$$

Отсюда

$$\operatorname{cond}_\infty(L) = \left(1 + \frac{1}{\varepsilon}\right)^2, \quad \operatorname{cond}_\infty(R) = \frac{1}{\varepsilon^2}.$$

б) Воспользуемся LR-разложением с выбором главного элемента:

$$\begin{pmatrix} 1 & 1 \\ \varepsilon & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ l_{21} & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{pmatrix},$$

$$L = \begin{pmatrix} 1 & 0 \\ \varepsilon & 1 \end{pmatrix}, \qquad R = \begin{pmatrix} 1 & 1 \\ 0 & 1 - \varepsilon \end{pmatrix},$$

$$L^{-1} = \begin{pmatrix} 1 & 0 \\ -\varepsilon & 1 \end{pmatrix}, \quad R^{-1} = \begin{pmatrix} 1 & -\frac{1}{1 - \varepsilon} \\ 0 & \frac{1}{1 - \varepsilon} \end{pmatrix}.$$

Отсюда

$$\operatorname{cond}_\infty(L) = (1 + \varepsilon)^2, \quad \operatorname{cond}_\infty(R) = 2\left(1 + \frac{1}{1 - \varepsilon}\right). \qquad \rhd$$

5.76. Доказать, что для невырожденной матрицы A существуют матрицы перестановок P_1 и P_2, нижняя треугольная матрица L и верхняя треугольная матрица R такие, что $P_1AP_2 = LR$. Показать, что достаточно использовать одну из матриц P_i.

Указание. В матрице P_1A переставлены строки исходной матрицы A, а в матрице AP_2 — столбцы. Для того чтобы матрица имела LR-разложение, необходимо и достаточно, чтобы все ее ведущие подматрицы (в том числе и A) были невырожденные.

Если методом Гаусса получено некоторое приближенное решение $\tilde{\mathbf{x}}$, то можно выполнить следующий процесс уточнения. Найдем вектор невязки $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$ с удвоенным количеством значащих цифр и решим систему $A\mathbf{z} = \frac{\mathbf{r}}{\|\mathbf{r}\|}$. Положим $\tilde{\mathbf{x}} := \tilde{\mathbf{x}} + \|\mathbf{r}\|\mathbf{z}$. Процесс уточнения значительно экономичнее, чем решение исходного уравнения, так как LR-разложение матрицы A уже имеется. Уточнение можно повторять до тех пор, пока убывает норма вектора невязки.

5.77. Пусть вещественная матрица A симметрична и положительно определена. Записать формулы для решения системы $A\mathbf{x} = \mathbf{b}$, основанные на разложении $A = R^TR$ с верхней треугольной матрицей R.

$\lhd$ Определим элементы матрицы R. В силу формулы умножения матриц имеем
$$a_{ij} = r_{1i}r_{1j} + r_{2i}r_{2j} + \ldots + r_{ii}r_{ij} \quad \text{при} \quad i < j,$$
$$a_{ii} = r_{1i}^2 + r_{2i}^2 + \ldots + r_{ii}^2 \quad \text{при} \quad i = j.$$
Отсюда получаем формулы для определения r_{ij}:
$$r_{11} = \sqrt{a_{11}}, \quad r_{1j} = \frac{a_{1j}}{r_{11}} \quad (j > 1),$$
$$r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2} \quad (i > 1),$$
$$r_{ij} = \frac{a_{ij} - \sum\limits_{k=1}^{i-1} r_{ki}r_{kj}}{r_{ii}} \quad (j > i),$$
$$r_{ij} = 0 \quad (i > j).$$
Дальнейшее решение исходной системы сводится к последовательному решению двух систем с треугольными матрицами: $R^T\mathbf{y} = \mathbf{b}$ и $R\mathbf{x} = \mathbf{y}$.

Элементы вектора $\mathbf{y}$ определяем по рекуррентным формулам аналогично r_{ij}:
$$y_1 = \frac{b_1}{r_{11}}, \quad y_i = \frac{b_i - \sum\limits_{k=1}^{i-1} r_{ki}y_k}{r_{ii}} \quad (i > 1).$$
Окончательное решение $\mathbf{x}$ находим по формулам
$$x_n = \frac{y_n}{r_{nn}}, \quad x_i = \frac{y_i - \sum\limits_{k=i+1}^{n} r_{ik}y_k}{r_{ii}} \quad (i < n).$$
Описанный алгоритм часто называют *методом Холецкого.* $\rhd$

5.78. Показать, что реализации прямого и обратного хода метода Холецкого требуют по порядку $\frac{1}{3}n^3$ и n^2 арифметических действий соответственно.

5.79. Пусть A — вещественная симметричная матрица. Записать формулы для вычисления матричного разложения $A = R^T D R$ с верхней треугольной матрицей R и диагональной матрицей D с элементами $d_{ii} = \pm 1$.

Ответ: последовательно для $i = 1, \ldots, n$ вычислим

$$d_{ii} = \operatorname{sign}\left(a_{ii} - \sum_{k=1}^{i-1} |r_{ki}|^2 d_{kk}\right), \quad r_{ii} = \left|a_{ii} - \sum_{k=1}^{i-1} |r_{ki}|^2 d_{kk}\right|^{1/2},$$

$$r_{ij} = \frac{a_{ij} - \sum\limits_{k=1}^{i-1} r_{ki} r_{kj} d_{kk}}{r_{ii} d_{ii}} \quad \text{для } i < j\,.$$

В этих формулах, как обычно, если верхний индекс суммирования меньше нижнего, то сумму полагают равной нулю.

5.80. Для матрицы

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & -5 & -6 \\ 3 & -6 & 18 \end{pmatrix}$$

вычислить элементы разложения $A = R^T D R$ с верхней треугольной матрицей R и диагональной матрицей D с элементами $d_{ii} = \pm 1$.

Ответ: $D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, $\quad R = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 3 & 4 \\ 0 & 0 & 5 \end{pmatrix}$.

Среди точных методов, требующих для реализации порядка $O(n^3)$ действий, одним из наиболее устойчивых к вычислительной погрешности является *метод отражений*.

Пусть имеем некоторый единичный вектор $\mathbf{w} \in \mathbf{R}^n$, $\|\mathbf{w}\|_2 = 1$. Построим по нему следующую матрицу: $U = I - 2\mathbf{w}\,\mathbf{w}^T$, называемую *матрицей Хаусхолдера*. Здесь I — единичная матрица, $\Omega = \mathbf{w}\,\mathbf{w}^T$ — матрица с элементами $\omega_{ij} = w_i w_j$, являющаяся результатом произведения вектор-столбца $\mathbf{w}$ на вектор-строку $\mathbf{w}^T$.

5.81. Доказать, что матрица U является симметричной и ортогональной матрицей, т. е. $U = U^T$ и $U^T U = I$, и все ее собственные значения равны ± 1.

Указание. Симметричность U следует из явного вида U. Так как $(\mathbf{w}, \mathbf{w}) = 1$, то $\Omega\Omega|_{ij} = \sum\limits_{k=1}^{n} w_i w_k w_k w_j = \Omega|_{ij}$ и

$$UU = I - 4\Omega + 4\Omega\Omega = I, \text{ т. е. } U^2 = U^T U = I.$$

5.82. Показать, что $U\mathbf{w} = -\mathbf{w}$, а если вектор $\mathbf{v}$ ортогонален $\mathbf{w}$, то $U\mathbf{v} = \mathbf{v}$.

5.83. Показать, что образ $U\mathbf{y}$ произвольного вектора $\mathbf{y}$ является зеркальным отражением относительно гиперплоскости, ортогональной вектору $\mathbf{w}$.

$\lhd$ Представим $\mathbf{y}$ в виде $\mathbf{y} = (\mathbf{y}, \mathbf{w})\mathbf{w} + \mathbf{v}$. Тогда из 5.82 следует $U\mathbf{y} = -(\mathbf{y}, \mathbf{w})\mathbf{w} + \mathbf{v}$. $\rhd$

5.84. Для векторов единичной длины $\mathbf{y}$ и $\mathbf{e}$ найти вектор $\mathbf{w}$ такой, что $U\mathbf{y} = \mathbf{e}$, где $U = I - 2\mathbf{w}\,\mathbf{w}^T$.

$\lhd$ Заметим, что $\mathbf{w} = \pm\dfrac{\mathbf{y} - \mathbf{e}}{\sqrt{(\mathbf{y} - \mathbf{e}, \mathbf{y} - \mathbf{e})}}$. Действительно, $(I - 2\mathbf{w}\,\mathbf{w}^T)\,\mathbf{y} = \mathbf{y} - \xi = \mathbf{e}$, так как

$$\xi_i = \frac{2\sum\limits_{k=1}^{n}(y_i - e_i)(y_k - e_k)y_k}{(\mathbf{y} - \mathbf{e}, \mathbf{y} - \mathbf{e})} = \frac{2(y_i - e_i)(1 - (\mathbf{y}, \mathbf{e}))}{2 - 2(\mathbf{y}, \mathbf{e})} = y_i - e_i.$$

Так как преобразование U не меняет длины вектора, то для неединичного вектора $\mathbf{y}$ имеем $U\mathbf{y} = \alpha\mathbf{e}$, $\alpha = \|\mathbf{y}\|_2$, и искомыми являются векторы $\mathbf{w} = \pm\,\dfrac{\mathbf{y} - \alpha\mathbf{e}}{\|\mathbf{y} - \alpha\mathbf{e}\|_2}$. $\rhd$

5.85. (Метод отражений). Показать, что произвольная квадратная матрица A может быть приведена к верхнему треугольному виду в результате последовательного умножения слева на ортогональные матрицы Хаусхолдера.

У к а з а н и е. По векторам $\mathbf{y}_1 = (a_{1,1}, \ldots, a_{n,1})^T$ и $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ можно построить вектор $\mathbf{w}_1$ и соответствующую матрицу U_1 (см. 5.84) так, чтобы первый столбец матрицы $A^{(1)} = U_1 A$ был пропорционален вектору $\mathbf{e}_1 \in \mathbf{R}^n$, т. е. $U_1\mathbf{y_1} = \pm\alpha_1\mathbf{e}_1$. Вычислим $\alpha_1 = (a_{1,1}^2 + a_{2,1}^2 + \ldots + a_{n,1}^2)^{1/2}$ и определим $\tilde{\mathbf{w}}_1 = \left(\dfrac{a_{1,1}}{\alpha_1} + \operatorname{sign}(a_{1,1}), \dfrac{a_{2,1}}{\alpha_1}, \ldots, \dfrac{a_{n,1}}{\alpha_1}\right)^T$, $\mathbf{w}_1 = \dfrac{\tilde{\mathbf{w}}_1}{\|\tilde{\mathbf{w}}_1\|_2}$. Такой выбор знака и предварительная нормировка на α_1 гарантируют малость вычислительной погрешности и устойчивость алгоритма.

Далее в пространстве $\mathbf{R}^{n-1}$ по вектору $\mathbf{y}_2 = (a_{2,2}^{(1)}, \ldots, a_{2,n}^{(1)})^T$ построить матрицу U_2', отображающую его в вектор, коллинеарный $\mathbf{e}_2 = (1, 0, \ldots, 0)^T \in \mathbf{R}^{n-1}$. Затем определить $U_2 = \begin{pmatrix} 1 & 0 \\ 0 & U_2' \end{pmatrix}$ и рассмотреть матрицу $A^{(2)} = U_2 U_1 A$. И так далее. На k-м шаге имеем $U_k = \begin{pmatrix} I_{k-1} & 0 \\ 0 & U_k' \end{pmatrix}$. Таким образом, матрица отражений U_k строится по вектору $\mathbf{w}_k = \dfrac{\tilde{\mathbf{w}}_k}{\|\tilde{\mathbf{w}}_k\|_2}$, $\mathbf{w}_k \in \mathbf{R}^n$, где $\tilde{\mathbf{w}}_k = \left(0, \ldots, 0, \dfrac{a_{k,k}^{(k-1)}}{\alpha_k} + \operatorname{sign}(a_{k,k}^{(k-1)}), \dfrac{a_{k+1,k}^{(k-1)}}{\alpha_k}, \ldots, \dfrac{a_{n,k}^{(k-1)}}{\alpha_k}\right)^T$, и $\alpha_k = ((a_{k,k}^{(k-1)})^2 + \ldots + (a_{n,k}^{(k-1)})^2)^{1/2}$. В результате преобразований получится верхняя треугольная матрица $R = UA$, где $U = U_{n-1} \ldots U_1$. При практической реализации явное вычисление U_k не требуется, так

как $U_k A^{(k-1)} = A^{(k-1)} - 2\mathbf{w}_k \left(\mathbf{w}_k^T A^{(k-1)}\right)$. При этом изменяются только элементы $a_{i,j}^{(k-1)}$, $k \leqslant i, j \leqslant n$, матрицы $A^{(k-1)}$. Из условия $UU = I$ имеем $A = UR$. Таким образом, произвольная квадратная матрица A может быть представлена в виде произведения симметричной ортогональной матрицы U и верхней треугольной матрицы R.

Рассмотренный алгоритм позволяет привести систему линейных уравнений $A\mathbf{x} = \mathbf{b}$ к виду $R\mathbf{x} = U\mathbf{b}$, а затем найти ее решение обратным ходом метода Гаусса. Пусть решается задача с возмущенной правой частью $A\tilde{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$ и $\|\delta\mathbf{b}\| \ll \|\mathbf{b}\|$. Так как ортогональные преобразования не меняют евклидову норму векторов, то для приведенной системы $R\tilde{\mathbf{x}} = U\mathbf{b} + U\boldsymbol{\delta}\mathbf{b}$ имеем $\|U\boldsymbol{\delta}\mathbf{b}\| = \|\delta\mathbf{b}\| \ll \|\mathbf{b}\| = \|U\mathbf{b}\|$, и относительная погрешность правой части не увеличилась.

5.86. Показать, что реализации прямого и обратного хода метода отражений в общем случае требуют по порядку $\frac{4}{3}n^3$ и n^2 арифметических действий соответственно.

5.87. Записать формулы метода отражений для задачи $A\mathbf{x} = \mathbf{b}$, где

$$A = \begin{pmatrix} c_0 & -b_0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ -a_1 & c_1 & -b_1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & -a_2 & c_2 & -b_2 & \dots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & -a_{N-2} & c_{N-2} & -b_{N-2} & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -a_{N-1} & c_{N-1} & -b_{N-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & -a_N & c_N \end{pmatrix}.$$

Оценить вычислительные затраты алгоритма.

Матрицу A можно привести к виду $A = QR$, где $Q^{-1} = Q^T$ — ортогональная матрица, *методом вращений*, более простым по сравнению с методом отражений.

Элементарной матрицей вращений второго порядка (*матрицей Гивенса*) называют матрицу

$$G(\varphi) = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix},$$

зависящую от некоторого параметра — угла φ.

5.88. Найти элементарную матрицу вращений $G(\varphi)$, переводящую произвольный ненулевой вектор $(a_1, a_2)^T$ в вектор со второй нулевой компонентой: $(-\sqrt{a_1^2 + a_2^2}, 0)^T$.

Ответ: $\cos\varphi = -\dfrac{a_1}{\sqrt{a_1^2 + a_2^2}}$, $\quad \sin\varphi = \dfrac{a_2}{\sqrt{a_1^2 + a_2^2}}$.

5.89. Показать, что при умножении матрицы A слева на матрицу

$$G_{kl}(\varphi) = \begin{pmatrix} 1 & 0 & \dots & 0 & \dots \\ 0 & \cos\varphi & 0 & -\sin\varphi & 0 \\ \dots & 0 & 1 & 0 & \dots \\ 0 & \sin\varphi & 0 & \cos\varphi & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

($g_{kl} = \sin\varphi$, т. е. синусы и косинусы находятся на пересечении строк и столбцов с номерами k и l, остальные диагональные элементы равны единице) можно получить нуль на позиции элемента a_{kl}.

5.90. (Метод вращений). Показать, что произвольная квадратная матрица A может быть приведена к верхнему треугольному виду в результате последовательного умножения слева на ортогональные матрицы вращений.

Указание. $G_{n\,n-1}\dots G_{3\,2}G_{n\,1}\dots G_{3\,1}G_{2\,1}A = R$.

5.91. Показать, что реализации прямого и обратного хода метода вращений в общем случае требуют по порядку $2\,n^3$ и n^2 арифметических действий соответственно.

5.92. Записать формулы метода вращений для задачи $A\mathbf{x} = \mathbf{b}$, где A — матрица из 5.87. Оценить вычислительные затраты алгоритма.

Рассмотренные методы отражений и вращений применяют не только при построении QR-разложения матрицы A, но и для приведения A к специальному виду: $(2p+1)$-диагональному, блочному диагональному, Хессенбергову. На основании данных разложений удается построить эффективные численные методы решения систем линейных уравнений, а также методы вычисления инвариантных подпространств и решения задачи на собственные значения.

5.4. Линейные итерационные методы

Рассмотрим класс итерационных методов решения систем линейных алгебраических уравнений, основанный на сжимающем свойстве оператора перехода. Различные постановки задачи минимизации нормы оператора перехода приводят к различным алгоритмам расчета.

Метод простой итерации. Преобразуем систему линейных алгебраических уравнений

$$A\,\mathbf{x} = \mathbf{b} \tag{5.3}$$

с невырожденной матрицей A к виду

$$\mathbf{x} = B\,\mathbf{x} + \mathbf{c}. \tag{5.4}$$

Если решение системы (5.4) находят как предел последовательности

$$\mathbf{x}^{k+1} = B\,\mathbf{x}^k + \mathbf{c}, \tag{5.5}$$

то такой процесс называют *методом простой итерации*, а матрицу B — *оператором перехода*. Справедливы следующие теоремы о сходимости метода.

Теорема 1. *Если* $\|B\| < 1$, *то система уравнений* (5.4) *имеет единственное решение и итерационный процесс* (5.5) *сходится к решению со скоростью геометрической прогрессии.*

Теорема 2. *Пусть система* (5.4) *имеет единственное решение. Итерационный процесс* (5.5) *сходится к решению системы* (5.4) *при любом начальном приближении тогда и только тогда, когда все собственные значения матрицы* B *по модулю меньше* 1.

Асимптотической скоростью сходимости $R_\infty(B)$ итерационного метода называют величину $R_\infty(B) = -\ln \rho(B)$, где $\rho(B)$ — спектральный радиус (максимальное по модулю собственное значение) оператора перехода B.

Рассмотрим общий способ перехода от системы (5.3) к системе (5.4). Всякая система

$$\mathbf{x} = \mathbf{x} - D\,(A\,\mathbf{x} - \mathbf{b}) \tag{5.6}$$

имеет вид (5.4) и при $\det(D) \neq 0$ равносильна системе (5.3). В то же время всякая система (5.4), равносильная (5.3), записывается в виде (5.6) с матрицей $D = (I - B)\,A^{-1}$.

Оптимальный линейный одношаговый метод. Для систем со знакоопределенными матрицами метод (5.5) обычно строят в виде

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\,\mathbf{x}^k = \mathbf{b}, \quad \text{т. е.} \quad B = I - \tau\,A, \;\; \mathbf{c} = \tau\,\mathbf{b}, \tag{5.7}$$

где τ — итерационный параметр. Так как точное решение $\mathbf{x}$ удовлетворяет уравнению (5.7), то для вектора ошибки $\mathbf{z}^k = \mathbf{x} - \mathbf{x}^k$ справедливы выражения

$$\mathbf{z}^{k+1} = (I - \tau A)\,\mathbf{z}^k, \quad \|\mathbf{z}^{k+1}\| \leqslant \|I - \tau A\|\,\|\mathbf{z}^k\|, \; k = 0, 1, 2, \ldots .$$

Итерационный параметр τ ищется из условия минимума нормы оператора перехода. Если $A = A^T > 0$ и выбрана евклидова векторная норма, то минимизационная задача

$$\min_\tau \Big(\max_{\lambda(A)} |1 - \tau\lambda(A)| \Big) = q$$

решается явно. Пусть известны точные границы спектра матрицы A, т. е. $\lambda(A) \in [m, M]$, тогда оптимальные значения соответственно равны

$$\tau = \frac{2}{m + M}\,, \quad q = \frac{M - m}{M + m} < 1$$

и справедлива оценка

$$\|\mathbf{x} - \mathbf{x}^k\|_2 \leqslant q^k \|\mathbf{x} - \mathbf{x}^0\|_2.$$

Оптимальный линейный N-шаговый метод. Будем считать, что в итерационном алгоритме

$$\frac{\mathbf{x}^{k+1}-\mathbf{x}^k}{\tau_{k+1}}+A\,\mathbf{x}^k=\mathbf{b} \tag{5.8}$$

допускается циклическое изменение (с периодом N) параметра τ в зависимости от номера итерации, т. е. $\tau_1,\tau_2,\ldots,\tau_N,\tau_1,\tau_2,\ldots$. В этом случае после N итераций для вектора ошибки имеем:

$$\mathbf{z}^{k+N}=\prod_{j=1}^{N}(I-\tau_jA)\,\mathbf{z}^k,\quad \|\mathbf{z}^{k+N}\|_2\leqslant\left\|\prod_{j=1}^{N}(I-\tau_jA)\right\|_2\|\mathbf{z}^k\|_2,\ k=0,1,2,\ldots.$$

Будем искать набор τ_j, $j=1,\ldots,N$, из условия минимума нормы оператора перехода после N итераций. Если $A=A^T>0$, то

$$\min_{\tau_j}\left\|\prod_{j=1}^{N}(I-\tau_jA)\right\|_2=\min_{\tau_j}\left(\max_{\lambda(A)}\left|\prod_{j=1}^{N}(1-\tau_j\lambda(A))\right|\right).$$

Пусть известны точные границы спектра матрицы A, т. е. $\lambda(A)\in[m,M]$, тогда оптимальные значения параметров равны обратным величинам корней многочлена Чебышёва степени N на отрезке $[m,M]$:

$$\tau_j^{-1}=\frac{M+m}{2}+\frac{M-m}{2}\cos\frac{\pi(2j-1)}{2N},$$

и справедлива оценка погрешности после N итераций

$$\|\mathbf{x}-\mathbf{x}^N\|_2\leqslant\frac{2q_1^N}{1+q_1^{2N}}\|\mathbf{x}-\mathbf{x}^0\|_2\leqslant2q_1^N\|\mathbf{x}-\mathbf{x}^0\|_2,\quad q_1=\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}.$$

При численной реализации N-шагового метода для устойчивости требуется специальным образом перемешивать значения параметров τ_j.

Недостатком рассмотренных оптимальных методов является требование информации о границах спектра матрицы A.

5.93. Пусть элементы матрицы B имеют вид $b_{kj}=\frac{1}{2}\cdot3^{-|k-j|}$. Доказать, что система $\mathbf{x}=B\mathbf{x}+\mathbf{c}$ имеет единственное решение и метод простой итерации сходится при любом начальном приближении.

Указание. $\|B\|_1=\|B\|_\infty<1$.

5.94. Найти все α, β, при которых метод простой итерации $\mathbf{x}^{k+1}=B\,\mathbf{x}^k+\mathbf{c}$, где

$$B=\begin{pmatrix}\alpha&\beta&0\\ \beta&\alpha&\beta\\ 0&\beta&\alpha\end{pmatrix},$$

сходится с произвольного начального приближения.

◁ Имеем $\det(B-\lambda I)=(\alpha-\lambda)(\alpha-\lambda-\sqrt{2}\beta)(\alpha-\lambda+\sqrt{2}\beta)=0$, $|\alpha|<1$, $|\alpha\pm\sqrt{2}\beta|<1$. ▷

5.95. Привести пример задачи $\mathbf{x} = B\mathbf{x} + \mathbf{c}$ такой, что у матрицы B есть собственное значение λ вне единичного круга, но метод (5.5) сходится при некотором начальном приближении.

$\lhd$ Имеем

$$B = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix},$$

$\lambda(B) = \frac{1}{2},\ \frac{3}{2}$; $\quad \mathbf{x}_{\lambda=1/2} = (1,-1)^T$; $\quad \mathbf{x}^0 - \mathbf{x} = t\,\mathbf{x}_{\lambda=1/2}$ $\quad$ при $\quad t \neq 0$. $\rhd$

5.96. Пусть матрица B в методе (5.5) имеет вид

$$B = \begin{pmatrix} \alpha & 4 \\ 0 & \beta \end{pmatrix} \quad 0 < \alpha, \beta < 1\,.$$

Показать, что величина ошибки $\mathbf{z}^k = \mathbf{x} - \mathbf{x}^k$ в норме $\|\cdot\|_\infty$ начинает монотонно убывать лишь с некоторого номера итерации N. Оценить N при $\alpha = \beta \approx 1$.

Ответ: $N \approx \frac{1}{1-\alpha}$.

5.97. Пусть все собственные значения матрицы A вещественные и положительные: $\lambda(A) > 0$. Доказать сходимость метода

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}$$

при $\tau = \|A\|^{-1}$ с любой матричной нормой, которая подчинена векторной.

$\lhd$ Собственные значения оператора перехода $B = I - \tau A$ имеют вид $\lambda(B) = 1 - \|A\|^{-1}\lambda(A)$. Так как $0 < \lambda(A) \leqslant \|A\|$, то $0 \leqslant \lambda(B) < 1$. $\rhd$

5.98. Доказать, что все собственные значения матрицы A размерности $n \times n$ принадлежат области комплексной плоскости $G(A)$, представляющей собой объединение кругов

$$G_i(A) = \{z : |z - a_{ii}| \leqslant R_i(A) = \sum_{j \neq i} |a_{ij}|\}, \quad i = 1, \ldots, n\,.$$

$\lhd$ Пусть λ — произвольное собственное значение матрицы A и $\mathbf{x}$ — соответствующий ему собственный вектор. Обозначим через x_i максимальную по модулю компоненту вектора $\mathbf{x}$. Если таких компонент несколько, то x_i — любая из них. Из равенства $A\,\mathbf{x} = \lambda\,\mathbf{x}$ следует соотношение $(\lambda - a_{ii})x_i = \sum\limits_{j \neq i} a_{ij}x_j$. Отсюда имеем

$$|\lambda - a_{ii}| \leqslant \sum_{j \neq i} |a_{ij}| \frac{|x_j|}{|x_i|} \leqslant \sum_{j \neq i} |a_{ij}| = R_i(A)\,.$$

Это утверждение называется *теоремой Гершгорина*. Имеется обобщение этого факта. $\rhd$

Теорема. *Если указанное объединение кругов $G(A)$ распадается на несколько связных частей, то каждая такая часть содержит столько собственных значений, сколько кругов ее составляют.*

5.99. Доказать, что все собственные значения матрицы A размерности $n \times n$ принадлежат области $G(A) \bigcap G(A^T)$.

Указание. Собственные значения матриц A и A^T совпадают.

5.100. Доказать, что у матрицы $\begin{pmatrix} 2 & 0{,}4 & 0{,}4 \\ 0{,}3 & 4 & 0{,}4 \\ 0{,}1 & 0{,}1 & 5 \end{pmatrix}$ все собственные значения вещественны и найти интервалы, которым они принадлежат.

Ответ: $1{,}6 \leqslant \lambda_1 \leqslant 2{,}4$, $3{,}5 \leqslant \lambda_2 \leqslant 4{,}5$, $4{,}8 \leqslant \lambda_3 \leqslant 5{,}2$.

5.101. Привести пример, демонстрирующий ложность утверждения: все собственные значения матрицы A размерности $n \times n$ принадлежат объединению кругов

$$|z - a_{ii}| \leqslant \min\{R_i(A), R_i(A^T)\}, \quad i = 1, \ldots, n\,.$$

Ответ: у матрицы $\begin{pmatrix} 0 & 0{,}1 \\ -40 & 5 \end{pmatrix}$ оба собственных значения $\lambda_1 = 1$ и $\lambda_2 = 4$ не принадлежат системе кругов: $|z| \leqslant 0{,}1$, $|z - 5| \leqslant 0{,}1$.

5.102. Доказать, что все собственные значения матрицы A размерности $n \times n$ принадлежат объединению кругов

$$|z - a_{ii}| \leqslant R_i^{\alpha}(A)\, R_i^{1-\alpha}(A^T), \quad i = 1, \ldots, n\,,$$

где α — произвольное число из отрезка $[0, 1]$ (*теорема Островского*).

5.103. Доказать, что все собственные значения матрицы A размерности $n \times n$ принадлежат объединению $\frac{n(n-1)}{2}$ *овалов Кассини*:

$$|z - a_{ii}||z - a_{jj}| \leqslant R_i(A)\, R_j(A), \quad i, j = 1, \ldots, n,\ i \neq j.$$

5.104. Доказать, что если для некоторого i и при всех j выполняются неравенства $|a_{ii} - a_{jj}| > R_i(A) + R_j(A)$, то в круге $|z - a_{ii}| \leqslant R_i(A)$ лежит точно одно собственное значение матрицы A.

5.105. Пусть $p_1, \ldots, p_n$ — положительные числа. Доказать, что собственные значения матрицы A принадлежат объединению кругов

$$|z - a_{ii}| \leqslant \frac{1}{p_i} \sum_{j \neq i} p_j |a_{ij}|, \quad i = 1, \ldots, n.$$

Указание. Пусть $S = \operatorname{diag}(p_1, \ldots, p_n)$ $(\det(S) \neq 0)$, тогда достаточно показать, что собственные значения матриц A и $S^{-1}AS$ совпадают.

5.106. С помощью 5.105 найти интервалы, которым принадлежат собственные значения матрицы

$$\begin{pmatrix} 7 & -16 & 8 \\ -16 & 7 & -8 \\ 8 & -8 & -5 \end{pmatrix}.$$

Указание. Точные собственные значения матрицы -9, -9 и 27.

5.107. Пусть $p_1, \ldots, p_n$ — положительные числа. Получить оценки для спектрального радиуса матрицы A:

$$\rho(A) \leqslant \min_{p_1,\ldots,p_n} \max_{1\leqslant i\leqslant n} \frac{1}{p_i} \sum_{j=1}^{n} p_j |a_{ij}|,$$

$$\rho(A) \leqslant \min_{p_1,\ldots,p_n} \max_{1\leqslant j\leqslant n} \frac{1}{p_j} \sum_{i=1}^{n} p_i |a_{ij}|.$$

5.108. Для матрицы $A = \begin{pmatrix} 1 & 1 \\ -1,5 & 2 \end{pmatrix}$ показать, что $\rho(A) \leqslant \min \|D^{-1}AD\|_\infty$, где минимум берется по всем матрицам $D = \operatorname{diag}(p_1, p_2)$ с положительными p_1, p_2.

5.109. Пусть A — матрица *простой структуры*, т. е. подобна диагональной ($A = QDQ^{-1}$, где столбцы $\mathbf{q}_i$ матрицы Q — собственные векторы матрицы A, а элементы диагональной матрицы D — соответствующие собственные значения, т. е. $d_{ii} = \lambda_i$), и все $\lambda(A) \in [m, M]$, $m > 0$. Доказать, что метод

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}$$

сходится при $0 < \tau < \frac{2}{M}$ с произвольного начального приближения.

$\lhd$ Пусть $\mathbf{z}^k$ — вектор ошибки на k-й итерации. Тогда

$$\mathbf{z}^{k+1} = (I - \tau A)\,\mathbf{z}^k = (Q\,Q^{-1} - \tau\,Q\,D\,Q^{-1})\;\mathbf{z}^k\,.$$

Умножим полученное выражение слева на Q^{-1} и сделаем замену $Q^{-1}\,\mathbf{z}^k = \tilde{\mathbf{z}}^k$. Тогда

$$\tilde{\mathbf{z}}^{k+1} = (I - \tau\,D)\,\tilde{\mathbf{z}}^k\,.$$

Здесь $B = I - \tau\,D$ — диагональная матрица, а ее собственные значения равны $\lambda(B) = 1 - \tau\,\lambda(A)$. Поэтому необходимым и достаточным условием сходимости метода является выполнение неравенства

$$|1 - \tau\,\lambda(A)| < 1 \qquad \forall \lambda(A) \in [m, M]\,,$$

откуда и следует искомый результат. $\rhd$

5.110. Пусть матрица системы $A\,\mathbf{x} = \mathbf{b}$ имеет вид

$$A = \begin{pmatrix} 2 & 0{,}3 & 0{,}5 \\ 0{,}1 & 3 & 0{,}4 \\ 0{,}1 & 0{,}1 & 4{,}8 \end{pmatrix}.$$

Доказать, что метод простой итерации $\mathbf{x}^{k+1} = (I - \tau A)\mathbf{x}^k + \tau\mathbf{b}$ при $0 < \tau < \frac{2}{5}$ сходится с произвольного начального приближения.

Указание. Воспользоваться аналогией с 5.100 и решением 5.109.

5.111. Пусть λ и $\mathbf{q}$ — собственное значение и соответствующий собственный вектор невырожденной матрицы простой структуры A, $\mathbf{x}^0$ — начальное приближение в методе простой итерации (5.7) для решения системы $A\mathbf{x} = \mathbf{b}$. Найти такое значение параметра метода, чтобы в разложении ошибки по собственным векторам коэффициент при векторе $\mathbf{q}$ на первой итерации был равен нулю.

Указание. Выписать оператор перехода для вектора ошибки за один шаг и получить $\tau = \frac{1}{\lambda}$.

5.112. Пусть для невырожденной матрицы простой структуры A порядка n известны все собственные значения $\lambda_1, \ldots, \lambda_n$. Построить итерационный метод (5.8) с переменными параметрами τ_k, который не более чем за n шагов приводил бы в точной арифметике к решению системы $A\mathbf{x} = \mathbf{b}$.

Указание. Разложить ошибку $\mathbf{x} - \mathbf{x}^k$ по базису $\{\mathbf{q}_i\}$ из собственных векторов матрицы A. Выбор $\tau_k = \lambda_k^{-1}$, $k = 1, \ldots, n$, обеспечивает на каждом шаге обнуление коэффициента при векторе $\mathbf{q}_k$ в разложении ошибки (см. 5.111).

5.113. Пусть в задаче $A\mathbf{x} = \mathbf{b}$ с матрицей простой структуры у матрицы A имеется одно отрицательное собственное значение $\lambda_1 \in [-2 - \varepsilon, -2 + \varepsilon]$, $\varepsilon = 0{,}01$, а остальные значения — положительные: $\lambda_i \in [1, 3]$, $i = 2, \ldots, n$. Предложить итерационный метод (5.8) для решения такой системы.

5.114. Для решения системы $\mathbf{x} = B\mathbf{x} + \mathbf{c}$ рассмотрим алгоритм с некоторым начальным приближением $\mathbf{x}^0$:

$$\mathbf{y}^{k+1} = B\mathbf{x}^k + \mathbf{c}, \quad \mathbf{x}^{k+1} = \alpha\mathbf{x}^k + (1 - \alpha)\mathbf{y}^{k+1}.$$

Пусть $B = B^T$ и $\lambda(B) \in [m, M]$, $m > 1$. Найти оптимальное значение итерационного параметра α.

$\lhd$ Имеем

$$\mathbf{x}^{k+1} = (\alpha I + (1 - \alpha)B)\mathbf{x}^k + (1 - \alpha)\mathbf{c},$$

$$\min_\alpha \varphi(\alpha) = \min_\alpha \max_\lambda |\alpha + (1 - \alpha)\lambda|, \; \alpha = \frac{m + M}{m + M - 2}. \qquad \rhd$$

5.115. Построить квадратную матрицу A размерности 31×31 с элементами $|a_{ij}| \leqslant 1$ и собственными значениями $|\lambda(A)| \leqslant 1$ такую, что $\|A^{30}\|_\infty \geqslant 10^9$.

Ответ: $a_{ij} = \begin{cases} 1 & \text{при} \quad i = j, \\ 1 & \text{при} \quad i+1 = j, \\ 0 & — \quad \text{иначе.} \end{cases}$

5.116. Пусть A — невырожденная матрица размерности $n \times n$ и X_0 — произвольная матрица размерности $n \times n$. Рассмотрим итерационный процесс

$$X_{k+1} = X_k + X_k(I - AX_k), \;\; k = 0, 1, \ldots .$$

Доказать, что $\lim\limits_{k\to\infty} X_k = A^{-1}$ тогда и только тогда, когда спектральный радиус матрицы $I - AX_0$ меньше 1. При этом $I - AX_k = (I - AX_0)^{2^k}$, $k = 0, 1, \ldots$. Доказать также, что если $AX_0 = X_0A$, то $AX_k = X_kA$ для всех k.

$\lhd$ Для приближений нетрудно получить равенство

$$I - AX_{k+1} = (I - AX_k)^2.$$

Пусть $X_k \to A^{-1}$. Тогда $I - AX_k \to 0$ и $(I - AX_0)^{2^k} \to 0$ при $k \to \infty$. Если допустить, что $\rho(I - AX_0) \geqslant 1$, то для собственного вектора $\mathbf{x}$, соответствующего собственному числу λ, $|\lambda| \geqslant 1$, вектор $(I - AX_0)^{2^k}\mathbf{x} = \lambda^{2^k}\mathbf{x}$ не стремится к нулю, т. е. имеет место противоречие.

Пусть теперь $\rho(I - AX_0) < 1$, тогда найдется (см. 5.41) норма матрицы $||\cdot||_*$, для которой $||I - AX_0||_* = q < 1$ и $||I - AX_k||_* \leqslant q^{2^k} \to 0$.

Для доказательства равенства $AX_k = X_kA$ при условии $AX_0 = X_0A$ воспользуемся индукцией. $\rhd$

5.117. При каких значениях параметра τ метод $\mathbf{x}^{k+1} = (I - \tau A)\mathbf{x}^k + \tau\mathbf{b}$ для системы уравнений $A\,\mathbf{x} = \mathbf{b}$ с матрицей:

$$1)\;\; A = \begin{pmatrix} 5 & 0,8 & 4 \\ 2,5 & 2 & 0 \\ 2 & 0,8 & 4 \end{pmatrix}; \quad 2)\;\; A = \begin{pmatrix} 2 & 1 & 0,5 \\ 3 & 5 & 1 \\ 1 & 3 & 3 \end{pmatrix};$$

$$3)\;\; A = \begin{pmatrix} 1 & 0,5 & 0,3 \\ 1 & 3 & 0 \\ 1 & 1 & 2 \end{pmatrix}; \quad 4)\;\; A = \begin{pmatrix} 3 & 1,2 & 0,8 \\ 1,4 & 2 & 0,1 \\ 0,6 & 0,4 & 1 \end{pmatrix}$$

сходится с произвольного начального приближения?

5.118. Пусть $A = A^T > 0$ и $\lambda(A) \in [m, M]$, $m > 0$. Записать наилучший по скорости сходимости в норме $\|\cdot\|_2$ итерационный процесс вида

$$\mathbf{x}^{k+1} = \mathbf{x}^k - P_1(A)(A\mathbf{x}^k - \mathbf{b}), \;\; P_1(t) = \alpha t + \beta.$$

5.119. Пусть приближения метода $\mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c}, ||B|| < 1$, сходятся к решению $\mathbf{x}$. Доказать, что

$$||\mathbf{x} - \mathbf{x}^k|| \leqslant ||(I - B)^{-1}||\;||\mathbf{x}^{k+1} - \mathbf{x}^k||.$$

Указание. Вывод оценки следует из равенства $\mathbf{x}-\mathbf{x}^k = (I-B)^{-1}(\mathbf{x}^{k+1}-\mathbf{x}^k)$.

5.120. Для приближений метода $\mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c}, ||B|| < 1$, доказать оценку $||\mathbf{x}^k|| \leqslant ||B||^k ||\mathbf{x}^0|| + \dfrac{||\mathbf{c}||}{1 - ||B||}$.

5.121. Пусть $A = I - B,\ b_{ij} \geqslant 0$. Доказать, что если все компоненты векторов $\mathbf{x}$ и $\mathbf{c}$ из задачи $A\mathbf{x} = \mathbf{c}$ неотрицательны, то приближения метода $\mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c},\ \mathbf{x}^0 = 0$, сходятся к $\mathbf{x}$.

$\lhd$ В силу того, что $A\mathbf{x} = \mathbf{c}$, $A = I - B$, неотрицательности решения $\mathbf{x}$ и элементов матрицы B, справедливо неравенство $\mathbf{x} \gg B^n\mathbf{c} + B^{n-1}\mathbf{c} + \cdots + \mathbf{c}$ для любого n (здесь использован знак $\gg$ для покомпонентного неравенства векторов; аналогичный смысл имеет знак $\ll$). С другой стороны, при $\mathbf{x}^0 = 0$ приближения удовлетворяют неравенствам $\mathbf{x}^0 \ll \mathbf{x}^1 = \mathbf{c} \ll \mathbf{x}, \ldots, \mathbf{x}^n \ll \mathbf{x}^{n+1} \ll B^n\mathbf{c} + B^{n-1}\mathbf{c} + \cdots + \mathbf{c} \ll \mathbf{x}$. Итак, последовательность $\{\mathbf{x}^n\}$ монотонно возрастает (монотонно возрастают все последовательности координат $\{x_i^n\}$), ограничена сверху в смысле $\ll$ вектором $\mathbf{x}$ и поэтому сходится. Переходя к пределу в равенстве $\mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c}$, убеждаемся в том, что ее предел совпадает с $\mathbf{x}$. $\rhd$

5.122. Пусть спектр матрицы A удовлетворяет условиям $0 < \delta \leqslant \mathrm{Re}\{\lambda(A)\} \leqslant 1\,,\ |\mathrm{Im}\{\lambda(A)\}| \leqslant 1$. Найти область значений вещественного параметра τ, при которых итерационный метод $\mathbf{x}^{k+1} = (I - \tau A)\mathbf{x}^k + \tau\,\mathbf{b}$ решения системы $A\,\mathbf{x} = \mathbf{b}$ сходится с произвольного начального приближения.

$\lhd$ По условию собственные значения λ оператора перехода $I - \tau\,A$ имеют вид

$$\lambda(I - \tau\,A) = 1 - \tau\,u - \mathrm{i}\,\tau\,v\,, \quad 0 < \delta \leqslant u \leqslant 1\,, \ \ |v| \leqslant 1\,.$$

Из условия сходимости $|\lambda|^2 = (1 - \tau\,u)^2 + \tau^2 v^2 < 1$ имеем неравенство $\tau < \dfrac{2\,u}{u^2 + v^2}$. Рассмотрим выражение

$$\min_{u,v} \frac{u}{u^2 + v^2} = \min_{u} \frac{u}{u^2 + 1} = \frac{\delta}{\delta^2 + 1}\ .$$

Отсюда следует ответ: $0 < \tau < \dfrac{2\,\delta}{\delta^2 + 1}$. $\rhd$

5.123. Исследовать сходимость метода $\mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c}$ для решения системы уравнений $\mathbf{x} = B\mathbf{x} + \mathbf{c}$ с матрицей

$$B = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} & \cdots & \frac{1}{2^n} & \frac{1}{2^{n+1}} \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{8} & \cdots & \frac{1}{2^{n-1}} & \frac{1}{2^n} \\ \frac{1}{8} & \frac{1}{4} & 0 & \frac{1}{4} & \cdots & \frac{1}{2^{n-2}} & \frac{1}{2^{n-1}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{2^{n+1}} & \frac{1}{2^n} & \frac{1}{2^{n-1}} & \cdots & \cdots & \frac{1}{4} & 0 \end{pmatrix}.$$

Ответ: метод сходится с произвольного начального приближения, так как $||B||_1 < 1$.

5.124. Построить сходящийся метод простой итерации (5.7) для системы уравнений с матрицей

$$A = \begin{pmatrix} 1 & 0,5 & 0 & 0 & \dots & 0 & 0 \\ 0 & 2 & 0,5 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0,5 & \dots & 0 & 0 \\ 0 & 0 & 0 & 2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0,5 \\ 0 & 0 & 0 & 0 & \dots & 0 & 2 \end{pmatrix}.$$

О т в е т: матрица положительно определена и имеет два кратных собственных числа $\lambda_1 = 1$ и $\lambda_2 = 2$, поэтому условие сходимости имеет вид: $0 < \tau < \frac{1}{2}$.

5.125. При каких условиях итерационный метод

$$\mathbf{x}^{k+1} = (2B^2 - I)\mathbf{x}^k + 2(B + I)\mathbf{c}$$

сходится быстрее метода простой итерации $\mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c}$?

$\lhd$ Метод простой итерации $\mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c}$ сходится при $|\lambda_i(B)| < 1$. Для погрешности $\mathbf{z}^k$ метода $\mathbf{x}^{k+1} = (2B^2 - I)\mathbf{x}^k + 2(B + I)\mathbf{c}$ справедливо равенство $\mathbf{z}^{k+1} = (2B^2 - I)\mathbf{z}^k$, $\lambda_i(2B^2 - I) = 2\lambda_i^2(B) - 1$, и этот итерационный метод сходится быстрее метода простой итерации, если спектр матрицы B расположен в подмножестве единичного круга комплексной плоскости, где функция $|2z^2 - 1|$ меньше функции $|z|$. В частности, если спектр матрицы B вещественный, то он должен принадлежать объединению интервалов $\left(-1, -\frac{1}{2}\right)$ и $\left(\frac{1}{2}, 1\right)$. $\rhd$

5.5. Вариационные методы

Класс вариационных методов строится как множество методов минимизации некоторых функционалов, минимум которых достигается на решении исходной системы линейных уравнений. Конкретный вид функционала и алгоритм минимизации определяют параметры итерационного процесса. Порядок сходимости рассматриваемых вариационных методов не хуже, чем у линейного одношагового метода. При этом для практической реализации данных методов не требуется знания границ m, M спектра матрицы A.

Метод наискорейшего градиентного спуска. Пусть $A = A^T > 0$. Расчетные формулы итерационного процесса имеют вид

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \tau_k(\mathbf{b} - A\mathbf{x}^k), \quad \tau_k = \frac{(\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, \mathbf{r}^k)}, \ k = 0, 1, \dots,$$

где $\mathbf{r}^k = \mathbf{b} - A\mathbf{x}^k$ — вектор невязки.

Отметим, что в приведенных формулах на каждой итерации требуется два умножения матрицы A на вектор.

5.126. Преобразовать формулы метода наискорейшего градиентного спуска так, чтобы на каждой итерации использовалось одно умножение матрицы A на вектор.

Ответ: пусть векторы $\mathbf{x}^k$ и $\mathbf{r}^k$ известны, тогда последовательно вычислим:

$$1)\ \mathbf{y} = A\mathbf{r}^k; 2)\ \tau_k = \frac{(\mathbf{r}^k, \mathbf{r}^k)}{(\mathbf{y}, \mathbf{r}^k)}; 3)\ \mathbf{x}^{k+1} = \mathbf{x}^k + \tau_k \mathbf{r}^k; 4)\ \mathbf{r}^{k+1} = \mathbf{r}^k - \tau_k \mathbf{y}.$$

Здесь на каждой итерации присутствует только одно умножение матрицы A на вектор, однако требуется хранить два вектора вместо одного.

5.127. Пусть $A = A^T > 0$ и $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$ — квадратичная функция. Доказать, что:

1) $F(\mathbf{x}) = ||\mathbf{x}^* - \mathbf{x}||_A^2 - ||\mathbf{x}^*||_A^2$, где $\mathbf{x}^*$ — точное решение системы $A\mathbf{x} = \mathbf{b}$;

2) равенство $F(\mathbf{x}^*) = \min\limits_{\mathbf{x}} F(\mathbf{x})$ выполняется тогда и только тогда, когда $\mathbf{x}^*$ — решение системы $A\mathbf{x} = \mathbf{b}$;

3) для градиента функции $F(\mathbf{x})$ справедлива формула

$$\operatorname{grad} F(\mathbf{x}) = 2(A\mathbf{x} - \mathbf{b}).$$

$\lhd$ 1) Преобразуем данное выражение

$$||\mathbf{x}^* - \mathbf{x}||_A^2 - ||\mathbf{x}^*||_A^2 = (A(\mathbf{x}^* - \mathbf{x}), \mathbf{x}^* - \mathbf{x}) - (A\mathbf{x}^*, \mathbf{x}^*) = \\ = (A\mathbf{x}, \mathbf{x}) - 2(A\mathbf{x}^*, \mathbf{x}) = F(\mathbf{x}).$$

2) Если $A > 0$, то $(A(\mathbf{x}^* - \mathbf{x}), \mathbf{x}^* - \mathbf{x}) > 0$ при $\mathbf{x} \neq \mathbf{x}^*$, поэтому функция $F(\mathbf{x})$ имеет минимум, и притом единственный, при $\mathbf{x} = \mathbf{x}^*$.

3) Последнее утверждение проверяется покомпонентным дифференцированием: $\dfrac{\partial F(\mathbf{x})}{\partial x_i}$. $\rhd$

5.128. Пусть решение системы $A\mathbf{x}^* = \mathbf{b}$ ищется как точка минимума функционала $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$ (см. 5.127) по следующему алгоритму:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \delta_k \operatorname{grad} F(\mathbf{x}^k),$$

где параметр δ_k выбирается из условия минимума величины

$$F(\mathbf{x}^k - \delta_k \operatorname{grad} F(\mathbf{x}^k)).$$

Доказать, что $2\delta_k \equiv \tau_k = \dfrac{(\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, \mathbf{r}^k)}$ и расчетные формулы совпадают с формулами наискорейшего градиентного спуска.

Указание. Подставив $\operatorname{grad} F(\mathbf{x}) = 2(A\mathbf{x} - \mathbf{b})$ в выражение для $\mathbf{x}^{k+1}$, получить

$$\mathbf{x}^{k+1} = \mathbf{x}^k + 2\delta_k(\mathbf{b} - A\mathbf{x}^k).$$

Далее из условия $F'_{\delta_k}(\mathbf{x}^{k+1}) = 0$ найти $2\delta_k \equiv \tau_k = \dfrac{(\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, \mathbf{r}^k)}$.

5.129. Показать, что на k-м шаге метода наискорейшего градиентного спуска минимизируется норма $\|\mathbf{z}^k\|_A = \sqrt{(A\mathbf{z}^k, \mathbf{z}^k)}$ вектора ошибки $\mathbf{z}^k = \mathbf{x} - \mathbf{x}^k$, где $\mathbf{x}$ — точное решение.

$\lhd$ Действительно, так как $\mathbf{z}^{k+1} = (I - \tau_k A)\mathbf{z}^k$, то

$$\|\mathbf{z}^{k+1}\|_A^2 = (A(I - \tau_k A)\mathbf{z}^k, (I - \tau_k A)\mathbf{z}^k) =$$
$$= \|\mathbf{z}^k\|_A^2 - 2\tau_k(A\mathbf{z}^k, A\mathbf{z}^k) + \tau_k^2(AA\mathbf{z}^k, A\mathbf{z}^k).$$

Отсюда после дифференцирования по τ_k находим, что минимум достигается при $\tau_k = \dfrac{(A\mathbf{z}^k, A\mathbf{z}^k)}{(AA\mathbf{z}^k, A\mathbf{z}^k)}$. Учитывая, что $A\mathbf{z}^k = \mathbf{b} - A\mathbf{x}^k = \mathbf{r}^k$, имеем $\tau_k = \dfrac{(\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, \mathbf{r}^k)}$.

Отметим, что минимизация евклидовой нормы $\|\mathbf{z}^k\| = \sqrt{(\mathbf{z}^k, \mathbf{z}^k)}$ вектора ошибки приводит к неконструктивным формулам для параметра $\tau_k = \dfrac{(\mathbf{r}^k, \mathbf{z}^k)}{(\mathbf{r}^k, \mathbf{r}^k)}$, так как вектор $\mathbf{z}^k$ неизвестен. $\rhd$

5.130. Пусть $A = A^T > 0$ и $\lambda(A) \in [m, M]$. Доказать, что метод наискорейшего градиентного спуска для решения системы $A\,\mathbf{x} = \mathbf{b}$ сходится с произвольного начального приближения и верна оценка

$$\|\mathbf{z}^k\|_A \leqslant \left(\frac{M-m}{M+m}\right)^k \|\mathbf{z}^0\|_A, \quad \text{где } \mathbf{z}^k = \mathbf{x} - \mathbf{x}^k,\ \|\mathbf{z}\|_A^2 = (A\mathbf{z}, \mathbf{z}).$$

$\lhd$ Действительно, параметр τ_k минимизирует на k-м шаге норму $\|\mathbf{z}^k\|_A$, следовательно с параметром τ_0 оптимального линейного одношагового метода оценка не лучше:

$$\|\mathbf{z}^{k+1}\|_A = \min_{\tau_k} \|(I - \tau_k A)\mathbf{z}^k\|_A \leqslant \|(I - \tau_0 A)\mathbf{z}^k\|_A \leqslant$$
$$\leqslant \|I - \tau_0 A\|_A \|\mathbf{z}^k\|_A = \frac{M-m}{M+m}\,\|\mathbf{z}^k\|_A,$$

так как $A = A^T > 0$ и, учитывая 5.32, для произвольного τ_0 имеем $\|I - \tau_0 A\|_A = \|I - \tau_0 A\|_2$. $\rhd$

5.131. Пусть $A = A^T > 0$ и $\lambda(A) \in [m, M]$. Доказать следующую оценку скорости сходимости метода наискорейшего градиентного спуска

$$\|\mathbf{z}^k\|_2 \leqslant \left(1 - \frac{m}{M}\right)^k \|\mathbf{z}^0\|_2, \quad \text{где} \quad \mathbf{z}^k = \mathbf{x} - \mathbf{x}^k.$$

Указание. Ведем следующие обозначения: $A\,\mathbf{e}_1 = m\,\mathbf{e}_1$, $A\,\mathbf{e}_2 = M\,\mathbf{e}_2$. Пусть $\mathbf{z}^k = \mathbf{e}_1 + \gamma\,\mathbf{e}_2$, где $\gamma \neq 0$ — произвольный параметр. Тогда $\mathbf{r}^k = \mathbf{b} - A\mathbf{x}^k = A\,\mathbf{z}^k$ и $\tau_k = \dfrac{(\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, \mathbf{r}^k)} = \dfrac{m^2 + \gamma^2 M^2}{m^3 + \gamma^2 M^3}$. В результате подстановки имеем $\mathbf{z}^{k+1} = \dfrac{\gamma(M-m)}{m^3 + \gamma^2 M^3}\,(\gamma\, M^2 \mathbf{e}_1 - m^2 \mathbf{e}_2)$, что приводит к искомой

оценке для этого частного случая. Если в разложении ошибки $\mathbf{z}^k$ присутствуют векторы, отвечающие собственным значениям $\lambda(A) \in (m, M)$, то несложно показать, что для соответствующих компонент $\mathbf{z}^{k+1}$ множитель перехода не превосходит величины $1 - \frac{m}{M}$.

5.132. Пусть $A = A^T > 0$ и $\lambda(A) \in [m, M]$. Доказать следующую оценку скорости сходимости метода наискорейшего градиентного спуска:

$$\|\mathbf{z}^k\|_2 \leqslant \left(\frac{M-m}{M+m}\right)^k \sqrt{\frac{M}{m}} \|\mathbf{z}^0\|_2, \quad \mathbf{z}^k = \mathbf{x} - \mathbf{x}^k.$$

Метод минимальных невязок. Пусть $A = A^T > 0$. Расчетные формулы итерационного процесса имеют вид

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \tau_k(\mathbf{b} - A\mathbf{x}^k), \quad \tau_k = \frac{(A\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, A\mathbf{r}^k)}, \quad k = 0, 1, \ldots,$$

где $\mathbf{r}^k = \mathbf{b} - A\mathbf{x}^k$ — вектор невязки.

5.133. Показать, что на k-м шаге метода минимальных невязок минимизируется норма $\|\mathbf{z}^k\|_{A^2} = \sqrt{(A\mathbf{z}^k, A\mathbf{z}^k)}$ вектора ошибки $\mathbf{z}^k = \mathbf{x} - \mathbf{x}^k$.

$\lhd$ Действительно, так как $\mathbf{z}^{k+1} = (I - \tau_k A)\mathbf{z}^k$, то

$$\|\mathbf{z}^{k+1}\|^2_{A^2} = (A(I - \tau_k A)\mathbf{z}^k, A(I - \tau_k A)\mathbf{z}^k) =$$
$$= \|\mathbf{z}^k\|^2_{A^2} - 2\tau_k(A^2\mathbf{z}^k, A\mathbf{z}^k) + \tau_k^2(A^2\mathbf{z}^k, A^2\mathbf{z}^k).$$

Отсюда после дифференцирования по τ_k, учитывая, что $A\mathbf{z}^k = \mathbf{b} - A\mathbf{x}^k = \mathbf{r}^k$, находим: минимум достигается при $\tau_k = \frac{(A\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, A\mathbf{r}^k)}$. Итерационный алгоритм с таким набором параметров называется *методом минимальных невязок*, так как $\|\mathbf{z}^{k+1}\|^2_{A^2} = (A\mathbf{z}^{k+1}, A\mathbf{z}^{k+1}) = \|\mathbf{r}^{k+1}\|^2_2$. $\rhd$

5.134. Пусть $A = A^T > 0$ и $\lambda(A) \in [m, M]$. Доказать, метод минимальных невязок для системы $A\mathbf{x} = \mathbf{b}$ сходится с произвольного начального приближения и верна оценка

$$\|\mathbf{z}^k\|_{A^2} \leqslant \left(\frac{M-m}{M+m}\right)^k \|\mathbf{z}^0\|_{A^2}, \quad \mathbf{z}^k = \mathbf{x} - \mathbf{x}^k, \ \|\mathbf{z}\|^2_{A^2} = (A\mathbf{z}, A\mathbf{z}).$$

У к а з а н и е. См. решение 5.130, учитывая, что $\|I - \tau_0 A\|_{A^2} = \|I - \tau_0 A\|_2$.

5.135. Пусть $A + A^T > 0$ и $\mu = \frac{\lambda_{\min}(A + A^T)}{2}$, $\sigma = \|A\|_2$. Показать, что метод минимальных невязок для системы $A\mathbf{x} = \mathbf{b}$ сходится с произвольного начального приближения и верна оценка

$$\|\mathbf{r}^{k+1}\|_2 \leqslant \left(1 - \frac{\mu^2}{\sigma^2}\right)^{1/2} \|\mathbf{r}^k\|_2.$$

$\lhd$ Так как $\mathbf{z}^{k+1} = (I - \tau_k A)\mathbf{z}^k$, где $\tau_k = \dfrac{(A\mathbf{r}^k, \mathbf{r}^k)}{(A\mathbf{r}^k, A\mathbf{r}^k)}$, то

$$\|\mathbf{r}^{k+1}\|_2^2 = \|\mathbf{z}^{k+1}\|_{A^2}^2 = (A(I - \tau_k A)\mathbf{z}^k,$$

$$A(I - \tau_k A)\mathbf{z}^k) = \|\mathbf{r}^k\|_2^2 - 2\tau_k(A\mathbf{r}^k, \mathbf{r}^k) + \tau_k^2(A\mathbf{r}^k, A\mathbf{r}^k) = \|\mathbf{r}^k\|_2^2 - \frac{(A\mathbf{r}^k, \mathbf{r}^k)^2}{(A\mathbf{r}^k, A\mathbf{r}^k)}.$$

Отсюда, учитывая неравенства

$$(A\mathbf{r}^k, A\mathbf{r}^k) \leqslant \|A\|_2^2 \|\mathbf{r}^k\|_2^2 \leqslant \sigma^2 \|\mathbf{r}^k\|_2^2,$$

$$(A\mathbf{r}^k, \mathbf{r}^k) = \left(\frac{A + A^T}{2}\,\mathbf{r}^k, \mathbf{r}^k\right) + \left(\frac{A - A^T}{2}\,\mathbf{r}^k, \mathbf{r}^k\right) = \left(\frac{A + A^T}{2}\,\mathbf{r}^k, \mathbf{r}^k\right) \geqslant \mu \|\mathbf{r}^k\|_2^2,$$

имеем требуемую оценку. $\rhd$

5.136. Пусть $\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n$ — базис пространства $\mathbf{R}^n$. Доказать сходимость с произвольного начального приближения следующего итерационного метода (*метода оптимального координатного спуска*) решения невырожденной системы уравнений $A\mathbf{x} = \mathbf{b}$:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \frac{(\mathbf{b} - A\mathbf{x}^k, A\mathbf{e}_j)}{||A\mathbf{e}_j||_2^2}\,\mathbf{e}_j, \quad j = \arg\max_l \frac{|(\mathbf{b} - A\mathbf{x}^k, A\mathbf{e}_l)|}{||A\mathbf{e}_l||_2}.$$

5.6. Неявные методы

Скорость сходимости рассмотренных итерационных процессов зависела от отношения $\frac{m}{M}$ границ спектра матрицы $A = A^T > 0$, т. е. от обусловленности задачи. Для «улучшения» исходной задачи можно перейти к некоторой эквивалентной системе $B^{-1}A\mathbf{x} = B^{-1}\mathbf{b}$ при условии невырожденности матрицы B

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + B^{-1}A\mathbf{x}^k = B^{-1}\mathbf{b}. \tag{5.9}$$

Метод спектрально-эквивалентных операторов. Пусть $A = A^T > 0$. Перепишем итерационный алгоритм (5.9) в следующем виде:

$$B\,\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}, \tag{5.10}$$

который также называют *обобщенным методом простой итерации* или *методом с предобусловливателем* B.

Неявный двухслойный итерационный алгоритм (5.10) требует на каждом шаге решения задач вида $B\mathbf{y} = \mathbf{f}$ и совпадает с рассмотренными выше методами при $B = I$. Известно, что алгоритм (5.10) сходится при $B > \frac{\tau}{2}A$, $\tau > 0$. Если дополнительно $B = B^T > 0$ и $m_1 B \leqslant A \leqslant M_1 B$, то при $\tau = \dfrac{2}{m_1 + M_1}$ метод сходится со скоростью геометрической прогрессии

с показателем $q = \dfrac{M_1 - m_1}{M_1 + m_1}$. Неявные методы с переменными τ типа минимальных невязок и наискорейшего градиентного спуска строятся аналогично и имеют скорость сходимости не хуже, чем у неявного оптимального линейного одношагового метода.

При удачном выборе оператора B можно принципиально улучшить скорость сходимости соответствующих итерационных процессов, однако необходимо учитывать трудоемкость нахождения $\mathbf{y} = B^{-1}\mathbf{f}$. Например, при $B = A$, $\tau = 1$ метод (5.10) сойдется за одну итерацию, но потребует решения исходной задачи $A\mathbf{x} = \mathbf{b}$.

Методы релаксации. Рассмотрим неявные методы с диагональной или треугольной матрицей B. Представим матрицу системы $A\,\mathbf{x} = \mathbf{b}$ в виде $A = L + D + R$, где D — диагональная матрица, L и R — соответственно левая нижняя и правая верхняя треугольные матрицы с нулевыми диагоналями (строго нижняя и строго верхняя треугольные матрицы). Будем предполагать, что все диагональные элементы исходной матрицы a_{ii} отличны от нуля, следовательно, любая матрица вида $D + \omega\, L$ с произвольным параметром ω обратима.

Методы релаксации описывают формулой (5.10) с матрицей $B = D + \omega L$. Здесь итерационный параметр ω называется *параметром релаксации*. Методы *Якоби* ($\omega = 0, \tau = 1$), *Гаусса—Зейделя* ($\omega = \tau = 1$) и *верхней релаксации* (в англоязычной литературе — SOR) ($\omega = \tau$) удобно представить соответственно в виде

$$D(\mathbf{x}^{k+1} - \mathbf{x}^k) + A\mathbf{x}^k = \mathbf{b},$$
$$(D + L)(\mathbf{x}^{k+1} - \mathbf{x}^k) + A\mathbf{x}^k = \mathbf{b},$$
$$(D + \omega L)\,\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\omega} + A\mathbf{x}^k = \mathbf{b}.$$

В случае $A = A^T > 0 (R = L^T)$ используют также *симметричный метод релаксации* (в англоязычной литературе — SSOR):

$$(D + \omega L)\,\frac{\mathbf{x}^{k+1/2} - \mathbf{x}^k}{\omega} + A\mathbf{x}^k = \mathbf{b},$$
$$(D + \omega R)\,\frac{\mathbf{x}^{k+1} - \mathbf{x}^{k+1/2}}{\omega} + A\mathbf{x}^{k+1/2} = \mathbf{b}.$$

5.137. Для решения системы $A\mathbf{x} = \mathbf{b}$ с матрицей

$$A = \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & \beta \\ 0 & \beta & \alpha \end{pmatrix}$$

применяются методы Якоби и Гаусса—Зейделя. Для каждого алгоритма найти все значения параметров α, β, обеспечивающие сходимость с произвольного начального приближения.

$\lhd$ Оператор перехода B (см. (5.5)) в методе Якоби имеет вид $B = -D^{-1}(L+R)$. Рассмотрим задачу на собственные значения $B\,\mathbf{x} = \lambda\,\mathbf{x}$, т. е. $-D^{-1}(L+R)\,\mathbf{x} = \lambda\,\mathbf{x}$. Перепишем последнее уравнение в эквивалентной форме $(L + \lambda D + R)\,\mathbf{x} = 0$, откуда имеем $\det(L + \lambda D + R) = 0$. Непосредственно вычисляя, находим

$$\det\begin{pmatrix} \alpha\lambda & \beta & 0 \\ \beta & \alpha\lambda & \beta \\ 0 & \beta & \alpha\lambda \end{pmatrix} = \alpha\lambda\,(\alpha^2\lambda^2 - 2\beta^2) = 0\,.$$

Следовательно, $\lambda_1 = 0$, $\lambda_{2,3}^2 = \dfrac{2\beta^2}{\alpha^2}$. Отсюда получаем ответ $\left|\dfrac{\beta}{\alpha}\right| < \dfrac{1}{\sqrt{2}}$.

Оператор перехода B в методе Гаусса—Зейделя имеет вид $B = -(D+L)^{-1}\,R$. Рассмотрим задачу на собственные значения $B\,\mathbf{x} = \lambda\,\mathbf{x}$. Имеем

$$-(D+L)^{-1}\,R\,\mathbf{x} = \lambda\,\mathbf{x},\ (\lambda\,L + \lambda\,D + R)\,\mathbf{x} = 0,\ \det(\lambda\,L + \lambda\,D + R) = 0\,.$$

В результате непосредственных вычислений имеем

$$\det\begin{pmatrix} \alpha\lambda & \beta & 0 \\ \beta\lambda & \alpha\lambda & \beta \\ 0 & \beta\lambda & \alpha\lambda \end{pmatrix} = \alpha\lambda^2\,(\alpha^2\lambda - 2\beta^2) = 0\,.$$

Следовательно, $\lambda_{1,2} = 0$, $\lambda_3 = 2\dfrac{\beta^2}{\alpha^2}$. Отсюда получаем ответ $\left|\dfrac{\beta}{\alpha}\right| < \dfrac{1}{\sqrt{2}}$.

В данном случае области сходимости методов совпадают. $\rhd$

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

$\lhd$ Запишем матричные представления операторов перехода

$$B_{\mathrm{J}} = \begin{pmatrix} 0 & -\dfrac{a_{12}}{a_{11}} \\ -\dfrac{a_{21}}{a_{22}} & 0 \end{pmatrix}, \quad B_{\mathrm{GZ}} = \begin{pmatrix} 0 & -\dfrac{a_{12}}{a_{11}} \\ 0 & \dfrac{a_{12}a_{21}}{a_{11}a_{22}} \end{pmatrix}.$$

Отсюда имеем следующие формулы для собственных значений:

$$\lambda_{1,2}^{\mathrm{J}} = \pm\sqrt{\frac{a_{12}a_{21}}{a_{11}a_{22}}}\,, \qquad \lambda_1^{\mathrm{GZ}} = 0, \quad \lambda_2^{\mathrm{GZ}} = \frac{a_{12}a_{21}}{a_{11}a_{22}}\,,$$

приводящие к искомому утверждению. $\rhd$

5.139. Пусть невырожденная матрица A обладает свойством диагонального преобладания, т. е. для всех i справедливо неравенство

$$\sum_{j\neq i} |a_{ij}| \leqslant q|a_{ii}|, \quad 0 \leqslant q < 1\,.$$

Доказать, что для вектора ошибки в методе Гаусса—Зейделя имеет место неравенство

$$\|\mathbf{x} - \mathbf{x}^k\|_\infty \leqslant q^k\,\|\mathbf{x} - \mathbf{x}^0\|_\infty\,.$$

$\lhd$ Обозначим вектор ошибки через $\mathbf{z}^k$. Для этого вектора имеет место соотношение $(D+L)\,\mathbf{z}^{k+1}+R\,\mathbf{z}^k=0$. Пусть $\|\mathbf{z}^{k+1}\|_\infty=|z_l^{k+1}|$. Запишем l-е уравнение

$$\sum_{j=1}^{l-1} a_{lj}\,z_j^{k+1}+a_{ll}\,z_l^{k+1}+\sum_{j=l+1}^{n} a_{lj}\,z_j^k=0$$

и решим его относительно z_l^{k+1}. Имеем

$$z_l^{k+1}=-\sum_{j=1}^{l-1}\frac{a_{lj}}{a_{ll}}\,z_j^{k+1}-\sum_{j=l+1}^{n}\frac{a_{lj}}{a_{ll}}\,z_j^k\,.$$

Отсюда получаем

$$\|\mathbf{z}^{k+1}\|_\infty=|z_l^{k+1}|\leqslant\alpha\,\|\mathbf{z}^{k+1}\|_\infty+\beta\,\|\mathbf{z}^k\|_\infty,$$

где

$$\alpha=\sum_{j=1}^{l-1}\left|\frac{a_{lj}}{a_{ll}}\right|,\quad \beta=\sum_{j=l+1}^{n}\left|\frac{a_{lj}}{a_{ll}}\right|.$$

Найденное соотношение можно переписать в виде

$$\|\mathbf{z}^{k+1}\|_\infty\leqslant\frac{\beta}{1-\alpha}\,\|\mathbf{z}^k\|_\infty.$$

По условию $\alpha+\beta\leqslant q<1$, следовательно,

$$\frac{\beta}{1-\alpha}\leqslant\frac{q-\alpha}{1-\alpha}\leqslant\frac{q-q\alpha}{1-\alpha}\leqslant q\,,$$

откуда имеем искомую оценку. $\rhd$

5.140. Исследовать сходимость метода Гаусса—Зейделя для матриц размерности $n\times n$ с элементами:

$$1)\quad a_{kj}=3^{-|k-j|},\quad 2)\quad a_{kj}=\begin{cases}2 & \text{при}\quad k=j,\\ -1 & \text{при}\quad |k-j|=1,\\ 0 & \text{при}\quad |k-j|>1.\end{cases}$$

Ответ: метод сходится в обоих случаях.

5.141. Показать, что выполнение неравенства $0<\tau<2$ является необходимым для сходимости метода верхней релаксации.

$\lhd$ Если формулу метода релаксации

$$(D+\tau\,L)\,\mathbf{x}^{k+1}+[\tau\,R+(\tau-1)\,D]\,\mathbf{x}^k=\tau\mathbf{b}$$

умножить слева на матрицу D^{-1}, то оператор перехода можно записать в следующем виде:

$$B=(I+\tau\,M)^{-1}\,((1-\tau)\,I+\tau\,N)\,.$$

Здесь I — единичная, $M=D^{-1}L$ и $N=D^{-1}R$ — строго нижняя и верхняя треугольные матрицы соответственно. Рассмотрим характеристический

многочлен $d(\lambda) = \det(B - \lambda I)$. По теореме Виета имеет место равенство $(-1)^n\, d(0) = \prod\limits_{i=1}^{n} \lambda_i(B)$. Так как у треугольных матриц M и N на главной диагонали расположены нули, то $d(0) = \det(B) = (1-\tau)^n$. Отсюда для спектрального радиуса оператора перехода получаем оценку

$$\rho(B) = \max_i |\lambda_i(B)| \geqslant \left| \prod_{i=1}^{n} \lambda_i(B) \right|^{1/n} = |\det(B)|^{1/n} = |1-\tau|\,,$$

которая в силу необходимого неравенства $\rho(B) < 1$ приводит к искомому ответу. $\triangleright$

5.142. Пусть матрица A простой структуры имеет собственные значения $\lambda(A) \in [m, M],\ m > 0$. Доказать, что при любом положительном значении итерационного параметра τ сходится метод следующего вида:

$$\frac{\mathbf{x}^{k+1} - \mathbf{x}^k}{\tau} + A\left(\frac{\mathbf{x}^{k+1} + \mathbf{x}^k}{2}\right) = \mathbf{b}\,.$$

Определить оптимальное значение τ_{opt}.

$\triangleleft$ Используя эквивалентную форму записи метода

$$\left(I + \tfrac{\tau}{2} A\right)\mathbf{x}^{k+1} = \left(I - \tfrac{\tau}{2} A\right)\mathbf{x}^k + \tau\mathbf{b}$$

и общность системы собственных векторов матриц слева и справа, выразим собственные значения оператора перехода B через собственные значения исходной матрицы

$$\lambda(B) = \frac{1 - \tau\,\dfrac{\lambda(A)}{2}}{1 + \tau\,\dfrac{\lambda(A)}{2}}\,.$$

Отсюда следует сходимость метода при $\tau > 0$. Для определения τ_{opt} рассмотрим следующую минимаксную задачу:

$$\min_{\tau>0} \max_{\lambda \in \left[\frac{m}{2}, \frac{M}{2}\right]} \frac{|1-\tau\lambda|}{1+\tau\lambda}\,.$$

Функция $f(\lambda) = \dfrac{1-\tau\lambda}{1+\tau\lambda}$ при $\lambda > 0$ и фиксированном $\tau > 0$ является убывающей, поэтому максимального значения функция $|f(\lambda)|$ достигает на границе отрезка: при $\lambda = \dfrac{m}{2}$ и (или) при $\lambda = \dfrac{M}{2}$. Можно убедиться, что минимум по τ имеет место в случае равенства $\left|f\left(\dfrac{m}{2}\right)\right| = \left|f\left(\dfrac{M}{2}\right)\right|$, которое приводит к уравнению для оптимального параметра

$$\frac{1 - \tau_{\text{opt}}\dfrac{m}{2}}{1 + \tau_{\text{opt}}\dfrac{m}{2}} = -\,\frac{1 - \tau_{\text{opt}}\dfrac{M}{2}}{1 + \tau_{\text{opt}}\dfrac{M}{2}}.$$

Решая это уравнение, имеем $\tau_{\text{opt}} = \dfrac{2}{\sqrt{mM}}$. $\triangleright$

5.143. При каких $\alpha \in [0,1]$ для матрицы A из 5.142 метод

$$\frac{\mathbf{x}^{k+1}-\mathbf{x}^k}{\tau} + A\left(\alpha\mathbf{x}^{k+1} + (1-\alpha)\mathbf{x}^k\right) = \mathbf{b}$$

сходится при любом $\tau > 0$?

$\lhd$ Используя идею решения 5.142, запишем условие сходимости метода

$$\max_{\lambda\in\left[\frac{m}{2},\frac{M}{2}\right]} \left|\frac{1-\tau(1-\alpha)\lambda}{1+\tau\alpha\lambda}\right| < 1 \quad \forall\tau > 0\,.$$

Сделав замену $t = \tau\,\lambda > 0$, получим неравенство

$$|1-t(1-\alpha)| < 1 + t\alpha\,.$$

Если выражение под знаком модуля неотрицательно, то получаем верное, в силу условия, неравенство $-t < 0$. Поэтому содержательным является другой случай: $t(1-\alpha) - 1 < 1 + t\alpha$. Из этого неравенства имеем $-\frac{2}{t} < 2\alpha - 1$, что, так как $t > 0$, приводит к ответу $\alpha \geqslant \frac{1}{2}$. $\rhd$

5.144. Невырожденная система $A\mathbf{x} = \mathbf{b}$ с матрицей $A = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$ решается методом Гаусса—Зейделя. Доказать, что:

1) если $|a| > 1$, то для некоторого начального приближения итерационный процесс не сходится;

2) если $|a| < 1$, то итерации сходятся при любом начальном приближении.

$\lhd$ Спектральный радиус матрицы перехода в методе Гаусса—Зейделя равен $|a|$. Если начальное приближение таково, что начальная погрешность $\mathbf{z}^0$ имеет ненулевую вторую координату z_2^0, то $z_1^k = -a^{2k-1}z_2^0$, $z_2^k = a^{2k}z_2^0$ и метод не сходится при $|a| > 1$. $\rhd$

5.145. Построить пример системы уравнений третьего порядка, для которой метод Якоби сходится, а метод Гаусса—Зейделя расходится.

5.146. Построить пример системы уравнений третьего порядка, для которой метод Гаусса—Зейделя сходится, а метод Якоби расходится.

5.147. Доказать, что обобщенный метод простой итерации

$$B\,\frac{\mathbf{x}^{k+1}-\mathbf{x}^k}{\tau} + A\mathbf{x}^k = \mathbf{b}\,, \quad A = A^T > 0\,, \ \det(B) \neq 0\,, \ \tau > 0\,,$$

сходится при условии $B - \frac{\tau}{2}A > 0$ (т. е. $(B\mathbf{x},\mathbf{x}) > \frac{\tau}{2}(A\mathbf{x},\mathbf{x})\ \forall\mathbf{x} \neq 0$).

$\lhd$ Из уравнения для ошибки

$$B\,\frac{\mathbf{z}^{k+1}-\mathbf{z}^k}{\tau} + A\mathbf{z}^k = 0$$

следует, что

$$\mathbf{z}^{k+1} = (I - \tau B^{-1}A)\mathbf{z}^k\,, \quad A\mathbf{z}^{k+1} = (A - \tau AB^{-1}A)\mathbf{z}^k\,. \tag{5.11}$$

Вычислим скалярное произведение, используя симметрию A,

$$(A\mathbf{z}^{k+1}, \mathbf{z}^{k+1}) = (A\mathbf{z}^k, \mathbf{z}^k) - 2\tau \left(\left[B - \frac{\tau}{2} A \right] B^{-1} A\mathbf{z}^k, B^{-1} A\mathbf{z}^k \right) . \qquad (5.12)$$

Из первого соотношения (5.11) имеем $B^{-1}A\mathbf{z}^k = -\dfrac{\mathbf{z}^{k+1} - \mathbf{z}^k}{\tau}$, что позволяет переписать (5.12) в виде

$$||\mathbf{z}^{k+1}||_A^2 - ||\mathbf{z}^k||_A^2 + \frac{2}{\tau} \left(\left[B - \frac{\tau}{2} A \right] (\mathbf{z}^{k+1} - \mathbf{z}^k), \mathbf{z}^{k+1} - \mathbf{z}^k \right) = 0 ,$$

где $||\mathbf{u}||_A = (A\mathbf{u}, \mathbf{u})^{1/2}$.

В силу конечномерности векторного пространства условие $B - \frac{\tau}{2} A > 0$ равносильно условию $B - \frac{\tau}{2} A \geqslant \varepsilon I$ с некоторым $\varepsilon > 0$ (здесь через I обозначена единичная матрица). Имеем

$$||\mathbf{z}^{k+1}||_A^2 - ||\mathbf{z}^k||_A^2 + 2\varepsilon\tau^{-1}||\mathbf{z}^{k+1} - \mathbf{z}^k||_2^2 \leqslant 0 \quad \forall k \geqslant 0 .$$

Из этого неравенства следует монотонное убывание и ограниченность последовательности $\{||\mathbf{z}^k||_A^2\}$, следовательно, сходимость $\|\mathbf{z}^k\|_A$ к некоторой величине $d \geqslant 0$. Переходя к пределу в данном неравенстве, получаем $\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_2 \to 0$, поэтому $\lim\limits_{k\to\infty} ||\mathbf{z}^{k+1} - \mathbf{z}^k||_2^2 = \lim\limits_{k\to\infty} ||\mathbf{x}^{k+1} - \mathbf{x}^k||_2^2 = 0$. Таким образом, метод сходится к некоторому $\mathbf{x}^\infty$. Из вида итерационного процесса следует неравенство

$$\|\mathbf{b} - A\mathbf{x}^k\|_2 \leqslant \frac{1}{\tau} \|B\|_2 \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2,$$

переходя в котором к пределу, убеждаемся, что $\mathbf{x}^\infty$ — решение уравнения $A\mathbf{x} = \mathbf{b}$, т. е. последовательность приближений $\{\mathbf{x}^k\}$ сходится к $\mathbf{x} = A^{-1}\mathbf{b}$. $\triangleright$

5.148. Пусть $A = A^T > 0$. Доказать, что метод релаксации сходится с произвольного начального приближения при $\tau \in (0, 2)$.

Указание. Использовать утверждение 5.147 при $B = D + \tau L$.

5.149. Пусть $B = L + R$, где L — нижняя треугольная матрица с нулями на диагонали, R — верхняя треугольная матрица. Пусть далее $||B||_\infty < 1$, так что итерационный процесс $\mathbf{x}^{k+1} = B\mathbf{x}^k + \mathbf{c}$ сходится. Доказать, что метод $\mathbf{x}^{k+1} = L\mathbf{x}^{k+1} + R\mathbf{x}^k + \mathbf{c}$ также сходится.

Указание. Пусть $||B||_\infty = \max\limits_i \sum\limits_j |b_{ij}| = q < 1$, $q_{1i} = \sum\limits_{j<i} |b_{ij}|$, $q_{2i} = \sum\limits_{j\geqslant i} |b_{ij}|$. Доказать, что для погрешности итерационного метода $\mathbf{x}^{k+1} = L\mathbf{x}^{k+1} + R\mathbf{x}^k + \mathbf{c}$ справедлива оценка

$$||\mathbf{z}^{k+1}||_\infty \leqslant \max_i \frac{q_{2i}}{1 - q_{1i}} ||\mathbf{z}^k||_\infty \leqslant \max_i \frac{q - q_{1i}}{1 - q_{1i}} ||\mathbf{z}^k||_\infty \leqslant q||\mathbf{z}^k||_\infty.$$

5.150. Для системы уравнений

$$4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1} = h^2 f_{ij},\ i,j = 1,2,...,n-1;\ nh = 1;$$
$$u_{0,i} = u_{i,0} = u_{n,i} = u_{i,n} = 0,\ \ i = 0,1,...,n\,,$$

записать расчетные формулы и найти асимптотическую скорость сходимости следующих итерационных методов: 1) метода Якоби; 2) метода Гаусса—Зейделя; 3) метода релаксации с оптимальным параметром релаксации; 4) симметричного метода релаксации с оптимальным параметром релаксации.

Ответ: спектральный радиус оператора перехода, асимптотическая скорость сходимости и оптимальный параметр таковы:

1) $\rho(B) = \cos\pi h$, $R_\infty(B) = \pi^2\frac{h^2}{2}$;
2) $\rho(B) = \cos^2\pi h$, $R_\infty(B) = \pi^2 h^2$;
3) $\rho(B) = \frac{1-\sin\pi h}{1+\sin\pi h}$, $R_\infty(B) = 2\pi h$, $\omega_{\text{opt}} = \frac{2}{1+\sin\pi h}$;
4) $\rho(B) = \dfrac{1-\sin\frac{\pi h}{2}}{1+\sin\frac{\pi h}{2}}$, $R_\infty(B) = \pi h$, $\omega_{\text{opt}} = \dfrac{2}{1+\sin\left(\frac{\pi h}{2}\right)}$.

5.151. Исследовать сходимость метода Якоби для решения системы уравнений с матрицей

$$A = \begin{pmatrix} 2 & -0,2 & 0,3 & 0,4 \\ 0,3 & -3 & 1 & -1,4 \\ 0,4 & 0,8 & 4 & 2,4 \\ -0,5 & 1,2 & -2,5 & -5 \end{pmatrix}.$$

Указание. Матрица имеет диагональное преобладание.

5.152. Найти все α, β, при которых метод Гаусса—Зейделя является сходящимся для системы уравнений с матрицей:

$$1)\ \begin{pmatrix} \alpha & 0 & \beta \\ 0 & \alpha & 0 \\ \beta & 0 & \alpha \end{pmatrix};\ 2)\ \begin{pmatrix} \alpha & \beta & 0 \\ \beta & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix};\ 3)\ \begin{pmatrix} \alpha & \alpha & 0 \\ \alpha & \beta & \beta \\ 0 & \beta & \alpha \end{pmatrix}.$$

Указание. См. решение 5.137.

Ответ: для случаев 1) и 2) имеем условие $|\beta| < |\alpha|$; 3) таких α и β не существует, так как имеется собственное значение оператора перехода $\lambda = \frac{\alpha^2+\beta^2}{\alpha\beta}$, модуль которого больше единицы.

5.153. Пусть матрицы A_i, $i = 1,2$, простой структуры имеют собственные значения $\lambda(A_i) \in [m, M]$, $m > 0$ и $A_1A_2 = A_2A_1$, $A = A_1 + A_2$. Доказать, что при любом положительном значении параметра τ сходится итерационный метод решения системы уравнений $A\mathbf{x} = \mathbf{b}$ следующего

вида:

$$\frac{\mathbf{x}^{k+1/2}-\mathbf{x}^k}{\tau}+A_1\mathbf{x}^{k+1/2}+A_2\mathbf{x}^k=\mathbf{b},$$
$$\frac{\mathbf{x}^{k+1}-\mathbf{x}^{k+1/2}}{\tau}+A_1\mathbf{x}^{k+1/2}+A_2\mathbf{x}^{k+1}=\mathbf{b}.$$

Определить оптимальное значение τ_{opt}.

$\lhd$ Обозначим $\mathbf{z}^k=\mathbf{x}-\mathbf{x}^k$, $\mathbf{z}^{k+1/2}=\mathbf{x}-\mathbf{x}^{k+1/2}$, где $\mathbf{x}$ — решение системы $A\mathbf{x}=\mathbf{b}$. Тогда

$$\mathbf{z}^{k+1}=(I+\tau A_2)^{-1}(I-\tau A_1)(I+\tau A_1)^{-1}(I-\tau A_2)\mathbf{z}^k\equiv P\mathbf{z}^k.$$

Матрица перехода P подобна матрице

$$B=(I-\tau A_1)(I+\tau A_1)^{-1}(I-\tau A_2)(I+\tau A_2)^{-1}.$$

Коммутирующие матрицы простой структуры A_1 и A_2 имеют общую полную систему собственных векторов. Это дает представления $A_i=QD_iQ^{-1}$, $i=1,2$, с диагональными матрицами D_i и совпадение собственных значений матриц A_i и D_i. Отсюда получаем оценку для спектрального радиуса матрицы B:

$$\rho(B)=\rho((I-\tau D_1)(I+\tau D_1)^{-1}(I-\tau D_2)(I+\tau D_2)^{-1})=$$
$$=\max_i\left|\frac{1-\tau\lambda_i(A_1)}{1+\tau\lambda_i(A_1)}\,\frac{1-\tau\lambda_i(A_2)}{1+\tau\lambda_i(A_2)}\right|\leqslant\max_{m\leqslant t\leqslant M}\left(\frac{1-\tau t}{1+\tau t}\right)^2.$$

Оптимальное значение $\tau_{\text{opt}}=\frac{1}{\sqrt{mM}}$, при этом $\rho(B)\leqslant\left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right)^2$. $\rhd$

5.154. Доказать сходимость итерационного метода из 5.153, если матрицы A_1, A_2 удовлетворяют следующим условиям:

$(A_i\mathbf{x},\mathbf{x})>0$ для $i=1,2$, и $\forall x\neq 0$, но не обязательно $A_1A_2=A_2A_1$.

5.155. Пусть матрицы A_i, $i=1,2$, простой структуры имеют собственные значения $\lambda(A_i)\in[m,M]$, $m>0$ и $A_1A_2=A_2A_1$, $A=A_1+A_2$. Доказать, что при любом положительном значении параметра τ сходится итерационный метод решения системы уравнений $A\mathbf{x}=\mathbf{b}$ следующего вида:

$$\frac{\mathbf{x}^{k+1/2}-\mathbf{x}^k}{\tau}+A_1\mathbf{x}^{k+1/2}+A_2\mathbf{x}^k=\mathbf{b},$$
$$\frac{\mathbf{x}^{k+1}-\mathbf{x}^{k+1/2}}{\tau}+A_2(\mathbf{x}^{k+1}-\mathbf{x}^k)=0.$$

Определить оптимальное значение τ_{opt}.

$\lhd$ Обозначим $\mathbf{z}^k=\mathbf{x}-\mathbf{x}^k$, $\mathbf{z}^{k+1/2}=\mathbf{x}-\mathbf{x}^{k+1/2}$, где $\mathbf{x}$ — решение системы $A\mathbf{x}=\mathbf{b}$. Тогда

$$\mathbf{z}^{k+1}=(I+\tau A_2)^{-1}(I+\tau A_1)^{-1}(I+\tau^2A_1A_2)\mathbf{z}^k\equiv B\mathbf{z}^k.$$

Коммутирующие матрицы простой структуры A_1 и A_2 имеют общую полную систему собственных векторов и представимы в виде $A_i = QD_iQ^{-1}$ с диагональными матрицами D_i, у которых те же спектры, что и A_i: $\lambda(D_i) = \lambda(A_i)$. В таком случае для спектрального радиуса матрицы B получаем следующую оценку:

$$\rho(B) = \rho\left((I+\tau D_2)^{-1}(I+\tau D_1)^{-1}(I+\tau^2 D_1 D_2)\right) =$$

$$= \max_i \frac{1+\tau^2\lambda_i(A_1)\lambda_i(A_2)}{(1+\tau\lambda_i(A_1))(1+\tau\lambda_i(A_2))} \leqslant \max_{t\in[m,M]} \frac{1+\tau^2t^2}{(1+\tau t)^2} < 1 \ \forall\tau > 0.$$

Так как матрица A невырождена (система имеет единственное решение) и все собственные значения оператора перехода лежат в единичном круге, то итерационный процесс сходится к решению задачи $A\mathbf{x} = \mathbf{b}$.

Рассмотрим оптимизационную задачу (см. 5.153). Имеем

$$\rho(B) \leqslant \min_{\tau>0} \max_{t\in[m,M]} \frac{1+\tau^2t^2}{(1+\tau t)^2} = \min_{\tau>0} \max\left\{\frac{1+\tau^2m^2}{(1+\tau m)^2}, \frac{1+\tau^2M^2}{(1+\tau M)^2}\right\}$$

Максимальное значение на отрезке функция достигает в одной из концевых точек, так как ее производная по t равна $\left(\frac{1+\tau^2t^2}{(1+\tau t)^2}\right)' = \frac{2\tau(t\tau-1)}{(1+\tau t)^4}$, и $t = \frac{1}{\tau}$ — точка локального минимума. Из явного вида минимизируемых функций следует, что τ_{opt} — решение следующего уравнения: $\frac{1+\tau^2m^2}{(1+\tau m)^2} = \frac{1+\tau^2M^2}{(1+\tau M)^2}$. Отсюда имеем $\tau_{\text{opt}} = \frac{1}{\sqrt{mM}}$, при этом $\rho(B) \leqslant \frac{M+m}{(\sqrt{M}+\sqrt{m})^2}$. $\triangleright$

5.156. Доказать сходимость итерационного процесса из 5.155, если матрицы A_1, A_2 удовлетворяют следующим условиям: $(A_i\mathbf{x}, \mathbf{x}) > 0$ для $i = 1, 2$, и $\forall x \neq 0$, но не обязательно $A_1A_2 = A_2A_1$.

5.157. Показать, что если матрица $A = M - N$ вырожденная, то нельзя получить оценку $\rho(M^{-1}N) < 1$ ни для какой невырожденной матрицы M.

$\triangleleft$ Имеем $A = M - N = M(I - M^{-1}N)$. Если $\rho(M^{-1}N) < 1$, то существует $(I - M^{-1}N)^{-1}$, как следствие существует $A^{-1} = (I - M^{-1}N)^{-1}M^{-1}$. $\triangleright$

5.158. Пусть $A = M - N$ и итерации $M\mathbf{x}^{k+1} = N\mathbf{x}^k + \mathbf{b}$ сходятся при произвольном начальном приближении. Доказать, что $\rho(M^{-1}N) < 1$.

Указание. Предположив, что $\rho(M^{-1}N) \geqslant 1$, выбрать такое начальное приближение $\mathbf{x}^0$, что погрешность $\mathbf{z}^0 = \mathbf{x} - \mathbf{x}^0$ пропорциональна собственному вектору матрицы $M^{-1}N$, соответствующему собственному значению λ такому, что $|\lambda| \geqslant 1$.

5.159. Пусть решаются задачи $A_i\mathbf{x} = \mathbf{b}_i$, $i = 1, 2$, где

$$A_1 = \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 1 & -\frac{3}{4} \\ -\frac{1}{12} & 1 \end{pmatrix}$$

и B_1 и B_2 — соответствующие этим матрицам операторы перехода в итерационном методе Якоби. Показать, что $\rho(B_1) > \rho(B_2)$, т. е. опровергнуть мнение о том, что относительное усиление диагонального преобладания влечет за собой более быструю сходимость метода Якоби.

Ответ: $\rho(B_1) = \frac{1}{2},\ \rho(B_2) = \frac{1}{4}$.

5.7. Проекционные методы

Эффективными методами решения системы линейных алгебраических уравнений большой размерности $A\mathbf{x}^* = \mathbf{b}$ являются итерационные методы проекционного типа. На каждом шаге такого метода реализуется *проекционный алгоритм*: в зависимости от текущего приближения $\mathbf{x} \in \mathbf{R}^n$ и номера итерации выбирают два m-мерных ($m \leqslant n$) подпространства $\mathcal{K}$ и $\mathcal{L}$; следующее приближение $\hat{\mathbf{x}}$ к точному решению $\mathbf{x}^*$ ищут в виде $\hat{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$, $\delta\mathbf{x} \in \mathcal{K}$, из условия $\mathbf{r} \perp \mathcal{L}$, $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$.

Таким образом, основная идея данного подхода заключается в построении вектора поправки $\delta\mathbf{x}$ из подпространства $\mathcal{K}$, обеспечивающего ортогональность вектора невязки $\mathbf{r}$ подпространству $\mathcal{L}$. Различные правила выбора подпространств $\mathcal{K}$ и $\mathcal{L}$ приводят к различным расчетным формулам.

5.160. Показать, что метод Гаусса—Зейделя решения систем линейных уравнений является проекционным методом.

$\lhd$ Определим $\mathcal{K} = \mathcal{L} = \{\mathbf{e}_i\}$ для $i = 1, ..., n$, где $\mathbf{e}_i$ — естественный i-й базисный вектор пространства $\mathbf{R}^n$. Тогда последовательно находим $\hat{\mathbf{x}} = \mathbf{x} + c_i\mathbf{e}_i$ и $(\mathbf{b} - A(\mathbf{x} + c_i\mathbf{e}_i), \mathbf{e}_i) = 0$. Отсюда имеем $c_i = \dfrac{b_i - \sum\limits_{j=1}^{n} a_{ij}x_j}{a_{ii}}$ при известных компонентах x_j, $j = i, i+1, \ldots, n$ и найденных x_j, $j = 1, 2, \ldots, i-1$. Таким образом, за n шагов проекционного алгоритма имеем $\mathbf{x}^{k+1} = \mathbf{x}^k + \sum\limits_{i=1}^{n} c_i\mathbf{e}_i$, что соответствует шагу метода Гаусса—Зейделя:

$$a_{ii}(x_i^{k+1} - x_i^k) + \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} + \sum_{j=i}^{n} a_{ij}x_j^k = b_i, \quad i = 1, \ldots, n. \qquad \rhd$$

Пусть текущие подпространства $\mathcal{K} = \operatorname{span}\{\mathbf{k}_1, \ldots, \mathbf{k}_m\}$ и $\mathcal{L} = \operatorname{span}\{\mathbf{l}_1, \ldots, \mathbf{l}_m\}$ являются линейными оболочками наборов базисных векторов $\mathbf{k}_i$ и $\mathbf{l}_i$. Определим соответствующие им матрицы $K = (\mathbf{k}_1 \ldots \mathbf{k}_m)$ и $L = (\mathbf{l}_1 \ldots \mathbf{l}_m)$ размерности $n \times m$. Положим $\hat{\mathbf{x}} = \mathbf{x} + K\mathbf{c}$. Тогда условие ортогональности приводит к следующей системе относительно искомого вектора коэффициентов $\mathbf{c}$:

$$L^T AK\mathbf{c} = L^T\mathbf{r}, \quad \mathbf{r} = \mathbf{b} - A\mathbf{x}.$$

Если матрица $L^T AK$ невырождена, то формула для очередного приближения имеет вид

$$\hat{\mathbf{x}} = \mathbf{x} + K(L^T AK)^{-1}L^T\mathbf{r}.$$

На практике большинство алгоритмов не требует нахождения явного вида матриц $B = L^T AK$ и B^{-1}.

5.161. Пусть либо $A = A^T > 0$ и $\mathcal{L} = \mathcal{K}$, либо $\det(A) \neq 0$ и $\mathcal{L} = A\mathcal{K}$. Показать, что для произвольных базисов $\{\mathbf{k}_i\}$ и $\{\mathbf{l}_i\}$ матрица $B = L^T AK$ невырождена.

Указание. Представить матрицу L в виде $L = KG$ с некоторой невырожденной матрицей G (преобразование базисов) для $\mathcal{L} = \mathcal{K}$ либо в виде $L = AKG$ для $\mathcal{L} = A\mathcal{K}$.

5.162. (Проекционная теорема). Показать, что для произвольного вектора $\mathbf{z}$ вектор $\hat{\mathbf{k}}$ является решением следующей задачи минимизации

$$\min_{\mathbf{k}\in\mathcal{K}}(\mathbf{z}-\mathbf{k}, \mathbf{z}-\mathbf{k})$$

тогда и только тогда, когда $(\mathbf{z}-\hat{\mathbf{k}}, \mathbf{v}) = 0$ для $\forall \mathbf{v} \in \mathcal{K}$.

$\lhd$ Рассмотрим разложение $\mathbf{z} = P\mathbf{z} + (\mathbf{z} - P\mathbf{z})$, где $P\mathbf{z} \in \mathcal{K}$, $\mathbf{z} - P\mathbf{z} \in \mathcal{K}^\perp$. В этом случае P называют *оператором ортогонального проектирования на* $\mathcal{K}$. Тогда

$$(\mathbf{z}-\mathbf{k}, \mathbf{z}-\mathbf{k}) = \|\mathbf{z}-\mathbf{k}\|^2 = \|\mathbf{z}-P\mathbf{z}+P\mathbf{z}-\mathbf{k}\|^2 = \|\mathbf{z}-P\mathbf{z}\|^2 + \|P\mathbf{z}-\mathbf{k}\|^2,$$

т. е. $\|\mathbf{z}-\mathbf{k}\|^2 \geqslant \|\mathbf{z}-P\mathbf{z}\|^2$, равенство возможно лишь при $\hat{\mathbf{k}} = P\mathbf{z}$. $\rhd$

5.163. Пусть $A = A^T > 0$ и $\mathcal{L} = \mathcal{K}$. Показать, что вектор $\hat{\mathbf{x}}$ является результатом проекционного алгоритма тогда и только тогда, когда

$$E(\hat{\mathbf{x}}) = \min_{\tilde{\mathbf{x}}\in\mathbf{x}+\mathcal{K}} E(\tilde{\mathbf{x}}), \quad \text{где } E(\tilde{\mathbf{x}}) = (A(\mathbf{x}^*-\tilde{\mathbf{x}}), \mathbf{x}^*-\tilde{\mathbf{x}}) \text{ и } A\mathbf{x}^* = \mathbf{b}.$$

$\lhd$ Из 5.162 следует, что решение задачи $\min\limits_{\mathbf{k}\in\mathcal{K}}(\mathbf{z}-\mathbf{k}, \mathbf{z}-\mathbf{k})_A$ для $\mathbf{z} = \mathbf{x}^* - \mathbf{x}$, где $(\mathbf{u}, \mathbf{v})_A = (A\mathbf{u}, \mathbf{v})$ эквивалентно нахождению вектора $\mathbf{k}$ из условия $(\mathbf{z}-\mathbf{k}, \mathbf{v})_A = 0 \quad \forall \mathbf{v} \in \mathcal{K}$, что соответствует определению проекционного алгоритма

$$(\mathbf{z}-\mathbf{k}, \mathbf{v})_A = (A(\mathbf{x}^* - (\mathbf{x}+\mathbf{k})), \mathbf{v}) = (\mathbf{b} - A\hat{\mathbf{x}}, \mathbf{v}) = 0 \quad \forall \mathbf{v} \in \mathcal{K}.$$

Такой подход к аппроксимации вектора $\mathbf{x}^*$ вектором $\hat{\mathbf{x}}$ называется *методом Галеркина*: $(\mathbf{b} - A\hat{\mathbf{x}}, \mathbf{v}) = 0 \quad \forall \mathbf{v} \in \mathcal{K}$. $\rhd$

5.164. Пусть A невырождена и $\mathcal{L} = A\mathcal{K}$. Показать, что вектор $\hat{\mathbf{x}}$ является результатом проекционного алгоритма тогда и только тогда, когда

$$E(\hat{\mathbf{x}}) = \min_{\tilde{\mathbf{x}}\in\mathbf{x}+\mathcal{K}} E(\tilde{\mathbf{x}}), \quad \text{где } E(\tilde{\mathbf{x}}) = (A(\mathbf{x}^*-\tilde{\mathbf{x}}), A(\mathbf{x}^*-\tilde{\mathbf{x}})) \text{ и } A\mathbf{x}^* = \mathbf{b}.$$

Указание. Решение совпадает с решением 5.163 с точностью до замены скалярного произведения на $(\mathbf{u}, \mathbf{v})_{A^TA} = (A^TA\mathbf{u}, \mathbf{v}) = (A\mathbf{u}, A\mathbf{v})$. При этом по условию $\mathbf{v} \in \mathcal{K}$, $A\mathbf{v} \in \mathcal{L}$. Такой подход к аппроксимации вектора $\mathbf{x}^*$ вектором $\hat{\mathbf{x}}$ называется *методом Петрова—Галеркина*: $(\mathbf{b} - A\hat{\mathbf{x}}, \mathbf{v}) = 0$ $\forall \mathbf{v} \in A\mathcal{K}$.

Одномерные проекционные методы. В простейшем случае в качестве базовых пространств $\mathcal{K}$ и $\mathcal{L}$ выбирают одномерные подпространства.

5.165. Показать, что проекционный алгоритм при $\mathcal{K} = \mathcal{L} = \{\mathbf{r}\}$, где $\mathbf{r} = \mathbf{b} - A\mathbf{x}$, соответствует методу наискорейшего градиентного спуска.

$\lhd$ Пространства $\mathcal{K}$ и $\mathcal{L}$ одномерны, следовательно, $\hat{\mathbf{x}} = \mathbf{x} + \tau\mathbf{r}$, и τ определяется из условия ортогональности $(\mathbf{b} - A(\mathbf{x} + \tau\mathbf{r}), \mathbf{r}) = 0$. Отсюда имеем $(\mathbf{r} - \tau A\mathbf{r}, \mathbf{r}) = 0$ и $\tau = \dfrac{(\mathbf{r}, \mathbf{r})}{(A\mathbf{r}, \mathbf{r})}$. $\rhd$

5.166. Показать, что проекционный алгоритм при $\mathcal{K} = \{\mathbf{r}\}$ и $\mathcal{L} = \{A\mathbf{r}\}$, где $\mathbf{r} = \mathbf{b} - A\mathbf{x}$, соответствует методу минимальных невязок.

О т в е т: в обозначениях 5.165 имеем $\tau = \dfrac{(A\mathbf{r}, \mathbf{r})}{(A\mathbf{r}, A\mathbf{r})}$.

5.167. Показать, что шаг проекционного метода при $\mathcal{K} = \{A^T\mathbf{r}\}$ и $\mathcal{L} = \{AA^T\mathbf{r}\}$ имеет вид $\hat{\mathbf{x}} = \mathbf{x} + \tau\mathbf{r}$, где $\mathbf{r} = \mathbf{b} - A\mathbf{x}$, $\tau = \dfrac{(A^T\mathbf{r}, A^T\mathbf{r})}{(AA^T\mathbf{r}, AA^T\mathbf{r})}$.

У к а з а н и е. Рассмотреть задачу $A^TA\mathbf{x} = A^T\mathbf{b}$.

5.168. Построить проекционный метод для пространств $\mathcal{K} = \mathcal{L} = \mathrm{span}\{\mathbf{r}, A\mathbf{r}\}$ и исследовать его сходимость.

5.169. Построить проекционный метод для пространств $\mathcal{K} = \mathrm{span}\{\mathbf{r}, A\mathbf{r}\}$ и $\mathcal{L} = A\mathcal{K}$ и исследовать его сходимость.

Проекционные методы в пространствах Крылова. Пусть пространства $\mathcal{L}$ зависят от номера итерации и $\mathcal{L}^1 \subset \mathcal{L}^2 \subset \ldots \subset \mathcal{L}^m \subset \ldots \subset \mathcal{L}^n = \mathbf{R}^n$. Тогда точное решение системы будет получено не позже, чем за n шагов. Если же цепочка $\mathcal{L}^m$ задается некоторым оптимальным образом, то можно рассчитывать, что требуемая точность $\|\mathbf{x}^* - \mathbf{x}^m\| \leqslant \varepsilon$, где $A\mathbf{x}^* = \mathbf{b}$, будет достигнута значительно раньше.

Эффективные алгоритмы удается построить, если в качестве $\mathcal{K}^m$ выбрать пространство Крылова $\mathcal{K}^m = \mathrm{span}\{\mathbf{r}, A\mathbf{r}, ..., A^{m-1}\mathbf{r}\}$ порядка m. При этом пространство $\mathcal{L}^m$ определяется как $\mathcal{L}^m = \mathcal{K}^m$ или как $\mathcal{L}^m = A\mathcal{K}^m$.

Метод сопряженных градиентов. Пусть $A = A^T > 0$. Построим проекционный метод для пары пространств

$$\mathcal{K}^m = \{\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{m-1}\mathbf{r}^0\}, \quad \mathcal{L}^m = \mathcal{K}^m, \quad \mathbf{r}^0 = \mathbf{b} - A\mathbf{x}^0,$$

при этом очередное приближение найдем в виде $\mathbf{x}^m = \mathbf{x}^0 + \sum\limits_{i=1}^{m} c_i A^{i-1}\mathbf{r}^0$, а коэффициенты c_i определим из условия $\mathbf{r}^m = (\mathbf{b} - A\mathbf{x}^m) \perp \mathcal{L}^m$. Такая форма алгоритма требует для нахождения c_i решения системы линейных уравнений. Рассмотрим эквивалентную, но более удобную с практической точки зрения, реализацию этого алгоритма.

Пусть в пространстве $\mathcal{K}^m = \{\mathbf{k}_1, \ldots, \mathbf{k}_m\}$ известен A-ортогональный базис, т. е. $(A\mathbf{k}_i, \mathbf{k}_j) = 0$ при $i \neq j$ и $\mathbf{k}_1 = \mathbf{r}^0$. Тогда $\mathbf{x}^m = \mathbf{x}^0 + \sum\limits_{i=1}^{m} \alpha_i \mathbf{k}_i$ и $\mathbf{r}^m = \mathbf{b} - A\mathbf{x}^m = \mathbf{b} - A\left(\mathbf{x}^0 + \sum\limits_{i=1}^{m} \alpha_i \mathbf{k}_i\right)$. В этом случае из условия $\mathbf{r}^m \perp \mathcal{L}^m$ имеем формулы для определения коэффициентов

$$(\mathbf{r}^m, \mathbf{k}_j) = (\mathbf{r}^0, \mathbf{k}_j) - \alpha_j (A\mathbf{k}_j, \mathbf{k}_j) = 0, \quad \alpha_j = \frac{(\mathbf{r}^0, \mathbf{k}_j)}{(A\mathbf{k}_j, \mathbf{k}_j)}\,,\ j = 1, \ldots, m.$$

Заметим, что $\mathbf{x}^m = \mathbf{x}^{m-1} + \alpha_m \mathbf{k}_m$. Отсюда следует, что $\mathbf{r}^m = \mathbf{r}^{m-1} - \alpha_m A\mathbf{k}_m$ и $\alpha_m = \dfrac{(\mathbf{r}^{m-1}, \mathbf{k}_m)}{(A\mathbf{k}_m, \mathbf{k}_m)}$. Для вычислений такая рекуррентная форма записи предпочтительнее.

Построим соответствующий рекуррентный алгоритм для определения $\{\mathbf{k}_i\}$, так как стандартная процедура типа Грама—Шмидта, требующая хранения всех элементов базиса $\{\mathbf{k}_i\}_{i=1}^m$, в данном случае оказывается существенно менее эффективна. Имеем

$$\operatorname{span}\Big\{\{\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{m-1}\mathbf{r}^0\}, A^m\mathbf{r}^0\Big\} = \operatorname{span}\Big\{\{\mathbf{k}_1, \ldots, \mathbf{k}_m\}, \mathbf{k}_{m+1}\Big\}.$$

Отсюда следует, что $\mathbf{k}_{m+1} = A^m\mathbf{r}^0 + \sum\limits_{i=1}^{m} \tilde{\beta}_i \mathbf{k}_i$. Так как

$$\mathbf{r}^m = \mathbf{r}^0 - A\sum_{i=1}^{m} c_i A^{i-1}\mathbf{r}^0 = \mathbf{r}^0 - \sum_{i=1}^{m} c_i A^i \mathbf{r}^0,$$

то при $\mathbf{r}^m \neq 0$ и $c_m \neq 0$ вектор $\mathbf{k}_{m+1}$ можно искать в виде $\mathbf{k}_{m+1} = \mathbf{r}^m + \sum\limits_{i=1}^{m} \beta_i \mathbf{k}_i$. Из условия $\mathbf{r}^m \perp \mathcal{L}^m$ следует равенство $(\mathbf{r}^m, A\mathbf{k}_i) = 0$ при $i < m$. Отсюда и из A-ортогональности векторов $\mathbf{k}_i$ имеем $\beta_i = 0$ при $i < m$, следовательно, $\mathbf{k}_{m+1} = \mathbf{r}^m + \beta_m \mathbf{k}_m$ и $\beta_m = -\dfrac{(\mathbf{r}^m, A\mathbf{k}_m)}{(A\mathbf{k}_m, \mathbf{k}_m)}$. Приведем формулы рекуррентного пересчета для очередного приближения $\mathbf{x}^m$ и базисного вектора $\mathbf{k}_m$:

$$\mathbf{x}^m = \mathbf{x}^{m-1} + \alpha_m \mathbf{k}_m, \quad \alpha_m = \frac{(\mathbf{r}^{m-1}, \mathbf{k}_m)}{(A\mathbf{k}_m, \mathbf{k}_m)}\,,$$

$$\mathbf{k}_{m+1} = \mathbf{r}^m + \beta_m \mathbf{k}_m, \quad \beta_m = -\frac{(\mathbf{r}^m, A\mathbf{k}_m)}{(A\mathbf{k}_m, \mathbf{k}_m)}\,,\ \mathbf{k}_1 = \mathbf{r}^0.$$

На шаге m данного метода минимизируется A-норма вектора ошибки на подпространствах Крылова $\mathcal{K}^m$, поэтому с точки зрения проекционных методов метод сопряженных градиентов является (см. 5.129) обобщением метода наискорейшего градиентного спуска. Метод сопряженных градиентов минимизирует значение функционала $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$ на векторах вида $\mathbf{x}^m = \mathbf{x}^0 + \sum\limits_{i=1}^{m} c_i A^{i-1}\mathbf{r}^0$ относительно c_i.

5.170. Показать, что для метода сопряженных градиентов для матриц $A = A^T > 0$ имеет место следующая оценка скорости сходимости:

$$\|\mathbf{z}^N\|_A \leqslant 2\left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right)^N \|\mathbf{z}^0\|_A,$$

где $\mathbf{z}^N = \mathbf{x}^* - \mathbf{x}^N$, $A\mathbf{x}^* = \mathbf{b}$ и $\lambda(A) \in [m, M]$.

$\lhd$ Сравним задачи минимизации ошибки, соответствующие методу сопряженных градиентов (см. 5.163) и оптимальному линейному N-шаговому методу (5.8). Принимая во внимание, что оптимальный N-шаговый процесс представляет собой метод простой итерации с чебышёвским набором параметров, найдем приближение $\mathbf{x}_{ch}^k$. Имеем

$$\frac{\mathbf{x}_{ch}^k - \mathbf{x}_{ch}^{k-1}}{\tau_k} + A\mathbf{x}_{ch}^{k-1} = \mathbf{b}, \quad k = 1, \ldots, N.$$

Отсюда следует, что

$$\mathbf{x}_{ch}^1 = \mathbf{x}^0 + \tau_1\mathbf{r}^0, \quad \mathbf{x}_{ch}^k = \mathbf{x}_{ch}^{k-1} + \tau_k(\mathbf{b} - A\mathbf{x}_{ch}^{k-1}).$$

По индукции, предполагая, что

$$\mathbf{x}_{ch}^{k-1} = \mathbf{x}^0 + \sum_{i=1}^{k-1} \tilde{c}_i A^{i-1}\mathbf{r}^0,$$

находим $\mathbf{x}_{ch}^N = \mathbf{x}^0 + \sum\limits_{i=1}^{N} \hat{c}_i A^{i-1}\mathbf{r}^0$. Учитывая, что $\mathbf{z}_{ch}^N = \prod\limits_{i=1}^{N}(1 - \tau_i A)\mathbf{z}^0$, получаем

$$\|\mathbf{z}_{ch}^N\|_A \leqslant \|\prod_{i=1}^{N}(1-\tau_i A)\|_A \|\mathbf{z}^0\|_A = \|\prod_{i=1}^{N}(1-\tau_i A)\|_2 \|\mathbf{z}^0\|_A \leqslant$$
$$\leqslant \frac{2q^N}{1+q^{2N}}\|\mathbf{z}^0\|_A \leqslant 2q^N\|\mathbf{z}^0\|_A, \quad q = \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}.$$

По определению приближение $\mathbf{x}_{cg}^N$ метода сопряженных градиентов имеет вид: $\mathbf{x}_{cg}^N = \mathbf{x}^0 + \sum\limits_{i=1}^{N} c_i A^{i-1}\mathbf{r}^0$. Отсюда следует, что приближения $\mathbf{x}_{ch}^N$ и $\mathbf{x}_{cg}^N$ могут отличаться только коэффициентами.

Так как вектор $\mathbf{x}_{cg}^N$ является решением задачи минимизации $\min\limits_{\mathbf{x}^N - \mathbf{x}^0 \in \mathcal{K}} (A\mathbf{z}^N, \mathbf{z}^N)$, где $\mathbf{z}^N = \mathbf{x}^* - \mathbf{x}^N$, то справедливо неравенство

$$(A\mathbf{z}_{cg}^N, \mathbf{z}_{cg}^N) = \min_{\mathbf{x}^N - \mathbf{x}^0 \in \mathcal{K}} (A\mathbf{z}^N, \mathbf{z}^N) \leqslant (A\mathbf{z}_{ch}^N, \mathbf{z}_{ch}^N)$$

и требуемая оценка. $\rhd$

5.171. Показать, что в методе сопряженных градиентов необходимыми и достаточными условиями минимума функционала $F(\mathbf{x}^m) = (A\mathbf{x}^m, \mathbf{x}^m) - 2(\mathbf{b}, \mathbf{x}^m)$ для любого $m \geqslant 1$ являются равенства $(\mathbf{r}^m, \mathbf{r}^j) = 0$, $j = 0, 1, \ldots, m-1$.

5.172. Показать, что в методе сопряженных градиентов для любого $m \geqslant 2$ имеют место соотношения ортогональности $(A\mathbf{r}^m, \mathbf{r}^j) = 0$, $j = 0, 1, \ldots, m-2$.

5.173. Получить следующие эквивалентные формулы метода сопряженных градиентов:

$$\alpha_m = \frac{(\mathbf{r}^{m-1}, \mathbf{r}^{m-1})}{(A\mathbf{k}_m, \mathbf{k}_m)}, \ \mathbf{x}^m = \mathbf{x}^{m-1} + \alpha_m \mathbf{k}_m, \ \mathbf{r}^m = \mathbf{r}^{m-1} - \alpha_m A\mathbf{k}_m,$$

$$\beta_m = \frac{(\mathbf{r}^m, \mathbf{r}^m)}{(\mathbf{r}^{m-1}, \mathbf{r}^{m-1})}, \ \mathbf{k}_{m+1}\mathbf{r}^m + \beta_m \mathbf{k}_m.$$

$\lhd$ Так как $\mathbf{k}_m = \mathbf{r}^{m-1} + \beta_{m-1}\mathbf{k}_{m-1}$ и $(\mathbf{r}^{m-1}, \mathbf{k}_{m-1}) = 0$, то $\alpha_m = \frac{(\mathbf{r}^{m-1}, \mathbf{r}^{m-1})}{(A\mathbf{k}_m, \mathbf{k}_m)}$. Далее, если $\mathbf{r}^{m-1} \neq 0$, то $\alpha_m \neq 0$ и $A\mathbf{k}_m = -\frac{(\mathbf{r}^m - \mathbf{r}^{m-1})}{\alpha_m}$. Таким образом, $(A\mathbf{k}_m, \mathbf{r}^m) = -\frac{(\mathbf{r}^m, \mathbf{r}^m)}{\alpha_m}$, следовательно, $\beta_m = \frac{(\mathbf{r}^m, \mathbf{r}^m)}{(\mathbf{r}^{m-1}, \mathbf{r}^{m-1})}$. $\rhd$

5.174. Доказать эквивалентную запись метода сопряженных градиентов:

$$\gamma_m = \frac{(\mathbf{r}^m, \mathbf{r}^m)}{(A\mathbf{r}^m, \mathbf{r}^m)}, \ \rho_m = \left[1 - \frac{\gamma_m}{\gamma_{m-1}} \frac{(\mathbf{r}^m, \mathbf{r}^m)}{(\mathbf{r}^{m-1}, \mathbf{r}^{m-1})} \frac{1}{\rho_{m-1}}\right]^{-1},$$

$$\mathbf{x}^{m+1} = \rho_m(\mathbf{x}^m + \gamma_m \mathbf{r}^m) + (1 - \rho_m)\mathbf{x}^{m-1},$$

$$\mathbf{r}^{m+1} = \rho_m(\mathbf{r}^m - \gamma_m A\mathbf{r}^m) + (1 - \rho_m)\mathbf{r}^{m-1},$$

где $\mathbf{r}^0 = \mathbf{b} - A\mathbf{x}^0$, $\mathbf{x}^{-1} = 0$ и $\rho_0 = 1$.

Обобщенный метод минимальных невязок. Далее будут полезны следующие обозначения: для множества вещественных прямоугольных матриц

$$\mathbf{R}^{m\times n} = \{A : \ a_{ij} \in \mathbf{R}^1, \ 1 \leqslant i \leqslant m, \ 1 \leqslant j \leqslant n\}$$

и для вещественных векторов

$$\mathbf{R}^n = \{\mathbf{v} : v_i \in \mathbf{R}^1, \ 1 \leqslant i \leqslant n\}.$$

Рассматриваемый ниже алгоритм GMRES (General Minimum Residual Method) предназначен для решения разреженных невырожденных линейных систем большой размерности. При этом симметрия и положительная определенность матрицы системы не предполагается, т. е. решается невырожденная система общего вида $A\,\mathbf{x} = \mathbf{b}$, $A \in \mathbf{R}^{n\times n}$, $\mathbf{b} \in \mathbf{R}^n$.

Важным элементом метода является использование пространств Крылова m-го порядка: $\mathcal{K}^m = \operatorname{span}\{\mathbf{r}^0, A\mathbf{r}^0, ..., A^{m-1}\mathbf{r}^0\}$, где $\mathbf{r}^0 = \mathbf{b} - A\,\mathbf{x}^0$, $\mathbf{x}^0$ — начальное приближение.

Для каждого m построим в пространстве $\mathcal{K}^m$ ортонормированный базис $\{\mathbf{k}_1, ..., \mathbf{k}_m\}$ рекуррентным образом: очередной вектор $\mathbf{k}_{m+1}$ определим из условия ортогональности вектора $A\mathbf{k}_m$ уже найденным векторам $\mathbf{k}_1, ..., \mathbf{k}_m$. Это удобно сделать с помощью следующего алгоритма:

1. $\mathbf{k}_1 = \dfrac{\mathbf{r}^0}{\|\mathbf{r}^0\|_2}$.
2. Цикл для $j = 1, 2, \ldots, m$.
3. Вычисляем скалярные произведения: $h_{ij} = (A\mathbf{k}_j, \mathbf{k}_i)$ для $i = 1, \ldots, j$.
4. Строим вектор: $\mathbf{z}_{j+1} = A\mathbf{k}_j - \sum\limits_{i=1}^{j} h_{ij}\mathbf{k}_i$.
5. Вычисляем его норму: $h_{j+1,j} = \|\mathbf{z}_{j+1}\|_2$.
6. Если $h_{j+1,j} = 0$, то останавливаемся.
7. Строим очередной вектор базиса: $\mathbf{k}_{j+1} = \dfrac{\mathbf{z}_{j+1}}{h_{j+1,j}}$.
8. Конец цикла по j.

Этот процесс требует хранения всех предыдущих элементов базиса. Будем их хранить в виде матриц $K_m = (\mathbf{k}_1 \ \ldots \ \mathbf{k}_m)$, $K_m \in \mathbf{R}^{n\times m}$, столбцами которых являются найденные $\mathbf{k}_j$.

Проанализируем информацию об исходной матрице A, полученную на основе этого алгоритма ортогонализации. Построим ортогональную матрицу $Q = (K_m\, K_{n-m}) \in \mathbf{R}^{n\times n}$, где $K_{n-m} = (\mathbf{k}_{m+1} \ \ldots \ \mathbf{k}_n)$. Учитывая, что в результате применения алгоритма вычислены только матрица K_m и вектор $\mathbf{k}_{m+1}$, имеем

$$H = Q^T AQ = (K_m\, K_{n-m})^T A(K_m\, K_{n-m}) =$$
$$= \begin{pmatrix} K_m^T AK_m & K_m^T AK_{n-m} \\ K_{n-m}^T AK_m & K_{n-m}^T AK_{n-m} \end{pmatrix} = \begin{pmatrix} H_m & H_{m,n-m} \\ H_{n-m,m} & H_{n-m} \end{pmatrix},$$

где $H_m \in \mathbf{R}^{m\times m}$ — некоторая *верхняя хессенбергова матрица* (т. е. $h_{ij} = 0$ при $i > j+1$). Отметим, что, в силу сохранения структуры H_m при любом m, в матрице $H_{n-m,m}$ имеется единственный ненулевой элемент $h_{m+1,m}$, который расположен в ее правом верхнем углу.

Итак, матрицы $H_{m,n-m}$ и H_{n-m} неизвестны, мы знаем только блоки H_m и $H_{n-m,m}$.

Определим m-е приближение к решению $\mathbf{x}$ по формуле

$$\mathbf{x}^m = \mathbf{x}^0 + K_m\mathbf{c}_m, \quad \mathbf{c}_m \in \mathbf{R}^m\,;$$

тогда для невязок $\mathbf{r}^m = \mathbf{b} - A\,\mathbf{x}^m$ справедливо соотношение

$$\mathbf{r}^m = \mathbf{r}^0 - AK_m\mathbf{c}_m\,.$$

Здесь неизвестным является вектор $\mathbf{c}_m \in \mathbf{R}^m$. Будем его искать из условия минимума евклидовой нормы невязки $\mathbf{r}^m$: $\|\mathbf{r}^0 - AK_m\mathbf{c}_m\|_2 \to \min$. Из полученного выше представления матрицы A имеем

$$\|\mathbf{r}^0 - AK_m\mathbf{c}_m\|_2 = \|\mathbf{r}^0 - (QHQ^T)K_m\mathbf{c}_m\|_2 = \|Q^T\mathbf{r}^0 - HQ^TK_m\mathbf{c}_m\|_2\,.$$

Последнее равенство справедливо в силу сохранения ортогональным преобразованием евклидовой длины вектора. Согласно определению вектора $\mathbf{k}_1$ и его ортогональности всем остальным векторам базиса, имеем $Q^T\mathbf{r}^0 = \|\mathbf{r}^0\|_2\mathbf{e}_1$, где $\mathbf{e}_1 = (1, 0, 0, \ldots, 0)^T \in \mathbf{R}^n$. Кроме того, вектор $Q^TK_m\mathbf{c}_m \in \mathbf{R}^n$, по определению матрицы Q, можно представить в виде

$(\mathbf{c}_m, 0, 0, \ldots, 0)^T \in \mathbf{R}^n$, поэтому последнее равенство для невязки можно переписать в виде

$$\|Q^T\mathbf{r}^0 - HQ^TK_m\mathbf{c}_m\|_2 = \left\|\|\mathbf{r}^0\|_2\mathbf{e}_1 - \begin{pmatrix} H_m & H_{m,n-m} \\ H_{n-m,m} & H_{n-m} \end{pmatrix}\begin{pmatrix} \mathbf{c}_m \\ 0 \end{pmatrix}\right\|_2 = \left\|\|\mathbf{r}^0\|_2\mathbf{e}_1 - \begin{pmatrix} H_m \\ H_{n-m,m} \end{pmatrix}\mathbf{c}_m\right\|_2 .$$

В последнем выражении фигурируют только известные (вычисленные ранее) блоки.

Уберем из вектора, стоящего под знаком нормы, нулевые компоненты т. е. компоненты с номерами $i \geqslant m+2$. Для этого определим матрицу $\hat{H}_m \in \mathbf{R}^{(m+1)\times m}$ как

$$\hat{H}_m = \begin{pmatrix} H_m \\ 0\,0\,\ldots\,0\,h_{m+1,m} \end{pmatrix}$$

и рассмотрим ее QR-разложение: $\hat{H}_m = U_mR_m$, где $U_m \in \mathbf{R}^{(m+1)\times m}$, $U_m^TU_m = I \in \mathbf{R}^{m\times m}$, $R_m \in \mathbf{R}^{m\times m}$. Имеем

$$\left\|\|\mathbf{r}^0\|_2\mathbf{e}_1 - \begin{pmatrix} H_m \\ H_{n-m,m} \end{pmatrix}\mathbf{c}_m\right\|_2 = \|\|\mathbf{r}^0\|_2\tilde{\mathbf{e}}_1 - \hat{H}_m\mathbf{c}_m\|_2 = \|\|\mathbf{r}^0\|_2\tilde{\mathbf{e}}_1 - U_mR_m\mathbf{c}_m\|_2,$$

где $\tilde{\mathbf{e}}_1 = (1, 0, 0, \ldots, 0)^T \in \mathbf{R}^{m+1}$. Минимум полученного выражения по всем векторам $\mathbf{c}_m$ достигается на векторе, удовлетворяющем уравнению

$$R_m\mathbf{c}_m = \|\mathbf{r}^0\|_2U_m^T\tilde{\mathbf{e}}_1$$

(невырожденность R_m следует из невырожденности H_m). Окончательно имеем

$$\mathbf{x}^m = \mathbf{x}^0 + \|\mathbf{r}^0\|_2K_mR_m^{-1}U_m^T\tilde{\mathbf{e}}_1.$$

Таким образом, обобщенный метод минимальных невязок можно сформулировать в следующем виде для $m = 1, 2, \ldots$:

1. Вычисляем матрицу $\hat{H}_m$.
2. Находим ее QR-разложение: $\hat{H}_m = U_mR_m$.
3. Решаем систему $R_m\mathbf{c}_m = U_m^T\tilde{\mathbf{e}}_1$.
4. Находим приближение $\mathbf{x}^m = \mathbf{x}^0 + \|\mathbf{r}^0\|_2K_mR_m^{-1}U_m^T\tilde{\mathbf{e}}_1$.

Алгоритм завершается, когда норма вектора невязки $\mathbf{r}^m$ становится достаточно малой. Ограничения по памяти и накопление вычислительной погрешности при решении промежуточных задач могут приводить к необходимости перезапуска (restart) алгоритма на шаге M с новым начальным вектором $\mathbf{x}^0_{\text{new}} = \mathbf{x}^M$.

Наиболее трудоемкой процедурой в методе является QR-разложение матрицы $\hat{H}_m$. В общем случае для этого требуется $O(m^3)$ арифметических операций, однако так как $\hat{H}_m$ — хессенбергова матрица, можно сократить требуемый объем до $O(m^2)$ действий.

Оценки скорости сходимости для алгоритмов такого типа малоинформативны и поэтому редко применяются на практике. Рассмотрим пример. Пусть матрица системы A диагонализуема и имеет только вещественные собственные значения, т. е. $A = S^{-1}\Lambda S$, где $\det(S) \neq 0$, $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Определим величину

$$q_m = \min_{\substack{p(x)\in P_m(x)\\ p(0)=1}} \max_{k=1,2,\ldots,n} |p(\lambda_k)|\,.$$

Тогда справедлива оценка

$$\|\mathbf{r}^m\|_2 \leqslant q_m \operatorname{cond}_2(S)\|\mathbf{r}^0\|_2\,.$$

Если $A = A^T > 0$, то $S^{-1} = S^T$ ($\operatorname{cond}_2(S) = 1$) и q_m такое же, как в оценке для метода сопряженных градиентов. Отметим различие в оценках: в методе сопряженных градиентов оценивается ошибка в A-норме, а здесь невязка — в евклидовой норме.

5.175. Система n уравнений $A\mathbf{x} = \mathbf{b}$ с матрицей

$$A = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0\\ 0 & 0 & 1 & \ldots & 0 & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & 0 & \ldots & 0 & 1\\ 1 & 0 & 0 & \ldots & 0 & 0 \end{pmatrix}.$$

и вектором $\mathbf{b} = (0, 0, \ldots, 0, 1)^T$ решается обобщенным методом минимальных невязок с начальным вектором $\mathbf{x}^0 = (0, 0, \ldots, 0, 0)^T$. Найти пространства Крылова $\mathcal{K}^m = \operatorname{span}\{\mathbf{r}^0, A\mathbf{r}^0, ..., A^{m-1}\mathbf{r}^0\}$ для $m \geqslant 1$ и определить количество итераций, необходимое для нахождения точного решения $\mathbf{x} = (1, 0, 0, \ldots, 0)^T$.

О т в е т: обозначим через $\mathbf{e}_i$ вектор с единственной ненулевой i-й компонентой; тогда пространства Крылова имеют вид

$$\mathcal{K}^m = \operatorname{span}\{\mathbf{e}_n, \mathbf{e}_{n-1}, \ldots, \mathbf{e}_{n-m+1}\}, \quad 1 \leqslant m \leqslant n,$$

а последовательные приближения определяются как

$$\mathbf{x}^0 = \mathbf{x}^1 = \cdots = \mathbf{x}^{n-1} = (0, 0, \ldots, 0, 0)^T, \quad \mathbf{x}^n = \mathbf{e}_1,$$

т. е. метод сходится ровно за n итераций.

5.8. Некорректные системы линейных уравнений

Пусть требуется решить систему линейных уравнений с матрицей A размерности $m \times n$

$$\begin{pmatrix} a_{11} & \ldots & a_{1n}\\ \ldots & \ldots & \ldots\\ a_{m1} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1\\ \ldots\\ x_n \end{pmatrix} = \begin{pmatrix} b_1\\ \ldots\\ b_m \end{pmatrix}.$$

Рассмотрим три случая: 1) $m = n$, $\det(A) \neq 0$; 2) $m < n$ и строки линейно независимы, т. е. $\operatorname{rank}(A) = m$; 3) $m > n$ и $\operatorname{rank}(A) = n$.

В случае 1) задача невырождена и вектор $\mathbf{x} = A^{-1}\mathbf{b}$ является точным решением. Для вектора невязки $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ имеем $\|\mathbf{r}\| = 0$.

В случае 2) задача недоопределена. Исходная система имеет целое подпространство решений размерности $n-m$. Для каждого решения имеем $\|\mathbf{r}\| = 0$. В случае 3) система переопределена и если она несовместна, то точного решения не существует, т. е. для произвольного $\mathbf{x} \in \mathbf{R}^n$ имеем $\|\mathbf{b} - A\mathbf{x}\| = \|\mathbf{r}\| > 0$. Представляют интерес методы решения переопределенных задач. Поэтому, если не оговаривается иное, считаем, что $m > n$ и $\operatorname{rank}(A) = n$. Для задач такого рода Гаусс предложил считать решением вектор $\mathbf{x}$, минимизирующий евклидову норму вектора невязки $\min\limits_{\mathbf{y}} \|\mathbf{b}-A\mathbf{y}\|_2 = \|\mathbf{b}-A\mathbf{x}\|_2$. Рассмотрим некоторые методы решения данной минимизационной задачи, называемой *задачей наименьших квадратов* (ЗНК).

Метод нормального уравнения. Рассматривают следующую, называемую *нормальной*, систему уравнений $A^TA\mathbf{x} = A^T\mathbf{b}$ с квадратной матрицей A^TA размерности $n \times n$. Отсюда находят вектор $\mathbf{x}$.

5.176. Показать, что нормальное уравнение имеет единственное решение.

$\lhd$ Действительно, $A^TA = (A^TA)^T$ и $(A^TA\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) > 0$, если только $A\mathbf{x} \neq 0$. Но $A\mathbf{x} \neq 0$ для всякого $\mathbf{x} \neq 0$, так как $\operatorname{rank}(A) = n$. Следовательно, матрица A^TA невырождена и нормальное уравнение имеет единственное решение. $\rhd$

5.177. Показать, что вектор $\mathbf{x}$ — решение задачи наименьших квадратов $\min\limits_{\mathbf{y}} \|\mathbf{b} - A\mathbf{y}\|_2$ тогда и только тогда, когда $\mathbf{x}$ — решение системы $A^TA\mathbf{x} = A^T\mathbf{b}$.

$\lhd$ Из 5.176 следует существование и единственность такого вектора $\mathbf{x}$, что $A^T(\mathbf{b} - A\mathbf{x}) = 0$. Рассмотрим $\mathbf{y} = \mathbf{x} + \Delta\mathbf{x}$. В этом случае $\|\mathbf{b} - A\mathbf{y}\|_2^2 = (\mathbf{b} - A(\mathbf{x} + \Delta\mathbf{x}), \mathbf{b} - A(\mathbf{x} + \Delta\mathbf{x})) = (\mathbf{b} - A\mathbf{x}, \mathbf{b} - A\mathbf{x}) + (A\Delta\mathbf{x}, A\Delta\mathbf{x}) + 2(A\Delta\mathbf{x}, \mathbf{b}-A\mathbf{x}) = \|\mathbf{b}-A\mathbf{x}\|_2^2 + (A\Delta\mathbf{x}, A\Delta\mathbf{x}) + 2(\Delta\mathbf{x}, A^T(\mathbf{b}-A\mathbf{x}))$. Следовательно, минимум достигается при $\Delta\mathbf{x} = 0$, т. е. на векторе $\mathbf{y} = \mathbf{x}$. $\rhd$

Метод QR-разложения. Метод нормального уравнения прост в реализации, однако в приближенной арифметике неустойчив для почти вырожденных задач большой размерности. Например, в случае квадратной матрицы $A = A^T$ имеем $\operatorname{cond}_2(A^TA) = \operatorname{cond}_2^2(A)$. Поэтому численное решение может сильно отличаться от точного. Метод, основанный на QR-разложении матрицы A, более устойчив к вычислительной погрешности. Разложение $A = QR$ при $Q^TQ = I$, $\det R \neq 0$ можно построить методом отражений или методом вращений.

5.178. Пусть известно представление $A = QR$, $Q^TQ = I$, $\det R \neq 0$. Показать, что решение $\mathbf{x}$ задачи наименьших квадратов является решением системы $R\mathbf{x} = Q^T\mathbf{b}$.

$\lhd$ Из метода нормального уравнения следует, что $\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b}$, поэтому

$$\mathbf{x} = (R^TQ^TQR)^{-1}R^TQ^T\mathbf{b} = (R^TR)^{-1}R^TQ^T\mathbf{b} = R^{-1}R^{-T}R^TQ^T\mathbf{b} = R^{-1}Q^T\mathbf{b}.$$

Таким образом, искомый вектор $\mathbf{x}$ является решением системы $R\mathbf{x} = Q^T\mathbf{b}$. $\rhd$

Формально этот метод более трудоемкий, но, построив однажды QR-разложение, можно быстро решать задачи с различными правыми частями.

Вырожденные задачи. Задача наименьших квадратов называется *вырожденной*, если $\operatorname{rank}(A) < n$. При численном решении вырожденных и почти вырожденных систем требуется изменить постановку задачи и соответственно применять другие методы. Рассмотрим следующий пример: найти решение при $m = n = 2$

$$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Для данного уравнения имеется семейство решений $\mathbf{x} = (1 - x_2, x_2)^T$, $\mathbf{r} = \mathbf{b} - A\mathbf{x} = (0, 1)^T$. Можно выбрать решения как с нормами порядка единицы, так и со сколь угодно большими. Однако для возмущенной задачи

$$\begin{pmatrix} 1 & 1 \\ 0 & \varepsilon \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

формально близкой к исходной при малых ε, имеется единственное решение $(1 - \varepsilon^{-1}, \varepsilon^{-1})^T$ с большой нормой порядка ε^{-1}. Это означает, что сколь угодно малое возмущение элементов матрицы может существенно изменить структуру и норму решения.

5.179. Пусть $\operatorname{rank}(A) = r$, $A \in \mathbf{R}^{m\times n}$, $m \geqslant n$ и $r < n$. Показать, что множество векторов $\mathbf{x}$, минимизирующих $\|\mathbf{b} - A\mathbf{x}\|_2$, образует $(n - r)$-мерное линейное подпространство.

$\lhd$ Пусть вектор $\mathbf{z} \in \ker(A)$ и $\dim(\ker(A)) = n - r$. Тогда $A\mathbf{z} = 0$, и если $\mathbf{x}$ минимизирует $\|\mathbf{b} - A\mathbf{x}\|_2$, то $\mathbf{x} + \mathbf{z}$ также минимизирует невязку, так как $\|\mathbf{b} - A(\mathbf{x} + \mathbf{z})\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2$. $\rhd$

5.180. Пусть ранг матрицы A в точной арифметике равен $r < n$ и первые r столбцов линейно независимы. Показать, что матрицу можно привести

к виду

$$A = QR = Q\begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix},$$

$$Q \in \mathbf{R}^{m\times n}, Q = (\mathbf{q}_1 \ldots \mathbf{q}_r\, \mathbf{q}_{r+1} \ldots \mathbf{q}_n) = (Q_1\, Q_2), Q^TQ = I,$$

$$\det R_{11} \neq 0, \quad R_{11} \in \mathbf{R}^{r\times r}, \quad R_{12} \in \mathbf{R}^{r\times(n-r)},$$

и для задачи наименьших квадратов с матрицей A имеется семейство решений

$$\mathbf{x} = \left(R_{11}^{-1}(Q_1^T\mathbf{b} - R_{12}\mathbf{x}_2), \mathbf{x}_2\right)^T.$$

Здесь $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)^T, \mathbf{x}_1 \in \mathbf{R}^r, \mathbf{x}_2 \in \mathbf{R}^{n-r}$.

$\lhd$ Искомое разложение $A = QR$, где $Q \in \mathbf{R}^{m\times n}$, а R имеет указанный в условии вид, можно построить, например, методом ортогонализации Грама—Шмидта. Решим задачу наименьших квадратов. Построим по матрице Q ортогональную матрицу $U \in \mathbf{R}^{m\times m}$ следующим образом. Дополним векторы $\mathbf{q}_i$ до базиса в пространстве $\mathbf{R}^m$ некоторой линейно независимой системой $\mathbf{p}_j$ и проведем ортогонализацию столбцов матрицы $B = (Q\,P)$ методом Грама—Шмидта. Для полученной матрицы U имеем $U = (Q\,\tilde{Q}) = (\mathbf{q}_1 \ldots \mathbf{q}_m)$. Матрица U ортогональная, поэтому ортогональны векторы $(\mathbf{q}_i, \mathbf{q}_j) = 0$, $i \neq j$, и $\tilde{Q}^TQ$ — нулевая матрица. Отсюда следует цепочка равенств:

$$\|\mathbf{b} - A\mathbf{x}\|_2^2 = \|U^T(\mathbf{b} - QR\mathbf{x})\|_2^2 = \left\|\begin{pmatrix} Q^T \\ \tilde{Q}^T \end{pmatrix}(\mathbf{b} - QR\mathbf{x})\right\|_2^2 =$$

$$= \left\|\begin{matrix} Q^T\mathbf{b} - R\mathbf{x} \\ \tilde{Q}^T(\mathbf{b} - QR\mathbf{x}) \end{matrix}\right\|_2^2 = \left\|\begin{matrix} Q^T\mathbf{b} - R\mathbf{x} \\ \tilde{Q}^T\mathbf{b} \end{matrix}\right\|_2^2 =$$

$$= \|Q_1^T\mathbf{b} - R_{11}\mathbf{x}_1 - R_{12}\mathbf{x}_2\|_2^2 + \|Q_2^T\mathbf{b}\|_2^2 + \|\tilde{Q}^T\mathbf{b}\|_2^2,$$

приводящая к ответу. $\rhd$

Метод QR-разложения с выбором главного столбца. Преобразуем исходную ЗНК так, чтобы первые r столбцов полученной матрицы $\tilde{A}$ были линейно независимы. Для этого в процессе вычислений переставим столбцы в матрице A так, что $\tilde{A} = AP = QR$, где P — некоторая матрица перестановок. Отсюда (см. 5.180) найдем решение задачи наименьших квадратов. Цель соответствующих перестановок — получить в матрице $R = \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}$ как можно лучше обусловленный блок R_{11} и как можно меньшие по модулю элементы в R_{22}. В приближенных вычислениях блок R_{22} обычно отличен от нуля, хотя исходная задача могла быть неполного ранга.

Вычисления проводятся на основе стандартного QR-разложения, например, методом отражений. На k-м шаге ($k = 1, \ldots, n$) в матрице $A^{(k)}$ выбирают столбец с номером j_k, $k \leqslant j_k \leqslant n$, с наибольшей величиной

$\max\limits_{k\leqslant j\leqslant n}\left(\sum\limits_{i=k}^{m}\left(a_{ij}^{(k)}\right)^2\right)^{1/2}$. Если таких столбцов несколько, то берут произвольный из них. В матрице $A^{(k)}$ найденный столбец j_k переставляют с k-м столбцом. Далее реализуют очередной шаг QR-разложения.

5.181. Оценить величину элемента r_{nn} в методе QR-разложения с выбором главного столбца для $A=\begin{pmatrix}1 & -1 & \dots & -1\\ 0 & 1 & -1 & \dots\\ \dots & \dots & \dots & \dots\\ 0 & \dots & 0 & 1\end{pmatrix}$, $A\in R^{n\times n}$.

Метод сингулярного (*SVD*) **разложения.** Метод применяют для решения гарантированно наилучшим образом плохо обусловленных и вырожденных задач.

Теорема. *Пусть A — матрица размерности $m\times n$, $m\geqslant n$. Тогда справедливо сингулярное разложение $A=U\Sigma V^T$, где:*

U — ортогональная матрица размерности $m\times n$, $U^TU=I$;

V — ортогональная матрица размерности $n\times n$;

Σ — диагональная матрица размерности $n\times n$ с элементами $\sigma_1\geqslant\sigma_2\geqslant\ldots\geqslant\sigma_n\geqslant 0$.

Столбцы $\mathbf{u}_1,\ldots,\mathbf{u}_n$ матрицы U называют *левыми сингулярными векторами* матрицы A, столбцы $\mathbf{v}_1,\ldots,\mathbf{v}_n$ матрицы V — *правыми сингулярными векторами*, величины σ_i — *сингулярными числами*.

Построив *SVD*-разложение, можно установить, является ли задача вырожденной ($\sigma_n=0$), невырожденной ($\sigma_n\neq 0$) или «хорошей» ($\frac{\sigma_n}{\sigma_1}$ не слишком мало).

Если $m<n$, то сингулярное разложение строят для матрицы A^T. Если $m=n$ и $A=A^T$, то сингулярные числа $\sigma_i=|\lambda_i|$, т. е. с точностью до знака совпадают с собственными числами, сингулярные векторы $\mathbf{v}_i$ являются соответствующими собственными векторами.

Геометрическая интерпретация SVD-разложения. Рассмотрим оператор A, переводящий элемент $\mathbf{x}\in\mathbf{R}^n$ в элемент $\mathbf{y}\in\mathbf{R}^m$. Единичная сфера под действием A переходит в эллипсоид. Векторы $\mathbf{u}_i$ задают полуоси эллипсоида, $\mathbf{v}_i$ — их прообразы, σ_i — коэффициенты удлинения векторов $\mathbf{v}_i$.

Алгебраическая интерпретация SVD-разложения. Рассмотрим оператор A, переводящий элемент $\mathbf{x}\in\mathbf{R}^n$ в элемент $\mathbf{y}\in\mathbf{R}^m$. В этом случае в пространстве $\mathbf{R}^n$ существует базис $\mathbf{v}_1,\ldots,\mathbf{v}_n$, а в пространстве $\mathbf{R}^m$ — векторы $\mathbf{u}_1,\ldots,\mathbf{u}_n$ такие, что матрица оператора A имеет диагональный вид, т. е. для произвольного вектора $\mathbf{x}=\sum\limits_{i=1}^{n}\beta_i\mathbf{v}_i$ имеем $\mathbf{y}=A\mathbf{x}=\sum\limits_{i=1}^{n}\sigma_i\beta_i\mathbf{u}_i$. Иначе говоря, всякая матрица A становится диаго-

нальной, если в области определения и в области значений подходящим образом выбраны ортогональные системы координат.

5.182. Найти сингулярное разложение матрицы A размерности $n \times n$

$$A = (1, 2, \ldots, n)^T (1, 1, \ldots, 1).$$

5.183. Показать, что если A — матрица полного ранга ($\operatorname{rank}(A) = n$), то решение $\mathbf{x}$ задачи наименьших квадратов $\min\limits_{\mathbf{y}} \|\mathbf{b} - A\mathbf{y}\|_2$ имеет вид $\mathbf{x} = V\Sigma^{-1}U^T\mathbf{b}$.

$\lhd$ Из метода нормального уравнения следует, что решение $\mathbf{x}$ можно представить в виде $\mathbf{x} = (A^TA)^{-1}A^T\mathbf{b} = (V\Sigma U^TU\Sigma V^T)^{-1}V\Sigma U^T\mathbf{b} = V\Sigma^{-1}U^T\mathbf{b}$. $\rhd$

5.184. Пусть матрица $A = (U_1\, U_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix} (V_1\, V_2)^T$ имеет ранг $r < n$. Показать, что пространство решений $\mathbf{x}$ задачи наименьших квадратов имеет вид $\mathbf{x} = V_1\Sigma_1^{-1}U_1^T\mathbf{b} + V_2\mathbf{z}$ с произвольным $\mathbf{z} \in \mathbf{R}^{n-r}$. Норма $\mathbf{x}$ минимальна при $\mathbf{z} = 0$. Здесь $U_1 \in \mathbf{R}^{m\times r}$, $\Sigma_1 \in \mathbf{R}^{r\times r}$, $V_1 \in \mathbf{R}^{n\times r}$.

Указание. См. 5.180, учитывая, что $V_1^TV_2\mathbf{z} = 0$ для любого вектора $\mathbf{z}$. Норма $\mathbf{x}$ минимальна при $\mathbf{z} = 0$, так как векторы $V_1\Sigma_1^{-1}U_1^T\mathbf{b}$ и $V_2\mathbf{z}$ ортогональны.

5.185. Пусть столбцы $\mathbf{u}_i$, $\mathbf{v}_i$, $i = 1, \ldots, n$, матриц $U = (\mathbf{u}_1 \ldots \mathbf{u}_n)$, $V = (\mathbf{v}_1 \ldots \mathbf{v}_n)$ такие, что $A = U\Sigma V^T = \sum\limits_{i=1}^{n} \sigma_i\mathbf{u}_i\mathbf{v}_i^T$. Доказать, что $\min\limits_{A'_k} \|A - A'_k\|_2$ по всем матрицам $A'_k \in \mathbf{R}^{m\times n}$ ранга $k < n$ равен σ_{k+1} и достигается на матрице $A_k = \sum\limits_{i=1}^{k} \sigma_i\mathbf{u}_i\mathbf{v}_i^T = U_k\Sigma_kV_k^T$.

$\lhd$ По построению матрица A_k имеет ранг k. Покажем, что $\|A - A_k\|_2 = \sigma_{k+1}$. Действительно,

$$\|A - A_k\|_2 = \left\| U \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{n-k} \end{pmatrix} V^T \right\|_2 \leqslant \|U\|_2 \left\| \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{n-k} \end{pmatrix} \right\|_2 \|V^T\|_2 \leqslant \sigma_{k+1}.$$

Но $\|A - A_k\|_2 \geqslant \|(A - A_k)\mathbf{v}_{k+1}\|_2 = \sigma_{k+1}\|U\mathbf{e}_{k+1}\| = \sigma_{k+1}$, где $\mathbf{e}_{k+1} = (0, \ldots, 0, 1, 0, \ldots, 0)^T$. Следовательно, $\|A - A_k\|_2 = \sigma_{k+1}$. Покажем, что не существует матрицы B ранга k, более близкой к A. Так как $\dim(\ker(B)) = n - k$ и $(n - k + k + 1) > n$, то существует ненулевое пересечение подпространств $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\}$ и $\ker(B)$. Рассмотрим вектор $\mathbf{h} = \sum\limits_{i=1}^{k+1} c_i\mathbf{v}_i$ из этого пересечения с нормой $\|\mathbf{h}\|_2^2 = \sum\limits_{i=1}^{k+1} c_i^2 = 1$. Имеем $\|A - B\|_2 \geqslant \|(A - B)\mathbf{h}\|_2 = \|A\mathbf{h}\|_2 = \|U\Sigma V^T\mathbf{h}\|_2 = \|\Sigma(V^T\mathbf{h})\|_2$, поскольку ортогональная матрица U не изменяет длины векторов. Далее,

так как $V^TV = I$, то $V^T\mathbf{v}_i = \mathbf{e}_i$. Таким образом, $V^T\mathbf{h} = \sum\limits_{i=1}^{k+1} c_i\mathbf{e}_i$ и $\Sigma V^T\mathbf{h} = \sum\limits_{i=1}^{k+1} \sigma_i c_i \mathbf{e}_i$. В результате получаем

$$\|\Sigma V^T\mathbf{h}\|_2 \geqslant \sigma_{k+1}\left(\sum_{i=1}^{k+1} c_i^2\right)^{1/2} = \sigma_{k+1}\|\mathbf{h}\|_2 = \sigma_{k+1}. \qquad \triangleright$$

Из 5.185 следует правило решения задачи наименьших квадратов в приближенной арифметике. В реальных вычислениях все σ_i получатся (с учетом машинной точности) отличными от нуля, поэтому зафиксируем некоторое значение ε_0. Будем считать, что величины $\sigma_i < \varepsilon_0$ при $i = k+1, \ldots, n$ соответствуют погрешности вычислений, следовательно, можно заменить исходную задачу задачей с матрицей $A_k = U_k\Sigma_k V_k^T$. Такой способ усечения матрицы A является оптимальным в том смысле, что полученная матрица A_k наиболее близка к A в норме $\|\cdot\|_2$.

5.9. Проблема собственных значений

Пусть S — произвольная невырожденная матрица. Говорят, что матрицы одинаковой размерности A и $B = S^{-1}AS$ *подобны*, а матрица S осуществляет подобие.

Теорема. *Для произвольной вещественной матрицы $A \in \mathbf{R}^{n\times n}$ найдется вещественная ортогональная матрица $Q \in \mathbf{R}^{n\times n}$ такая, что*

$$Q^TAQ = \begin{pmatrix} R_{11} & R_{12} & \ldots & R_{1m} \\ 0 & R_{22} & \ldots & R_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & R_{mm} \end{pmatrix} \equiv R.$$

Здесь каждый диагональный блок $R_{ii}\,(m \leqslant n)$ представляет собой либо вещественное собственное значение, либо 2×2-матрицу, отвечающую сопряженной паре комплексных собственных значений. Матрицу R называют *действительной формой Шура*, из которой можно легко определить собственные векторы и собственные числа исходной матрицы.

Теорема 1 (*закон инерции*). *Каждая матрица $A = A^T$ подобна некоторой диагональной матрице вида* $\operatorname{diag}(I_\pi, -I_\nu, O_\xi)$ *с единичными матрицами I_π, I_ν и нулевой O_ξ, где π, ν, ξ соответственно число положительных, отрицательных и нулевых собственных чисел матрицы A.*

Теорема 2 (*критерий Сильвестра*). *Число ν отрицательных собственных значений невырожденной симметричной матрицы A равно числу перемен знаков последовательности главных миноров Δ_k, $1 \leqslant k \leqslant n$. Для положительной определенности матрицы $A = A^T$ необходимо и достаточно выполнение неравенств $\Delta_k > 0$ при $1 \leqslant k \leqslant n$. Для отрицательной определенности необходимо и достаточно, чтобы знаки главных миноров чередовались, при этом $\Delta_1 < 0$.*

5.186. Пусть $B = S^{-1}AS$. Доказать, что матрицы A и B имеют одни и те же собственные значения, а вектор $\mathbf{e}_A$ является собственным вектором A тогда и только тогда, когда $\mathbf{e}_B = S^{-1}\mathbf{e}_A$ — собственный вектор B.

$\lhd$ Собственные значения B находим из условия

$$\det(B - \lambda I) = \det(S^{-1}AS - \lambda S^{-1}S) =$$

$$= \det(S^{-1}(A - \lambda I)S) = \det(S^{-1})\det(A - \lambda I)\det(S) = \det(A - \lambda I).$$

Собственные значения совпадают, так как равны характеристические многочлены. Условие $A\mathbf{e} = \lambda\,\mathbf{e}$ равносильно следующему: $S^{-1}ASS^{-1}\mathbf{e} = \lambda S^{-1}\mathbf{e}$, или $BS^{-1}\mathbf{e} = \lambda S^{-1}\mathbf{e}$, т. е. собственные векторы $\mathbf{e}_A$ и $\mathbf{e}_B$ связаны соотношением $\mathbf{e}_B = S^{-1}\mathbf{e}_A$. $\rhd$

5.187. Предположим, что матрица $A \in \mathbf{R}^{n\times n}$ — симметричная и положительно определенная. Показать, что:

1) существует единственная симметричная и положительно определенная матрица X такая, что $A = X^2$;

2) если $X_0 = I$, $X_{k+1} = \dfrac{X_k + AX_k^{-1}}{2}$, то X_k стремится к $\sqrt{A}$, где $\sqrt{A}$ означает матрицу X из 1).

5.188. Пусть A — симметричная матрица размерности $n \times n$, $\lambda \in \mathbf{R}^1$, $\mathbf{x} \in \mathbf{R}^n$ — соответственно произвольное число и вектор, причем $||\mathbf{x}||_2 = 1$. Доказать, что существует собственное значение λ_k матрицы A, для которого $|\lambda_k - \lambda| \leqslant ||A\mathbf{x} - \lambda\mathbf{x}||_2$.

$\lhd$ Пусть $\{\mathbf{q}_k\}$, $k = 1, \ldots, n$, — полная ортонормированная система собственных векторов матрицы A, $\mathbf{x} = \sum\limits_{k=1}^{n} c_k\mathbf{q}_k$. Тогда $||A\mathbf{x} - \lambda\mathbf{x}||_2^2 =$

$= \sum\limits_{k=1}^{n} (\lambda_k - \lambda)^2 c_k^2 \geqslant \min\limits_k (\lambda_k - \lambda)^2$. $\rhd$

5.189. Показать, что для максимального и минимального собственных значений симметричной матрицы A справедливы следующие оценки: $\lambda_{\min}(A) \leqslant \min\limits_{1\leqslant i\leqslant n} a_{ii}$, $\lambda_{\max}(A) \geqslant \max\limits_{1\leqslant i\leqslant n} a_{ii}$.

$\lhd$ Имеем

$$\lambda_{\min}(A) = \min_{||\mathbf{x}||_2=1} (A\mathbf{x}, \mathbf{x}) \leqslant (A\mathbf{e}_i, \mathbf{e}_i) = a_{ii},$$

$$\lambda_{\max}(A) = \max_{||\mathbf{x}||_2=1} (A\mathbf{x}, \mathbf{x}) \geqslant (A\mathbf{e}_i, \mathbf{e}_i) = a_{ii},$$

где $\mathbf{e}_i$ — вектор с i-й компонентой 1 и остальными компонентами 0. $\rhd$

5.190. Доказать, что у вещественной трехдиагональной матрицы

$$a_{ij} = \begin{cases} b_i & \text{при} \quad i = j, \\ a_i & \text{при} \quad i = j + 1, \\ c_i & \text{при} \quad i + 1 = j, \\ 0 & — \quad \text{иначе}, \end{cases}$$

все собственные значения вещественные, если $a_{i+1}c_i > 0$, $i = 1, 2, ..., n - 1$.

У к а з а н и е. Пусть диагональная матрица D определена следующим образом:

$$D = \operatorname{diag}(d_1, \ldots, d_n), \;\; d_1 = 1,\; d_i = d_{i-1}\sqrt{\frac{c_{i-1}}{a_i}}.$$

Тогда $B = DAD^{-1}$ — симметричная матрица, вещественный спектр которой совпадает со спектром подобной матрицы A.

5.191. Доказать, что для трехдиагональной матрицы из 5.190 верно неравенство $|\lambda_k(A)| < 1\ \forall k$, если $|a_i| + |b_i| + |c_i| \leqslant 1\ \forall i$, $a_1 = c_n = 0$, и если хотя бы для одного значения индекса i неравенство строгое, а $a_{i+1}c_i \neq 0,\ i = 1, 2, ..., n-1$.

5.192. Пусть A и B — матрицы размерности $m \times n$ и $n \times m$ соответственно, $m \geqslant n$; $P_C(\lambda) = \det(\lambda I - C)$ — характеристический многочлен квадратной матрицы C. Доказать справедливость равенства

$$P_{AB}(\lambda) = \lambda^{m-n} P_{BA}(\lambda)\,.$$

$\lhd$ Рассмотрим следующие тождества для блочных матриц размерности $(m+n) \times (m+n)$:

$$\begin{pmatrix} A\,B & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} A\,B & A\,B\,A \\ B & B\,A \end{pmatrix},$$

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix} \begin{pmatrix} 0 & 0 \\ B & B\,A \end{pmatrix} = \begin{pmatrix} A\,B & A\,B\,A \\ B & B\,A \end{pmatrix}.$$

Здесь I — единичная матрица соответствующей размерности. Поскольку блочная матрица $K = \begin{pmatrix} I & A \\ 0 & I \end{pmatrix}$ размерности $(m+n) \times (m+n)$ невырожденная, имеем

$$\begin{pmatrix} I & A \\ 0 & I \end{pmatrix}^{-1} \begin{pmatrix} A\,B & 0 \\ B & 0 \end{pmatrix} \begin{pmatrix} I & A \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ B & B\,A \end{pmatrix}.$$

Таким образом, две матрицы размерности $(m+n) \times (m+n)$

$$O_1 = \begin{pmatrix} A\,B & 0 \\ B & 0 \end{pmatrix}, \quad O_2 = \begin{pmatrix} 0 & 0 \\ B & B\,A \end{pmatrix}$$

подобны: $K^{-1}O_1K = O_2$. Собственные значения матрицы O_1 — это собственные значения матрицы $A\,B$ вместе с n нулями, а собственные значения матрицы O_2 — собственные значения матрицы $B\,A$ вместе с m нулями. Поскольку характеристические многочлены подобных матриц совпадают

$$P_{O_2}(\lambda) = \det(\lambda\,I - O_2) = \det(\lambda\,K^{-1}K - K^{-1}O_1K) =$$
$$= \det K^{-1} \det(\lambda\,I - O_1) \det K = P_{O_1}(\lambda)\,,$$

отсюда следует утверждение задачи.

Из рассмотренного решения следует совпадение соответствующих жордановых клеток матриц AB и BA. $\rhd$

5.193. Доказать, что для квадратных матриц A, B одинаковой размерности спектры матриц AB и BA совпадают.

5.194. Доказать, что если A, B — симметричные матрицы размерности $n \times n$, то необходимым и достаточным условием равенства $AB = BA$ является существование базиса в пространстве $\mathbf{R}^n$, составленного из общих собственных векторов матриц A и B.

5.195. Доказать, что если матрицы A и B коммутируют, то существует собственное значение $\lambda(AB)$, равное произведению собственных значений $\lambda(A)\lambda(B)$.

5.196. Доказать, что если A — симметричная и положительно определенная матрица, а B — симметричная матрица, то все собственные значения $\lambda(AB)$ матрицы AB вещественные.

У к а з а н и е. Воспользоваться тем, что AB подобна симметричной матрице $A^{1/2}BA^{1/2}$.

5.197. Доказать, что если A, B — симметричные и положительно определенные матрицы, то все собственные значения $\lambda(AB)$ матрицы AB положительные.

У к а з а н и е. Воспользоваться 5.196.

5.198. Пусть A — симметризуемая матрица, т. е. существует невырожденная матрица T такая, что TAT^{-1} — симметричная матрица. Доказать, что система собственных векторов матрицы A образует базис.

У к а з а н и е. Воспользоваться тем, что если $TAT^{-1}\mathbf{e} = \lambda\mathbf{e}$, то $T^{-1}\mathbf{e}$ — собственный вектор матрицы A, соответствующий тому же собственному значению λ. Доказать, что из полноты системы векторов $\{\mathbf{e}_i\}$ следует полнота системы $\{T^{-1}\mathbf{e}_i\}$.

5.199. Доказать, что если A — симметричная и положительно определенная матрица, а B — симметричная матрица, то система собственных векторов матрицы AB образует базис.

У к а з а н и е. Воспользоваться 5.196 и 5.198.

5.200. Доказать, что если A, B — симметричные и положительно определенные, коммутирующие матрицы, то матрица AB положительно определена.

$\lhd$ Все собственные значения AB положительны в силу 5.197. Из коммутируемости A и B следует симметрия AB, а критерием положительной определенности симметричной матрицы является положительность ее собственных значений. $\rhd$

5.201. Доказать положительную определенность матрицы

$$A = \begin{pmatrix} 0,5 & 1 & 1 & 1 & \dots & 1 & 1 \\ 1 & 2,5 & 3 & 3 & \dots & 3 & 3 \\ 1 & 3 & 4,5 & 5 & \dots & 5 & 5 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 3 & 5 & 7 & \dots & \frac{1}{2}(4n-7) & 2n-3 \\ 1 & 3 & 5 & 7 & \dots & 2n-3 & \frac{1}{2}(4n-3) \end{pmatrix}.$$

$\lhd$ Обозначим матрицу A размерности $n \times n$ через A_n. Пусть далее левая треугольная матрица P_n размерности $n \times n$ определена равенством

$$P_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -2 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ -2 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Тогда

$$P_n A_n = \begin{pmatrix} 0,5 & 1 & \dots & 1 \\ 0 & & & \\ \dots & & A_{n-1} & \\ 0 & & & \end{pmatrix}.$$

Таким образом, $\det A_k = 0,5 \det A_{k-1} = 2^{-k} > 0$ для любого главного минора $\det A_k$, $k = 1, \dots, n$. Согласно критерию Сильвестра положительной определенности, симметричная матрица $A_n = A$ положительно определена. $\rhd$

5.202. Доказать положительную определенность матрицы:

$$1)\ A = \begin{pmatrix} 2 & -1 & \frac{1}{2} & -\frac{1}{3} \\ -1 & 3 & -1 & -\frac{1}{2} \\ \frac{1}{2} & -1 & 4 & 2 \\ -\frac{1}{3} & -\frac{1}{2} & 2 & 5 \end{pmatrix};\ 2)\ A = \begin{pmatrix} 12 & -6 & 3 & -2 \\ -6 & 18 & -6 & 6 \\ 3 & -6 & 24 & 15 \\ -2 & 6 & 15 & 20 \end{pmatrix}.$$

У к а з а н и е. 1) Используя теорему Гершгорина, доказать положительность всех собственных значений симметричной матрицы A. 2) Использовать критерий Сильвестра положительной определенности симметричной матрицы и прямое вычисление ее главных миноров.

5.203. Пусть матрицы $A, A^T \in \mathbf{R}^{n \times n}$ имеют строгое диагональное преобладание и положительные диагональные элементы. Доказать, что матрица A положительно определена, т. е. $(A\mathbf{x}, \mathbf{x}) > 0\ \forall \mathbf{x} \neq 0$.

$\lhd$ Симметричная часть $S = \frac{A + A^T}{2}$ матрицы A имеет положительную и строго доминирующую диагональ, поэтому ее собственные значения положительны и S положительно определена, а вместе с ней положительно определена и матрица A. $\rhd$

5.204. Построить пример симметричной положительно определенной матрицы размерности 3×3, трехдиагональная часть которой не является положительно определенной.

Ответ: $A = \begin{pmatrix} 1 & q & \frac{1}{2} \\ q & 1 & q \\ \frac{1}{2} & q & 1 \end{pmatrix}$ при $q = \sqrt{\frac{1+\alpha}{2}}$, $\alpha \in \left(0, \frac{1}{2}\right)$.

5.205. Пусть $A = A^T > 0$. Доказать, что если $\lambda_{\max}(A) = a_{kk}$ при некотором k, где $1 \leqslant k \leqslant n$, то $a_{ik} = a_{kj} = 0$ при всех $i \neq k,\ j \neq k$.

5.206. Пусть $A_n(a, b)$ — вещественная трехдиагональная матрица размерности $n \times n$:

$$a_{ij} = \begin{cases} a & \text{при} \quad i = j, \\ b & \text{при} \quad i = j+1, i = j-1, \\ 0 & \text{иначе.} \end{cases}$$

Доказать следующие равенства:

1) $\det A_{n+1}(a, b) = a \det A_n(a, b) - b^2 \det A_{n-1}(a, b),\ \ n \geqslant 2;$

2) $\det A_n(a, b) = \left(\left(\frac{a}{2} + \sqrt{\frac{a^2}{4} - b^2} \right)^{n+1} - \right.$

$$\left. - \left(\frac{a}{2} - \sqrt{\frac{a^2}{4} - b^2} \right)^{n+1} \right) \Big/ \left(2\sqrt{\frac{a^2}{4} - b^2} \right),\ n \geqslant 1;$$

3) $\det A_n(a, b) = \sum\limits_{k=0}^{[n/2]} C_{n+1}^{2k+1} \left(\frac{a^2}{4} - b^2 \right)^k \left(\frac{a}{2} \right)^{n-2k},\ \ k \geqslant 1.$

5.207. Пусть матрица $A_n(a, b)$ определена, как в 5.206. Найти все ее собственные значения и собственные векторы.

5.208. Пусть матрица $A_n(a, b)$ определена, как в 5.206. Доказать, что она положительно определена тогда и только тогда, когда $a - 2\,|b|\,\cos\frac{\pi}{n+1} > 0$.

5.209. Матрица Уилкинсона при $\varepsilon = 0$

$$A = \begin{pmatrix} 20 & 20 & 0 & 0 & \dots & 0 & 0 \\ 0 & 19 & 20 & 0 & \dots & 0 & 0 \\ 0 & 0 & 18 & 20 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 2 & 20 \\ \varepsilon & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

имеет наименьшее по модулю собственное значение, равное 1. Как оно изменится при $\varepsilon = 20^{-19} \cdot 20! \approx 5 \cdot 10^{-7}$?

$\lhd$ Характеристическое уравнение для возмущенной матрицы Уилкинсона имеет вид

$$\det(A - \lambda I) = (20 - \lambda)(19 - \lambda) \cdots (1 - \lambda) - 20^{19} \cdot \varepsilon = 0.$$

Свободный член в этом уравнении равен 0, следовательно, наименьшее собственное значение также равно 0. $\rhd$

5.210. Пусть

$$A_n(\alpha) = \begin{pmatrix} 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \\ \alpha & 0 & \dots & 0 \end{pmatrix}$$

— матрица размерности $n \times n$. Доказать, что характеристическое уравнение матрицы $A_n(\alpha)$ имеет вид $\lambda^n = \alpha$. Сравнить собственные числа близких матриц $A_{20}(2^{-20})$ и $A_{20}(0)$.

Ответ: $\lambda_k(2^{-20}) = 0{,}5\mathrm{e}^{\pi \mathrm{i} \frac{k-1}{10}}, k = 1, \dots, 20$ и $\lambda_k(0) = 0$, $k = 1, \dots, 20$.

Степенной метод. Алгоритм вычисления максимального по модулю собственного значения λ матрицы A имеет вид

$$\mathbf{x}^{k+1} = A\mathbf{x}^k, \;\; \lambda^{k+1} = \frac{(\mathbf{x}^{k+1}, \mathbf{x}^k)}{(\mathbf{x}^k, \mathbf{x}^k)}, \;\; \mathbf{x}^k \neq 0; \;\; k = 0, 1, 2, \dots .$$

При его практической реализации на каждом шаге нормируют текущий вектор: $\mathbf{x}^k$ заменяют на $\dfrac{\mathbf{x}^k}{\|\mathbf{x}^k\|}$.

5.211. Пусть A — матрица простой структуры (собственные векторы $\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n$ матрицы образуют базис в $\mathbf{C}^n$). Пусть далее $|\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant ... \geqslant |\lambda_n|$ и L — линейная оболочка $\mathbf{q}_2, \mathbf{q}_3, ..., \mathbf{q}_n$. Доказать, что для степенного метода при условии $\mathbf{x}^0 \notin L$ справедлива оценка $\lambda^k = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)$.

5.212. Доказать, что если в условии 5.211 матрица A является симметричной, то для степенного метода справедлива оценка $\lambda^k = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right)$.

5.213. Пусть матрица A размерности $n \times n$ имеет n различных собственных значений. Предположим, что $\mathbf{x}^0$ принадлежит линейной оболочке некоторых собственных векторов $\mathbf{q}_{i_1}, \mathbf{q}_{i_2}, \dots, \mathbf{q}_{i_t}$, но не принадлежит никакой их линейной подоболочке. К какому собственному значению матрицы сходятся итерации степенного метода в точной арифметике и с какой скоростью?

Ответ: к максимальному по модулю λ из λ_{i_k}, $1 \leqslant k \leqslant t$. При численных расчетах, как правило, имеется ненулевой (порядка машинной точности) коэффициент c_1 в разложении $\mathbf{x}^0$ по векторам $\mathbf{q}_i$, поэтому метод сходится к максимальному по модулю из всех собственных значений λ_1. Сходимость к следующему по абсолютной величине λ_2 (см. 5.211) можно обеспечить только постоянным исключением каким-либо способом вектора $\mathbf{q}_1$ из очередного приближения $\mathbf{x}^k$. На этой идее основан рассматриваемый далее метод итерирования подпространств.

Метод обратной итерации. Этот метод, по сути соответствующий степенному методу для матрицы A^{-1}, можно применять для вычисления наименьшего по модулю собственного значения λ:

$$\mathbf{x}^k := \frac{\mathbf{x}^k}{\|\mathbf{x}^k\|},\ A\mathbf{x}^{k+1} = \mathbf{x}^k,\ \lambda^k = \frac{(\mathbf{x}^k, \mathbf{x}^{k+1})}{(\mathbf{x}^{k+1}, \mathbf{x}^{k+1})}\,.$$

При этом на каждом шаге алгоритма требуется решать систему $A\mathbf{x}^{k+1} = \mathbf{x}^k$.

Степенной метод и метод обратной итерации можно также применять к матрице $A - cI$, что позволяет влиять на сходимость. Например, если с высокой точностью известно приближение $\tilde{\lambda}$ к некоторому собственному значению λ, то метод обратной итерации с параметром $c = \tilde{\lambda}$ обычно сходится за несколько итераций. Скорость сходимости существенно замедляется при вычислении одного из группы близких собственных значений.

5.214. Пусть собственные значения симметричной матрицы A удовлетворяют цепочке неравенств $\lambda_1 < \lambda_2 < \ldots < \lambda_n$. Выяснить, к какому собственному значению λ_s в зависимости от параметра c сходится итерационный процесс

$$\mathbf{x}^{k+1} = (A - cI)\mathbf{x}^k,\ \lambda^k = c + \frac{(\mathbf{x}^{k+1}, \mathbf{x}^k)}{(\mathbf{x}^k, \mathbf{x}^k)}\,.$$

Найти скорость сходимости.

Ответ: к λ_s, для которого справедливо $|\lambda_s - c| = \max\limits_i |\lambda_i - c|$. Скорость сходимости равна $O(q^{2n})$, где $q = \max\limits_{i \neq s} \dfrac{|\lambda_i - c|}{|\lambda_s - c|}$. Отсюда следует, что процесс сходится к λ_1 при $c > \dfrac{\lambda_1 + \lambda_n}{2}$ или к λ_n при $c < \dfrac{\lambda_1 + \lambda_n}{2}$.

5.215. В условии 5.214 выбрать постоянную c так, чтобы итерационный процесс с наилучшей скоростью сходился к λ_1 (или к λ_n).

Ответ: как следует из 5.214, оптимальное значение c_s для $s = 1, n$ является решением следующей минимаксной задачи: $\min\limits_c \max\limits_{i \neq s} \dfrac{|\lambda_i - c|}{|\lambda_s - c|}$. Так как рассматриваемая функция линейная по λ_i, то модуль имеет максимальное значение в граничных точках. Например, при $s = 1$ это соответствует $\min\limits_c \max \left\{ \dfrac{|\lambda_2 - c|}{|\lambda_1 - c|},\ \dfrac{|\lambda_n - c|}{|\lambda_1 - c|} \right\}$. Можно показать (например, графически), что оп-

тимальное значение $c_1 = \frac{\lambda_2 + \lambda_n}{2}$. Аналогично $c_n = \frac{\lambda_1 + \lambda_{n-1}}{2}$. Скорость сходимости степенного метода при оптимальном сдвиге зависит от $\lambda_1, \ldots, \lambda_n$ и не может стать сколь угодно высокой за счет параметра сдвига.

5.216. Пусть собственные значения симметричной матрицы A удовлетворяют цепочке неравенств $\lambda_1 < \lambda_2 < \ldots < \lambda_n$. Выяснить, к какому собственному значению λ_t сходится в зависимости от параметра c метод обратной итерации со сдвигом

$$(A - cI)\mathbf{x}^{k+1} = \mathbf{x}^k,\ \lambda^k = c + \frac{(\mathbf{x}^k, \mathbf{x}^{k+1})}{(\mathbf{x}^{k+1}, \mathbf{x}^{k+1})}.$$

Найти скорость сходимости.

Ответ: к λ_t, для которого справедливо $|\lambda_t - c| = \min\limits_i |\lambda_i - c|$. Скорость сходимости равна $O(q^{2^n})$, где $q = \frac{|\lambda_t - c|}{\min\limits_{i \neq t} |\lambda_i - c|}$. Отсюда следует, что процесс в зависимости от значения c может сходиться к любому λ_t. При этом $q \to 0$, если $c \to \lambda_t$.

Скорость сходимости метода обратной итерации со сдвигом можно значительно повысить, если изменять значение сдвига от шага к шагу. Рассмотрим функцию $R_A(\mathbf{x}) = \frac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}$, называемую *отношением Рэлея*.

5.217. Пусть $A = A^T > 0$. Доказать, что

$$\lambda_{\max}(A) = \max_{\mathbf{x} \neq 0} R_A(\mathbf{x}), \qquad \lambda_{\min}(A) = \min_{\mathbf{x} \neq 0} R_A(\mathbf{x}).$$

Указание. Пусть λ_i — i-е собственное значение и $A\mathbf{q}_i = \lambda_i \mathbf{q}_i$. Из условия $A = A^T$ следует, что собственные векторы образуют базис и можно считать, что $(\mathbf{q}_i, \mathbf{q}_j) = \delta_i^j$. При этом $\lambda_i = \frac{(A\mathbf{q}_i, \mathbf{q}_i)}{(\mathbf{q}_i, \mathbf{q}_i)}$. Представив произвольный вектор $\mathbf{x}$ в виде разложения по собственным векторам, получим требуемый результат.

5.218. Пусть $A = A^T > 0$. Доказать, что $\forall \mathbf{x} \in \mathbf{R}^n$ и $\forall \mu \in \mathbf{R}^1$ имеет место свойство минимальности невязки

$$\|(A - R_A(\mathbf{x})\, I)\mathbf{x}\|_2 \leqslant \|(A - \mu\, I)\mathbf{x}\|_2.$$

$\lhd$ Рассмотрим квадратичную по μ функцию $\|(A - \mu I)\mathbf{x}\|_2^2 = (A\mathbf{x}, A\mathbf{x}) - 2\mu(A\mathbf{x}, \mathbf{x}) + \mu^2(\mathbf{x}, \mathbf{x})$, минимум которой достигается в точке $\mu = R_A(\mathbf{x})$. $\rhd$

Неравенство из 5.218 показывает, что наилучший сдвиг для метода обратной итерации, который можно получить из найденного приближения к собственному вектору, есть отношение Рэлея $R_A(\mathbf{x}^k)$. При таком выборе сдвига сходимость к собственному вектору, если она есть (см. 5.211),

является кубической: $\lim\limits_{k\to\infty}\left|\dfrac{\varphi_{k+1}}{\varphi_k^3}\right| \leqslant 1$, где φ_k — угол между собственным вектором $\mathbf{x}$ и его приближением $\mathbf{x}^k$.

5.219. Пусть метод обратной итерации с постоянным сдвигом сходится к собственному значению λ_c матрицы $A = A^T$. Показать, что начиная с некоторого k выполняется оценка

$$|\lambda_c - \lambda^k| \leqslant \frac{1}{\|\mathbf{x}^{k+1}\|_2},$$

т. е. величина $\|\mathbf{x}^{k+1}\|_2^{-1}$ характеризует скорость сходимости итерационного процесса.

$\lhd$ Для приближений $\mathbf{x}^k$, $\mathbf{x}^{k+1}$ метода обратной итерации справедливо выражение

$$R_A(\mathbf{x}^{k+1}) = \frac{(A\mathbf{x}^{k+1}, \mathbf{x}^{k+1})}{(\mathbf{x}^{k+1}, \mathbf{x}^{k+1})} = c + \frac{(\mathbf{x}^k, \mathbf{x}^{k+1})}{(\mathbf{x}^{k+1}, \mathbf{x}^{k+1})} = \lambda^k.$$

Отсюда и из 5.218, 5.188 получаем, что

$$1 = \|\mathbf{x}^k\|_2 = \|(A - cI)\mathbf{x}^{k+1}\|_2 \geqslant \|(A - R_A(\mathbf{x}^{k+1})I)\mathbf{x}^{k+1}\|_2 =$$
$$= \|(A - \lambda^k I)\mathbf{x}^{k+1}\|_2 \geqslant \min_i |\lambda_i - \lambda^k| \|\mathbf{x}^{k+1}\|_2.$$

Так как метод сходится к λ_c, то начиная с некоторого k имеем $\min\limits_i |\lambda_i - \lambda^k| = |\lambda_c - \lambda^k|$. Отсюда следует искомая оценка. $\rhd$

5.220. Предположим, что матрица $A \in \mathbf{R}^{n\times n}$ — симметричная и положительно определенная. Рассмотрим следующие итерации: $A_0 = A$, для $k = 1, 2, \ldots$ строим $A_{k-1} = R_k R_k^T$ (разложение Холецкого) и определяем $A_k = R_k^T R_k$. Здесь R_k — верхняя треугольная матрица. Показать, что:

1) эти итерации сходятся;

2) если матрица $A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$, $a \geqslant c$, имеет собственные значения $\lambda_1 \geqslant \lambda_2 > 0$, то матрицы A_k сходятся к матрице $\operatorname{diag}(\lambda_1, \lambda_2)$.

Рассмотрим методы нахождения нескольких (всех) собственных значений и поиска инвариантных подпространств.

Подпространство $\mathbf{H} \subset \mathbf{R}^n$ называется *инвариантным подпространством* матрицы A, если $A\mathbf{H} \subset \mathbf{H}$. В качестве $\mathbf{H}$ можно взять, например, подпространство $\operatorname{span}\{\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_m}\}$, являющееся линейной оболочкой собственных векторов $\mathbf{e}_{i_j}$.

5.221. Пусть A — матрица размерности $n \times n$ и $X = (\mathbf{x}_1 \ldots \mathbf{x}_m)$, $\mathbf{x}_i \in \mathbf{R}^n$ — произвольная матрица с линейно независимыми столбцами. Показать, что подпространство $\operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ тогда и только тогда

инвариантно относительно A, когда найдется такая матрица B размерности $m \times m$, что $AX = XB$. В случае $m = n$ собственные значения матрицы B являются собственными значениями матрицы A.

У к а з а н и е. Инвариантность подпространства означает, что существуют константы c_j такие, что $A\mathbf{x}_i = \sum_{j=1}^{m} c_j \mathbf{x}_j$. В данном случае $A\mathbf{x}_i = \sum_{j=1}^{n} b_{ji}\mathbf{x}_j$. Если $m = n$, то $B = X^{-1}AX$ и матрицы подобны. Если $A\mathbf{x}_i = \lambda_i \mathbf{x}_i$, то $B = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$.

Ортогональная итерация (итерирование подпространств). Рассмотрим следующий итерационный алгоритм нахождения m-мерного инвариантного подпространства матрицы A, образованного линейной комбинацией собственных векторов, отвечающих m наибольшим по модулю собственным значениям: возьмем произвольную матрицу $V_0 \in \mathbf{R}^{n\times m}$ с m ортонормированными столбцами и последовательно для $k = 0, 1, \ldots$ будем вычислять для матриц $P_{k+1} = AV_k$ разложения вида $P_{k+1} = V_{k+1}R_{k+1}$, т. е. проводить ортонормировку столбцов P_{k+1}.

Известно, что если модули всех собственных значений различны и матрица A не вырождена, то метод сходится и столбцы матрицы V_∞ задают базис в искомом подпространстве. Действительно, рассмотрим A^kV_0 — образ векторов-столбцов исходной матрицы $V_0 = (\mathbf{v}_1 \ \ldots \ \mathbf{v}_m)$. Все векторы-столбцы образа, оставаясь линейно независимыми, при $k \to \infty$ сходятся к старшему собственному вектору, что соответствует методу простой итерации нахождения наибольшего по модулю собственного значения. Чтобы в результате итераций имела место сходимость к ортонормированному базису в подпространстве, образованном первыми собственными векторами, необходимо на каждом шаге проводить ортогонализацию: исключать из вектора $A^k\mathbf{v}_2$ составляющую $A^k\mathbf{v}_1$, из вектора $A^k\mathbf{v}_3$ — составляющие $A^k\mathbf{v}_1$, $A^k\mathbf{v}_2$, и т. д. Из 5.211 следует, что скорость сходимости характеризуется величинами $\left|\frac{\lambda_2}{\lambda_1}\right|, \left|\frac{\lambda_3}{\lambda_2}\right|, \ldots$.

QR-алгоритм. Модифицируем ортогональную итерацию так, чтобы она допускала сдвиги и обращения матрицы, как в методе обратной итерации. Это приводит к идее QR-алгоритма — наиболее популярного метода вычисления всех собственных значений и векторов матрицы не слишком большой размерности. Пусть задана матрица A размерности $n \times n$. Положим $A_0 = A$ и вычислим $A_0 = Q_0R_0$, где Q_0 — ортогональная (в комплексном случае — унитарная) матрица, R_0 — верхняя треугольная матрица. Далее определим $A_1 = R_0Q_0$, т. е. перемножим полученные в результате разложения матрицы в обратном порядке. Таким образом, на каждом шаге вычисляется QR-разложение матрицы $A_k = Q_kR_k$ и находится

$A_{k+1} = R_k Q_k$. Отметим, что все полученные матрицы A_k подобны A_0. Результатом является «почти верхняя треугольная» предельная матрица A_∞, для которой несложно вычисляются собственные значения. В случае невырожденной матрицы QR-разложение с положительными элементами r_{ii} треугольной матрицы R единственно, поэтому в дальнейшем будем полагать для произвольной матрицы $r_{ii} \geqslant 0$.

При практическом использовании метода сначала проводят масштабирование (уравновешивание) матрицы, сближающее ее норму со спектральным радиусом, а затем приводят ее к верхней форме Хессенберга H ($h_{ij} = 0$ при $i > j+1$), которая инвариантна относительно QR-итераций. Само же разложение используют со сдвигами, т. е. применяют к матрицам вида $A_k = H_k - c_k I$.

5.222. Показать, что собственные значения матриц A и A_k из QR-алгоритма при $k = 1, 2, \ldots$ совпадают.

$\lhd$ Так как $A_{k+1} = R_k Q_k = Q_k^T (Q_k R_k) Q_k = Q_k^T A_k Q_k$, то из $Q_k^{-1} = Q_k^T$ следует, что матрицы A_{k+1} и A_k ортогонально подобны и имеют одинаковые собственные значения. $\rhd$

5.223. Пусть A_k — матрица из QR-алгоритма, а V_k — из метода ортогональной итерации для $V_0 = I$. Показать, что $A_k = V_k^T A V_k$.

У к а з а н и е. Равенство $A_k = V_k^T A V_k$ проверить по индукции.

5.224. Доказать, что если A — нормальная матрица ($A^T A = A A^T$), то последовательность треугольных матриц R_k из QR-алгоритма сходится к диагональной матрице.

$\lhd$ Рассмотрим две соседние матрицы QR-алгоритма, обозначая их для простоты через A и B. Переход от A к B описывается формулами

$$A = Q\,R\,, \quad B = R\,Q\,. \tag{5.13}$$

Пусть $\mathbf{b}^i, \mathbf{a}^i, \mathbf{r}^i$ — столбцы, $\mathbf{b}_i$, $\mathbf{a}_i$, $\mathbf{r}_i$ — строки соответственно матриц B, A, R ($i = 1, \ldots, n$). Так как ортогональные преобразования не меняют евклидову длину вектора, то из (5.13) следует: $\|\mathbf{a}^i\|_2 = \|\mathbf{r}^i\|_2$, $\|\mathbf{r}_i\|_2 = \|\mathbf{b}_i\|_2$ ($i = 1, \ldots, n$). Из нормальности матриц A и B имеем: $\|\mathbf{a}^i\|_2 = \|\mathbf{a}_i\|_2$, $\|\mathbf{b}^i\|_2 = \|\mathbf{b}_i\|_2$ ($i = 1, \ldots, n$). Положим

$$\Delta_m = \sum_{i=1}^{m} \|\mathbf{b}_i\|_2^2 - \|\mathbf{a}_i\|_2^2, \quad m = 1, \ldots, n-1.$$

Тогда

$$\begin{aligned} \Delta_1 &= \|\mathbf{b}_1\|_2^2 - \|\mathbf{a}_1\|_2^2 = \|\mathbf{r}_1\|_2^2 - \|\mathbf{r}^1\|_2^2 = |r_{12}|^2 + \cdots + |r_{1n}|^2\,, \\ \Delta_m &= \sum_{i=1}^{m} \|\mathbf{r}_i\|_2^2 - \sum_{i=1}^{m} \|\mathbf{r}^i\|_2^2 = \sum_{i=1}^{m} \sum_{j=m+1}^{n} |r_{ij}|^2\,, \quad m = 2, \ldots, n-1. \end{aligned} \tag{5.14}$$

Если теперь составить для каждого k величины $\delta_m^{(k)} = \sum\limits_{i=1}^{m} \|\mathbf{a}_i^{(k)}\|_2^2$, где $\mathbf{a}_i^{(k)}$ — строки матрицы A_k и $m = 1, \ldots, n-1$, то получим $n-1$ последовательностей $\delta_m^{(k)}$. Из (5.14) следует, что каждая из этих последовательностей монотонно возрастает и каждая ограничена (например, общим значением квадрата евклидовой нормы матриц A_k). Поэтому последовательности $\delta_m^{(k)}$ сходятся. Соответствующие последовательности $\Delta_m^{(k)}$, где $\Delta_m^{(k)} = \delta_m^{(k+1)} - \delta_m^{(k)}$, сходятся к нулю, вместе с тем сходятся к нулю все наддиагональные элементы матриц R_k.

Так как $\|\mathbf{r}_1^{(k)}\|_2^2 = \|\mathbf{a}_1^{(k+1)}\|_2^2$, то

$$|r_{11}^{(k)}|^2 = \delta_1^{(k+1)} - |r_{12}^{(k)}|^2 - \cdots - |r_{1n}^{(k)}|^2 .$$

По соглашению $r_{11}^{(k)} \geqslant 0$ для всех k, поэтому последовательность $r_{11}^{(k)}$ имеет предел. Точно так же из равенств

$$|r_{22}^{(k)}|^2 = \|\mathbf{r}_2^{(k)}\|_2^2 - |r_{23}^{(k)}|^2 - \cdots - |r_{2n}^{(k)}|^2 =$$
$$= \|\mathbf{a}_2^{(k+1)}\|_2^2 - |r_{23}^{(k)}|^2 - \cdots - |r_{2n}^{(k)}|^2 = \delta_2^{(k+1)} - \delta_1^{(k+1)} - |r_{23}^{(k)}|^2 - \cdots - |r_{2n}^{(k)}|^2$$

следует, что сходится последовательность $r_{22}^{(k)}$. Продолжая рассуждать аналогичным образом, установим существование предела матричной последовательности R_k. $\triangleright$

5.225. Доказать, что нормальная матрица A вида $A = Q\,D$, где Q — ортогональная матрица, D — диагональная матрица с неотрицательными элементами, с точностью до симметричной перестановки строк и столбцов является блочно-диагональной. При этом каждый диагональный блок только скалярным множителем отличается от ортогональной матрицы соответствующего порядка.

$\triangleleft$ Из равенства $A\,A^T = A^T A$ следует, что $D^2 = Q\,D^2 Q^T$, или $D^2 Q = Q\,D^2$. Отсюда получаем поэлементное равенство $q_{ij}(d_{ii}^2 - d_{jj}^2) = 0 \ \forall i, j$, т. е. $q_{ij} = 0$, если $d_{ii} \neq d_{jj}$. Матрица перестановок P, группирующая равные диагональные элементы матрицы D: $D \to \tilde{D} = P^T D P$, приводит Q к блочно-диагональному виду: $Q \to \tilde{Q} = P^T Q P$. Но тогда и матрица $\tilde{A} = P^T A P = (P^T Q P)(P^T D P) = \tilde{Q}\tilde{D}$ — блочно-диагональная, причем каждый диагональный блок есть произведение одноименного блока ортогональной матрицы Q на число d, отвечающее этому блоку. $\triangleright$

5.226. Исследовать применение QR-алгоритма для матрицы

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

$\triangleleft$ Имеем: $A_0 = A_1 = \ldots = A_k \quad \forall k$. $\triangleright$

5.227. Пусть $\lambda_1, \ldots, \lambda_n$ — собственные значения комплексной матрицы $A = F + \mathrm{i}\,G$, где F и G — вещественные матрицы. Показать, что числа $\lambda_1, \ldots, \lambda_n$, $\bar{\lambda}_1, \ldots, \bar{\lambda}_n$ составляют спектр вещественной матрицы удвоенного порядка $A_R = \begin{pmatrix} F & -G \\ G & F \end{pmatrix}$.

У к а з а н и е. С каждым собственным вектором $\mathbf{z}$ матрицы A ассоциировано двумерное инвариантное подпространство матрицы A_R. Действительно, представим $\mathbf{z}$ в виде $\mathbf{z} = \mathbf{x} + \mathrm{i}\,\mathbf{y}$ с вещественными векторами $\mathbf{x}, \mathbf{y}$, и пусть $\lambda = \mu + \mathrm{i}\,\nu$ — соответствующее собственное значение. Комплексное равенство $A\,\mathbf{z} = \lambda\,\mathbf{z}$ эквивалентно каждой из следующих систем вещественных равенств:

$$\begin{cases} F\mathbf{x} - G\mathbf{y} = \mu\mathbf{x} - \nu\mathbf{y}, \\ G\mathbf{x} + F\mathbf{y} = \nu\mathbf{x} + \mu\mathbf{y}; \end{cases} \qquad \begin{cases} F(-\mathbf{y}) - G\mathbf{x} = \mu(-\mathbf{y}) - \nu\mathbf{x}, \\ G(-\mathbf{y}) + F\mathbf{x} = \nu(-\mathbf{y}) + \mu\mathbf{x}, \end{cases}$$

означающих, что подпространство, являющееся линейной оболочкой векторов $\mathbf{z}_R^1 = (\mathbf{x}, \mathbf{y})^T$, $\mathbf{z}_R^2 = (-\mathbf{y}, \mathbf{x})^T$, инвариантно относительно матрицы A_R. Нетрудно проверить, что они линейно независимы.

QR-алгоритм со сдвигом. Скорость сходимости QR-алгоритма можно существенно повысить, если применять разложение к матрицам вида $A_k = H_k - c_k\,I$. Положим $A_0 = A$, выберем сдвиг c_0 вблизи некоторого собственного значения $\lambda_k(A)$ и вычислим разложение $A_0 - c_0 I = Q_0 R_0$. Найдем матрицу $A_1 = R_0 Q_0 + c_0 I$. Повторим процедуру для матриц A_k, $k = 1, 2, \ldots$. Если c_k окажется равным некоторому собственному значению, то у матрицы R_k на диагонали стоит нуль, поэтому можно найти соответствующий собственный вектор и задачу для A_{k+1} сформулировать как задачу на единицу меньшей размерности.

5.228. Показать, что собственные значения матриц A и A_k из QR-алгоритма со сдвигом при $k = 1, 2, \ldots$ совпадают.

$\lhd$ Так как

$$A_{k+1} = R_k Q_k + c_k I = Q_k^T (Q_k R_k + c_k I) Q_k = Q_k^T A_k Q_k,$$

то матрицы A_k и A_{k+1} ортогонально подобны. $\rhd$

5.229. Показать, что QR-алгоритм со сдвигами c_k для начальной матрицы в форме Хессенберга

$$H_k - c_k I = Q_k R_k\,, \quad H_{k+1} = R_k Q_k + c_k I$$

порождает последовательность H_k хессенберговых ортогонально подобных матриц.

5.230. При каком из следующих сдвигов:

1) базовый алгоритм: $c_k \equiv 0$;

2) сдвиги по Рэлею: $c_k = h_{nn}^{(k)}$ где $h_{nn}^{(k)}$ — правый нижний элемент матрицы H_k;

3) сдвиги по Уилкинсону: c_k выбирается как одно из собственных значений матрицы $\begin{pmatrix} h_{n-1,n-1}^{(k)} & h_{n-1,n}^{(k)} \\ h_{n,n-1}^{(k)} & h_{n,n}^{(k)} \end{pmatrix}$ (правая нижняя подматрица второго порядка матрицы H_k), матрица вида $H_0 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$ не меняется в процессе QR-итераций?

Ответ: случаи 1 и 3.

Глава 6
Решение нелинейных уравнений

Итерационные методы вычисления изолированного (отделенного от других) корня z уравнения $f(x) = 0$, как правило, требуют указания какой-либо области D, содержащей этот единственный корень, и алгоритма нахождения очередного приближения x_{n+1} по уже имеющимся x_n, ..., x_{n-k}.

Широко используемые способы отделения корней — графический и табличный — базируются на свойствах гладкости функции; в случае, когда $f(x)$ является алгебраическим полиномом степени n, существуют аналитические подходы.

Если $f(x)$ — непрерывная, то вещественный корень z принадлежит любому отрезку, на концах которого эта функция имеет значения разных знаков. Деля отрезок пополам, получаем универсальный метод вычисления корня (метод бисекции). Этот подход не требует знания хорошего начального приближения. Если оно имеется, то для гладких функций используют более эффективные методы.

Пусть отыскивается единственный на отрезке $[a, b]$ корень z уравнения $f(x) = 0$ в предположении непрерывности функции $f(x)$. Если в его окрестности функция представляется в виде $f(x) = (x - z)^p g(x)$, где p — натуральное число, а $g(x)$ — ограниченная функция такая, что $g(z) \neq 0$, то число p называют *кратностью* корня. Если $p = 1$, то корень называют *простым*. При нечетном p функция $f(x)$ меняет знак на $[a, b]$, т. е. $f(a)f(b) < 0$, а при четном p — нет.

Итерационный метод решения порождает последовательность приближений x_n, которая сходится к корню: $\lim\limits_{n\to\infty} |x_n - z| = 0$. Величину $e_n = |x_n - z|$ называют *абсолютной ошибкой* на n-й итерации. Итерационный метод имеет *порядок* m (или *скорость сходимости* m), если m — наибольшее положительное число, для которого существует такая конечная постоянная $q > 0$, что

$$\lim_{n\to\infty} \sup \frac{e_{n+1}}{e_n^m} \leqslant q < \infty.$$

Постоянную q называют *константой асимптотической ошибки*, ее обычно оценивают через производные функции $f(x)$ в точке $x = z$. При $m = 1\,(q \in (0, 1))$ сходимость называют *линейной* (иногда говорят, что в этом случае метод сходится со скоростью геометрической прогрессии, знаменатель которой q), при $1 < m < 2$ — *сверхлинейной*, при $m = 2$ — *квадратичной* и т. д. Из сходимости с порядком $m > 1$ следует оценка $e_{n+1} \leqslant q_n e_n$, $q_n \to 0$ при $n \to \infty$. При этом $e_{n+1} \leqslant e_0 \prod\limits_{i=0}^{n} q_i$. Иногда скорость сходимости может замедляться при приближении к искомому решению, что соответствует $q_n \to 1$, но $e_n \to 0$ при $n \to \infty$. Таким свойством обладают методы с *полиномиальной* скоростью сходимости $e_n \leqslant e_0(1+\alpha n e_0^l)^{-1/l}$.

Данная оценка верна, например, если $e_{n+1} \leqslant (1 - \alpha e_n^l) e_n$ с некоторыми $l \geqslant 1$ и $0 < \alpha \leqslant e_0^{-l}$. Для методов с полиномиальной скоростью сходимости число итераций n, необходимое для достижения ошибки порядка ε имеет асимптотику $n \approx \varepsilon^{-l}$, что существенно ограничивает их применение для расчетов с высокой точностью.

Особое внимание в теории решения нелинейных уравнений уделяется методам со сверхлинейной скоростью сходимости. При практических расчетах традиционно применяют методы с квадратичной сходимостью, так как итерационные процессы более высокого порядка ($m > 2$) обычно требуют серьезного увеличения вычислительных затрат.

6.1. Метод простой итерации и смежные вопросы

Исходное уравнение $f(x) = 0$ часто заменяют эквивалентным ему уравнением $x = \varphi(x)$. Эту замену можно сделать, положив, например,

$$\varphi(x) = x + \psi(x) f(x),$$

где $\psi(x)$ — произвольная непрерывная знакопостоянная функция.

Метод простой итерации. Выберем некоторое начальное приближение $x_0 \in [a, b]$ к корню z, дальнейшие приближения будем вычислять по формуле

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, 2, \ldots .$$

Известно, что последовательность x_n сходится к z, если отображение $y = \varphi(x)$ является *сжимающим (сжимающим экспоненциально)*, т. е. при некотором $0 \leqslant q < 1$ выполнено условие

$$\rho\left(\varphi(x_1), \varphi(x_2)\right) \leqslant q\, \rho\left(x_1, x_2\right),$$

либо *слабо сжимающим (сжимающим полиномиально)*, т. е. при некоторых $\alpha > 0, l \geqslant 1$ выполнено условие

$$\rho\left(\varphi(x_1), \varphi(x_2)\right) \leqslant \frac{\rho\left(x_1, x_2\right)}{\left(1 + \alpha \rho^l\left(x_1, x_2\right)\right)^{1/l}}.$$

Здесь $\rho\left(x_1, x_2\right)$ — расстояние между точками x_1 и x_2. Сходимость последовательности x_n гарантируется, если оценка сжатия выполняется либо для всех точек $x_{1,2} \in \mathbf{R}^1$, либо для точек $x_{1,2} \in [a, b]$ при условии, что $\varphi(x) \in [a, b]$, $\forall x \in [a, b]$.

Метод секущих. Пусть x_{n-1} и x_n — последовательные приближения к корню. Заменим кривую $y = f(x)$ прямой, проходящей через точки $(x_{n-1}, f(x_{n-1}))$ и $(x_n, f(x_n))$. В качестве следующего приближения к корню возьмем точку пересечения этой прямой с осью абсцисс. Расчетная формула принимает вид

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}\, f(x_n).$$

Метод хорд. Пусть $f(a)f(b) < 0$. Идея метода (его еще называют методом *ложного положения*) состоит в замене кривой $y = f(x)$ хордами, проходящими через концы отрезков, в которых $f(x)$ имеет противоположные знаки. Метод хорд требует, чтобы один конец отрезка, на котором ищется корень, был неподвижен. В качестве неподвижного конца x_0 выбирают конец отрезка, для которого знак $f(x)$ совпадает со знаком второй производной $f''(x)$. Расчетная формула имеет вид

$$x_{n+1} = x_n - \frac{x_n - x_0}{f(x_n) - f(x_0)}\ f(x_n).$$

Метод парабол. Пусть x_{n-2}, x_{n-1} и x_n — три последовательных приближения к корню. Заменим кривую $y = f(x)$ параболой, проходящей через точки $(x_{n-2}, f(x_{n-2}))$, $(x_{n-1}, f(x_{n-1}))$ и $(x_n, f(x_n))$. В качестве следующего приближения к корню возьмем ближайшую к x_n точку пересечения параболы с осью абсцисс. Этот подход исключительно эффективен для нахождения корней многочлена как с действительными, так и с комплексными коэффициентами.

При нахождении кратных корней ($p > 1$) для большинства алгоритмов характерно замедление скорости сходимости.

Методы решения систем нелинейных уравнений. Рассмотрим некоторые из этих методов. Будем считать, что заданная система

$$\begin{cases} f_1(x^1, x^2, ..., x^m) = 0, \\ f_2(x^1, x^2, ..., x^m) = 0, \\ \dots\dots\dots\dots\dots\dots \\ f_m(x^1, x^2, ..., x^m) = 0 \end{cases}$$

записывается в операторной форме как $\mathbf{F}(\mathbf{x}) = 0$, $\mathbf{x} = (x^1, ..., x^m)^T$. Тогда для решения задачи можно пытаться построить некоторое сжимающее отображение $\mathbf{G}(\mathbf{x})$ и применить *метод простой итерации* $\mathbf{x}_{n+1} = \mathbf{G}(\mathbf{x}_n)$. В многомерном случае также возможна следующая модификация метода:

$$\begin{cases} x^1_{n+1} = g_1(x^1_n, x^2_n, ..., x^m_n), \\ x^2_{n+1} = g_2(x^1_{n+1}, x^2_n, ..., x^m_n), \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ x^m_{n+1} = g_m(x^1_{n+1}, ..., x^{m-1}_{n+1}, x^m_n). \end{cases}$$

Оператор $\mathbf{G}(\mathbf{x})$ можно взять, например, в виде $\mathbf{G}(\mathbf{x}) = \mathbf{x} - \tau\mathbf{F}(\mathbf{x})$, τ — параметр. Это соответствует обобщению метода простой итерации для систем линейных уравнений $\frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{\tau} + \mathbf{F}(\mathbf{x}_n) = 0$ на случай нелинейных задач.

Если из функции $\mathbf{F}$ удается выделить некоторую главную линейную часть, т. е. получить представление $\mathbf{F}(\mathbf{x}) = \mathbf{B}\mathbf{x} + \mathbf{H}(\mathbf{x})$, то можно использовать *метод Пикара* $\mathbf{B}\mathbf{x}_{n+1} + \mathbf{H}(\mathbf{x}_n) = 0$ или его модификацию

$$\mathbf{B}\,\frac{\mathbf{x}_{n+1} - \mathbf{x}_n}{\tau} + \mathbf{F}(\mathbf{x}_n) = 0.$$

Обобщением метода Гаусса—Зейделя и метода Якоби на нелинейный случай является *метод покоординатного спуска*: компоненты очередного приближения $\mathbf{x}_{n+1}$ определяются при решении одномерных нелинейных уравнений следующих систем соответственно:

$$\begin{cases} f_1(x^1_{n+1}, x^2_n, \dots, x^m_n) = 0, \\ f_2(x^1_{n+1}, x^2_{n+1}, \dots, x^m_n) = 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ f_m(x^1_{n+1}, x^2_{n+1}, \dots, x^m_{n+1}) = 0, \end{cases} \qquad \begin{cases} f_1(x^1_{n+1}, x^2_n, \dots, x^m_n) = 0, \\ f_2(x^1_n, x^2_{n+1}, \dots, x^m_n) = 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ f_m(x^1_n, \dots, x^{m-1}_n, x^m_{n+1}) = 0. \end{cases}$$

Иногда нахождение корней системы уравнений удобно свести к нахождению минимума некоторого функционала $\mathbf{\Phi}(\mathbf{x})$, например $\mathbf{\Phi}(\mathbf{x}) = \|\mathbf{F}(\mathbf{x})\|^2$. Напомним, что таким же образом строились методы решения линейных систем. Для минимизации $\mathbf{\Phi}(\mathbf{x})$ можно рассмотреть следующий нестационарный процесс, который называют *методом установления*:

$$\frac{d}{dt}\,\mathbf{x}(t) + \operatorname{grad}\mathbf{\Phi}(\mathbf{x}(t)) = 0.$$

При $\operatorname{grad}\mathbf{\Phi}(\mathbf{x}) \neq 0$ имеем

$$\frac{d}{dt}\,\mathbf{\Phi}(\mathbf{x}) = \Big(\operatorname{grad}\mathbf{\Phi}(\mathbf{x}), \frac{d\mathbf{x}}{dt}\Big) = -\Big(\operatorname{grad}\mathbf{\Phi}(\mathbf{x}), \operatorname{grad}\mathbf{\Phi}(\mathbf{x})\Big) < 0\ ,$$

т. е. вдоль траектории $\mathbf{x}(t)$ значение $\mathbf{\Phi}(\mathbf{x})$ не возрастает.

Другой возможный нестационарный процесс, решение которого также устанавливается в точке минимума, имеет вид

$$\frac{d^2\mathbf{x}}{dt^2} + \gamma\,\frac{d\mathbf{x}}{dt} + \operatorname{grad}\mathbf{\Phi}(\mathbf{x}) = 0,\ \gamma > 0.$$

В данном случае вдоль траектории не возрастает значение функционала $\frac{1}{2}\left(\frac{d\mathbf{x}}{dt}\,, \frac{d\mathbf{x}}{dt}\right) + \mathbf{\Phi}(\mathbf{x})$. При численной реализации операторы производных заменяют соответствующими разностными аналогами.

6.1. Пусть уравнение $f(x) \equiv x - \varphi(x) = 0$ имеет корень z, для его вычисления применяется метод простой итерации $x_{n+1} = \varphi(x_n)$, функция $\varphi(x)$ имеет непрерывную производную в открытой окрестности z и $|\varphi'(z)| \leqslant q < 1$. Доказать, что метод сходится к z при выборе начального приближения x_0 из некоторой окрестности корня.

$\lhd$ Функция $\varphi'(x)$ непрерывна и $|\varphi'(z)| < 1$, следовательно, $|\varphi'(x)| < 1$ в некоторой окрестности $Q_\delta = [z-\delta, z+\delta]$. Возьмем произвольную точку $x_0 \in Q_\delta$. По теореме Лагранжа имеем $\varphi(x_0) = \varphi(z) + \varphi'(\xi)(x_0 - z)$, где точка $\xi \in Q_\delta$. Так как $z = \varphi(z)$, то $|\varphi(x_0) - z| = |\varphi'(\xi)(x_0 - z)| < |x_0 - z|$, т. е. функция φ отображает отрезок Q_δ в себя. Далее, так как

$$x_{n+1} - z = \varphi(x_n) - \varphi(z), \quad n = 0, 1, 2, \dots\ ,$$

то по теореме Лагранжа для каждого n существует такое ξ_n, что

$$x_{n+1} - z = (x_n - z)\,\varphi'(\xi_n).$$

Последовательно применяя указанную теорему, получаем

$$x_{n+1} - z = (x_n - z)\,\varphi'(\xi_n) = (x_{n-1} - z)\,\varphi'(\xi_n)\varphi'(\xi_{n-1}) = \ldots$$
$$\ldots = (x_0 - z)\,\varphi'(\xi_n)\varphi'(\xi_{n-1})\ldots\varphi'(\xi_0),$$

где $\xi_n, \xi_{n-1}, \ldots, \xi_0 \in Q_\delta$. Так как $|\varphi'(\xi_i)| \leqslant q$, $i = 0, 1, 2, \ldots n$, то

$$|x_{n+1} - z| \leqslant |x_0 - z|\,q^{n+1}.$$

При $q < 1$ правая часть этого неравенства стремится к нулю, т. е. последовательность x_n сходится к корню z. $\triangleright$

6.2. Дополнительно к условиям 6.1 потребуем $\varphi'(z) = 0$, непрерывность и ограниченность $\varphi''(x)$. Доказать, что для метода $x_{n+1} = \varphi(x_n)$ в окрестности корня z верна квадратичная оценка сходимости

$$|z - x_{n+1}| \leqslant C(z - x_n)^2.$$

$\triangleleft$ Оценим (см. решение 6.1) скорость сходимости метода для $x_n \in Q_\delta$:

$$x_{n+1} - z = \varphi(x_n) - \varphi(z) = \varphi(z + (x_n - z)) - \varphi(z) =$$
$$= \varphi(z) + (x_n - z)\varphi'(z) + \frac{1}{2}\,(x_n - z)^2\,\varphi''(\xi_n) - \varphi(z) =$$
$$= \frac{\varphi''(\xi_n)}{2}\,(x_n - z)^2,\ \xi_n \in [x_n,\, z].$$ $\triangleright$

6.3. Пусть на некотором отрезке $Q_\delta = [a - \delta, a + \delta]$ функция $\varphi(x)$ удовлетворяет условию Липшица $|\varphi(x') - \varphi(x'')| \leqslant q|x' - x''|$ с константой $q < 1$ и в точке a выполняется неравенство $|a - \varphi(a)| \leqslant (1 - q)\delta$. Показать, что на отрезке Q_δ уравнение $f(x) \equiv x - \varphi(x) = 0$ имеет единственный корень z и последовательность $x_n = \varphi(x_{n-1})$ сходится к корню z для произвольного $x_0 \in Q_\delta$.

$\triangleleft$ Пусть $x_0 \in Q_\delta$, т. е. $|a - x_0| \leqslant \delta$. Тогда

$$|\varphi(x_0) - a| = |\varphi(x_0) - \varphi(a) + \varphi(a) - a| \leqslant$$
$$\leqslant |\varphi(x_0) - \varphi(a)| + |\varphi(a) - a| \leqslant q|x_0 - a| + (1 - q)\delta \leqslant \delta.$$

Таким образом, функция $\varphi(x)$ отображает Q_δ в себя и является, по условию Липшица, сжимающей с константой q. Применяя принцип сжимающих отображений, завершаем решение. $\triangleright$

6.4. Пусть функция $f'(y)$ непрерывна и $\left|\dfrac{f(x_n)}{f'(y)}\right| \leqslant \varepsilon$ для всех $y \in [x_n - \varepsilon, x_n + \varepsilon]$. Доказать, что для некоторого $z \in [x_n - \varepsilon, x_n + \varepsilon]$ справедливо равенство $f(z) = 0$.

$\triangleleft$ По теореме Лагранжа имеем $f(x_n + t) = f(x_n) + f'(y)t$, отсюда следует $\dfrac{f(x_n + t)}{f'(y)} = \dfrac{f(x_n)}{f'(y)} + t$. Выражение в правой части равенства неотрицательно при $t = \varepsilon$ и неположительно при $t = -\varepsilon$. Из условия следует, что

производная функции не меняет знака при $t \in [-\varepsilon, \varepsilon]$, поэтому приходим к выводу, что если оба значения $f(x_n - \varepsilon)$ и $f(x_n + \varepsilon)$ отличны от нуля, то они имеют разные знаки.

Установленный факт лежит в основе точки зрения, согласно которой в качестве критерия остановки итерационного метода для нахождения простых корней условие $\left|\frac{f(x_n)}{f'(x_n)}\right| \leqslant \varepsilon$ предпочтительнее, чем условие $|f(x_n)| \leqslant \varepsilon$. $\triangleright$

6.5. Пусть уравнение $f(x) = 0$ имеет корень на отрезке $[a, b]$, причем функция $f(x)$ дифференцируема, а производная $f'(x)$ знакопостоянна на этом отрезке. Построить равносильное уравнение вида $x = \varphi(x)$, для которого на $[a, b]$ выполнено условие $|\varphi'(x)| \leqslant q < 1$.

$\triangleleft$ Для определенности будем считать, что $f'(x) > 0$. Пусть $0 < m \leqslant f'(x) \leqslant M$. Заменим исходное уравнение $f(x) = 0$ равносильным

$$x = \varphi(x), \quad \varphi(x) = x - \lambda f(x), \quad \lambda > 0.$$

Подберем параметр λ так, чтобы на $[a, b]$ выполнялось неравенство

$$0 \leqslant \varphi'(x) = 1 - \lambda f'(x) \leqslant q < 1.$$

При $\lambda = \frac{1}{M}$ получаем $q = 1 - \frac{m}{M} < 1$. $\triangleright$

6.6. Построить итерационный процесс вычисления всех корней уравнения $f(x) \equiv x^3 + 3x^2 - 1 = 0$ методом простой итерации.

$\triangleleft$ Табличным способом выделим отрезки, на концах которых функция $f(x)$ имеет разные знаки:

x	-3	-2	-1	0	1	2	3
$\operatorname{sign} f(x)$	$-$	$+$	$+$	$-$	$+$	$+$	$+$

Таким образом, корни исходного уравнения лежат на отрезках $[-3, -2]$, $[-1, 0]$, $[0, 1]$, для каждого из которых построим свой итерационный процесс.

Так как на $[-3, -2]$ имеем $x \neq 0$, то исходное уравнение можно разделить на x^2. В результате получаем равносильное уравнение

$$x = \varphi(x), \quad \varphi(x) = \frac{1}{x^2} - 3.$$

Итерационный процесс для нахождения первого корня: $x_{n+1} = \frac{1}{x_n^2} - 3$. Сходимость имеет место для всех начальных приближений x_0 из этого отрезка, так как для $x \in [-3, -2]$ имеет место оценка

$$|\varphi'(x)| = \left|-\frac{2}{x^3}\right| \leqslant \frac{1}{4} < 1, \quad \varphi(x) \in [-3, -2].$$

Для двух других отрезков исходное уравнение представим следующим образом $x^2(x + 3) - 1 = 0$, при этом $x + 3 \neq 0$. Если $x_0 \in [-1, 0]$, то

определим итерационный процесс в виде $x_{n+1} = -\frac{1}{\sqrt{x_n+3}}$; если $x_0 \in [0,1]$, то в виде $x_{n+1} = \frac{1}{\sqrt{x_n+3}}$. Можно показать, что в процессе итераций соответствующие отрезки отображаются в себя, поэтому (см. 6.1) сходимость построенных итерационных процессов следует из оценки

$$|\,\varphi'(x)\,| = \frac{1}{2}\left|\frac{1}{\sqrt{x+3}}\right|^3 < 1. \qquad \triangleright$$

6.7. Определить область начальных приближений x_0, для которых итерационный процесс $x_{n+1} = \frac{x_n^3+1}{20}$ сходится.

$\triangleleft$ Решаем уравнение $x^3 - 20x + 1 = 0$, имеющее три различных вещественных корня: $z_1 < z_2 < z_3$. В зависимости от выбора начального приближения x_0 итерационный процесс либо расходится, либо сойдется к одному из корней z_i, $i = 1, 2, 3$.

Запишем формулу итерационного процесса в виде

$$x_{n+1} - x_n = \frac{x_n^3 - 20x_n + 1}{20}\,,$$

или, что тоже самое, — в эквивалентной форме:

$$x_{n+1} - x_n = \frac{(x_n - z_1)(x_n - z_2)(x_n - z_3)}{20}\,.$$

Отсюда имеем, что при $x_n < z_1$ справедливо $x_{n+1} - x_n < 0$, и последовательность x_n монотонно убывает. Это означает расходимость итерационного процесса при $x_0 < z_1$, так как $x_n < x_0 < z_i$, $i = 1, 2, 3$. Аналогично показывается, что при $z_3 < x_0$ выполняются неравенства $z_i < x_n < x_{n+1}$, и метод расходится.

Точки $x_0 = z_1$, $x_0 = z_2$ и $x_0 = z_3$ являются неподвижными, а отображение $x_{n+1} = \frac{x_n^3+1}{20}$ монотонно. Отсюда следует, что для $z_1 < x_0 < z_2$ имеем $z_1 < x_n < x_{n+1} < z_2$. Таким образом, последовательность x_n монотонно возрастает и ограничена сверху, следовательно, итерационный процесс сходится к точке z_2. Аналогично доказывается, что для $z_2 < x_0 < z_3$ имеем $z_2 < x_{n+1} < x_n < z_3$, т. е. метод сходится к точке z_2. $\triangleright$

6.8. Оценить скорость сходимости итерационного процесса $x_{n+1} = x_n - x_n^3 + x_n^4$ к корню $z = 0$ при малых x_0.

6.9. Пусть z — простой корень уравнения $f(x) = 0$. Оценить скорость сходимости метода хорд в его окрестности.

$\triangleleft$ Представим сходящийся метод хорд как частный случай метода простой итерации:

$$x = \varphi(x), \quad \varphi(x) = x - \frac{f(x)}{f(x) - f(x_0)}\,(x - x_0).$$

Вблизи простого корня z уравнения $f(x)=0$ имеем

$$x_{n+1}-z=(x_n-z)\,\varphi'(z)+\frac{1}{2}\,(x_n-z)^2\,\varphi''(\xi),\quad \xi\in[x_n,z],$$

где

$$\varphi'(z)=1+\frac{f'(z)}{f(x_0)}\,(z-x_0)=$$

$$=\frac{f(z)-f'(z)(z-x_0)+\dfrac{f''(\eta)}{2}\,(z-x_0)^2+f'(z)(z-x_0)}{f(x_0)}=$$

$$=\frac{(z-x_0)^2}{2}\,\frac{f''(\eta)}{f(x_0)}=\frac{(z-x_0)^2}{2}\,\frac{f''(\eta)}{f(z)+f'(\xi)(z-x_0)},\quad \eta,\xi\in[x_0,z].$$

Если начальное приближение x_0 взять в окрестности корня, для которой справедливо $|\varphi'(z)|\leqslant q<1$, то, учитывая 6.1, приходим к выводу, что метод хорд для x_1 из достаточно малой окрестности $[z-\delta,z+\delta]$ имеет линейную скорость сходимости. $\triangleright$

6.10. Пусть z — простой корень уравнения $f(x)=0$. Оценить скорость сходимости метода секущих в его окрестности.

$\triangleleft$ Преобразуем расчетную формулу метода секущих

$$x_{n+1}=x_n-\frac{x_n-x_{n-1}}{f(x_n)-f(x_{n-1})}\,f(x_n)$$

к виду

$$x_{n+1}-z=x_n-z-\frac{((x_n-z)-(x_{n-1}-z))f(z+(x_n-z))}{f(z+(x_n-z))-f(z+(x_{n-1}-z))}\,.$$

Разложим $f(z+(x_n-z))$ и $f(z+(x_{n-1}-z))$ в ряды Тейлора в точке z и подставим в последнюю формулу, учитывая, что $f(z)=0$. Имеем

$$x_{n+1}-z=x_n-z-\frac{(x_n-z)f'(z)+0,5(x_n-z)^2\,f''(z)+\dots}{f'(z)+0,5((x_n-z)+(x_{n-1}-z))f''(z)+\dots}=$$

$$=(x_n-z)\left(1-\frac{1+0,5(x_n-z)\,\dfrac{f''(z)}{f'(z)}+\dots}{1+0,5(x_n-z)\,\dfrac{f''(z)}{f'(z)}+0,5(x_{n-1}-z)\,\dfrac{f''(z)}{f'(z)}+\dots}\right)=$$

$$=0,5(x_n-z)\,(x_{n-1}-z)\,\frac{f''(z)}{f'(z)}+O((x_n-z)^3).$$

Опустив члены более высокого порядка малости, для ошибки получаем уравнение

$$x_{n+1}-z=C\,(x_n-z)\,(x_{n-1}-z),\quad C=\frac{1}{2}\,\frac{f''(z)}{f'(z)}\,,\ f'(z)\neq 0.$$

Предположим, что скорость сходимости определяется соотношением $x_{n+1} - z = A\,(x_n - z)^m$, в котором значения A и m пока неизвестны. Тогда $x_n - z = A\,(x_{n-1} - z)^m$, откуда $x_{n-1} - z = A^{-1/m}\,(x_n - z)^{1/m}$. Подставим эти соотношения в уравнение для ошибки

$$A(x_n - z)^m = C\,(x_n - z)\,A^{-1/m}\,(x_n - z)^{1/m}.$$

Приравнивая степени и коэффициенты многочленов, получаем два уравнения с двумя неизвестными

$$m = 1 + \frac{1}{m}, \quad 1 = C\,A^{-(1+\frac{1}{m})}.$$

Из первого уравнения находим показатель скорости сходимости метода секущих $m = 0,5\,(1 + \sqrt{5}) \approx 1,618$. При этом константа A равна $\left(\frac{1}{2}\,\frac{f''(z)}{f'(z)}\right)^{\frac{1}{m}}$. ▷

6.11. Доказать, что все корни уравнения

$$f(x) \equiv a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 = 0$$

расположены в кольце $\frac{|a_0|}{b + |a_0|} \leqslant |z| \leqslant 1 + \frac{c}{|a_n|}$, где $c = \max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\}$, $b = \max\{|a_1|, |a_2|, \ldots, |a_n|\}$.

◁ Для корней $|z| > 1$ имеем

$$a_n z^n = -(a_{n-1}z^{n-1} + \ldots + a_1 z + a_0), \quad |z| \leqslant \sum_{i=1}^{n} \frac{|a_{n-i}|}{|a_n|}\,\frac{1}{|z|^{i-1}} \leqslant \frac{c}{|a_n|}\,\frac{|z|}{|z| - 1},$$

откуда $|z| - 1 \leqslant \frac{c}{|a_n|}$ и $|z| \leqslant 1 + \frac{c}{|a_n|}$.

Если теперь $|a_0| > 0$, то все корни уравнения отличны от нуля. Делая замену $u = \frac{1}{z}$, приходим к уравнению $a_0 u^n + a_1 u^{n-1} + \ldots + a_n = 0$. Из предыдущей оценки следует $|u| \leqslant 1 + \frac{b}{|a_0|}$, или $|z| \geqslant \frac{|a_0|}{b + |a_0|}$. ▷

6.12. Доказать, что если при $x = a$ имеют место неравенства $f(a) > 0\,, f'(a) > 0\,, \ldots\,, f^{(n)}(a) > 0\,$, то уравнение

$$f(x) \equiv a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$$

не имеет действительных корней, бóльших a.

У к а з а н и е. Использовать формулу Тейлора для многочлена $f(x)$ степени n

$$f(x) = f(a) + \sum_{k=1}^{n} \frac{1}{k!}\,f^{(k)}(a)(x - a)^k.$$

6.13. Найти границы действительных корней уравнения

$$x^4 - 35\,x^3 + 380\,x^2 - 1350\,x + 1000 = 0\,.$$

Указание. Применяя 6.12, получить, что положительные корни расположены на $[0, 74, 22]$, а отрицательных корней нет.

6.14. Пусть $x_{n+1} = \sqrt{x_n + 2}$. Доказать, что $\lim\limits_{n\to\infty} x_n = 2$ для любого $x_0 \geqslant -2$.

$\lhd$ Пусть $\varphi(x) = \sqrt{x+2}$ и $x_0 \geqslant -2$. Так как $\varphi(x_0) \geqslant 0$, то сходимость метода достаточно доказать для $x_0 \geqslant 0$. Рассмотрим отрезок $Q = [0, 2+x_0]$. Несложно проверить, что $x_0 \in Q$ и $\varphi(x_n) \in Q$, $\quad |\varphi'(x)| < 1 \quad \forall x, x_n \in Q$. Следовательно, приближения $x_{n+1} = \varphi(x_n)$ сходятся к $x = 2$ — единственной неподвижной точке отображения $\varphi(x)$. $\rhd$

6.15. Доказать, что итерационный процесс $x_{n+1} = \cos x_n$ сходится для любого начального приближения $x_0 \in \mathbf{R}^1$.

$\lhd$ При любом $x_0 \in \mathbf{R}^1$ имеем $x_1 \in [-1, 1]$. Для этого отрезка выполнены условия сходимости метода простой итерации $x_{n+1} = \cos x_n$. $\rhd$

6.16. Исследовать сходимость метода простой итерации $x_{n+1} = x_n^2 - 2x_n + 2$ в зависимости от выбора начального приближения x_0.

$\lhd$ Уравнение $x = x^2 - 2x + 2$ имеет два корня $z_1 = 1$ и $z_2 = 2$. Пусть $\varphi(x) = x^2 - 2x + 2$, тогда при $x \in \left(\frac{1}{2}, \frac{3}{2}\right)$ имеем $\varphi(x) \in \left(\frac{1}{2}, \frac{3}{2}\right)$, $|\varphi'(x)| < 1$. Поэтому при $x_0 \in \left(\frac{1}{2}, \frac{3}{2}\right)$ приближения сходятся к z_1. Дальнейший анализ проводим аналогично решению 6.7, используя, в частности, эквивалентную запись итерационного метода в виде $x_{n+1} - x_n = (x_n - 2)(x_n - 1)$. $\rhd$

Ответ: метод сходится к $z_1 = 1$ при $x_0 \in (0, 2)$. Если $x_0 = 0$ или $x_0 = 2$, то метод сходится к $z_2 = 2$. Для остальных начальных приближений метод расходится.

6.17. Для уравнения $x = 2^{x-1}$, имеющего два корня $z_1 = 1$ и $z_2 = 2$, рассмотрим метод простой итерации. Исследовать его сходимость в зависимости от выбора начального приближения x_0.

$\lhd$ Пусть $\varphi(x) = 2^{x-1}$, тогда $\varphi'(x) < 1$ при $x \in (-\infty, x^*)$, где $x^* \in (1, 2)$ — решение уравнения $2^{x-1} \ln 2 = 1$. Функция $\varphi(x)$ отображает промежуток $(-\infty, x^*)$ в себя. Таким образом, при $x_0 \in (-\infty, x^*)$ метод сходится к $z = 1$. При $x_0 < 2$ приближения x_n монотонно убывают и при некотором n попадают в $(-\infty, x^*)$ (ср. с решением 6.14), поэтому сходятся к $z = 1$. При $x_0 > 2$ приближения стремятся к ∞. $\rhd$

6.18. Доказать, что метод простой итерации для решения уравнения $x = \varphi(x)$ сходится при любом начальном приближении:

1) $\varphi(x) = \alpha \sin^2 x + \beta \cos^2 x + \gamma$, где $|\alpha - \beta| < 1$;

2) $\varphi(x) = a\,\mathrm{e}^{-b\,x^2} + c$, где $b \geqslant 0$, $2a^2 b < \mathrm{e}$.

Указание. 2) Найти максимальное значение $|\varphi'(x)|$, $\varphi(x) = a\,\mathrm{e}^{-b\,x^2} + c$, и убедиться, что оно при указанных условиях меньше 1.

6.19. Уравнение $x + \ln x = 0$, имеющее корень $z \approx 0.6$, предлагается решать одним из следующих методов простой итерации: 1) $x_{n+1} = -\ln x_n$; 2) $x_{n+1} = \mathrm{e}^{-x_n}$; 3) $x_{n+1} = \dfrac{x_n + \mathrm{e}^{-x_n}}{2}$; 4) $x_{n+1} = \dfrac{3x_n + 5\mathrm{e}^{-x_n}}{8}$. Исследовать сходимость этих методов и сравнить их скорости.

6.20. Пусть функция $\varphi'(x)$ непрерывна на отрезке $[z - \delta, z + \delta]$, где z — единственная неподвижная точка для $\varphi(x)$. Может ли метод простой итерации сходиться к z, если $|\varphi'(z)| = 1$? Может ли он расходиться в этом случае?

Указание. Рассмотреть два примера. 1) $\varphi(x) = \sin x$, $z = 0$, $|\varphi'(z)| = 1$, метод сходится с любого начального приближения. 2) $\varphi(x) = x^2 + x$, $z = 0$, $|\varphi'(z)| = 1$, метод расходится, если $x_0 > 0$.

6.21. Показать, что для всякого a существует единственное решение $z(a, \varepsilon)$ уравнения $x + \varepsilon \sin x + a = 0$ при $|\varepsilon| \leqslant 1$.

$\lhd$ Воспользовавшись заменой $y = x + a$, преобразуем уравнение к виду $y + \varepsilon \sin(y - a) = 0$ и обозначим $\varphi(y) = -\varepsilon \sin(y - a)$. Функция $\varphi(y)$ отображает множество $\mathbf{R}^1$ в отрезок $Q_\varepsilon = [-|\varepsilon|, |\varepsilon|]$, поэтому можно считать, что $y \in Q_\varepsilon$. При $|\varepsilon| < 1$ имеем $|\varphi'(y)| \leqslant |\varepsilon| < 1$, поэтому $\varphi(y)$ является сжимающим отображением на отрезке Q_ε, следовательно, имеет единственную неподвижную точку $y^* = z(a, \varepsilon)$.

Пусть теперь $|\varepsilon| = 1$. Тогда для произвольных y_1, y_2 имеем

$$\varphi(y_2) - \varphi(y_1) = \int\limits_{y_1}^{y_2} \varphi'(t)dt = -\varepsilon \int\limits_{y_1}^{y_2} \cos(t - a)dt.$$

Так как $y_1, y_2 \in Q_\varepsilon$, то $\cos(t - a)$ может обратиться в единицу лишь в конечном числе точек, поэтому

$$\left| \int\limits_{y_1}^{y_2} \cos(t - a)dt \right| \leqslant q|y_2 - y_1|, \quad q < 1.$$

Таким образом, $|\varphi(y_2) - \varphi(y_1)| \leqslant q|y_2 - y_1|$, т. е. $\varphi(y)$ задает сжимающее отображение, поэтому имеет единственную неподвижную точку. $\rhd$

6.22. Найти область сходимости метода простой итерации для следующих уравнений: 1)$x = \mathrm{e}^{2x} - 1$; 2)$x = \frac{1}{2} - \ln x$; 3)$x = \operatorname{tg} x$.

6.23. Записать расчетные формулы метода парабол и найти корни уравнения $2x + \lg x = -0,5$ с точностью 10^{-2}.

6.24. Пусть z — p-кратный корень уравнения $f(x) = 0$. Оценить скорость сходимости метода секущих в его окрестности.

6.25. Пусть отображение $\varphi : \mathbf{R}^n \to \mathbf{R}^n$ имеет единственную неподвижную точку $\mathbf{z} = \varphi(\mathbf{z})$ и непрерывно дифференцируемо в некоторой ее окрестности.

1) Доказать, что если все собственные значения его якобиана $\varphi'(\mathbf{x})$ в точке $\mathbf{z}$ по модулю больше 1, то метод простой итерации не сходится.

2) Известно, что хотя бы одно собственное значение якобиана $\varphi'(\mathbf{z})$ по модулю больше 1. Можно ли утверждать, что для всех приближений $\mathbf{x}_0$, достаточно близких к $\mathbf{z}$, верна оценка $\|\mathbf{z} - \varphi(\mathbf{x}_0)\| < \|\mathbf{z} - \mathbf{x}_0\|$?

$\lhd$ 1) Введем обозначение: $A = \varphi'(\mathbf{z})$. Собственные значения матрицы A больше единицы по модулю, поэтому существует A^{-1} со спектральным радиусом $\rho(A^{-1}) < 1$. Известно (см. 5.41), что для любого $\varepsilon > 0$ найдется норма $||\cdot||_*$ такая, что $||A^{-1}||_* \leqslant \rho(A^{-1}) + \varepsilon$, следовательно, $||A^{-1}||_* = q < 1$ при достаточно малом $\varepsilon > 0$. Далее, если $||B||_* = \alpha < \frac{1}{q}$, то существует $(A + B)^{-1}$ и $||(A + B)^{-1}||_* < \frac{q}{1 - \alpha q}$. При достаточно малом α получаем $||(A + B)^{-1}||_* = q_1 < 1$.

Пусть теперь $U(\mathbf{z})$ — такая окрестность $\mathbf{z}$, что $||\varphi'(\mathbf{x}) - \varphi'(\mathbf{z})||_* < \alpha$ для $\mathbf{x} \in U(\mathbf{z})$. Допустим, что начальное приближение $\mathbf{x}_0 \in U(\mathbf{z})$, метод сходится и все $\mathbf{x}_n \in U(\mathbf{z})$. Тогда для погрешности $\mathbf{e}_n = \mathbf{x}_n - \mathbf{z}$ имеем $\mathbf{e}_{n+1} = \varphi'(\mathbf{y})\mathbf{e}_n$, $\mathbf{y} \in U(\mathbf{z})$, поэтому для любого n справедливо равенство

$$\mathbf{e}_n = (A + (\varphi'(\mathbf{y}) - \varphi'(\mathbf{z})))^{-1}\mathbf{e}_{n+1},$$

откуда следует, что

$$||\mathbf{e}_n||_* \leqslant q_1||\mathbf{e}_{n+1}||_*.$$

Поэтому $||\mathbf{e}_0||_* \leqslant q_1^n||\mathbf{e}_n||_*$, т. е. в пределе получаем $||\mathbf{e}_0||_* = 0$, что противоречит произвольности выбора начального приближения из окрестности $U(\mathbf{z})$.

2) Приведенная в условии оценка неверна, что следует из разложения функции $\mathbf{z}$ в ряд Тейлора в точке $\mathbf{x}_0 = \mathbf{z} + \mathbf{e}_0$:

$$\mathbf{z} - \varphi(\mathbf{x}_0) = \varphi(\mathbf{z}) - \varphi(\mathbf{x}_0) = \varphi'(\mathbf{z})\mathbf{e}_0 + o(\|\mathbf{e}_0\|),$$

где $\mathbf{e}_0$ имеет достаточно малую норму и пропорционален собственному вектору, отвечающему собственному значению, которое по модулю больше единицы. $\rhd$

6.2. Метод Ньютона. Итерации высшего порядка

Метод Ньютона. В случае одного уравнения формула *метода Ньютона* имеет вид

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\,.$$

Метод состоит в замене дуги кривой $y = f(x)$ касательной к ней в процессе каждой итерации. Это видно из уравнения касательной, проведенной в точке $(x_n, f(x_n))$:

$$y - f(x_n) = f'(x_n)(x - x_n),$$

из которого следует формула итерационного процесса, если положить $y = 0$ и $x = x_{n+1}$.

Метод Ньютона соответствует методу простой итерации $\frac{x_{n+1} - x_n}{\tau_n} + f(x_n) = 0$ с оптимальным, в некотором смысле, переменным параметром τ_n. Действительно, пусть z — изолированный простой (т. е. $f'(z) \neq 0$) корень, пусть также z и все x_n принадлежат некоторому отрезку $[a, b]$. Тогда

$$z - x_{n+1} = z - x_n + \tau_n f(x_n) - \tau_n f(z) = (1 - \tau_n f'(\xi_n))(z - x_n),$$

следовательно, при $\tau_n = \frac{1}{f'(\xi_n)}$ метод сходится за одну итерацию. Точка ξ_n неизвестна, поэтому на текущем шаге выбираем $\tau_n = \frac{1}{f'(x_n)}$, при этом верна оценка

$$|z - x_{n+1}| \leqslant \max_{\xi \in [a,b]} |1 - \tau_n f'(\xi)|\,|z - x_n|.$$

Рассмотрим случай системы m нелинейных уравнений

$$\mathbf{F}(\mathbf{x}) = 0\,,$$

где $\mathbf{x} = (x^1, \ldots, x^m)^T$, $\mathbf{F} = (f_1, \ldots, f_m)^T$. Будем предполагать отображение $\mathbf{F} : \mathbf{R}^m \to \mathbf{R}^m$ непрерывно дифференцируемым в некоторой окрестности решения $\mathbf{z}$, так что

$$\mathbf{F}'(\mathbf{x}) = \left[\frac{\partial f_i}{\partial x^j}\right], \quad 1 \leqslant i, j \leqslant m\,.$$

В предположении обратимости этого оператора метод Ньютона можно записать в виде

$$\mathbf{x}_{n+1} = \mathbf{x}_n - (\mathbf{F}'(\mathbf{x}_n))^{-1}\,\mathbf{F}(\mathbf{x}_n).$$

Введем обозначение: $\Omega_a = \{\mathbf{x} : \|\mathbf{x} - \mathbf{z}\| < a\}$, где $\|\cdot\|$ — норма в $\mathbf{R}^m$. Пусть при некоторых $a, a_1, a_2 : 0 < a,\ 0 < a_1, a_2 < \infty$, выполнены следующие условия:

1) $\|(\mathbf{F}'(\mathbf{x}))^{-1}\mathbf{y}\| \leqslant a_1\|\mathbf{y}\|$ при $\mathbf{x} \in \Omega_a$ и $\forall\,\mathbf{y}$;

2) $\|\mathbf{F}(\mathbf{u}_1) - \mathbf{F}(\mathbf{u}_2) - \mathbf{F}'(\mathbf{u}_2)(\mathbf{u}_1 - \mathbf{u}_2)\| \leqslant a_2\|\mathbf{u}_1 - \mathbf{u}_2\|^2$ при $\mathbf{u}_1, \mathbf{u}_2 \in \Omega_a$.

Обозначим также $c = a_1 a_2$, $b = \min(a, c^{-1})$.

Теорема. *При условиях* 1, 2 *и* $\mathbf{x}_0 \in \Omega_b$ *метод Ньютона сходится с оценкой погрешности*

$$\|\mathbf{x}_n - \mathbf{z}\| \leqslant c^{-1} \left(c\|\mathbf{x}_0 - \mathbf{z}\|\right)^{2^n},$$

т. е. квадратично.

Условия теоремы гарантируют, что корень $\mathbf{z}$ простой. В случае двукратного корня ($p = 2$) метод Ньютона сходится линейно; скорость сходимости замедляется при повышении кратности.

Интерполяционные методы построения итераций высшего порядка. Пусть $x_n, ..., x_{n-m+1}$ — набор из m приближений к корню z функции $f(x)$. Тогда в качестве очередного приближения x_{n+1} целесообразно выбрать ближайший к x_n нуль интерполяционного многочлена $L_m(x)$, построенного по узлам $x_n, ..., x_{n-m+1}$. Это требует нахождения корней многочлена $L_m(x)$. Как следствие, широкое применение имеют только алгоритмы при $m = 2, 3$, т. е. метод секущих и метод парабол.

Чтобы избежать проблем, связанных с решением алгебраического уравнения $L_m(x) = 0$, естественно интерполировать обратную к $y = f(x)$ функцию $x = F(y)$ по узлам $y_{n-i} = f(x_{n-i})$, $i = 0, ..., m-1$, и в качестве очередного приближения взять значение полученного интерполяционного многочлена в нуле. Линейная обратная интерполяция ($m = 2$) соответствует методу секущих, но уже при $m = 3$ прямая и обратная интерполяция приводят к различным алгоритмам.

Метод Чебышёва. Пусть z — простой корень уравнения $f(x) = 0$ и $F(y)$ — обратная к $f(x)$ функция. Тогда $x \equiv F(f(x))$ и $z = F(0)$. Разложим $F(0)$ в ряд Тейлора в окрестности некоторой точки y

$$F(0) = F(y) + \sum_{k=1}^{m} F^{(k)}(y)\,\frac{(-y)^k}{k!} + \dots .$$

Приблизим значение $F(0)$ значением частичной суммы в точке $y = f(x)$

$$z = F(0) \approx \varphi_m(x) = x + \sum_{k=1}^{m} (-1)^k F^{(k)}(f(x))\,\frac{(f(x))^k}{k!},$$

что соответствует замене функции F многочленом φ_m, производные которого совпадают с соответствующими производными F в точке $y = f(x)$. Итерационный метод вида $x_{n+1} = \varphi_m(x_n)$ имеет порядок сходимости $m+1$.

δ^2-процесс Эйткена. Вычислим по имеющемуся приближению x_n значения $x_{n+1} = \varphi(x_n)$ и $x_{n+2} = \varphi(x_{n+1})$. Так как в малой окрестности простого корня z имеются представления

$$x_{n+1} - z \approx \varphi'(z)(x_n - z), \quad x_{n+2} - z \approx \varphi'(z)(x_{n+1} - z),$$

то из данных соотношений получаем

$$\varphi'(z) \approx \frac{x_{n+2} - x_{n+1}}{x_{n+1} - x_n}\,, \quad z \approx \frac{x_{n+2} - \varphi'(z)x_{n+1}}{1 - \varphi'(z)} \approx \frac{x_{n+2}x_n - x_{n+1}^2}{x_{n+2} - 2x_{n+1} + x_n}\,.$$

Таким образом, за следующее после x_n приближение разумно принять

$$x_{n+1} = \frac{x_n\varphi(\varphi(x_n)) - \varphi(x_n)\varphi(x_n)}{\varphi(\varphi(x_n)) - 2\varphi(x_n) + x_n} = \varphi(\varphi(x_n)) - \frac{(\varphi(\varphi(x_n)) - \varphi(x_n))^2}{\varphi(\varphi(x_n)) - 2\varphi(x_n) + x_n}\,.$$

Известно, что если процесс $x_{n+1} = \varphi(x_n)$ имел линейную скорость сходимости, то данная модификация имеет скорость сходимости более высокого порядка, но возможно, только сверхлинейную. Применение рассмотренной модификации, например, к квадратично сходящейся последовательности формально не приводит к повышению порядка сходимости. Данное преобразование является частным случаем (при $\varphi_1 = \varphi_2 = \varphi$) метода Стеффенсона—Хаусхолдера—Островского построения итерационной функции φ_3 более высокого порядка по известным φ_1 и φ_2:

$$\varphi_3(x) = \frac{x\varphi_1(\varphi_2(x)) - \varphi_1(x)\varphi_2(x)}{x - \varphi_1(x) - \varphi_2(x) + \varphi_1(\varphi_2(x))}\,.$$

6.26. Построить итерационный метод Ньютона для вычисления $\sqrt[p]{a}$, $a > 0$, где p — положительное вещественное число.

$\lhd$ Значение $\sqrt[p]{a}$ является корнем уравнения

$$f(x) \equiv x^p - a = 0.$$

Для этого уравнения метод Ньютона имеет вид

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^p - a}{px_n^{p-1}} = \frac{p-1}{p}\,x_n + \frac{a}{px_n^{p-1}}\,.$$

Для $p = 2$ получаем $x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right)$. $\rhd$

6.27. Пусть уравнение $f(x) = 0$ имеет на отрезке $[a,\, b]$ простой корень, причем $f(x)$ — трижды непрерывно дифференцируемая функция. Показать, что при этих условиях метод Ньютона имеет квадратичную скорость сходимости.

$\lhd$ Метод Ньютона имеет вид $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$. Обозначим через z искомый корень. Тогда z — корень уравнения $x = \varphi(x)$, $\varphi(x) = x - \frac{f(x)}{f'(x)}$. Таким образом, можно рассматривать метод Ньютона как частный случай метода простой итерации, для которого

$$\varphi'(x) = \frac{f(x)\,f''(x)}{(f'(x))^2}\,, \quad \text{следовательно,} \quad \varphi'(z) = 0.$$

Согласно 6.1, найдется такая окрестность корня Q_δ, что $\varphi(Q_\delta) \subset Q_\delta$. Оценим скорость сходимости метода Ньютона, используя разложение в ряд Тейлора в окрестности точки z

$$x_{n+1} - z = \varphi(x_n) - \varphi(z) = \frac{1}{2}\,(x_n - z)^2\,\varphi''(\xi),\ \xi \in [x_n,\, z].$$

Итак, вблизи корня метод Ньютона имеет квадратичную скорость сходимости. $\triangleright$

6.28. Пусть уравнение $f(x) = 0$ имеет на отрезке $[a, b]$ корень z кратности p, причем $f(x)$ — дважды непрерывно дифференцируемая функция.

Показать, что при этих условиях метод Ньютона сходится со скоростью геометрической прогрессии со знаменателем $\dfrac{p-1}{p}$.

$\triangleleft$ Поступая так же, как и в случае простого корня 6.27, получим $x_{n+1} - z = (x_n - z)\,\varphi'(z) + 0,5(x_n - z)^2\,\varphi''(\xi)$, где $\xi \in [x_n,\, z]$. Однако в случае $p > 1$ в выражении

$$\varphi'(x) = \frac{f(x)\,f''(x)}{\left(f'(x)\right)^2}$$

при $x = z$ содержится неопределенность «нуль на нуль», так как z — одновременно корень уравнения $f'(x) = 0$. Оценим $\varphi'(x)$.

Функция $f(x)$ в окрестности корня z кратности p ведет себя как $a\,(x-z)^p + o(|x-z|^p)$, где a — ненулевая константа. Тогда в малой окрестности корня

$$\varphi'(x) = \frac{f(x)\,f''(x)}{\left(f'(x)\right)^2} = \frac{a\,(x-z)^p\,ap(p-1)\,(x-z)^{p-2}}{a^2p^2\,(x-z)^{2p-2}} + o(1),$$

$$\varphi'(z) = \frac{p-1}{p} < 1.$$

Отсюда следует, что чем выше кратность корня, тем медленнее сходимость. $\triangleright$

6.29. Пусть уравнение $f(x) = 0$ имеет на отрезке $[a, b]$ корень z кратности p, причем $f(x)$ — дважды непрерывно дифференцируемая функция. Построить модификацию метода Ньютона, имеющую квадратичную скорость сходимости.

$\triangleleft$ Требуемую модификацию будем искать в виде

$$x_{n+1} = x_n - \alpha\,\frac{f(x_n)}{f'(x_n)}$$

и подберем параметр α так, чтобы имела место квадратичная сходимость. Рассмотрим данную модификацию как специальный случай метода простой итерации $x_{n+1} = \varphi(x)$, для которого выполнено $z = \varphi(z)$, причем

вблизи корня

$$\varphi'(x) = 1 - \alpha + \alpha\,\frac{f(x)f''(x)}{\bigl(f'(x)\bigr)^2} = 1 - \alpha + \alpha\,\frac{p-1}{p} + o(1),$$

$$\varphi'(z) = \frac{p-\alpha}{p}\,.$$

Для обеспечения квадратичной сходимости параметр α надо подобрать таким, чтобы $\varphi'(z) = 0$, что и выполняется при $\alpha = p$. $\triangleright$

6.30. Построить метод Ньютона для вычисления значения a^{-1} так, чтобы расчетные формулы не содержали операций деления. Определить область сходимости метода при $a > 0$.

$\triangleleft$ Искомое число является корнем уравнения $\frac{1}{ax} - 1 = 0$. Для этого уравнения метод Ньютона имеет вид: $x_{n+1} = 2\,x_n - a\,x_n^2$, или $x_{n+1} = x_n(2 - a\,x_n)$.

Если $x_0 = 0$ или $x_0 = \frac{2}{a}$, то сходимость к корню не имеет места, так как все x_n равны нулю. Если $x_0 < 0$, то сходимости также не будет, поскольку все x_n останутся отрицательными. Если взять $x_0 > \frac{2}{a}$, то также все $x_n < 0$.

Из вида итерационного процесса следует, если $x_n \in \left(0, \frac{1}{a}\right]$, то $x_{n+1} \in \left(0, \frac{1}{a}\right]$, если же $x_n \in \left[\frac{1}{a}, \frac{2}{a}\right)$, то $x_{n+1} \in \left[0, \frac{1}{a}\right)$. Пусть $x_n \in \left(0, \frac{1}{a}\right]$. Тогда из равенства $x_{n+1} - x_n = x_n(1 - ax_n)$ получаем, что $x_{n+1} > x_n$, а из условия $x_{n+1} = x_n(2 - ax_n)$, что $x_{n+1} \leqslant \frac{1}{a}$. Так как итерационный процесс имеет две неподвижные точки 0 и $\frac{1}{a}$, то приближения сходятся к $\frac{1}{a}$.

Таким образом, сходимость к корню имеет место, если начальное приближение берется из интервала $\left(0, \frac{2}{a}\right)$. $\triangleright$

6.31. Пусть уравнение $f(x) = 0$ имеет на отрезке $[a, b]$ корень z неизвестной кратности $p > 1$, причем $f(x)$ — трижды непрерывно дифференцируемая функция. Построить модификацию метода Ньютона с квадратичной скоростью сходимости и предложить способ численной оценки величины кратности корня.

$\triangleleft$ Для уравнения $g(x) \equiv \frac{f(x)}{f'(x)} = 0$ корень z — простой, следовательно, для уравнения $g(x) = 0$ метод Ньютона выглядит так:

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} = x_n - \frac{f(x_n)f'(x_n)}{\bigl(f'(x_n)\bigr)^2 - f(x_n)f''(x_n)}$$

и имеет квадратичный порядок сходимости.

В окрестности z функция $f(x) \approx a\,(x-z)^p$, поэтому

$$g(x) = \frac{f(x)}{f'(x)} \approx \frac{a\,(x-z)^p}{ap\,(x-z)^{p-1}} = \frac{1}{p}\,(x-z).$$

Из двух соседних итераций для x_1 и x_2 имеем систему приближенных уравнений

$$g(x_1) \approx \frac{1}{p}\,(x_1-z), \quad g(x_2) \approx \frac{1}{p}\,(x_2-z).$$

Отсюда получаем оценку для кратности p корня z:

$$p \approx \frac{x_2-x_1}{g(x_2)-g(x_1)}\,.$$

Такой способ оценивания p можно применять на каждой итерации. $\triangleright$

6.32. Для решения уравнения $x^3 - x = 0$ применяют метод Ньютона. При каком начальном приближении он сходится и к какому корню?

О т в е т: обозначим области сходимости метода Ньютона

$$x_{n+1} = \varphi(x_n), \quad \varphi(x) = \frac{2x^3}{3x^2-1}$$

к корням $z = -1, 0, +1$ через X_-, X_0, X_+ соответственно. Кроме того, определим последовательности точек $\{x_n^{\pm}\}$ для $n \geqslant 0$ следующими условиями:

$$\varphi(x_{n+1}^{\pm}) = x_n^{\pm}, \quad x_0^{\pm} = \pm\frac{1}{\sqrt{3}},$$

для элементов которых справедливы неравенства

$$-\frac{1}{\sqrt{3}} = x_0^- < x_1^+ < x_2^- < \cdots < -\frac{1}{\sqrt{5}} < 0 < \frac{1}{\sqrt{5}} < \cdots < x_2^+ < x_1^- < x_0^+ = \frac{1}{\sqrt{3}}$$

и существуют пределы

$$\lim_{k\to\infty} x_{2k}^- = \lim_{k\to\infty} x_{2k-1}^+ = -\frac{1}{\sqrt{5}}, \qquad \lim_{k\to\infty} x_{2k}^+ = \lim_{k\to\infty} x_{2k-1}^- = \frac{1}{\sqrt{5}}.$$

Тогда

$$X_- = \left(-\infty, x_0^-\right) \cup \bigcup_{k=1}^{\infty} \left[\left(x_{2k-1}^+, x_{2k}^-\right) \cup \left(x_{2k-1}^-, x_{2(k-1)}^+\right)\right],$$

$$X_0 = \left(-\frac{1}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right),$$

$$X_+ = \left(x_0^+, \infty\right) \cup \bigcup_{k=1}^{\infty} \left[\left(x_{2(k-1)}^-, x_{2k-1}^+\right) \cup \left(x_{2k}^+, x_{2k-1}^-\right)\right].$$

Кроме того, если $x_0 = x_n^{\pm}$, $n \geqslant 0$, то метод не определен, а при $x_0 = \pm\frac{1}{\sqrt{5}}$ имеем $x_1 = \pm\frac{1}{\sqrt{5}}$, т. е. метод «зацикливается».

Таким образом, области сходимости к корням $z = \pm 1$ являются объединениями перемежающихся открытых интервалов, разделенных точками зацикливания метода.

6.33. Доказать, что если на отрезке $[a,b]$ функция $f'(x)$ не обращается в нуль, функция $f''(x)$ непрерывна и не меняет знака, кроме того, выполнены условия

$$f(a)f(b)<0\,,\quad \max\left[\left|\frac{f(a)}{f'(a)}\right|,\left|\frac{f(b)}{f'(b)}\right|\right]\leqslant b-a\,,$$

то метод Ньютона для решения уравнения $f(x)=0$ сходится при любом $x_0\in[a,b]$.

6.34. Указать область сходимости метода решения уравнения $x=\frac{1}{a}$, не содержащего операций деления:

$$x_{n+1}=(1+C)\,x_n-a\,C\,x_n^2\,,$$

в зависимости от параметра $C\neq 0$.

6.35. Рассматривается метод Ньютона вычисления $\sqrt{a}$ при $1\leqslant a\leqslant 4$, x_0 полагают равным значению многочлена наилучшего равномерного приближения для $\sqrt{a}$ на $[1,4]$: $x_0=Q_1^0(a)=\frac{17}{24}+\frac{a}{3}$. Доказать справедливость оценки $|x_4-\sqrt{a}|\leqslant 0,5\cdot 10^{-25}$.

6.36. Для нахождения $a^{1/3}$ используют итерационный процесс

$$x_{n+1}=A\,x_n+B\,\frac{a}{x_n^2}+C\,\frac{a^2}{x_n^5}\,.$$

Найти значения параметров A,B,C, обеспечивающие максимальный порядок сходимости.

6.37. Записать формулы метода Чебышёва для функции $f(x)=x^p-a$.

$\lhd$ Обратная к f функция имеет вид $F(y)=(a+y)^{1/p}$, а производные F определяются формулой

$$F^{(k)}(y)=x^{1-kp}\prod_{j=0}^{k-1}\left(\frac{1}{p}-j\right).$$

Таким образом,

$$\varphi_m(x)=x+x\sum_{k=1}^{m-1}\frac{1}{k!}\left(\frac{a-x^p}{px^p}\right)^k\prod_{j=0}^{k-1}(1-jp).$$

В частности, $\varphi_2(x)=\left(\frac{x}{p}\right)\left(p-1+\frac{a}{x^p}\right)$. При $p=2$ получаем формулу Ньютона—Херона $x_{n+1}=\varphi_2(x_n)$ для приближенного вычисления квадратных корней.

Если $p=-1$, то $\varphi_m(x)=x\sum\limits_{k=0}^{m-1}(1-ax)^k$. В этом случае итерационный процесс $x_{n+1}=\varphi_m(x_n)$ при $|1-ax|<1$ сходится к решению уравнения $x-\frac{1}{a}=0$. Данный метод позволяет находить значение $\frac{1}{a}$ с произвольной точностью, не используя операцию деления. $\rhd$

6.38. Показать, что метод вычисления $a^{1/p}$

$$x_{n+1} = \varphi(x_n),\ \varphi(x) = x\,\frac{(p-1)x^p + (p+1)a}{(p+1)x^p + (p-1)a}$$

имеет третий порядок.

6.39. Определить порядок сходимости метода

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{f''(x_n)(f(x_n))^2}{2(f'(x_n))^3}\,.$$

О т в е т: порядок сходимости $m = 3$.

6.40. Определить порядок сходимости модифицированного метода Ньютона $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_0)}$.

6.41. Определить порядок сходимости метода

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{f\left(x_n - (f'(x_n))^{-1} f(x_n)\right)}{f'(x_n)}\,.$$

О т в е т: порядок сходимости $m = 3$.

6.42. Для нахождения простого нуля z функции $f(x) \in C^{(4)}$ используют итерационный метод

$$x_{n+1} = 0,5\,(y_{n+1} + v_{n+1})\,,$$

где

$$y_{n+1} = x_n + \frac{f(x_n)}{f'(x_n)}\,,\quad v_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}\,,\quad g(x) = \frac{f(x)}{f'(x)}\,.$$

Доказать, что если метод сходится, то скорость сходимости — кубичная.

6.43. Для нахождения нуля z функции $f(x)$ используют итерационный метод

$$x_{n+1} = g(x_n)\,,\quad g(x) = x - \frac{(f(x))^2}{f(x + f(x)) - f(x)}\,.$$

Исследовать поведение функции $g(x)$ в окрестности корня z.

6.44. Записать расчетную формулу метода Ньютона для системы уравнений:

1) $\begin{cases} \sin(x+y) - 1,3x = 0,1, \\ x^2 + y^2 = 1; \end{cases}$ 2) $\begin{cases} x^{10} + y^{10} = 1024, \\ \mathrm{e}^x - \mathrm{e}^y = 1. \end{cases}$

6.45. Указать начальное приближение и оценить число итераций в методе Ньютона, требующихся для достижения точности 10^{-3} при решении системы уравнений

$$\begin{cases} x^3 - y^2 = 1, \\ xy^3 - y = 4. \end{cases}$$

6.46. Проверить, что $\mathbf{z} = (1,1,1)^T$ — одно из решений системы уравнений $\mathbf{F}(\mathbf{x}) = 0$, где $\mathbf{F} : \mathbf{R}^3 \to \mathbf{R}^3$ имеет вид

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 x_2^3 + x_2 x_3 - x_1^4 - 1 \\ x_2 + x_2^2 + x_3 - 3 \\ x_2 x_3 - 1 \end{bmatrix}.$$

Сходится ли метод Ньютона к $\mathbf{z}$ при достаточно близких начальных приближениях?

6.47. Для решения нелинейной краевой задачи

$$y'' = f(x,y) \quad \text{при} \quad x \in (0,X), \quad y(0) = a, \quad y(X) = b$$

рассматривается система нелинейных алгебраических уравнений с параметром $h = \frac{X}{N}$:

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = f(x_k, y_k),\ k = 1,2,\ldots,N-1, \quad y_0 = a,\ \ y_N = b.$$

Здесь y_k — приближения к значениям $y(kh)$. Записать расчетные формулы метода Ньютона для решения приведенной системы. Указать способ их реализации: 1) $f(x,y) = x^2 + y^3$; 2) $f(x,y) = y^2 \exp(x)$; 3) $f(x,y) = \cos x \sin y$.

Глава 7
Элементы теории разностных схем

На первых этапах практического решения обыкновенных дифференциальных уравнений и уравнений в частных производных применялись методы, в которых приближенное решение строилось в виде некоторой аналитической формулы. В настоящее время наибольшее распространение получили сеточные, вариационно- и проекционно-разностные методы, позволяющие получать либо приближенные значения решения на некотором множестве точек, либо приближенное разложение решения по некоторой системе базисных функций.

В главе излагаются базовые понятия общей теории численного решения дифференциальных уравнений. Рассматриваются различные способы перехода от дифференциальных задач к разностным и некоторые точные алгоритмы решения полученных уравнений.

7.1. Основные определения

Постановки задач. Пусть в области D с границей Γ задана дифференциальная задача

$$L\,u = f \quad \text{в} \quad D \tag{7.1}$$

с граничным условием

$$l\,u = \varphi \quad \text{на} \quad \Gamma\,. \tag{7.2}$$

Здесь L и l — дифференциальные операторы; f и φ — заданные элементы, u — искомый элемент некоторых линейных нормированных пространств F, Φ и U соответственно.

Если одной из переменных является время t, то наиболее часто рассматривают области вида

$$D(t, \mathbf{x}) = d(\mathbf{x}) \times [t_0, T]\,,$$

где t — время, $\mathbf{x} = (x_1, \ldots, x_n)$ — совокупность пространственных координат. Это означает, что решение ищется в пространственной области $d(\mathbf{x})$ на отрезке времени $[t_0, T]$. В этом случае условия, заданные при $t = t_0$, называют *начальными*, а условия, заданные на границе $\Gamma(\mathbf{x})$ области $d(\mathbf{x})$, — *граничными* или *краевыми*.

Задачу только с начальными условиями называют *задачей Коши*. Задачу с начальными и граничными условиями называют *смешанной краевой задачей*.

Для решения дифференциальных задач часто используют разностный метод.

Разностный метод. Для его применения определяют некоторую *сетку* — конечное множество точек (узлов) $\overline{D}_h = D_h \bigcup \Gamma_h$, принадлежащее области $\overline{D} = D \bigcup \Gamma$. Как правило, $\Gamma_h \subset \Gamma$. Будем рассматривать только

сетки, узлами которых являются все точки пересечения заданных наборов параллельных прямых (плоскостей), причем по каждой переменной выбирается свой, как правило, постоянный шаг. Сетки по времени и пространству обычно определяют независимо. Сеточный параметр h является, в общем случае, вектором, компоненты которого состоят из шагов сетки по каждой переменной. Для изучения свойств разностных схем вводится понятие величины шага сетки, в качестве которого принимается какая-либо *сеточная норма* вектора h, например,

$$||h||_\infty = \max_{1\leqslant i\leqslant n} h_i \quad \text{или} \quad ||h||_2 = \left(\sum_{i=1}^{n} h_i^2\right)^{1/2},$$

где n — число независимых переменных в дифференциальной задаче. Чтобы избежать новых и ненужных обозначений, в приводимых ниже оценках под h понимается величина шага сетки.

Если $X \subset Y$ и функция v определена на множестве Y, то ее *следом* на множестве X называют функцию, определенную на X и совпадающую там с v. Если функция v определена на некотором множестве Y, содержащем Y_h, то ее след на Y_h будем обозначать $(v)_h$. Часто пространства F_h, Φ_h и U_h определяют как пространства следов функций из F, Φ и U на D_h, Γ_h и $\overline{D}_h$ соответственно. При этом задаются *согласованные* нормы пространств, т. е. для достаточно гладких функций $v \in U$ выполняется соотношение

$$\lim_{h\to 0} \| (v)_h \|_{U_h} = \| v \|_U .$$

Все производные, входящие в уравнение и краевые условия, заменяют *разностными аппроксимациями*. При записи этих аппроксимаций в каждом внутреннем узле сетки берут одно и то же количество соседних узлов, образующих строго определенную конфигурацию, называемую *шаблоном*. В результате дифференциальные операторы L и l заменяются разностными L_h и l_h.

Для нахождения приближенного решения задачи (7.1), (7.2) определим *разностную схему* — семейство разностных задач, зависящих от параметра h:

$$L_h u_h = f_h \quad \text{в} \quad D_h , \tag{7.3}$$

$$l_h u_h = \varphi_h \quad \text{на} \quad \Gamma_h . \tag{7.4}$$

Решение разностной схемы u_h, называемое *разностным*, принимается в качестве приближенного решения дифференциальной задачи.

Аппроксимация. Говорят, что разностная схема (7.3), (7.4) *аппроксимирует* с порядком аппроксимации $p = \min(p_1, p_2)$ дифференциальную задачу (7.1), (7.2), если для любых достаточно гладких функций u, f, φ из соответствующих пространств существуют такие постоянные h_0, c_1, p_1, c_2 и p_2, что для всех $h \leqslant h_0$ выполняются неравенства

$$\|L_h(u)_h - (Lu)_h\|_{F_h} + \|(f)_h - f_h\|_{F_h} \leqslant c_1 h^{p_1},$$

$$\|l_h(u)_h - (lu)_h\|_{\Phi_h} + \|(\varphi)_h - \varphi_h\|_{\Phi_h} \leqslant c_2 h^{p_2},$$

причем c_1, p_1, c_2 и p_2 не зависят от h.

Выражения, стоящие под знаком норм, называют *погрешностями аппроксимации.*

Оператор L_h из (7.3) *локально аппроксимирует* в точке x_i дифференциальный оператор L из (7.1), если для достаточно гладкой функции $u \in U$ существуют такие положительные постоянные h_0, c и p, не зависящие от h, что при всех $h \leqslant h_0$ справедливо неравенство

$$\left| \left(L_h(u)_h - (Lu)_h \right)\big|_{x=x_i} \right| \leqslant c\, h^p .$$

Число p при этом называют *порядком* аппроксимации. Аналогично определяют порядок локальной аппроксимации оператора l_h.

Также используется понятие аппроксимации на решении, позволяющее строить схемы более высокого порядка точности на фиксированном шаблоне. Говорят, что разностная схема (7.3), (7.4) *аппроксимирует на решении* u с порядком аппроксимации $p = \min(p_1, p_2)$ дифференциальную задачу (7.1), (7.2), если существуют такие постоянные h_0, c_1, p_1, c_2 и p_2, что для всех $h \leqslant h_0$ выполняются неравенства

$$\|L_h(u)_h - f_h\|_{F_h} \leqslant c_1 h^{p_1}, \quad \|l_h(u)_h - \varphi_h\|_{\Phi_h} \leqslant c_2 h^{p_2},$$

причем c_1, p_1, c_2 и p_2 не зависят от h. Предполагается, что при этом выполнены условия нормировки

$$\lim_{h\to 0} \| f_h \|_{F_h} = \| f \|_F , \quad \lim_{h\to 0} \| \varphi_h \|_{\Phi_h} = \| \varphi \|_\Phi .$$

Порядки аппроксимаций обычно оценивают с помощью разложения в ряды Тейлора. Порядок аппроксимации разностной схемы может быть разным по разным переменным. Если погрешность аппроксимации стремится к нулю при любом законе стремления шагов по различным переменным к нулю, то такую аппроксимацию называют *безусловной.* Если же погрешность аппроксимации стремится к нулю при одних законах убывания шагов и не стремится к нулю при других, то аппроксимацию называют *условной.*

Устойчивость. Разностная схема (7.3), (7.4) *устойчива*, если решение системы разностных уравнений существует, единственно и непрерывно зависит от входных данных f_h, φ_h, причем эта зависимость равномерна относительно величины шага сетки. Это означает, что для любого $\varepsilon > 0$ существуют не зависящие от h величины h_0 и $\delta = \delta(\varepsilon)$ такие, что для произвольных функций $u_h^{(i)}$, $i = 1, 2$, являющихся решениями (7.3), (7.4), из неравенств $h \leqslant h_0$, $\left\| f_h^{(1)} - f_h^{(2)} \right\|_{F_h} \leqslant \delta$, $\left\| \varphi_h^{(1)} - \varphi_h^{(2)} \right\|_{\Phi_h} \leqslant \delta$ следует, что $\left\| u_h^{(1)} - u_h^{(2)} \right\|_{U_h} \leqslant \varepsilon$.

Линейная схема устойчива, если

$$\left\| u_h^{(1)} - u_h^{(2)} \right\|_{U_h} \leqslant c_1 \left\| f_h^{(1)} - f_h^{(2)} \right\|_{F_h} + c_2 \left\| \varphi_h^{(1)} - \varphi_h^{(2)} \right\|_{\Phi_h} ,$$

где c_1 и c_2 — постоянные, не зависящие от $h \leqslant h_0$. Это означает, что ε и δ здесь связаны линейно.

Устойчивость называют *безусловной*, если указанные неравенства выполняются при произвольном соотношении шагов по различным переменным. Если же для выполнения неравенств шаги должны удовлетворять дополнительным соотношениям, то устойчивость называют *условной*.

Непрерывную зависимость по f_h (равномерную относительно h) называют устойчивостью *по правой части*, а непрерывную зависимость по φ_h — устойчивостью *по граничным условиям*. Если рассматривается смешанная краевая задача, то устойчивость по граничному условию при $t = t_0$ называют устойчивостью *по начальным данным*.

Сходимость. Решение u_h разностной схемы (7.3), (7.4) *сходится* к решению u дифференциальной задачи (7.1), (7.2), если существуют такие постоянные h_0, c и p, что для всех $h \leqslant h_0$ выполнено неравенство

$$\|(u)_h - u_h\|_{U_h} \leqslant c\,h^p,$$

причем c и p не зависят от h. Число p называют *порядком* сходимости разностной схемы; при этом говорят, что разностное решение u_h имеет порядок точности p.

Теорема Филиппова (о связи аппроксимации, устойчивости и сходимости). *Пусть выполнены следующие условия:*

1) *операторы L, l и L_h, l_h — линейные;*

2) *решение u дифференциальной задачи* (7.1), (7.2) *существует и единственно;*

3) *разностная схема* (7.3), (7.4) *аппроксимирует дифференциальную задачу* (7.1), (7.2) *с порядком p;*

4) *разностная схема* (7.3), (7.4) *устойчива.*

Тогда решение разностной схемы u_h сходится к решению u дифференциальной задачи с порядком не ниже p.

$\lhd$ Операторы L и L_h линейные, поэтому

$$L_h(u_h - (u)_h) = L_h u_h - L_h(u)_h =$$
$$= f_h - L_h(u)_h \pm (Lu)_h = ((Lu)_h - L_h(u)_h) + (f_h - (f)_h)\,.$$

Отсюда имеем уравнение

$$L_h(u_h - (u)_h) = ((Lu)_h - L_h(u)_h) + (f_h - (f)_h)\,.$$

Аналогично для краевых условий находим

$$l_h(u_h - (u)_h) = ((lu)_h - l_h(u)_h) + (\varphi_h - (\varphi)_h)\,.$$

Решение разностной задачи устойчиво, поэтому по определению для линейных задач получаем

$$\|u_h - (u)_h\|_{U_h} \leqslant c_1(\|(Lu)_h - L_h(u)_h\|_{F_h} + \|f_h - (f)_h\|_{F_h}) +$$
$$+c_2(\|(lu)_h - l_h(u)_h\|_{\Phi_h} + \|\varphi_h - (\varphi)_h\|_{\Phi_h}) \leqslant ch^p\,.$$

Это неравенство означает сходимость с порядком p. Теорема доказана. $\rhd$

Если порядок аппроксимации на решении выше p, то для получения более точной оценки доказательство теоремы можно модифицировать. Для этого в первой системе равенств доказательства не следует добавлять $\pm(Lu)_h$, а применить сразу оценку устойчивости к величине $f_h - L_h(u)_h$ из определения аппроксимации на решении. Аналогичное следует проделать и для краевых условий.

Для многомерных задач порядок аппроксимации по разным переменным может быть неодинаковым, поэтому порядки сходимости по разным переменным также могут быть различными. Если аппроксимация и (или) устойчивость разностной схемы условные, то сходимость имеет место только при тех соотношениях между шагами сетки по разным переменным, при которых выполнены условия аппроксимации и (или) устойчивости. В классе задач с решениями конечной гладкости требование устойчивости является необходимым условием сходимости.

7.2. Методы построения разностных схем

Метод неопределенных коэффициентов. Пусть имеется некоторый шаблон (несколько расположенных группой узлов сетки) и требуется найти разностный оператор L_h, локально аппроксимирующий дифференциальный оператор L в узле x_i. В этом случае в выражении $(\,L_h(u)_h - (Lu)_h\,)|_{x=x_i}$ оператор L_h берут с неопределенными коэффициентами. Для нахождения искомых коэффициентов с помощью формулы Тейлора строят разложения в точке x_i для всех значений функции $u(x)$, входящих в выражение $L_h(u)_h$ и группируют множители при $u(x_i), u'(x_i), u''(x_i), \ldots$. Далее, последовательно обнуляя найденные множители, приходят к системе линейных алгебраических уравнений, решая которую находят коэффициенты разностной схемы. Порядок аппроксимации и главный член погрешности определяется после подстановки найденных коэффициентов в первый ненулевой множитель при соответствующей производной функции $u(x)$ в точке x_i.

Рассмотрим пример. Пусть для задачи

$$L\,u \equiv u'' = f(x)\,, \quad 0 < x < 1\,,\ u(0) = u(1) = 0$$

на равномерной сетке $\overline{D}_h = \{x_i = i\,h,\ i = 0, \ldots, N,\ N\,h = 1\}$ требуется построить схему методом неопределенных коэффициентов на трехточечном шаблоне.

Будем строить оператор L_h в виде

$$(L_h u_h)_i = \frac{a\,u_{i+1} + b\,u_i + c\,u_{i-1}}{h^2}\,.$$

Запишем разложения по формуле Тейлора для достаточно гладкой функции $u(x)$ в точке $x = x_i$:

$$u(x_i \pm h) = u(x_i) \pm h\,u'(x_i) + \frac{h^2}{2!}\,u''(x_i) \pm \frac{h^3}{3!}\,u'''(x_i) +$$

$$+ \frac{h^4}{4!}\,u^{(4)}(x_i) \pm \frac{h^5}{5!}\,u^{(5)}(x_i) + \frac{h^6}{6!}\,u^{(6)}(\xi_i^{\pm}).$$

Подставим полученные выражения в формулу для $L_h(u)_h$ и сгруппируем множители при одинаковых производных $u(x)$ (или, что то же самое, — степенях h)

$$L_h(u)_h\,|_{x=x_i} = \frac{1}{h^2}\Big[(a+b+c)u(x_i) + h(a-c)u'(x_i) + \\ + \frac{h^2}{2!}(a+c)u''(x_i) + \frac{h^3}{3!}(a-c)u^{(3)}(x_i) + \frac{h^4}{4!}(a+c)u^{(4)}(x_i) + \\ + \frac{h^5}{5!}(a-c)u^{(5)}(x_i) + \frac{h^6}{6!}\left(a\,u^{(6)}(\xi_i^+) + c\,u^{(6)}(\xi_i^-)\right)\Big].$$

По определению локальной аппроксимации

$$L_h(u)_h\,|_{x=x_i} = u''(x_i) + O(h^p), \quad p > 0,$$

откуда имеем систему уравнений

$$a+b+c=0, \quad a-c=0, \quad \frac{a+c}{2}=1,$$

решая которую, получим

$$L_h(u)_h\,|_{x=x_i} \equiv \frac{u(x_{i+1}) - 2\,u(x_i) + u(x_{i-1})}{h^2} = u''(x_i) + \frac{h^2}{12}\,u^{(4)}(x_i) + O(h^4),$$

т.е. L_h локально аппроксимирует оператор второй производной L в точке $x = x_i$ со вторым порядком.

Запишем разностный аналог рассматриваемой задачи:

$$\frac{u_{i+1} - 2\,u_i + u_{i-1}}{h^2} = f_i\,, \quad 0 < i < N\,, \quad u_0 = u_N = 0.$$

Отметим, что здесь u_i — приближение к решению $u(x_i)$.

Интегро-интерполяционный метод. В качестве примера опишем применение этого метода к построению разностной схемы на равномерной сетке $\overline{D}_h = \{x_i = i\,h,\ i = 0, \ldots, N;\ N\,h = 1\}$ для задачи

$$L\,u \equiv -u'' + p(x)\,u = f(x)\,, \quad 0 < x < 1\,, \quad 0 \leqslant p(x) \leqslant p_1\,, \\ u'(0) = \alpha_1 u(0) + \beta_1\,, \quad u(1) = 0\,.$$

Введем обозначение $\omega(x) = u'(x)$ и перепишем исходное уравнение в виде $\omega'(x) = p(x)\,u(x) - f(x)$. Проинтегрируем в пределах от $x_{i-1/2}$ до $x_{i+1/2}$ ($x_{i\pm1/2} = x_i \pm h/2$):

$$\omega(x_{i+1/2}) - \omega(x_{i-1/2}) = \int\limits_{x_{i-1/2}}^{x_{i+1/2}} [p(x)\,u(x) - f(x)]\,dx\,.$$

Полученное равенство служит основой для построения разностных схем. Заменим интеграл в правой части, например, по квадратурной формуле прямоугольников

$$\int\limits_a^b \varphi(x)\,dx = (b-a)\,\varphi\left(\frac{a+b}{2}\right) + O((b-a)^3).$$

Разделив обе части на h, получим:

$$\frac{\omega(x_{i+1/2}) - \omega(x_{i-1/2})}{h} = p(x_i)u(x_i) - f(x_i) + O(h^2).$$

Так как $u'(x) = \omega(x)$, то на отрезке $[x_i, x_{i+1}]$ имеем

$$u(x_{i+1}) - u(x_i) = h\,\omega(x_{i+1/2}) + O(h^3).$$

Аналогичное выражение

$$u(x_i) - u(x_{i-1}) = h\,\omega(x_{i-1/2}) + O(h^3)$$

справедливо на отрезке $[x_{i-1}, x_i]$. Поэтому дискретный аналог исходного уравнения принимает вид

$$-\,\frac{u_{i+1} - 2\,u_i + u_{i-1}}{h^2} + p_i u_i = f_i\,, \quad 0 < i < N\,,$$

где u_i обозначает приближение к точному решению $u(x_i)$, в то время как $p_i = p(x_i)$ и $f_i = f(x_i)$ — значения известных функций в узлах сетки.

Для аппроксимации краевого условия третьего рода проинтегрируем исходное уравнение от 0 до $\frac{h}{2}$:

$$u'\left(\frac{h}{2}\right) - u'(0) = \int\limits_0^{h/2} [p(x)\,u(x) - f(x)]\,dx\,.$$

Далее опять воспользуемся формулой прямоугольников для интеграла и заменим $u'(0)$ на $\alpha_1 u(0) + \beta_1$, а $u'\left(\frac{h}{2}\right)$ на $\frac{u(h) - u(0)}{h} + O(h^2)$. В результате получим

$$\frac{u(h) - u(0)}{h} - \alpha_1 u_0 - \beta_1 + O(h^2) = \frac{h}{2}\left[p\left(\frac{h}{4}\right)\,u\left(\frac{h}{4}\right) - f\left(\frac{h}{4}\right)\right] + O(h^3)\,.$$

Левая часть равенства содержит слагаемое $O(h^2)$, поэтому в его правой части значения функций в точке $x = \frac{h}{4}$ можно заменить их значениями в точке $x = 0$, сохранив тот же порядок аппроксимации $O(h^2)$. В результате получим $\frac{u_1 - u_0}{h} = \alpha_1 u_0 + \beta_1 + \frac{h}{2}\,(p(0)u_0 - f(0))\,.$

Окончательная разностная схема имеет вид

$$-\,\frac{u_{i+1} - 2\,u_i + u_{i-1}}{h^2} + p_i u_i = f_i,\ 0 < i < N,\ \frac{u_1 - u_0}{h} = \overline{\alpha}_1 u_0 + \overline{\beta}_1,\ u_N = 0,$$

где новые коэффициенты принимают значения $\overline{\alpha}_1 = \alpha_1 + \frac{h\,p(0)}{2}\,,\ \overline{\beta}_1 = \beta_1 - \frac{h\,f(0)}{2}\,.$

Интегральное тождество Марчука. Для задачи

$$L\,u \equiv -(k(x)\,u')' + p(x)\,u = f(x), \quad 0 < x < 1,\ u(0) = u(1) = 0,$$

у которой переменные коэффициенты удовлетворяют условиям $0 < k_0 \leqslant k(x) \leqslant k_1$, $0 \leqslant p(x) \leqslant p_1$, и $k(x), p(x), f(x)$ могут иметь конечное число разрывов первого рода, построение разностной схемы основывается на интегральном тождестве, которому удовлетворяет решение исходной

задачи

$$-\frac{u(x_{i+1})-u(x_i)}{\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}}+\frac{u(x_i)-u(x_{i-1})}{\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}}+\int\limits_{x_{i-1/2}}^{x_{i+1/2}}[p(x)\,u-f(x)]dx=$$

$$=-\frac{1}{\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}}\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}\int\limits_{x_{i+1/2}}^{x}[p(\xi)\,u(\xi)-f(\xi)]d\xi+$$

$$+\frac{1}{\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}}\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}\int\limits_{x_{i-1/2}}^{x}[p(\xi)\,u(\xi)-f(\xi)]d\xi.$$

Докажем это тождество. Введем обозначение $\omega(x)=k(x)\,u'(x)$ и перепишем исходное уравнение в виде

$$\omega'(x)=p(x)\,u(x)-f(x).$$

Проинтегрируем уравнение от $x_{i-1/2}$ до $x_{i+1/2}$ ($x_{i\pm1/2}=x_i\pm h/2$)

$$\omega(x_{i+1/2})-\omega(x_{i-1/2})=\int\limits_{x_{i-1/2}}^{x_{i+1/2}}[p(x)\,u(x)-f(x)]\,dx.$$

Для нахождения $\omega(x_{i\pm1/2})$ поступим следующим образом. Проинтегрируем уравнение для $\omega'(x)$ от $x_{i-1/2}$ до x

$$k(x)u'(x)=\omega(x_{i-1/2})+\int\limits_{x_{i-1/2}}^{x}[p(\xi)\,u(\xi)-f(\xi)]\,d\xi.$$

Разделим это выражение на $k(x)$ и проинтегрируем от x_{i-1} до x_i. В результате получим

$$u(x_i)-u(x_{i-1})=\omega(x_{i-1/2})\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}+\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}\int\limits_{x_{i-1/2}}^{x}[p(\xi)\,u(\xi)-f(\xi)]\,d\xi.$$

Отсюда находим явное выражение для

$$\omega(x_{i-1/2})=\frac{1}{\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}}\left\{u(x_i)-u(x_{i-1})-\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}\int\limits_{x_{i-1/2}}^{x}[p(\xi)\,u(\xi)-f(\xi)]\,d\xi\right\}.$$

Аналогичное выражение для $\omega(x_{i+1/2})$ найдем, заменив в полученной формуле индекс i на $i+1$. Теперь, используя $\omega(x_{i\pm1/2})$, приходим к искомому интегральному тождеству.

Рассмотрим следующий пример. Пусть коэффициенты в исходном уравнении имеют вид

$$k(x)=\begin{cases}1 & \text{при}\quad 0\leqslant x\leqslant\frac{1}{2},\\ 2 & \text{при}\quad \frac{1}{2}<x\leqslant 1,\end{cases}\qquad p(x)\equiv 0.$$

Запишем интегральное тождество Марчука:

$$-\frac{u(x_{i+1})-u(x_i)}{\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}}+\frac{u(x_i)-u(x_{i-1})}{\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}}-\int\limits_{x_{i-1/2}}^{x_{i+1/2}}f(x)dx=$$

$$=\frac{1}{\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}}\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}\int\limits_{x_{i+1/2}}^{x}f(\xi)d\xi-\frac{1}{\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}}\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}\int\limits_{x_{i-1/2}}^{x}f(\xi)d\xi.$$

Предположим (для удобства), что точка $x=\frac{1}{2}$ — узел сетки при любом h, т. е. $h=\frac{1}{N}$, $N=2K$. При этом $i=\frac{N}{2}$ — соответствующее значение индекса i. Вычислим величины

$$t_i=\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}=\begin{cases}h & \text{при} \quad 0\leqslant i<\frac{N}{2}\,,\\ \frac{h}{2} & \text{при} \quad \frac{N}{2}\leqslant i<N.\end{cases}$$

Заменим по формуле прямоугольников интеграл в левой части тождества

$$\int\limits_{x_{i-1/2}}^{x_{i+1/2}}f(x)dx\approx h\,f(x_i)\equiv h\,f_i.$$

Теперь рассмотрим выражения в правой части тождества. Одно из них, например,

$$\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}\int\limits_{x_{i+1/2}}^{x}f(\xi)d\xi,$$

применяя квадратурную формулу прямоугольников, запишем в виде

$$\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}\int\limits_{x_{i+1/2}}^{x}f(\xi)d\xi=\frac{h}{k(x_{i+1/2})}\int\limits_{x_{i+1/2}}^{x_{i+1/2}}f(\xi)d\xi+O(h^3)=O(h^3).$$

Множитель при рассматриваемом интеграле в тождестве равен $O(h^{-1})$, поэтому все выражение для гладких функций имеет порядок $O(h^2)$ и его можно отбросить. Аналогично можно поступить и с другим выражением в правой части равенства.

Окончательный результат можно записать так:

$$-\frac{a_iu_{i+1}-b_iu_i+c_iu_{i-1}}{h^2}=f_i,\quad 0<i<N\,,u_0=u_N=0,$$

где коэффициенты определяются по следующим формулам:

$$b_i=a_i+c_i,$$

$$a_i=c_i=1\quad\text{при}\quad 1\leqslant i<\frac{N}{2}\,,$$

$$a_i=c_i=2\quad\text{при}\quad \frac{N}{2}<i<N,$$

$$a_i=2,c_i=1\quad\text{при}\quad i=\frac{N}{2}\,.$$

Метод Ритца. Пусть требуется найти решение дифференциального уравнения $Lu = f$ в гильбертовом пространстве U, учитывающем краевые условия. Пусть оператор L является самосопряженным и положительно определенным относительно скалярного произведения $(\cdot,\cdot)$, т. е. $\forall u, v \in U$

$$(Lu, v) = (u, Lv) \text{ и } (Lv, v) \geqslant \delta(v, v),\ \delta > 0.$$

Тогда решение исходной задачи сводят к поиску элемента $u \in U$, минимизирующего функционал

$$J(v) = (Lv, v) - 2\,(f, v) \equiv a(v, v) - 2(f, v),$$

где $a(u, v)$ — билинейная форма, как правило, получаемая в результате интегрирования по частям с учетом краевых условий выражения (Lu, v) для $u, v \in U$.

Чтобы определить приближения к элементу u, строят последовательность конечномерных подпространств $U_h \subset U$ с известными базисами $\{\varphi_j^h, j = 0, 1, \ldots, N\}$ и в каждом U_h находят элемент $u_h = \sum\limits_{j=0}^{N} \alpha_j \varphi_j^h$, минимизирующий $J(v)$. Из условий минимума функционала $J(v)$ на элементе $u_h \in U_h$ имеем

$$\frac{\partial J(u_h)}{\partial \alpha_i} = 0\,, \quad i = 0, 1, \ldots, N,$$

откуда следует система линейных алгебраических уравнений $A\alpha = \mathbf{b}$ для определения вектора коэффициентов α, где $a_{ij} = a\left(\varphi_j^h, \varphi_i^h\right)$, $b_i = \left(f, \varphi_i^h\right)$, $i, j = 0, 1, \ldots, N$. Если последовательность U_h полна в U (т. е. $\forall v \in U$ существует последовательность $\{v_h \in U_h\}$ такая, что $\|v - v_h\|_U \to 0$ при $h \to 0$), то $\lim\limits_{h\to 0} \|u - u_h\|_U = 0$.

В качестве базисных элементов φ_j^h в простейшем случае используются кусочно-линейные функции. Например, для произвольной сетки $x_0 < x_1 < \cdots < x_N$ эти функции имеют вид

$$\varphi_0^h(x) = \begin{cases} \dfrac{x_1 - x}{x_1 - x_0} & \text{при} \quad x_0 \leqslant x \leqslant x_1, \\ 0 & \text{при} \quad x_1 \leqslant x \leqslant x_N; \end{cases}$$

$$\varphi_N^h(x) = \begin{cases} 0 & \text{при} \quad x_0 \leqslant x \leqslant x_{N-1}, \\ \dfrac{x - x_{N-1}}{x_N - x_{N-1}} & \text{при} \quad x_{N-1} \leqslant x \leqslant x_N; \end{cases}$$

$$\varphi_j^h(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}} & \text{при} \quad x_{j-1} \leqslant x \leqslant x_j\,, \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j} & \text{при} \quad x_j \leqslant x \leqslant x_{j+1}\,, \\ 0 & \text{при остальных } x \end{cases}$$

для $j = 1, \ldots, N-1$. Если меры носителей базисных функций много меньше меры исходной области (как в рассмотренном случае), то метод Ритца часто называют *методом конечных элементов*.

Воспользуемся методом Ритца для решения дифференциального уравнения

$$L\,u \equiv -(k(x)\,u')' + p(x)\,u = f(x), \quad 0 < x < 1,$$

с краевыми условиями $u(0) = u(1) = 0$ и коэффициентами $k(x) = 1 + x,\ p(x) = 1$.

Возьмем пространство функций

$$U = \left\{ u(x) \ : \ \int_0^1 \left[\left(u'(x)\right)^2 + u^2(x) \right] \, dx < \infty, \ u(0) = u(1) = 0 \right\}$$

со скалярным произведением $(u, v) = \int_0^1 u(x)v(x)\, dx$ и поставим в соответствие исходной дифференциальной задаче с краевыми условиями задачу минимизации на пространстве U функционала

$$J(v) = \int_0^1 \left[k(x) \left(v'(x)\right)^2 + p(x)v^2(x) - 2f(x)v(x) \right] dx \equiv a(v, v) - 2(f, v).$$

Определим последовательность конечномерных подпространств $U_h \subset U$ как последовательность линейных оболочек

$$\operatorname{span}\{\varphi_1^h(x), \varphi_2^h(x), \ldots, \varphi_{N-1}^h(x)\}$$

полных наборов кусочно-линейных базисных функций $\varphi_j^h(x) \in U$ на равномерной сетке ($x_{j+1} - x_j = h,\ N\,h = 1$), и будем искать приближенное решение u_h в виде

$$u_h = \sum_{j=1}^{N-1} \alpha_j \varphi_j^h.$$

В силу краевых условий $u_h(0) = u_h(1) = 0$ (так как $u_h \in U_h \subset U$) функции φ_0^h и φ_N^h в представлении u_h отсутствуют; поэтому формально можно считать, что соответствующие коэффициенты α_0 и α_N равны нулю.

Найдем выражения для матричных элементов a_{ij} системы $A\alpha = \mathbf{b}$:

$$a_{ij} = a\left(\varphi_j^h, \varphi_i^h\right) = \int_0^1 \left[k(x) (\varphi_j^h)' (\varphi_i^h)' + p(x)\, \varphi_j^h \varphi_i^h \right] dx, \quad i, j = 1, 2, \ldots, N-1.$$

Так как при $i = 1, 2, \ldots, N-1$

$$(\varphi_i^h)' = \begin{cases} \frac{1}{h} & \text{при} \quad x_{i-1} < x < x_i, \\ -\frac{1}{h} & \text{при} \quad x_i < x < x_{i+1}, \\ 0 & \text{при } x \notin [x_{i-1}, x_{i+1}], \end{cases}$$

в результате непосредственных вычислений имеем

$$a_{ij} = \begin{cases} -\frac{1}{h}\left[1 + \frac{x_{i-1} + x_i}{2}\right] + \frac{h}{6} & \text{при} \quad j = i - 1, \\ \frac{2}{h}\left[1 + x_i\right] + \frac{2h}{3} & \text{при} \quad j = i, \\ -\frac{1}{h}\left[1 + \frac{x_{i+1} + x_i}{2}\right] + \frac{h}{6} & \text{при} \quad j = i + 1, \\ 0 & \text{при остальных } j. \end{cases}$$

Для компонент вектора правой части получим

$$b_i = \left(f, \varphi_i^h\right) = \int\limits_0^1 f\,\varphi_i^h dx = \int\limits_{x_{i-1}}^{x_{i+1}} f\,\varphi_i^h dx \approx f(x_i) \int\limits_{x_{i-1}}^{x_{i+1}} \varphi_i^h dx = h\,f_i.$$

Разделив обе части уравнения на h, окончательно имеем

$$-\,\frac{a_i\alpha_{i+1} - b_i\alpha_i + c_i\alpha_{i-1}}{h^2} + \frac{\alpha_{i+1} + 4\,\alpha_i + \alpha_{i-1}}{6} = f_i, \quad 0 < i < N,$$

где коэффициенты определяются формулами

$$c_i = 1 + \frac{x_{i-1} + x_i}{2}\,, \quad a_i = 1 + \frac{x_{i+1} + x_i}{2}\,, \quad b_i = a_i + c_i.$$

Для корректного замыкания системы в ней следует положить, как отмечено выше, $\alpha_0 = \alpha_N = 0$.

Метод Галеркина. В отличие от метода Ритца метод Галеркина не требует самосопряженности и положительной определенности оператора L из задачи $Lu = f,\ u \in U$.

Для нахождения приближенного решения в каждом из конечномерных подпространств U_h отыскивают элемент u_h такой, что для любого $v \in U_h$ справедливо равенство $(Lu_h - f, v) = 0$, которое обычно записывают в более удобной, следующей из интегрирования по частям, форме $a(u_h, v) = (f, v)$. Соответствующие коэффициенты α_j разложения u_h по базису подпространства U_h определяют в результате решения системы уравнений, имеющей тот же вид, что и в методе Ритца. Однако обосновать сходимость метода Галеркина удается для более широкого класса задач.

Рассмотрим применение метода Галеркина для несамосопряженной задачи

$$L\,u \equiv -u'' + r(x)\,u' = f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0,$$

с коэффициентом $r(x) = 3\,x^2$. Определим пространства U, U_h и скалярное произведение $(\cdot,\cdot)$, как в примере на метод Ритца. Решение будем искать в виде

$$u_h = \sum_{j=1}^{N-1} \alpha_j \varphi_j^h,$$

где неизвестные коэффициенты α_j найдем из системы линейных алгебраических уравнений $A\alpha = \mathbf{b}$, в которой

$$a_{ij} = a\left(\varphi_j^h, \varphi_i^h\right), \quad b_i = \left(f, \varphi_i^h\right), \quad i, j = 1, 2, \ldots, N-1.$$

Вычислим матричные элементы:

$$a_{ij} = a\left(\varphi_j^h, \varphi_i^h\right) = \int\limits_0^1 \left[(\varphi_j^h)'(\varphi_i^h)' + r(x)\,(\varphi_j^h)'\varphi_i^h\right] dx.$$

Первое слагаемое в этой формуле имеет вид

$$\int\limits_{x_{i-1}}^{x_{i+1}} (\varphi_j^h)'(\varphi_i^h)' dx = \begin{cases} -\frac{1}{h} & \text{при} \quad j = i-1, \\ \frac{2}{h} & \text{при} \quad j = i, \\ -\frac{1}{h} & \text{при} \quad j = i+1, \\ 0 & \text{при} \quad |j-i| > 1. \end{cases}$$

Для второго слагаемого в результате несложных вычислений получаем

$$\int\limits_{x_{i-1}}^{x_{i+1}} 3\,x^2 (\varphi_j^h)'\,\varphi_i^h dx = \begin{cases} -\frac{1}{4}\,[2\,x_{i-1}^2 + (x_{i-1}+x_i)^2] & \text{при} \quad j = i-1, \\ h\,x_i & \text{при} \quad j = i, \\ \frac{1}{4}\,[2\,x_{i+1}^2 + (x_{i+1}+x_i)^2] & \text{при} \quad j = i+1, \\ 0 & \text{при} \quad |j-i| > 1. \end{cases}$$

Определив f_i, как в предыдущем примере, запишем окончательный результат в виде

$$-\frac{\alpha_{i+1} - 2\,\alpha_i + \alpha_{i-1}}{h^2} + \frac{a_i\alpha_{i+1} + b_i\alpha_i + c_i\alpha_{i-1}}{h} = f_i\,, \quad 0 < i < N,$$

$$\alpha_0 = \alpha_N = 0,$$

где коэффициенты определяют по формулам

$$c_i = -\,\frac{1}{4}\,[2\,x_{i-1}^2 + (x_{i-1}+x_i)^2]\,,$$

$$b_i = h\,x_i, \quad a_i = \frac{1}{4}\,[2\,x_{i+1}^2 + (x_{i+1}+x_i)^2]\,.$$

В случае постоянного коэффициента $r(x) \equiv r$ при производной u' в исходном уравнении, второе слагаемое в левой части линейной системы имеет вид $r\ \frac{\alpha_{i+1} - \alpha_{i-1}}{2h}$.

Метод аппроксимации функционала. В этом методе минимизируемый функционал $J(v)$ заменяют приближенным функционалом $J_h(\varphi)$. Пусть на отрезке $[a,b]$ введена сетка $x_i, i = 0,1,\ldots,N$. Тогда производные в функционале заменяем конечными разностями, а интегралы — квадратурами. Например, используя составную формулу прямоугольников,

$$\text{интеграл} \quad \int\limits_a^b (\varphi')^2\,dx \quad \text{заменяем на} \quad \sum_{i=1}^{N} \left(\frac{\varphi_i - \varphi_{i-1}}{h}\right)^2 h.$$

Таким образом, приходим к задаче минимизации приближенного функционала $J_h(\varphi)$. Разностная схема получается приравниванием к нулю величин $\frac{\partial J_h}{\partial \varphi_i}$, $i = 0,1,\ldots,N$.

Краевая задача с достаточно гладким решением $u(x)$

$$L\,u \equiv -(k(x)\,u')' + p(x)\,u = f(x),\ 0 < x < 1,$$
$$0 < k_0 \leqslant k(x) \leqslant k_1,\ 0 \leqslant p(x) \leqslant p_1,\ u(0) = u(1) = 0,$$

эквивалентна задаче отыскания точки минимума $u \in U$ квадратичного функционала

$$J(v) = \int\limits_0^1 \left[k(x)\,(v')^2 + p(x)\,v^2\right] dx - 2\int\limits_0^1 f(x)\,v\,dx.$$

Введем, как и выше, равномерную сетку и на ней аппроксимируем $J(v)$, предварительно записав его в виде

$$J(v) = \sum_{i=1}^{N} \int\limits_{x_{i-1}}^{x_i} k(x)\,(v')^2\,dx + \sum_{i=1}^{N} \int\limits_{x_{i-1}}^{x_i} (p(x)\,v^2 - 2\,f(x)\,v)dx.$$

Далее аппроксимируем интегралы по формулам прямоугольников и трапеций соответственно

$$\int\limits_{x_{i-1}}^{x_i} k(x)\,(v')^2\,dx = k(x_{i-1/2})\left(\frac{v(x_i)-v(x_{i-1})}{h}\right)^2 h + O(h^3),$$

$$\int\limits_{x_{i-1}}^{x_i} (p(x)\,v^2 - 2\,f(x)\,v)dx =$$

$$= \frac{h}{2}\left[(p(x_{i-1})v^2(x_{i-1}) - 2f(x_{i-1})v(x_{i-1})) + (p(x_i)v^2(x_i) - 2f(x_i)v(x_i))\right] + O(h^3).$$

Таким образом, вместо $J(v)$ получаем функционал $J_h(\varphi)$:

$$J_h(\varphi) = \sum_{i=1}^{N} k(x_{i-1/2})\left(\frac{\varphi_i-\varphi_{i-1}}{h}\right)^2 h + \sum_{i=1}^{N-1} (p_i\varphi_i^2 - 2\,f_i\varphi_i)\,h,$$

где $\varphi = (\varphi_0, \varphi_1, \ldots, \varphi_N)$ — произвольная сеточная функция, удовлетворяющая условиям $\varphi_0 = \varphi_N = 0$. Приравнивая к нулю первые производные

$$\frac{\partial J_h(\varphi)}{\partial \varphi_i} = 0\,, \quad i = 1, \ldots, N-1\,,$$

получаем искомую разностную схему:

$$-\frac{1}{h}\left(k(x_{i+1/2})\,\frac{\varphi_{i+1}-\varphi_i}{h} - k(x_{i-1/2})\,\frac{\varphi_i-\varphi_{i-1}}{h}\right) + p_i\varphi_i = f_i,$$
$$0 < i < N,\ \varphi_0 = \varphi_N = 0.$$

Метод сумматорного тождества. Аналогично методу аппроксимации функционала, интегральное тождество $(Lu - f, v) = 0$ для любого $v \in U$ заменяют сумматорным тождеством $(L_h\varphi_h - f_h, v_h) = 0$ для любого $v_h \in U_h$. Так как в конечномерном пространстве размерности $N+1$ векторы $\mathbf{e}_k$, $k = 0, 1, \ldots, N$ образуют базис (k-я компонента вектора $\mathbf{e}_k$ равна

единице, остальные — нулю), то разностная схема получается из системы уравнений

$$(L_h\varphi_h - f_h, \mathbf{e}_k)=0, \quad k=0,1,\ldots,N.$$

Например, для задачи

$$L\,u \equiv -(k(x)u')'+p(x)\,u=f(x), 0<x<1,$$
$$0<k_0\leqslant k(x)\leqslant k_1, 0\leqslant p(x)\leqslant p_1,$$
$$k(0)\,u'(0)=\alpha_1 u(0)+\beta_1, -k(1)\,u'(1)=\alpha_2 u(1)+\beta_2,$$

справедливо интегральное тождество

$$I(u,v)\equiv\int\limits_0^1 (k\,u'v'+p\,u\,v-f\,v)\,dx+(\alpha_1 u(0)+\beta_1)v(0)+(\alpha_2 u(1)+\beta_2)v(1)=0,$$

где $v=v(x)$ — произвольная непрерывная на $[0,1]$ функция, имеющая квадратично-интегрируемую первую производную.

Для построения разностной схемы на равномерной сетке аппроксимируем интегральное тождество сумматорным тождеством для сеточных функций, например,

$$I_h(\varphi,\psi)=\sum_{i=1}^{N} k(x_{i-1/2})\left(\frac{\varphi_i-\varphi_{i-1}}{h}\right)\left(\frac{\psi_i-\psi_{i-1}}{h}\right)h+$$
$$+\sum_{i=1}^{N-1}(p_i\varphi_i-f_i)\psi_i\,h+(\overline{\alpha}_1\varphi_0+\overline{\beta}_1)\psi_0+(\overline{\alpha}_2\varphi_N+\overline{\beta}_2)\psi_N,$$

где $\psi=(\psi_0,\psi_1,\ldots,\psi_N)$ — произвольная сеточная функция. Коэффициенты $\overline{\alpha}_k$ и $\overline{\beta}_k (k=1,2)$ связаны с исходными коэффициентами следующими соотношениями:

$$\overline{\alpha}_1=\alpha_1+p_0\,\frac{h}{2}\,,\ \overline{\beta}_1=\beta_1-f_0\,\frac{h}{2}\,,$$

$$\overline{\alpha}_2=\alpha_2+p_N\,\frac{h}{2}\,,\ \overline{\beta}_2=\beta_2-f_N\,\frac{h}{2}\,.$$

Форма дополнительных слагаемых зависит от выбора квадратурной формулы для аппроксимации интеграла $\int\limits_0^1 (p\,u-f)\,v\,dx$. В данном случае мы воспользовались составной формулой трапеций, т. е.

$$\int\limits_{x_{i-1}}^{x_i}(p\,u-f)\,v\,dx=\frac{h}{2}\,[(p(x_{i-1})u(x_{i-1})-$$
$$-f(x_{i-1}))\,v(x_{i-1})+(p(x_i)u(x_i)-f(x_i))\,v(x_i)]+O(h^3).$$

Суммируя по всем $i=1,\ldots,N$, получаем вторую сумму в $I_h(\varphi,\psi)$, а оставшиеся слагаемые $\frac{h}{2}\,[(p_0\varphi_0-f_0)\,\psi_0+(p_N\varphi_N-f_N)\,\psi_N]$ изменят значения α_k и $\beta_k (k=1,2)$. Например, при $i=0$ имеем

$$(\alpha_1\varphi_0+\beta_1)\psi_0+\frac{h}{2}\,(p_0\varphi_0-f_0)\,\psi_0=(\overline{\alpha}_1\varphi_0+\overline{\beta}_1)\psi_0\,.$$

Полагая теперь $\psi=\mathbf{e}_k$, т. е. $\psi_i=\delta_i^k\,(0<k<N)$, и учитывая, что

$$\frac{\psi_i-\psi_{i-1}}{h}=\begin{cases}-\frac{1}{h} & \text{при} \quad i=k+1,\\ \frac{1}{h} & \text{при} \quad i=k,\\ 0 & \text{в остальных случаях,}\end{cases}$$

при $i=k\,(0<k<N)$ получаем

$$-\frac{1}{h}\left(k(x_{i+1/2})\,\frac{\varphi_{i+1}-\varphi_i}{h}-k(x_{i-1/2})\,\frac{\varphi_i-\varphi_{i-1}}{h}\right)+p_i\varphi_i=f_i.$$

Далее, если $\psi_i=\delta_i^0$, то имеем

$$k(x_{1/2})\,\frac{\varphi_1-\varphi_0}{h}=\overline{\alpha}_1\varphi_0+\overline{\beta}_1;$$

аналогично при $\psi_i=\delta_i^N$ находим

$$-k(x_{N-1/2})\,\frac{\varphi_N-\varphi_{N-1}}{h}=\overline{\alpha}_2\varphi_N+\overline{\beta}_2.$$

Последние три выражения приводят к системе из $N+1$ уравнения с $N+1$ неизвестным $(\varphi_0,\varphi_1,\ldots,\varphi_N)$, т. е. искомая разностная схема построена.

Метод построения точных разностных схем. Разностную схему называют *точной*, если ее решение совпадает с решением дифференциального уравнения в узлах сетки. На примере уравнения второго порядка

$$-(k(x)u')'=f(x),\,0<x<1,\,0<k_0\leqslant k(x)\leqslant k_1,\quad u(0)=u(1)=0$$

рассмотрим метод построения точной разностной схемы на равномерной сетке. Воспользуемся тем же подходом, что и в методе интегрального тождества Марчука. Проинтегрируем исходное уравнение по отрезку $[x_i,x]$ и результат разделим на $k(x)$. Имеем

$$u'(x)=\frac{k(x_i)u'(x_i)}{k(x)}-\frac{1}{k(x)}\int\limits_{x_i}^{x}f(\xi)d\xi.$$

Проинтегрировав последнее равенство по отрезкам $[x_{i-1},x_i]$, $[x_i,x_{i+1}]$ и умножив результаты соответственно на величины $\frac{1}{h}\,a_i$, $\frac{1}{h}\,a_{i+1}$, где $a_i=\left(\frac{1}{h}\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}\right)^{-1}$, получаем

$$a_i\,\frac{u(x_i)-u(x_{i-1})}{h}=k(x_i)u'(x_i)-\frac{a_i}{h}\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}\int\limits_{x_i}^{x}f(\xi)d\xi,$$

$$a_{i+1}\,\frac{u(x_{i+1})-u(x_i)}{h}=k(x_i)u'(x_i)-\frac{a_{i+1}}{h}\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}\int\limits_{x_i}^{x}f(\xi)d\xi.$$

Исключая $k(x_i)u'(x_i)$, имеем точную схему

$$-\frac{a_{i+1}(u_{i+1}-u_i)-a_i(u_i-u_{i-1})}{h^2}=f_i,$$

где

$$f_i=\frac{1}{h^2}\left(a_{i+1}\int\limits_{x_i}^{x_{i+1}}\frac{dx}{k(x)}\int\limits_{x_i}^{x}f(\xi)d\xi+a_i\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}\int\limits_{x}^{x_i}f(\xi)d\xi\right),$$

а коэффициенты a_i определены выше.

На практике реальная точность схемы определятся точностью вычисления интегралов в полученных формулах.

В случае $k(x)\equiv1$ коэффициенты $a_i=1$ при всех i, а выражения для f_i принимают вид

$$f_i=\frac{1}{h^2}\left(\int\limits_{x_i}^{x_{i+1}}\int\limits_{x_i}^{x}f(\xi)d\xi dx+\int\limits_{x_{i-1}}^{x_i}\int\limits_{x}^{x_i}f(\xi)d\xi dx\right).$$

7.1. Справедливы ли следующие равенства:

1) $\lim\limits_{h\to0}\dfrac{u(x+h)-2u(x)+u(x-h)}{h^2}=\lim\limits_{h\to0}\dfrac{\frac{u(x+2h)+u(x)}{2}-2u(x)+\frac{u(x)+u(x-2h)}{2}}{h^2}$,

2) $\lim\limits_{h\to0}\dfrac{u(x+h)-u(x-h)}{2h}=\lim\limits_{h\to0}\dfrac{\frac{u(x+2h)+u(x)}{2}-\frac{u(x)+u(x-2h)}{2}}{2h}$,

если $u(x)\in C^{(4)}$?

О т в е т: 1) нет; 2) да.

В задачах 7.2–7.7 следует обращать внимание на области определения искомого решения разностного уравнения и его правой части. В некоторых случаях более высокий порядок аппроксимации схемы может быть достигнут в результате выбора смещенных сеток ih и $ih\pm\frac{h}{2}$, $i=0,1,\ldots$.

7.2. Рассмотрим дифференциальную задачу

$$u'+a(x)u(x)=f(x),\quad x\in[0,1],\quad u(0)=0\,;\quad a(x),\ f(x)\in C^{(4)}[0,1].$$

Считая, что функции u_i и f_i определены в узлах $x_i=ih$, $h=\frac{1}{N}$, $i=0,\ldots,N$, найти порядок аппроксимации на решении разностной схемы:

1) $\dfrac{u_{i+1}-u_i}{h}+a_iu_i=f_i$, $0\leqslant i\leqslant N-1$, $\quad u_0=0$;

2) $\dfrac{u_{i+1}-u_{i-1}}{2h}+a_iu_i=f_i$, $1\leqslant i\leqslant N-1$, $\quad u_0=0, u_1=hf_0$,

где $a_i=a(x_i)$, $f_i=f(x_i)$.

О т в е т: 1) $O(h)$; 2) $O(h^2)$.

7.3. Рассмотрим дифференциальную задачу

$$u'+a(x)u(x)=f(x),\quad x\in[0,1],\quad u(0)=0;\quad a(x),\ f(x)\in C^{(4)}[0,1].$$

Считая, что функция u_i определена в узлах $x_i=ih$, $h=\frac{1}{N}$, $i=0,\dots,N$, а функция f_i — в узлах $x_{i+1/2}=\left(i+\frac{1}{2}\right)h$, $i=0,\dots,N-1$, найти порядок аппроксимации на решении разностной схемы

$$\frac{u_{i+1}-u_i}{h}+a_i\,\frac{u_{i+1}+u_i}{2}=f_i,\quad u_0=0, i=0,\dots,N-1,$$

где $a_i=a(x_{i+1/2})$, $f_i=f(x_{i+1/2})$.

Ответ: порядок аппроксимации равен $O(h^2)$; ответ не изменится, если использовать следующие аппроксимации для коэффициента и правой части уравнения: $a_i=\frac{a(x_{i+1})+a(x_i)}{2}$, $f_i=\frac{f(x_{i+1})+f(x_i)}{2}$.

7.4. Для дифференциальной задачи

$$-u''=f(x),\quad x\in[0,1],\quad u(0)=u(1)=0,$$

построить разностную схему второго порядка аппроксимации, которая при каждом h является системой линейных алгебраических уравнений с симметричной положительно определенной матрицей.

Указание. Рассмотреть схему из примера на метод неопределенных коэффициентов, для которой воспользоваться решением 2.86.

7.5. Для дифференциальной задачи

$$-u''=f(x),\quad x\in[0,1],$$
$$u(0)=a, u(1)=b,\quad u\in C^{(4)}[0,1],$$

на трехточечном шаблоне с переменными шагами сетки построить разностные схемы первого и второго порядка аппроксимации на решении.

Ответ: для произвольной неравномерной сетки

$$0=x_0<x_1<\dots<x_{N-1}<x_N=1,\quad h_i=x_i-x_{i-1},\quad 1\leqslant i\leqslant N,$$

схема

$$-\frac{2}{h_{i+1}+h_i}\left(\frac{u_{i+1}-u_i}{h_{i+1}}-\frac{u_i-u_{i-1}}{h_i}\right)=f_i,\quad 0<i<N$$

с краевыми условиями $u_0=a$, $u_N=b$ имеет на решении порядок аппроксимации $O(h)$ при $f_i=f(x_i)$ и порядок $O(h^2)$ — при

$$f_i=f(x_i)+\frac{1}{3}\,\frac{h_{i+1}-h_i}{h_{i+1}+h_i}\,(f(x_{i+1})-f(x_{i-1}))\,.$$

7.6. Для дифференциальной задачи

$$u'+c\,u=f(x),\quad c=\text{const},\quad u(0)=a,$$

интегро-интерполяционным методом на трехточечном шаблоне с постоянным шагом построить схему четвертого порядка аппроксимации.

Ответ: Для приближенного вычисления интеграла по отрезку $[x_{i-1},x_{i+1}]$ использовать формулу Симпсона, а для получения недостающего начального условия $u_1\approx u(h)$ применить формулу Тейлора (необходимые производные при $x=0$ можно получить, дифференцируя уравнение требуемое число раз).

7.7. Для дифференциальной задачи

$$-(k(x)u')'=1, \quad x\in[0,1],$$
$$u(0)=u(1)=0, \quad k(x)=\begin{cases}\frac{1}{2} & \text{при} \quad 0\leqslant x<\frac{1}{4},\\ 1 & \text{при} \quad \frac{1}{4}\leqslant x\leqslant 1,\end{cases}$$

построить разностную схему с помощью интегрального тождества Марчука, если точка разрыва $k(x)$ является узлом сетки.

7.8. Показать, что для дифференциальной задачи

$$L\,u\equiv-(k(x)u')'+p(x)u=f(x), \quad x\in[0,1], \quad u(0)=u'(1)=0,$$

переменные коэффициенты которой удовлетворяют условиям $0<k_0\leqslant k(x)\leqslant k_1$, $0\leqslant p_0\leqslant p(x)\leqslant p_1$, квадратичная часть $a(v,v)$ функционала $J(v)$ в методе Ритца удовлетворяет оценке снизу

$$a(v,v)=\int\limits_0^1 [k(x)(v'(x))^2+p(x)v^2(x)]\,dx \geqslant (k_0+p_0)\int\limits_0^1 v^2(x)\,dx \quad \forall v\in U.$$

У к а з а н и е. Вывести неравенство $\int\limits_0^1 v^2(x)\,dx \leqslant \int\limits_0^1 (v'(x))^2 dx$ для произвольной функции $v(x)\in U$, где

$$U=\left\{u(x)\ :\ \int\limits_0^1 \left[(u'(x))^2+u^2(x)\right]\,dx<\infty,\ u(0)=0\right\}.$$

7.9. Для произвольной функции $v\in U$, где

$$U=\left\{u(x)\ :\ \int\limits_0^1 \left[(u'(x))^2+u^2(x)\right]\,dx<\infty,\ u(0)=u(1)=0\right\}$$

показать справедливость неравенства

$$\int\limits_0^1 v^2(x)dx \leqslant \frac{1}{\pi^2}\int\limits_0^1 (v'(x))^2 dx.$$

У к а з а н и е. Воспользоваться решением спектральной задачи

$$-w''=\lambda w, \quad w(0)=w(1)=0.$$

7.10. Дана дифференциальная задача

$$-u''+c\,u=f(x), \quad x\in[0,1],$$
$$u(0)=u(1)=0, \quad c=\text{const}.$$

При каких c для решения этой задачи можно применять метод Ритца?
О т в е т: $c>-\pi^2$.

7.11. Для дифференциальной задачи

$$-u'' + u = f(x), \quad x \in [0,1], \quad u'(0) = u'(1) = 0,$$

построить разностную схему методом Ритца, взяв кусочно-линейные функции на равномерной сетке в качестве базисных.

Ответ: Функция $u(x)$ доставляет минимум функционалу

$$J(v) = \int\limits_0^1 \left((v')^2 + v^2 - 2f(x)v\right) dx$$

на пространстве

$$U = \left\{ u(x) \ : \ \int\limits_0^1 \left[(u'(x))^2 + u^2(x)\right] \, dx < \infty \right\}.$$

7.12. Показать, что для дифференциальной задачи

$$L\,u \equiv -(k(x)u')' + p(x)u = f(x), \quad x \in [0,1],$$
$$u(0) = 0, \ u'(1) + \alpha u(1) = \beta,$$

переменные коэффициенты которой удовлетворяют условиям $0 < k_0 \leqslant k(x) \leqslant k_1$, $0 \leqslant p(x) \leqslant p_1$, функционал в методе Ритца имеет вид

$$J(v) = \int\limits_0^1 (k(x)(v'(x))^2 + p(x)v^2(x) - 2f(x)v(x))\, dx + \\ + \alpha k(1)v^2(1) - 2\beta k(1)v(1).$$

Указание. Рассмотреть коэффициент при 2ε в неравенстве $J(u) \leqslant J(u + \varepsilon\, w)$, справедливом при ε любого знака и любом $w \in U$, где

$$U = \left\{ u(x) \ : \ \int\limits_0^1 \left[(u'(x))^2 + u^2(x)\right] \, dx < \infty, \ u(0) = 0 \right\}.$$

7.13. Для дифференциальной задачи

$$-\left(k(x)u'\right)' = 1, \quad x \in [0,1],$$
$$u(0) = u(1) = 0, \quad k(x) = \begin{cases} \frac{3}{2} & \text{при} \quad 0 \leqslant x < \frac{1}{4}, \\ 2 & \text{при} \quad \frac{1}{4} \leqslant x \leqslant 1, \end{cases}$$

построить разностную схему методом Ритца, взяв кусочно-линейные функции на равномерной сетке в качестве базисных и считая, что точка разрыва $k(x)$ является узлом сетки.

7.14. Пусть функция $u(x)$ доставляет минимум функционалу

$$J(v) = a(v,v) - 2(f,v) \equiv \int\limits_0^1 \left(k(x)\,(v')^2 + p(x)v^2 - 2f(x)v\right) dx,$$

где переменные коэффициенты удовлетворяют условиям $0<k_0\leqslant k(x)\leqslant k_1$, $0\leqslant p(x)\leqslant p_1$, на пространстве

$$U=\left\{u(x)\ :\ \int\limits_0^1\left[(u'(x))^2+u^2(x)\right]\,dx<\infty,\ u(0)=0\right\}.$$

Показать справедливость равенства $a(u,v)=(f,v)$ с произвольной функцией $v\in U$.

Если дополнительно функция $u(x)$ удовлетворяет уравнению

$$-\left(k(x)\,u'\right)'+p(x)\,u=f(x),$$

т. е. является достаточно гладкой, то краевое условие $u'(1)=0$ для нее выполняется автоматически (без включения в определение пространства U).

$\lhd$ Если функция u доставляет минимум функционалу $J(v)$ на пространстве U, то для произвольных величин — числа ε и функции $v\in U$ — имеем

$$J(u)\leqslant J(u+\varepsilon v)=J(u)+2\varepsilon\left[a(u,v)-(f,v)\right]+\varepsilon^2a(v,v),$$

т. е. $2\varepsilon\left[a(u,v)-(f,v)\right]+\varepsilon^2a(v,v)\ \geqslant\ 0$. В силу произвольности знака ε и положительности величины $a(v,v)$ при $v\neq 0$ (см. 7.8), отсюда следует $a(u,v)=(f,v)\ \forall v\in U$. Полученное равенство лежит в основе определения *слабого* (или *обобщенного*) *решения* дифференциальной задачи.

Если функция $u(x)$ — достаточно гладкая, то интегрирование по частям дает

$$0=a(u,v)-(f,v)=\int\limits_0^1(k(x)u'v'+p(x)uv-fv)\,dx=$$

$$=\int\limits_0^1 v\left[-(k(x)u')'+p(x)u-f\right]\,dx+k(1)u'(1)v(1)=k(1)u'(1)v(1).$$

Из этого равенства, в силу произвольности значения $v(1)$ и положительности $k(x)$, следует $u'(1)=0$. $\rhd$

7.15. Пусть функция $u(x)\in C^{(2)}[0,1]$ является решением дифференциальной задачи

$$-\left(k(x)\,u'\right)'+p(x)\,u=f(x),\quad x\in[0,1],\quad u(0)=u'(1)=0,$$

достаточно гладкие переменные коэффициенты которой удовлетворяют условиям $0<k_0\leqslant k(x)\leqslant k_1$, $0\leqslant p(x)\leqslant p_1$. Показать, что $u(x)$ доставляет единственный минимум функционалу $J(v)=a(v,v)-2(f,v)$ на пространстве

$$U=\left\{u(x)\ :\ \int\limits_0^1\left[(u'(x))^2+u^2(x)\right]\,dx<\infty,\ u(0)=0\right\}.$$

$\lhd$ Запишем квадратичный функционал, соответствующий исходной задаче,

$$J(v)=a(v,v)-2(f,v)\equiv\int\limits_0^1\Big(k(x)\,(v')^2+p(x)v^2-2f(x)v\Big)\,dx.$$

Функция $u(x)$ принадлежит U, зафиксируем ее и рассмотрим выражение $a(v-u, v-u) - a(u,u)$ как функционал от $v \in U$. Этот функционал имеет единственную точку минимума $v = u$, так как первое слагаемое неотрицательно и в силу 7.8 обращается в нуль только тогда, когда аргумент равен нулю. При этом второе слагаемое от v не зависит.

Раскрывая скобки, получим

$$a(v-u, v-u) - a(u,u) = a(v,v) - 2a(u,v) + a(u,u) - a(u,u) = \\ = a(v,v) - 2(f,v) \equiv J(v).$$

Выше было использовано равенство $a(u,v) = (f,v)$ из 7.14. $\triangleright$

7.16. В задаче

$$-(k(x)\,u')' + p(x)\,u = f(x), \quad x \in [0,1], \quad u(0) = u'(1) = 0,$$

переменные коэффициенты которой удовлетворяют условиям $0 < k_0 \leqslant k(x) \leqslant k_1$, $0 \leqslant p(x) \leqslant p_1$, методом Ритца (конечных элементов) построить аппроксимацию краевого условия $u'(1) = 0$, используя кусочно-линейные базисные функции на равномерной сетке.

$\triangleleft$ Запишем квадратичный функционал, соответствующий исходной задаче,

$$J(v) = \int\limits_0^1 \Big(k(x)\,(v')^2 + p(x)v^2 - 2f(x)v \Big)\,dx,$$

а приближенное решение будем искать в виде

$$u_h = \sum_{j=1}^{N} \alpha_j \varphi_j^h(x), \quad N\,h = 1.$$

Далее подставим u_h в J и рассмотрим систему

$$\frac{\partial J(u_h)}{\partial \alpha_i} = 0\,, \quad 1 \leqslant i \leqslant N.$$

Нас интересует последнее уравнение системы (при $i = N$)

$$a\,\alpha_{N-1} + b\,\alpha_N = c,$$

где

$$a = \int\limits_0^1 (k(x)\varphi'_{N-1}\varphi'_N + p(x)\varphi_{N-1}\varphi_N)\,dx,$$

$$b = \int\limits_0^1 \Big(k(x)\,(\varphi'_N)^2 + p(x)\varphi_N^2 \Big)\,dx, \quad c = \int\limits_0^1 f(x)\varphi_N dx.$$

Запишем формулу для $\varphi_N(x)$:

$$\varphi_N(x) = \begin{cases} 0 & \text{при} \quad 0 \leqslant x \leqslant x_{N-1} = 1-h, \\ \dfrac{x - x_{N-1}}{h} & \text{при} \quad x_{N-1} \leqslant x \leqslant x_N = 1. \end{cases}$$

Эта базисная функция отлична от нуля только на отрезке $[1-h, 1]$, поэтому область интегрирования сужается, т. е. потребуется только часть функции $\varphi_{N-1}(x)$:

$$\varphi_{N-1}(x) = \frac{1-x}{h} \quad \text{при} \quad 1-h \leqslant x \leqslant 1.$$

В случае постоянных коэффициентов $k(x) \equiv k,\ p(x) \equiv p$ величины a, b, c определяются так:

$$a = -\frac{k}{h} + \frac{p\,h}{6}, \quad b = \frac{k}{h} + \frac{p\,h}{3}, \quad c = \int\limits_{1-h}^{1} f(x)\,\frac{x-1+h}{h}\,dx.$$

Для сравнения приведем аппроксимацию второго порядка, построенную интегро-интерполяционным методом:

$$a_1\,u_{N-1} + b_1\,u_N = c_1,$$

где

$$a_1 = -\frac{k}{h}, \quad b = \frac{k}{h} + \frac{p\,h}{2}, \quad c_1 = \frac{h}{2}\,f_N. \qquad \triangleright$$

7.17. Для дифференциальной задачи

$$-\left(k(x)u'\right)' + p(x)\,u = f(x),\ x \in [0,1],\ u'(0) = u(1) = 0,$$

переменные коэффициенты которой удовлетворяют условиям

$$0 < k_0 \leqslant k(x) \leqslant k_1, \quad 0 \leqslant p(x) \leqslant p_1,$$

на равномерной сетке построить разностную схему методом аппроксимации функционала.

7.18. Показать, что решение разностной схемы

$$k(x_i)\,\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + \frac{k(x_{i+1}) - k(x_{i-1})}{2h}\,\frac{u_{i+1} - u_{i-1}}{2h} = 0,$$

$$0 < i < N, \quad u_0 = 1, \quad u_N = 0, \quad N\,h = 1,$$

построенной на равномерной сетке ($x_i = ih,\ 0 \leqslant i \leqslant N$), не сходится к решению дифференциальной задачи

$$\left(k(x)u'\right)' = 0, \quad x \in [0,1], \quad u(0) = 1, \quad u(1) = 0$$

в классе положительных кусочно-постоянных коэффициентов

$$k(x) = \begin{cases} k_1 & \text{при} \quad 0 < x < \xi, \\ k_2 & \text{при} \quad \xi < x < 1, \end{cases}$$

где ξ — иррациональное число, $\xi = x_n + \theta h,\ 0 < \theta < 1$.

$\triangleleft$ Запишем решение дифференциальной задачи

$$u(x) = \begin{cases} 1 - \alpha_0 x & \text{при} \quad 0 \leqslant x \leqslant \xi, \quad \alpha_0 = (\delta + (1-\delta)\xi)^{-1}, \\ \beta_0(1-x) & \text{при} \quad \xi \leqslant x \leqslant 1, \quad \beta_0 = \delta\alpha_0,\ \delta = \dfrac{k_1}{k_2}. \end{cases}$$

Точное решение разностной задачи имеет вид

$$u_i = \begin{cases} 1 - \alpha x_i & \text{при} \quad 0 \leqslant x_i \leqslant x_n, \\ \beta(1 - x_i) & \text{при} \quad x_{n+1} \leqslant x_i \leqslant 1, \end{cases}$$

где коэффициенты α и β можно получить из уравнений в точках x_n (слева от разрыва) и x_{n+1} (справа от разрыва):

$$\beta\,(1 - x_{n+1}) + \alpha\left[x_n + h\,\frac{5\delta - 1}{3\delta + 1}\right] - 1 = 0,$$

$$\beta\left[(1 - x_{n+1}) + h\,\frac{5 - \delta}{3 + \delta}\right] + \alpha x_n - 1 = 0.$$

Отсюда при $\delta = 5$ имеем $\alpha = 0$, $\beta = (1 - x_{n+1})^{-1}$; при $\delta = \frac{1}{5}$ получаем $\beta = 0, \quad \alpha = x_n^{-1}$. Здесь переход к пределу при $h \to 0$ (т. е. $x_n, x_{n+1} \to \xi$) не приводит к решению дифференциального уравнения.

В остальных случаях удобно представление $\beta = \mu\alpha$, где

$$\alpha = \Big(\mu + (1-\mu)x_n + h\,\frac{5\delta - 1}{3\delta + 1} - h\mu\Big)^{-1}, \ \mu = \frac{(3+\delta)(5\delta-1)}{(5-\delta)(3\delta+1)}\,.$$

Доопределив сеточную функцию u_i линейно между узлами, получим непрерывную функцию $\tilde{u}(x,h)$, совпадающую с u_i в узлах x_i. Найдем

$$\lim_{h\to 0} \tilde{u}(x,h) = \begin{cases} 1 - \alpha_1 x & \text{при} \quad 0 \leqslant x \leqslant \xi, \\ \beta_1(1-x) & \text{при} \quad \xi \leqslant x \leqslant 1. \end{cases}$$

При $h \to 0$ имеем

$$\lim_{h\to 0} \alpha = \alpha_1 = (\mu + (1-\mu)\xi)^{-1}, \quad \lim_{h\to 0} \beta = \beta_1 = \mu\alpha_1.$$

Совпадение коэффициентов α_1 с α_0 и β_1 с β_0 возможно только в случае равенства $\mu = \delta$, эквивалентного уравнению $(\delta - 1)^3 = 0$, т. е. только при $k_1 = k_2$. $\triangleright$

7.19. Для дифференциальной задачи

$$-\big(k(x)u'\big)' = 1, \quad x \in [0,1],$$
$$u(0) = u(1) = 0, \quad k(x) = \begin{cases} 1 & \text{при } 0 \leqslant x < \frac{\pi}{5}\,, \\ \frac{1}{3} & \text{при } \frac{\pi}{5} \leqslant x \leqslant 1, \end{cases}$$

построить разностную схему методом Галеркина, взяв кусочно-линейные функции на равномерной сетке в качестве базисных.

7.20. Для дифференциальной задачи

$$-u'' + a\,u' + p\,u = 1, \quad x \in [0,1],$$
$$a = \text{const}, \ p = \text{const} \geqslant 0, \quad u(0) = u(1) = 1,$$

построить разностную схему методом Галеркина, взяв кусочно-линейные функции в качестве базисных.

7.21. Для дифференциальной задачи

$$-\big(k(x)u'\big)' + a(x)\,u' + p(x)\,u = f(x), \quad x \in [0,1],$$
$$u(0) = u(1) = 0,$$

переменные коэффициенты которой удовлетворяют условиям

$$0 < k_0 \leqslant k(x) \leqslant k_1, \quad |a(x)| \leqslant a_1, \quad 0 \leqslant p(x) \leqslant p_1,$$

на равномерной сетке построить разностную схему методом сумматорного тождества.

7.22. Привести пример последовательности сеточных функций $\{\varphi_i^h\}, i = 0, 1, \ldots, N,\ N\,h = 1$ из семейства пространств $\{U_h\}$, которая сходилась бы при $h \to 0$ к некоторой функции $u \in U$, если $\|\varphi^h\| = \left(h \sum\limits_{i=0}^{N} \left(\varphi_i^h\right)^2\right)^{1/2}$, и расходилась, если $\|\varphi^h\| = \max\limits_i |\varphi_i^h|$.

Ответ: $u(x) = 1, \quad \varphi_i^h = \begin{cases} 1 & \text{при} \quad i \neq 0, \\ 1 + h^{-1/4} & \text{при} \quad i = 0. \end{cases}$

7.23. Сходится ли последовательность сеточных функций $\{\varphi_i^h\}$, $i = 0, 1, \ldots, N$, $N\,h = 1$, в норме $\|\varphi^h\| = \max\limits_i |\varphi_i^h|$ к функции $u(x)$ и с каким порядком, если

$$\varphi_i^h = \frac{1}{2}\left(u\left(x_i + \frac{h}{2}\right) + u\left(x_i - \frac{h}{2}\right)\right),\ \varphi_0^h = u(0),\ \varphi_N^h = u(1),\ x_i = ih\,,$$

а $u(x)$ принадлежит одному из пространств $C^{(k)}$, $k \geqslant 0$? Существуют ли функции $u(x)$, к которым $\{\varphi_i^h\}$ сходится с бесконечным порядком?

Ответ: порядок сходимости равен: $o(1)$ при $u \in C$, $O(h)$ при $u \in C^{(1)}$, $O(h^2)$ при $u \in C^{(k)}, k \geqslant 2$. Если $u(x) = \text{const}$, то порядок сходимости — бесконечный.

7.24. Для дифференциальной задачи

$$-\left(k(x)u'\right)' = f(x), \quad x \in [0, 1],$$
$$0 < k_0 \leqslant k(x) \leqslant k_1, \quad u(0) = u(1) = 0$$

построить на равномерной сетке схему четвертого порядка аппроксимации, заменяя в точной разностной схеме значения интегралов приближенными.

7.25. (Проекционная теорема в методе Ритца). Пусть u — точка минимума функционала $J(v) = (Lv, v) - 2(v, f) \equiv a(v, v) - 2(v, f)$ на U, U_h — замкнутое подпространство U. Доказать, что:

1) функция $u_h \in U_h$, на которой достигается минимум, удовлетворяет условию

$$a(u_h, z_h) = (f, z_h) \quad \forall z_h \in U_h\,.$$

В частности, если U_h совпадает с U, то $a(u, z) = (f, z)\ \ \forall z \in U$;

2) точка минимума u_h есть проекция u на U_h по отношению к энергетическому скалярному произведению $a(u, v)$ или, что то же, ошибка $u - u_h$ ортогональна U_h:

$$a(u - u_h, z_h) = 0 \quad \forall z_h \in U_h;$$

3) минимум $J(z_h)$ и минимум $a(u - z_h, u - z_h)$, где z_h пробегает подпространство U_h, достигаются на одной и той же функции u_h, так что

$$a(u - u_h, u - u_h) = \min_{z_h \in U_h} a(u - z_h, u - z_h).$$

$\lhd$ 1) Если u_h минимизирует $J(v)$ на U_h, то для произвольных $\varepsilon \in \mathbf{R}^1$ и $z_h \in U_h$ имеем

$$I(u_h) \leqslant I(u_h + \varepsilon\, z_h) = I(u_h) + 2\varepsilon[a(u_h, z_h) - (f, z_h)] + \varepsilon^2 a(z_h, z_h).$$

Отсюда получаем

$$0 \leqslant 2\varepsilon[a(u_h, z_h) - (f, z_h)] + \varepsilon^2 a(z_h, z_h).$$

Так как ε может иметь любой знак, а второе слагаемое строго положительно, то $a(u_h, z_h) = (f, z_h)$. В частности, если U_h совпадает с U, то имеем $a(u, z) = (f, z) \quad \forall z \in U$.

2) Второе утверждение следует из первого. Вычитая первое из полученных равенств из второго, так как $z_h \in U_h \subset U$, получим

$$a(u - u_h, z_h) = 0 \quad \forall z_h \in U_h.$$

3) Рассмотрим следующее выражение для произвольного z_h:

$$a(u - u_h - z_h, u - u_h - z_h) = a(u - u_h, u - u_h) - 2a(u - u_h, z_h) + a(z_h, z_h).$$

В силу предыдущего утверждения, второе слагаемое равно нулю, а третье неотрицательно, поэтому имеем

$$a(u - u_h, u - u_h) \leqslant a(u - u_h - z_h, u - u_h - z_h) \quad \forall z_h \in U_h.$$

Это неравенство обращается в равенство только при $a(z_h, z_h) = 0$, т. е. при $z_h = 0$, поэтому

$$a(u - u_h, u - u_h) = \min_{z_h \in U_h} a(u - z_h, u - z_h).$$

Существование и единственность $u_h \in U_h$ следует из замкнутости U_h. Если последовательность $v_h^n \in U_h$ фундаментальная, т. е. $a(v_h^n - v_h^m, v_h^n - v_h^m)$ стремится к нулю при $n, m \to \infty$, то существует элемент $v_h \in U_h$, для которого справедливо $a(v_h^n - v_h, v_h^n - v_h) \to 0$ при $n \to \infty$. Это имеет место всегда, если пространство U_h конечномерно. $\rhd$

7.26. Пусть функция $y(x)$ удовлетворяет условию

$$\|y''\|^2 = \int_0^1 [y''(x)]^2 dx < \infty \quad \text{и} \quad y_I(x) = \sum_{i=0}^{N} y(x_i)\varphi_i(x)$$

— ее линейный интерполянт, построенный на равномерной сетке $x_i = ih$, $0 \leqslant i \leqslant N$, $Nh = 1$. Доказать справедливость следующих неравенств

$$\|y' - y_I'\| \leqslant \frac{h}{\pi}\,\|y''\|, \quad \|y - y_I\| \leqslant \left(\frac{h}{\pi}\right)^2 \|y''\|.$$

$\lhd$ Рассмотрим какой-либо отрезок длины h, для простоты удобно взять — $[0, h]$. Построим на нем функцию

$$\Delta(x) = y(x) - y_I(x).$$

По предположению о гладкости $y(x)$ функция $\Delta(x)$ имеет конечный интеграл

$$\int_0^h (\Delta'')^2 dx = \int_0^h (y'')^2 dx < \infty,$$

также выполнены равенства $\Delta(0) = \Delta(h) = 0$, поэтому справедливо представление $\Delta(x)$ в виде ряда Фурье

$$\Delta(x) = \sum_{l=1}^{\infty} d_l \sin \frac{\pi l x}{h}\,.$$

В результате непосредственных вычислений имеем

$$\int_0^h [\Delta'(x)]^2 dx = \frac{h}{2} \sum_{l=1}^{\infty} \left(\frac{\pi l}{h}\right)^2 d_l^2, \quad \int_0^h [\Delta''(x)]^2 dx = \frac{h}{2} \sum_{l=1}^{\infty} \left(\frac{\pi l}{h}\right)^4 d_l^2.$$

Так как $l \geqslant 1$, то справедливо неравенство

$$\left(\frac{\pi l}{h}\right)^2 d_l^2 \leqslant \left(\frac{h}{\pi}\right)^2 \left(\frac{\pi l}{h}\right)^4 d_l^2\,,$$

поэтому, суммируя по l, получаем

$$\int_0^h [\Delta'(x)]^2 dx \leqslant \left(\frac{h}{\pi}\right)^2 \int_0^h [\Delta''(x)]^2 dx = \left(\frac{h}{\pi}\right)^2 \int_0^h [y'']^2 dx\,.$$

Последнее неравенство справедливо на каждом отрезке длины h, потому суммирование по всем i дает

$$\|y' - y_I'\|^2 \leqslant \left(\frac{h}{\pi}\right)^2 \|y''\|^2.$$

Аналогично получаем

$$\int_0^h [\Delta(x)]^2 dx = \frac{h}{2} \sum_{l=1}^{\infty} d_l^2 \leqslant \left(\frac{h}{\pi}\right)^4 \|y''\|^2\,, \text{т. е. } \|y - y_I\| \leqslant \left(\frac{h}{\pi}\right)^2 \|y''\|. \qquad \triangleright$$

7.3. Методы прогонки и стрельбы. Метод Фурье

Рассмотрим эффективные методы решения разностных уравнений, основанные на специальных свойствах оператора задачи.

Метод прогонки. Пусть требуется найти решение системы уравнений:

$$\begin{aligned} c_0 y_0 - b_0 y_1 &= f_0\,, \quad && i = 0\,, \\ -a_i y_{i-1} + c_i y_i - b_i y_{i+1} &= f_i\,, \quad && 1 \leqslant i \leqslant N-1\,, \\ -a_N y_{N-1} + c_N y_N &= f_N\,, \quad && i = N\,, \end{aligned} \tag{7.5}$$

или в векторном виде

$$A\,\mathbf{y} = \mathbf{f}\,,$$

где $\mathbf{y} = (y_0, y_1, \ldots, y_N)^T$ — вектор неизвестных, $\mathbf{f} = (f_0, f_1, \ldots, f_N)^T$ — заданный вектор правых частей, A — квадратная матрица размерности $(N+1) \times (N+1)$

$$A = \begin{pmatrix} c_0 & -b_0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ -a_1 & c_1 & -b_1 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & -a_2 & c_2 & -b_2 & \ldots & 0 & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & \ldots & -a_{N-2} & c_{N-2} & -b_{N-2} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & -a_{N-1} & c_{N-1} & -b_{N-1} \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & -a_N & c_N \end{pmatrix}$$

Основная идея метода состоит в представлении решения в виде

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}\,, \quad i = N-1, N-2, \ldots, 0\,, \tag{7.6}$$

для которого значения α_i, β_i и y_N вычисляются по коэффициентам исходной системы и правой части. Перепишем первое из уравнений (7.5) в виде (7.6). Имеем

$$y_0 = \alpha_1 y_1 + \beta_1\,, \quad \alpha_1 = \frac{b_0}{c_0}\,, \quad \beta_1 = \frac{f_0}{c_0}\,.$$

Затем к полученному соотношению добавим уравнение из (7.5) при $i = 1$:

$$\begin{aligned} y_0 &= \alpha_1 y_1 + \beta_1\,, \\ -a_1 y_0 + c_1 y_1 - b_1 y_2 &= f_1\,. \end{aligned} \tag{7.7}$$

Исключим из этой системы переменную y_0

$$(c_1 - a_1\alpha_1) y_1 - b_1 y_2 = f_1 + a_1\beta_1$$

и перепишем полученное соотношение в виде (7.6)

$$y_1 = \alpha_2 y_2 + \beta_2\,, \quad \alpha_2 = \frac{b_1}{c_1 - a_1\alpha_1}\,, \quad \beta_2 = \frac{f_1 + a_1\beta_1}{c_1 - a_1\alpha_1}\,.$$

Следующий шаг аналогичен предыдущему: возьмем последнее соотношение и добавим к нему уравнение из (7.5) при $i = 2$

$$\begin{aligned} y_1 &= \alpha_2 y_2 + \beta_2\,, \\ -a_2 y_1 + c_2 y_2 - b_2 y_3 &= f_2\,. \end{aligned}$$

Отличие этой пары уравнений от (7.7) состоит только в увеличении индексов на единицу, поэтому сразу можно написать результат шага

$$y_2 = \alpha_3 y_3 + \beta_3\,, \quad \alpha_3 = \frac{b_2}{c_2 - a_2\alpha_2}\,, \quad \beta_3 = \frac{f_2 + a_2\beta_2}{c_2 - a_2\alpha_2}\,.$$

Таким образом, добавляя каждый раз к последнему полученному соотношению вида (7.6) следующее уравнение из системы (7.5), найдем формулы для вычисления α_i, β_i

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i\alpha_i}\,, \quad \beta_{i+1} = \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i}\,.$$

Этот процесс закончится, когда мы придем к последнему уравнению системы (7.5), содержащему только два значения неизвестных:

$$y_{N-1} = \alpha_N y_N + \beta_N\,,$$
$$-a_N y_{N-1} + c_N y_N = f_N\,.$$

Исключая из этой системы y_{N-1}, получаем

$$y_N = \frac{f_N + a_N\beta_N}{c_N - a_N\alpha_N}\ ,$$

что формально соответствует β_{N+1}.

Полученные соотношения называют *формулами правой прогонки*. Сформулируем алгоритм решения системы (7.5). Рекуррентно вычислить прогоночные коэффициенты α_i, β_i:

$$\alpha_1 = \frac{b_0}{c_0}\,,\quad \alpha_{i+1} = \frac{b_i}{c_i - a_i\alpha_i}\ ,$$
$$\beta_1 = \frac{f_0}{c_0}\,,\quad \beta_{i+1} = \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i}\ ,$$

где i последовательно принимает значения $1, 2, \ldots, N-1$. Эту часть алгоритма называют *прямым ходом* прогонки.

Вычислить y_N:

$$y_N = \frac{f_N + a_N\beta_N}{c_N - a_N\alpha_N}\ .$$

Рекуррентно определить остальные компоненты вектора неизвестных

$$y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1}\,,\quad i = N-1, N-2, \ldots, 0\,.$$

Эту часть алгоритма называют *обратным ходом* прогонки. Данный метод является реализацией метода Гаусса решения систем линейных алгебраических уравнений с трехдиагональными матрицами.

Сформулируем достаточные условия корректности и устойчивости алгоритма.

Теорема. *Пусть коэффициенты системы* (7.5) *действительные и удовлетворяют условиям:* c_0, c_N, a_i, b_i *при* $i = 1, 2, \ldots, N-1$ *отличны от нуля и*

$$|c_i| \geqslant |a_i| + |b_i|\,,\quad i = 1, 2, \ldots, N-1,$$
$$|c_0| \geqslant |b_0|\,,\quad |c_N| \geqslant |a_N|\,,$$

причем хотя бы одно из неравенств является строгим. Тогда для формул метода прогонки справедливы следующие неравенства:

$$c_i - a_i\alpha_i \neq 0\,,\quad |\alpha_i| \leqslant 1\,,\quad i = 1, 2, \ldots, N,$$

гарантирующие корректность и устойчивость метода.

7.27. Для решения системы (7.5) вывести формулы метода прогонки, в которых последующие компоненты вектора неизвестных вычисляются через предыдущие:

$$y_{i+1} = \xi_{i+1}y_i + \eta_{i+1}\,,$$

а прогоночные коэффициенты — наоборот:

$$\xi_i = \varphi(\xi_{i+1}; A)\,, \quad \eta_i = \psi(\eta_{i+1}, \xi_{i+1}; A)\,.$$

Такие соотношения называют *формулами левой прогонки.*

Ответ:
$$\xi_N = \frac{a_N}{c_N}\,, \qquad \xi_i = \frac{a_i}{c_i - b_i\xi_{i+1}}\,, \qquad i = N-1, N-2, \ldots, 1\,,$$
$$\eta_N = \frac{f_N}{c_N}\,, \qquad \eta_i = \frac{f_i + b_i\eta_{i+1}}{c_i - b_i\xi_{i+1}}\,, \qquad i = N-1, N-2, \ldots, 0\,,$$
$$y_0 = \eta_0\,, \qquad y_{i+1} = \xi_{i+1}y_i + \eta_{i+1}\,, \quad i = 0, 1, \ldots, N-1\,.$$

7.28. Для случая коэффициентов системы (7.5)

$$a_N = b_0 = 0\,, \quad a_i = b_i = 1\,, \quad c_i = c\,, \quad i = 1, 2, \ldots, N-1\,,$$

комбинируя алгоритмы правой и левой прогонок, записать формулы для нахождения величины y_M, где $M = \dfrac{N+1}{2}$, N — нечетное.

Ответ: один из возможных вариантов имеет следующий вид:

$$\alpha_1 = 0\,, \qquad \alpha_{i+1} = \frac{1}{c - \alpha_i}\,, \qquad i = 1, 2, \ldots, M-1\,,$$
$$\beta_1 = \frac{f_0}{c_0}\,, \qquad \beta_{i+1} = (f_i + \beta_i)\alpha_{i+1}\,, \qquad i = 1, 2, \ldots, M-1\,,$$
$$\xi_{N-i+1} = \alpha_i\,, \qquad i = 1, 2, \ldots, M\,,$$
$$\eta_N = \frac{f_N}{c_N}\,, \qquad \eta_i = (f_i + \eta_{i+1})\alpha_{N-i+1}\,, \qquad i = N-1, N-2, \ldots, M\,,$$
$$y_M = \frac{\eta_M + \alpha_M\beta_M}{1 - \alpha_M^2}\,.$$

7.29. Можно ли применить метод прогонки, если коэффициенты системы (7.5) имеют вид $c_N = c_0 = 1, c_i = 2, i = 1, 2, \ldots, N-1$, $a_i = 1$, $i = 1, 2, \ldots, N$, $b_i = 1$, $i = 0, 1, \ldots, N-1$.

Ответ: нельзя, знаменатель в формуле для y_N обращается в нуль. Система является вырожденной, так как $A\mathbf{y} = \mathbf{0}$ при $y_i \equiv 1$.

7.30. Записать формулы для решения системы

$$-a_0y_{N-1} + c_0y_0 - b_0y_1 = f_0\,, \quad i = 0\,,$$
$$-a_iy_{i-1} + c_iy_i - b_iy_{i+1} = f_i\,, \quad 1 \leqslant i \leqslant N-1\,,$$
$$y_N = y_0\,.$$

Указание. Решение y_i представить в виде линейной комбинации сеточных функций u_i и v_i

$$y_i = u_i + y_0v_i\,, \quad 0 \leqslant i \leqslant N\,,$$

где u_i — решение неоднородной задачи

$$-a_iu_{i-1} + c_iu_i - b_iu_{i+1} = f_i\,, \quad 1 \leqslant i \leqslant N-1\,, \quad u_N = u_0\,;$$

v_i — решение однородной задачи

$$-a_iv_{i-1} + c_iv_i - b_iv_{i+1} = 0\,, \quad 1 \leqslant i \leqslant N-1\,, \quad v_N = v_0 = 1\,.$$

О т в е т: $\alpha_2 = \frac{b_1}{c_1}$, $\beta_2 = \frac{f_1}{c_1}$, $\gamma_2 = \frac{a_1}{c_1}$,

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i\alpha_i}\,, \quad \beta_{i+1} = \frac{f_i + a_i\beta_i}{c_i - a_i\alpha_i}\,, \quad \gamma_{i+1} = \frac{a_i\gamma_i}{c_i - a_i\alpha_i}\,,$$

где $i = 2, 3, \ldots, N$;

$$u_{N-1} = \beta_N\,, \qquad v_{N-1} = \alpha_N + \gamma_N\,,$$
$$u_i = \alpha_{i+1}u_{i+1} + \beta_{i+1}\,, \qquad v_i = \alpha_{i+1}y_{i+1} + \gamma_{i+1}\,,$$

где $i = N-2, N-3, \ldots, 1$;

$$y_0 = \frac{\beta_{N+1} + \alpha_{N+1}u_1}{1 - \gamma_{N+1} - \alpha_{N+1}v_1}\,, \quad y_i = u_i + y_0v_i\,, \quad 1 \leqslant i \leqslant N\,.$$

Данные соотношения называют *формулами циклической прогонки.*

7.31. Записать формулы пятиточечной прогонки для решения системы

$$\begin{aligned}
&c_0y_0 - d_0y_1 + e_0y_2 = f_0\,, && i = 0\,,\\
&-b_1y_0 + c_1y_1 - d_1y_2 + e_1y_3 = f_1\,, && i = 1\,,\\
&a_iy_{i-2} - b_iy_{i-1} + c_iy_i - d_iy_{i+1} + e_iy_{i+2} = f_i\,, && 2 \leqslant i \leqslant N-2\,,\\
&a_{N-1}y_{N-3} - b_{N-1}y_{N-2} + c_{N-1}y_{N-1} - d_{N-1}y_N = f_{N-1}\,, && i = N-1\,,\\
&a_Ny_{N-2} - b_Ny_{N-1} + c_Ny_N = f_N\,, && i = N\,.
\end{aligned}$$

У к а з а н и е. Решение y_i следует искать в виде

$$y_i = \alpha_{i+1}y_{i+1} - \beta_{i+1}y_{i+2} + \gamma_{i+1}\,, \quad 0 \leqslant i \leqslant N-2\,,$$
$$y_{N-1} = \alpha_Ny_N + \gamma_N\,.$$

О т в е т: формулы для прогоночных коэффициентов имеют вид

$$\alpha_1 = \frac{d_0}{c_0}\,, \quad \alpha_2 = \frac{1}{\Delta_1}\,(d_1 - \beta_1b_1)\,,$$
$$\alpha_{i+1} = \frac{1}{\Delta_i}\,[d_i + \beta_i(a_i\alpha_{i-1} - b_i)]\,, \quad i = 2, 3, \ldots, N-1\,;$$
$$\gamma_1 = \frac{f_0}{c_0}\,, \quad \gamma_2 = \frac{1}{\Delta_1}\,(f_1 + \gamma_1b_1)\,,$$
$$\gamma_{i+1} = \frac{1}{\Delta_i}\,[f_i - a_i\gamma_{i-1} - \gamma_i(a_i\alpha_{i-1} - b_i)]\,, \quad i = 2, 3, \ldots, N\,;$$
$$\beta_1 = \frac{e_0}{c_0}\,, \quad \beta_{i+1} = \frac{e_i}{\Delta_i}\,, \quad i = 1, 2, \ldots, N-2\,,$$

где $\Delta_1 = c_1 - \alpha_1b_1$, $\Delta_i = c_i - a_i\beta_{i-1} + \alpha_i(a_i\alpha_{i-1} - b_i)$, $2 \leqslant i \leqslant N$.

Формулы для решения таковы:

$$y_i = \alpha_{i+1}y_{i+1} - \beta_{i+1}y_{i+2} + \gamma_{i+1}\,, \quad i = N-2, N-3, \ldots, 0\,,$$
$$y_{N-1} = \alpha_Ny_N + \gamma_N\,, \quad y_N = \gamma_{N+1}\,.$$

Приведенный алгоритм является реализацией метода Гаусса решения систем с пятидиагональными матрицами.

Метод стрельбы. Идею этого подхода наиболее просто изложить в терминах дифференциальных уравнений. Пусть требуется решить краевую задачу

$$u'' - p(x)\,u = f(x)\,, 0 < x < 1\,, u(0) = a\,, u(1) = b\,.$$

Построим частное решение $u_1(x)$ неоднородного уравнения

$$u_1'' - p(x)\,u_1 = f(x)\,,$$

удовлетворяющее условию $u_1(0) = a$, и какое-либо нетривиальное частное решение $u_2(x) \neq 0$ однородного уравнения

$$u_2'' - p(x)\,u_2 = 0\,,$$

удовлетворяющее условию $u_2(0) = 0$. Решение исходной задачи будем искать в следующем виде:

$$u(x) = u_1(x) + C\,u_2(x),$$

где постоянная C определяется из условия

$$u_1(1) + C\,u_2(1) = b\,.$$

Применим близкую идею к решению системы (7.5). Будем искать решение y_i в виде

$$y_i = \delta\,u_i + (1-\delta)\,v_i\,,$$

где δ — параметр, подлежащий определению, а сеточные функции u_i и v_i удовлетворяют уравнениям

$$c_0u_0 - b_0u_1 = f_0\,, c_0v_0 - b_0v_1 = f_0 \quad \text{при } i = 0\,,$$
$$-a_iu_{i-1} + c_iu_i - b_iu_{i+1} = f_i\,, -a_iv_{i-1} + c_iv_i - b_iv_{i+1} = f_i \text{ при } 1 \leqslant i \leqslant N-1.$$

К этим системам для однозначного определения u_i и v_i необходимо добавить при $b_0 \neq 0$ начальные условия u_0 и v_0 ($u_0 \neq v_0$). Если $b_0 = 0$, то добавляют значения u_1 и v_1 ($u_1 \neq v_1$). Теперь можно последовательно определить $u_2, u_3, \ldots, u_N$ и $v_2, v_3, \ldots, v_N$. Неизвестный параметр δ найдем из уравнения

$$-a_N(\delta\,u_{N-1} + (1-\delta)\,v_{N-1}) + c_N(\delta\,u_N + (1-\delta)\,v_N) = f_N\,,$$

т. е.

$$\delta = \frac{f_N + a_Nv_{N-1} - c_Nv_N}{a_N(v_{N-1} - u_{N-1}) + c_N(u_N - v_N)}\,.$$

Метод стрельбы — хорошее дополнение к методу прогонки: области их корректности и устойчивости практически не пересекаются.

7.32. Для случая постоянных коэффициентов системы (7.5): $c_0 = 1$, $b_0 = 0$, $f_0 = 3$, $a_i = 1$, $c_i = \frac{26}{5}$, $b_i = 1$, $f_i = 0$, $a_N = 0$, $c_N = 1$, $f_N = 4$, найти решение методом стрельбы и проанализировать его устойчивость.

$\lhd$ Рассмотрим вспомогательные функции u_i и v_i. Из исходной системы имеем $u_0 = 3$. Так как $b_0 = 0$, положим $u_1 = \varphi$. Далее находим

$$u_{i+1} = \frac{26}{5}\,u_i - u_{i-1}\,, \quad i = 1, 2, \ldots, N-1\,.$$

Это решение можно представить в виде

$$u_i = \frac{5\,\varphi - 3}{24}\ 5^i + \frac{75 - 5\,\varphi}{24}\ 5^{-i}\,, \quad i = 0, 1, \ldots, N\,.$$

Аналогично, полагая $v_1 = \psi$, приходим к формуле

$$v_i = \frac{5\,\psi - 3}{24}\ 5^i + \frac{75 - 5\,\psi}{24}\ 5^{-i}\,, \quad i = 0, 1, \ldots, N\,.$$

Используя вычисленные u_N и v_N, определим δ из уравнения

$$\delta\, u_N + (1 - \delta)\, v_N = 4\,;$$

подставляя его значение в выражение $y_i = \delta\, u_i + (1 - \delta)\, v_i$, получаем

$$y_i = 3\ \frac{5^{N-i} - 5^{i-N}}{5^N - 5^{-N}} + 4\ \frac{5^i - 5^{-i}}{5^N - 5^{-N}}\ .$$

В данном случае алгоритм является вычислительно неустойчивым. Действительно, $\max\limits_i |u_i|$ и $\max\limits_i |v_i|$ растут, как 5^N. Поэтому малым возмущениям значений $u_1 = \varphi$ и $v_1 = \psi$ соответствуют большие возмущения в u_N и v_N, следовательно, и в величине δ. Для исходной системы выполнены достаточные условия корректности и устойчивости метода прогонки, который и является здесь предпочтительным для нахождения y_i. $\triangleright$

7.33. Методом стрельбы найти решение системы

$$\begin{aligned} y_0 - y_1 &= 0\,, && i = 0\,,\\ y_{i-1} - y_i + y_{i+1} &= 0\,, && 1 \leqslant i \leqslant N - 1\,,\\ y_N &= 1\,, && i = N\,. \end{aligned}$$

Проанализировать устойчивость и корректность метода.

$\triangleleft$ В исходной системе $b_0 \neq 0$, поэтому положим $y_0 = \varphi$. Далее находим

$$y_1 = y_0\,, \quad y_{i+1} = y_i - y_{i-1}\,, \quad 1 \leqslant i \leqslant N - 1\,.$$

Общая формула решения этой задачи Коши, зависящего от величины φ, имеет вид

$$y_i = \varphi \left[\cos\frac{i\,\pi}{3} + \frac{1}{\sqrt{3}} \sin\frac{i\,\pi}{3}\right], \quad 0 \leqslant i \leqslant N\,.$$

Для постоянной φ имеем уравнение

$$1 = y_N = \varphi \left[\cos\frac{N\,\pi}{3} + \frac{1}{\sqrt{3}} \sin\frac{N\,\pi}{3}\right],$$

которое однозначно разрешимо при $N \neq -1 + 3\,k\,, k = 2, 3, \ldots$. Это ограничение для N является условием применимости (корректности) метода стрельбы. Сам алгоритм является вычислительно устойчивым, так как корни характеристического уравнения

$$\mu^2 - \mu + 1 = 0$$

комплексно сопряжены и по модулю равны единице, следовательно, не приводят к росту возмущений начальных данных. $\triangleright$

7.34. Для краевой задачи

$$-u'' + u = f(x)\,, \quad 0 < x < 1\,, \quad u(0) + u'(0) = a\,, \quad u(1) + u'(1) = b\,,$$

построить трехточечную разностную схему порядка аппроксимации $O(h^2)$ и проанализировать устойчивость метода стрельбы для нахождения ее решения.

7.35. Для задачи

$$-u'' + u = f(x)\,, \quad 0 < x < 1\,, \quad u(0) = a\,, \quad \int\limits_0^1 u(x)\,dx = b\,,$$

построить разностную схему и предложить метод нахождения ее решения.

Метод Фурье (базисных функций). Рассмотрим метод решения системы линейных уравнений $A\mathbf{y} = \mathbf{f}$, $\mathbf{y}, \mathbf{f} \in \mathbf{R}^N$, при условии, что известны все собственные векторы и собственные значения матрицы A:

$$A\varphi^{(n)} = \lambda^{(n)}\varphi^{(n)},\ n = 1, ..., N,$$

и система $\{\varphi^{(n)}\}$ образует ортонормированный базис в пространстве $\mathbf{R}^N$. Будем искать решение в виде $\mathbf{y} = \sum\limits_{n=1}^{N} c_n\varphi^{(n)}$. Подставим данное разложение в исходную систему уравнений

$$A\left(\sum_{n=1}^{N} c_n\varphi^{(n)}\right) = \sum_{n=1}^{N} c_n\lambda^{(n)}\varphi^{(n)} = \mathbf{f}.$$

Умножая последнее равенство скалярно на $\varphi^{(m)}$, $m = 1, \ldots, N$, и учитывая ортонормированность базиса, получим

$$\left(\sum_{n=1}^{N} \lambda^{(n)} c_n\varphi^{(n)}, \varphi^{(m)}\right) = (\mathbf{f}, \varphi^{(m)})\,,$$

т. е. $c_m\lambda^{(m)} = (\mathbf{f}, \varphi^{(m)})$. Отсюда находим коэффициенты $c_m = \dfrac{(\mathbf{f}, \varphi^{(m)})}{\lambda^{(m)}}$, $m = 1, \ldots, N$, и затем вычисляем вектор $\mathbf{y}$.

Проблема нахождения собственных векторов и собственных значений в общем случае значительно сложнее решения системы линейных уравнений, поэтому данный метод применяют для задач с известными собственными векторами и собственными значениями. Например, для решения задач, возникающих при аппроксимации уравнений в частных производных.

7.36. Найти методом Фурье решение задачи

$$-\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = f_i,\ i = 1, ..., N-1,\ y_0 = y_N = 0,\ h = \frac{1}{N}.$$

У к а з а н и е. В данном случае собственные векторы и собственные значения можно найти аналитически:

$$\varphi_i^{(n)} = \sqrt{2}\sin(\pi nih), \quad \lambda^{(n)} = \frac{4}{h^2}\sin^2\left(\frac{\pi nh}{2}\right),\ n = 1, \ldots, N-1.$$

При этом $\varphi^{(n)}$ ортонормированы относительно стандартного скалярного произведения $(\varphi^{(n)}, \varphi^{(m)}) = \sum\limits_{i=1}^{N-1} \varphi_i^{(n)}\varphi_i^{(m)}h$.

7.37. Найти методом Фурье решение задачи

$$-\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} = f_i, \quad i = 1, ..., N-1,\ h = \frac{1}{N},$$

$$-\frac{2}{h^2}(y_1 - y_0) = f_0, \quad \frac{2}{h^2}(y_N - y_{N-1}) = f_N.$$

У к а з а н и е. Собственные векторы и собственные значения имеют вид

$$\varphi_i^{(0)} = 1,\ \varphi_i^{(N)} = \cos(\pi Nih) = (-1)^i,$$
$$\varphi_i^{(n)} = \sqrt{2}\cos(\pi nih),\ n = 1, \ldots N-1,$$
$$\lambda^{(0)} = 0,\ \lambda^{(n)} = \frac{4}{h^2}\sin^2\left(\frac{\pi nh}{2}\right),\ n = 1, \ldots N.$$

При этом $\varphi^{(n)}$ ортонормированы относительно следующего скалярного произведения:

$$(\varphi^{(n)}, \varphi^{(m)}) = \sum_{i=1}^{N-1} \varphi_i^{(n)}\varphi_i^{(m)}h + \frac{h}{2}\left(\varphi_0^{(n)}\varphi_0^{(m)} + \varphi_N^{(n)}\varphi_N^{(m)}\right).$$

Глава 8

Дифференциальные уравнения

В главе рассмотрено численное решение обыкновенных дифференциальных уравнений первого и второго порядка. Общая теория разностных схем применена для построения дискретных аналогов дифференциальных задач с начальными или краевыми условиями. Конкретизированы понятия аппроксимации, устойчивости и сходимости. Особое внимание уделено исследованию методов решения и оценкам погрешности.

8.1. Задача Коши

Конкретизируем в случае задачи Коши для обыкновенного дифференциального уравнения

$$y' = f(x, y), \tag{8.1}$$

$$y(x_0) = y_0, \tag{8.2}$$

общие понятия разностного метода. Пусть, для простоты, рассматривается равномерная сетка $x_k = x_0 + kh$, $k \geqslant 0$. Тогда *разностной схемой для задачи (8.1), (8.2)* называют семейство разностных уравнений

$$\frac{1}{h}\sum_{i=0}^{n} a_{-i} y_{k-i} = \sum_{i=0}^{n} b_{-i} f_{k-i}, \quad k = n, n+1, \ldots, \tag{8.3}$$

с известными начальными условиями $y_0 = y(x_0)$, $y_1, \ldots, y_{n-1}$, где a_{-i}, b_{-i} не зависят от h, $a_0 \neq 0$ и $f_{k-i} = f(x_{k-i}, y_{k-i})$.

Разностная задача (8.3) аппроксимирует на решении (8.1) дифференциальную на отрезке $[x_0, x_0 + X]$ с порядком p, если для функции погрешности

$$r_k^h = \frac{1}{h}\sum_{i=0}^{n} a_{-i} y(x_{k-i}) - \sum_{i=0}^{n} b_{-i} f(x_{k-i}, y(x_{k-i}))$$

справедлива оценка $\|r^h\|_{F_h} \leqslant ch^p$ и выполнено условие нормировки $\lim\limits_{h\to 0} \| f_h \|_{F_h} = \| f \|_F$. Напомним, что постоянные c и p не зависят от шага h.

В общем случае задача (8.3) — нелинейная система, поэтому аппроксимацию левой и правой частей уравнения (8.1) нужно рассматривать отдельно. При оценке порядка аппроксимации разностной схемы следует также учитывать порядок, с которым начальные условия аппроксимируют значения точного решения задачи (8.1), (8.2) в соответствующих узлах сетки. Если рассматривается только уравнение (8.1) без начального условия (8.2), то под разностной схемой понимают систему (8.3), а ее начальные условия во внимание не принимают.

Рассмотрим характеристическое уравнение для левой части разностной схемы (фактически для аппроксимации уравнения $y' = 0$):

$$F(\mu) \equiv \sum_{i=0}^{n} a_{-i}\mu^{n-i} = 0\,.$$

Схема называется *α-устойчивой*, если выполнено следующее условие: все корни характеристического уравнения принадлежат единичному кругу и на границе круга нет кратных корней. Это условие является необходимым. Можно показать, что для любой разностной схемы, не удовлетворяющей условию α-устойчивости, существует дифференциальное уравнение с бесконечно-дифференцируемой правой частью, для которого даже при отсутствии округлений и погрешностей в начальных данных, решение его разностного аналога не стремится к непрерывному решению при измельчении шага.

Если в задаче не приведен конкретный вид правой части, то устойчивость понимают в смысле α-устойчивости.

8.1. Показать, что необходимым и достаточным условием аппроксимации уравнения (8.1) разностными уравнениями (8.3) является выполнение равенств: $\sum\limits_{i=0}^{n} a_{-i} = 0,\ -\sum\limits_{i=0}^{n} ia_{-i} = 1,\ \sum\limits_{i=0}^{n} b_{-i} = 1\,.$

$\lhd$ Пусть $y(x)$ — произвольная гладкая функция. Тогда условия аппроксимации для левой и правой частей уравнения (8.1) означает справедливость соотношений в произвольном узле x_k, $k \geqslant n$:

$$\lim_{h\to 0} \frac{1}{h}\sum_{i=0}^{n} a_{-i}y_{k-i} = y'(x_k)\,,\ \lim_{h\to 0} \sum_{i=0}^{n} b_{-i}f(x_{k-i}, y_{k-i}) = f(x_k, y_k).$$

Согласно формуле Тейлора,

$$y(x - ih) = y(x) - ihy'(x) + O(h^2),$$
$$f(x - ih, y(x - ih)) = f(x, y(x)) + O(h).$$

Подставляем эти выражения в условия аппроксимации, имеем

$$\lim_{h\to 0} \left[\left(\frac{1}{h}\sum_{i=0}^{n} a_{-i}\right) y(x_k) - \left(\sum_{i=0}^{n} ia_{-i}\right) y'(x_k) + O(h)\right] = y'(x_k),$$

$$\lim_{h\to 0} \left[\left(\sum_{i=0}^{n} b_{-i}\right) f(x_k, y(x_k)) + O(h)\right] = f(x_k, y(x_k)),$$

откуда в силу произвольности функции $y(x)$ и следует необходимость и достаточность указанных в условии задачи равенств. $\rhd$

8.2. Проверить, аппроксимирует ли разностная схема уравнение (8.1):

1) $\frac{1}{h}(y_k - y_{k-1}) = f_{k-1}$; 2) $\frac{1}{h}(y_k - y_{k-1}) = \frac{1}{2}(f_k + f_{k-1})$;

3) $\frac{1}{h}(y_k - y_{k-1}) = \frac{1}{2}(3f_{k-1} - f_{k-2})$; 4) $\frac{1}{3h}(y_k - y_{k-3}) = f_{k-1}$;

5) $\frac{1}{8h}(y_k - 3y_{k-2} + 2y_{k-3}) = \frac{1}{2}(f_{k-1} + f_{k-2})$; 6) $\frac{1}{2h}(3y_k - 4y_{k-1} + y_{k-2}) = f_k$.

Указание. Использовать условия, сформулированные в 8.1.

Ответ: 1) да; 2) да; 3) да; 4) да; 5) нет; 6) да.

8.3. Для задачи $y'+y=x+1$, $y(0)=0$ рассматривается схема

$$\frac{y_{k+1}-y_k}{h}+\frac{y_{k+1}+y_k}{2}=\left(k+\frac{1}{2}\right)h+1, \quad y_0=0.$$

Каков порядок аппроксимации на решении данной схемы?

Ответ: второй.

8.4. Для задачи $y'+y=x+1$, $y(0)=0$ рассматривается схема

$$\frac{y_{k+1}-y_{k-1}}{2h}+y_k=kh+1, \quad y_0=0, \quad y_1=0.$$

Каков порядок аппроксимации на решении данной схемы? Можно ли его улучшить?

Ответ: первый; можно, если положить $y_1=h$, то порядок аппроксимации равен двум. В отличие от дифференциального случая для разностной задачи необходимы два начальных условия. Поэтому аппроксимация решения в точке $x=h$ — часть формальной аппроксимации дифференциального оператора L.

8.5. Пусть для решения задачи $y'+5y=5$, $y(0)=2$ построена следующая разностная схема:

$$\frac{y_{k+1}-y_{k-1}}{2h}+5y_k=5, \quad y_0=2,\ y_1=2-5h.$$

Исследовать ее аппроксимацию и сходимость.

$\lhd$ Схема имеет второй порядок аппроксимации на решении. Проанализируем сходимость. Несложно показать, что точные решения дифференциальной и разностной задач имеют вид

$$y(x)=\mathrm{e}^{-5x}+1,$$

$$y_k=1+C_1\mu_1^k+C_2\mu_2^k, \quad \mu_{1,2}=-5h\pm\sqrt{1+25h^2},\ |\mu_1|<1, |\mu_2|>1.$$

Так как коэффициенты C_1, C_2 находятся из начальных условий y_0, y_1:

$$1+C_1+C_2=2, \quad 1+C_1\mu_1+C_2\mu_2=2-5h,$$

то имеем $C_1, C_2\neq 0$. Следовательно, решение разностной задачи содержит растущую компоненту, и разностная схема на больших промежутках времени неверно отражает решение дифференциальной задачи, хотя схема α-устойчива и разностное решение сходится на любом конечном интервале к решению дифференциальной задачи. $\rhd$

8.6. Для задачи

$$y'+a(x)y=f(x), \quad y(0)=c$$

рассматривается схема

$$\frac{y_{k+1}-y_k}{h}+(\alpha_1 a(x_k)+\alpha_2 a(x_{k+1}))\,(\beta_1 y_k+\beta_2 y_{k+1})=\gamma_1 f(x_k)+\gamma_2 f(x_{k+1}),$$

$$y_0=c.$$

Какими следует выбрать α_k, β_k и γ_k, чтобы получить второй порядок аппроксимации на решении?

Ответ: все коэффициенты равны $\frac{1}{2}$.

8.7. Построить для уравнения (8.1) разностную схему с наивысшим порядком аппроксимации p на решении

$$\frac{y_k - y_{k-2}}{2h} = a_1 f_k + a_0 f_{k-1} + a_{-1} f_{k-2}.$$

Указание. Использовать метод неопределенных коэффициентов построения разностных схем, заменив f на y' и сдвинув (для удобства вычислений) индексы заменой $j = k-1$.

Ответ: $a_1 = a_{-1} = \frac{1}{6}, a_0 = \frac{2}{3}, p = 4$.

8.8. Исследовать устойчивость разностной схемы

$$\theta\ \frac{y_{k+1} - y_k}{h} + (1-\theta)\ \frac{y_k - y_{k-1}}{h} = f_k \quad \text{при} \quad \theta \in [0,1].$$

Ответ: схема устойчива при $\theta = 0$ и $1 \geqslant \theta \geqslant \frac{1}{2}$.

8.9. При каких a, b и c схема

$$\frac{1}{h}\,(y_k + a y_{k-1} - a y_{k-3} - y_{k-4}) = b f_{k-1} + c f_{k-2} + b f_{k-3}$$

для уравнения $y' = f$ имеет максимальный порядок аппроксимации на решении? Выполнено ли условие α-устойчивости?

$\lhd$ Учитывая необходимые условия аппроксимации (см. 8.1), запишем систему для определения коэффициентов

$$2a + 4 = 1, \quad 2b + c = 1, \quad 8 + a = 3b,$$

или $a = -\frac{3}{2}$, $b = \frac{13}{6}$, $c = -\frac{10}{3}$. При этом характеристическое уравнение имеет вид

$$(\mu^2 - 1)\ \left(\mu^2 - \frac{3}{2}\ \mu + 1\right) = 0,$$

т. е. условие α-устойчивости выполнено.

Без учета нормировки $\lim\limits_{h \to 0} \| f_h \|_{F_h} = \| f \|_F$ можно прийти к неверному ответу: $a = 28, b = 12, c = 36$, для которого условие α-устойчивости не выполнено. $\rhd$

8.10. Исследовать сходимость решения разностной схемы

$$\frac{\varphi_k - \varphi_{k-1}}{h} + l\psi_{k-1} = 0, \quad \varphi_0 = a, \quad h = \frac{1}{N}\ ,$$

$$\frac{\psi_k - \psi_{k-1}}{h} - l\varphi_{k-1} = 0, \quad \psi_0 = b, \quad k = 1, \ldots, N,$$

к решению дифференциальной задачи

$$u'+lv=0\,,\quad u(0)=a\,,$$
$$v'-lu=0\,,\quad v(0)=b$$

на отрезке $x\in[0,1]$ при $l=\mathrm{const}\neq 0$, используя решения обеих задач.

$\lhd$ Запишем дифференциальную задачу в виде

$$\mathbf{y}'=-A\mathbf{y},\quad \mathbf{y}(0)=\mathbf{d},$$

где

$$\mathbf{y}=\begin{pmatrix}u\\ v\end{pmatrix},\quad A=\begin{pmatrix}0 & l\\ -l & 0\end{pmatrix},\quad \mathbf{d}=\begin{pmatrix}a\\ b\end{pmatrix}.$$

Тогда

$$\mathbf{y}=\exp(-Ax)\,\mathbf{d}.$$

Так как $\lambda_{1,2}(A)=\pm\,\mathrm{i}\,l$, то, обозначив через X матрицу, столбцами которой являются собственные векторы матрицы A, получаем

$$\mathbf{y}=X\begin{pmatrix}\mathrm{e}^{-\lambda_1 x} & 0\\ 0 & \mathrm{e}^{-\lambda_2 x}\end{pmatrix}X^{-1}\mathbf{d}.$$

Для нахождения решения разностной задачи представим ее в виде

$$\mathbf{y}_k^h=A_h\mathbf{y}_{k-1}^h,\quad k=1,\,\ldots\,,N,\quad \mathbf{y}_0^h=\mathbf{d},$$

где

$$\mathbf{y}_k^h=\begin{pmatrix}\varphi_k\\ \psi_k\end{pmatrix},\quad A_h=I-hA=\begin{pmatrix}1 & -hl\\ hl & 1\end{pmatrix}.$$

Так как $\mathbf{y}_k^h=(A_h)^k\,\mathbf{y}_0^h$, то

$$\mathbf{y}_k^h=X\begin{pmatrix}(1-\mathrm{i}lh)^k & 0\\ 0 & (1+\mathrm{i}lh)^k\end{pmatrix}X^{-1}\mathbf{d}\,.$$

При нахождении $(A_h)^k$ использовано совпадение собственных векторов матриц A_h и A и связь между их собственными числами

$$\lambda(A_h)=1-h\lambda(A)\,.$$

Можно показать, что $\exp(\pm\mathrm{i}lx_k)-(1\pm\mathrm{i}lh)^k=O(h)$, и так как по условию $kh\leqslant 1$, то для $k=1,2,\ldots,N$ имеем $\|\mathbf{y}(x_k)-\mathbf{y}_k^h\|_\infty=O(h)$. Вводя в пространстве $\mathbf{Y}_h$ норму

$$\|\mathbf{y}_h\|_{\mathbf{Y}_h}=\max_{0\leqslant k\leqslant N}(\|\mathbf{y}_k^h\|_\infty),$$

приходим к следующей оценке сходимости решения разностной схемы к решению дифференциальной задачи

$$\|(\mathbf{y})_h-\mathbf{y}_h\|_{\mathbf{Y}_h}=O(h).$$

Таким образом, схема имеет первый порядок сходимости. $\rhd$

8.11. Для задачи $y'=y$, $y(0)=1$ рассмотрим схему

$$\frac{y_{k+1}-y_k}{h}=y_k,\quad y_0=1,\quad k\geqslant 0\,.$$

В разложении ошибки $y(x_N)-y_N=c_1h+c_2h^2+\cdots$ найти постоянную c_1 для $x_N=Nh=1$.

$\lhd$ Для разностной задачи имеем

$$y_N = (1+h)y_{N-1} = (1+h)^N y_0 = (1+h)^N,$$

а точное решение дифференциальной задачи при $x = x_N$ равно $y(x_N) = \exp(x_N)$. Пусть $x_N = N\,h = 1$, тогда

$$y(x_N) - y_N = \mathrm{e} - (1+h)^{1/h} = \mathrm{e} - \exp\left[\frac{1}{h}\ln(1+h)\right] =$$

$$= \mathrm{e}\,\left(1 - \exp\left[-\frac{h}{2} + O(h^2)\right]\right) = \frac{\mathrm{e}}{2}\,h + O(h^2). \qquad \rhd$$

Ответ: $c_1 = \frac{\mathrm{e}}{2}$.

8.12. Для задачи $y' = y$, $y(0) = 1$ рассмотрим схему

$$\frac{y_{k+1} - y_k}{h} = \frac{y_{k+1} + y_k}{2}, \quad y_0 = 1, \quad k \geqslant 0.$$

В разложении ошибки $y(x_N) - y_N = c_1 h + c_2 h^2 + \cdots$ найти постоянную c_1 для $x_N = Nh = 1$.

Ответ: $c_1 = 0$.

8.13. Для задачи $y' = y$, $y(0) = 1$ рассмотрим схему

$$\frac{y_{k+1} - y_{k-1}}{2h} = y_k, \quad y_0 = 1, \quad y_1 = \mathrm{e}^h, \quad k \geqslant 1.$$

В разложении ошибки $y(x_N) - y_N = c_1 h + c_2 h^2 + \cdots$ найти постоянную c_1 для $x_N = Nh = 1$.

Указание. Вывести формулу

$$y_k = y_0\left[\frac{\mu_2}{\mu_2 - \mu_1}\,\mu_1^k - \frac{\mu_1}{\mu_2 - \mu_1}\,\mu_2^k\right] + y_1\left[-\frac{1}{\mu_2 - \mu_1}\,\mu_1^k + \frac{1}{\mu_2 - \mu_1}\,\mu_2^k\right],$$

где $\mu_{1,2}$ — корни уравнения $\mu^2 + 2h\mu - 1 = 0$:

$$\mu_1 = -h + \sqrt{1+h^2} = 1 - h + \frac{h^2}{2} + O(h^4),\ \mu_2 = -\left(1 + h + \frac{h^2}{2}\right) + O(h^4).$$

Ответ: $c_1 = 0$.

8.14. Для задачи $y' = y$, $y(0) = 1$ рассмотрим схему

$$4\,\frac{y_{k+1} - y_{k-1}}{2h} - 3\,\frac{y_{k+1} - y_k}{h} = y_k, \quad y_0 = 1, \quad y_1 = \mathrm{e}^h, \quad k \geqslant 1.$$

В разложении ошибки $y(x_N) - y_N = c_1 h + c_2 h^2 + \cdots$ найти постоянные c_1 и c_2 для $x_N = Nh = 1$.

Ответ: эта схема неустойчива, сходимости нет.

8.15. Для задачи $y' + y = \cos 2x$, $y(0) = 0$, построить трехточечную разностную схему второго порядка сходимости.

Ответ: например,

$$\frac{y_{k+1} - y_{k-1}}{2h} + y_k = \cos(2hk), \quad y_0 = 0, \quad y_1 = h, \quad k \geqslant 1.$$

8.16. Для задачи $y' + 5y = \sin 2x$, $y(0) = 2$, построить двухточечную разностную схему второго порядка сходимости.

Ответ: например,

$$\frac{y_{k+1} - y_k}{h} + 5\,\frac{y_{k+1} + y_k}{2} = \frac{\sin(2h(k+1)) + \sin(2hk)}{2}\,, \quad y_0 = 2, \quad k \geqslant 0.$$

8.17. Для задачи $y' - y = \exp 2x$, $y(0) = 1$, построить трехточечную разностную схему второго порядка сходимости.

Ответ: например,

$$\frac{y_{k+1} - y_{k-1}}{2h} - y_k = \exp(2hk), \quad y_0 = 1, \quad y_1 = 1 + 2h, \quad k \geqslant 1.$$

8.18. Для задачи $y' - 2y = \exp x$, $y(0) = 1$, построить двухточечную разностную схему второго порядка сходимости.

Ответ: например,

$$\frac{y_{k+1} - y_k}{h} - (y_{k+1} + y_k) = \frac{\exp(h(k+1)) + \exp(hk)}{2}\,, \quad y_0 = 1, \quad k \geqslant 0.$$

8.19. Привести пример неустойчивой разностной схемы, аппроксимирующей уравнение $y' = f(x, y)$ строго: 1) с первым порядком; 2) со вторым порядком; 3) с третьим порядком.

Указание. Например, можно взять заведомо α-неустойчивую схему

$$4\,\frac{y_{k+1} - y_{k-1}}{2h} - 3\,\frac{y_{k+1} - y_k}{h} = c\,f_{k-1} + d\,f_k + e\,f_{k+1}$$

и методом неопределенных коэффициентов получить заданный порядок аппроксимации.

8.20. Найти главный член погрешности аппроксимации на решении и исследовать устойчивость разностной схемы

$$\frac{y_k - y_{k-2}}{2h} = \frac{f_k + 4f_{k-1} + f_{k-2}}{6}\,.$$

Ответ: $-\,\frac{h^4}{180}\,y^{(5)}(\xi)$, схема α-устойчива.

8.21. Найти главный член погрешности аппроксимации на решении и исследовать устойчивость разностной схемы

$$\frac{y_{k+1} - y_k}{h} = \frac{5\,f_k + 8\,f_{k+1} - f_{k+2}}{12}\,.$$

Ответ: $\frac{h^3}{24}\,y^{(4)}(\xi)$, схема α-устойчива.

8.22. Найти главный член погрешности аппроксимации на решении и исследовать устойчивость разностной схемы

$$\frac{y_{k+4} - y_k}{4h} = \frac{2\,f_{k+1} - f_{k+2} + 2\,f_{k+3}}{3}\,.$$

Ответ: $\frac{7\,h^4}{90}\,y^{(5)}(\xi)$, схема α-устойчива.

8.23. Найти главный член погрешности аппроксимации на решении и исследовать устойчивость разностной схемы:

$$\frac{y_k + 4y_{k-1} - 5y_{k-2}}{6h} = \frac{2f_{k-1} + f_{k-2}}{3}\,.$$

Методы Рунге—Кутты и Адамса. Один из наиболее популярных подходов к решению задачи Коши для уравнений первого порядка $y' = f(x, y)$, $y(x_0) = y_0$ заключается в следующем. Зафиксируем некоторые числа $\alpha_2, \ldots, \alpha_q$, $p_1, \ldots, p_q$, $\beta_{i,j}$, $0 < j < i \leqslant q$, и последовательно вычислим

$$k_1(h) = hf(x, y),$$
$$k_2(h) = hf(x + \alpha_2 h, y + \beta_{2,1} k_1(h)),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$k_q(h) = hf(x + \alpha_q h, y + \beta_{q,1} k_1(h) + \cdots + \beta_{q,q-1} k_{q-1}(h)).$$

Расчетная формула имеет вид

$$y(x + h) \approx z(h) = y(x) + \sum_{i=1}^{q} p_i k_i(h).$$

Обозначим погрешность метода на шаге через $\varphi(h) = y(x+h) - z(h)$. Если $f(x, y)$ — достаточно гладкая функция своих аргументов, то справедлива формула Тейлора

$$\varphi(h) = \sum_{i=0}^{s} \frac{\varphi^{(i)}(0)}{i!}\, h^i + \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!}\, h^{s+1},$$

где $0 < \theta < 1$. Выберем параметры метода α_i, p_i, $\beta_{i,j}$ так, что $\varphi'(0) = \cdots = \varphi^{(s)}(0) = 0$. Тогда величина s называется *порядком метода.*

8.24. Построить метод при $q = 1$ и записать формулу погрешности.

$\lhd$ Имеем

$$\varphi(h) = y(x + h) - y(x) - p_1 hf(x, y), \quad \varphi(0) = 0,$$
$$\varphi'(0) = (y'(x + h) - p_1 f(x, y))|_{h=0} = f(x, y)(1 - p_1),$$
$$\varphi''(h) = y''(x + h).$$

Равенство $\varphi'(0) = 0$ выполняется для всех гладких функций $f(x, y)$ только в случае $p_1 = 1$. Для погрешности этого метода на шаге получаем выражение

$$\varphi(h) = \frac{y''(x + \theta h)h^2}{2}\,. \qquad \rhd$$

8.25. Построить все методы при $q = 2$.

$\lhd$ Запишем расчетную формулу в виде

$$\varphi(h) = y(x + h) - y(x) - p_1 hf(x, y) - p_2 hf(\bar{x}, \bar{y}),$$

где $\bar{x} = x + \alpha_2 h$, $\bar{y} = \beta_{21} h f(x, y)$. Вычислим производные функции $\varphi(h)$:

$$\varphi'(h)=y'(x+h)-p_1 f(x,y)-p_2 f(\bar{x},\bar{y})-p_2 h(\alpha_2 f_x(\bar{x},\bar{y})+\beta_{21} f_y(\bar{x},\bar{y}) f(x,y)),$$
$$\varphi''(h)=y''(x+h)-2p_2(\alpha_2 f_x(\bar{x},\bar{y})+\beta_{21} f_y(\bar{x},\bar{y}) f(x,y))-p_2 h(\alpha_2^2 f_{xx}(\bar{x},\bar{y})+$$
$$+2\alpha_2\beta_{21} f_{xy}(\bar{x},\bar{y}) f(x,y)+\beta_{21}^2 f_{yy}(\bar{x},\bar{y})(f(x,y))^2),$$
$$\varphi'''(h)=y'''(x+h)-3p_2(\alpha_2^2 f_{xx}(\bar{x},\bar{y})+2\alpha_2\beta_{21} f_{xy}(\bar{x},\bar{y}) f(x,y))+$$
$$+\beta_{21}^2 f_{yy}(\bar{x},\bar{y})(f(x,y))^2)+O(h).$$

Согласно исходному дифференциальному уравнению

$$y'=f, \quad y''=f_x+f_y f, \quad y'''=f_{xx}+2f_{xy} f+f_{yy} f^2+f_y y''.$$

Подставим в выражения $\varphi(h)$, $\varphi'(h)$, $\varphi''(h)$, $\varphi'''(h)$ значение $h=0$; воспользовавшись этими соотношениями, получим

$$\begin{gathered}\varphi(0)=y-y=0,\\ \varphi'(0)=(1-p_1-p_2) f(x,y),\\ \varphi''(0)=(1-2p_2\alpha_2) f_x(x,y)+(1-2p_2\beta_{21}) f_y(x,y) f(x,y),\\ \varphi'''(0)=(1-3p_2\alpha_2^2) f_{xx}(x,y)+(2-6p_2\beta_{21}) f_{xy}(x,y) f(x,y)+\\ +(1-3p_2\beta_{21}^2) f_{yy}(x,y)(f(x,y))^2+f_y(x,y) y''(x).\end{gathered} \tag{8.4}$$

Соотношение $\varphi'(0) = 0$ выполняется при всех $f(x, y)$, если

$$1 - p_1 - p_2 = 0, \tag{8.5}$$

соотношение $\varphi''(0) = 0$ выполняется, если

$$1 - 2p_2\alpha_2 = 0 \quad \text{и} \quad 1 - 2p_2\beta_{21} = 0. \tag{8.6}$$

Таким образом, $\varphi(0) = \varphi'(0) = \varphi''(0) = 0$ при всех $f(x, y)$, если выполнены три соотношения (8.5), (8.6) относительно четырех параметров. Задавая произвольно один из параметров, получим различные методы Рунге—Кутты с $s = 2$. Например, при $p_1 = \frac{1}{2}$ получаем $p_2 = \frac{1}{2}$, $\alpha_2 = 1$, $\beta_{21} = 1$. При $p_1 = 0$ получаем $p_2 = 1$, $\alpha_2 = \frac{1}{2}$, $\beta_{21} = \frac{1}{2}$. В случае уравнения $y' = y$, согласно (8.4), имеем $\varphi'''(0) = y$ независимо от значений p_1, p_2, α_2, β_{21}. Отсюда следует, что нельзя построить формул Рунге—Кутты со значениями $q = 2$ и $s = 3$. $\triangleright$

8.26. Определить порядок метода s для следующей совокупности формул при $q = 3$:

$$k_1 = hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2}, y + \frac{k_1}{2}\right),$$

$$k_3 = hf(x + h, y - k_1 + 2k_2), \quad z(h) = y(x) + \frac{k_1 + 4k_2 + k_3}{6}.$$

Ответ: $s = 3$.

8.27. Определить порядок метода s для следующей совокупности формул при $q = 4$:

$$k_1 = hf(x,\, y), \quad k_2 = hf\left(x + \frac{h}{2},\, y + \frac{k_1}{2}\right),$$

$$k_3 = hf\left(x + \frac{h}{2},\, y + \frac{k_2}{2}\right), \quad k_4 = hf(x + h,\, y + k_3),$$

$$z(h) = y(x) + \frac{k_1 + 2k_2 + 2k_3 + k_4}{6}.$$

Ответ: $s = 4$.

8.28. Доказать, что погрешность метода на шаге $\varphi(h)$ имеет главный член, т. е. справедливо представление вида

$$\varphi(h) = \psi(x,\, y)h^{s+1} + O(h^{s+2}).$$

$\lhd$ Пусть в уравнении $y' = f$ функция $f(x, y)$ и все ее производные до порядка $s + 1$ включительно равномерно ограничены в области $G : \ x_0 \leqslant x \leqslant x_0 + X, \ -\infty < y < \infty$. Тогда также равномерно ограничены производные всех решений уравнения $y' = f$ до порядка $s + 2$ включительно. В этом случае согласно формуле Тейлора представление погрешности можно записать в уточненной форме

$$\varphi(h) = \frac{\varphi^{(s+1)}(0)}{(s+1)!}\, h^{s+1} + \frac{\varphi^{(s+2)}(\theta h)}{(s+2)!}\, h^{s+2}.$$

Отсюда имеем

$$\varphi^{(s+1)}(0) = y^{(s+1)}(0) - z^{(s+1)}(0).$$

Величины $y^{(s+1)}(0)$ и $z^{(s+1)}(0)$ явно выражаются через значения в точке (x, y) функции f и ее производных порядка не выше s. Правая часть равенства дифференцируема $s + 1$ раз, отсюда следует, что функция $\psi(x, y)$ дифференцируема в области G и ее производные ψ_x и ψ_y равномерно ограничены в этой области. Аналогично устанавливается, что величина $\varphi^{(s+2)}(\theta h)$ равномерно ограничена при $x_0 \leqslant x < x + h \leqslant x_0 + X$. Таким образом, искомое соотношение имеет место. $\rhd$

8.29. Найти главный член погрешности расчетной формулы

$$y^*_{j+1} = y_j + hf(x_j,\, y_j),$$

$$y_{j+1} = y_j + \frac{h}{2}\left(f(x_j,\, y_j) + f(x_{j+1},\, y^*_{j+1})\right).$$

Ответ: $(B - A)h^3$, где $B = \dfrac{f_y y''}{6}$, $A = \dfrac{f_{xx} + 2f_{xy}y' + f_{yy}(y')^2}{12}$.

8.30. Найти главный член погрешности расчетной формулы

$$y_{j+1/2} = y_j + \frac{h}{2}\, f(x_j,\, y_j),$$

$$y_{j+1} = y_j + hf\left(x_j + \frac{h}{2}\,,\, y_{j+1/2}\right).$$

Ответ: $\left(B + \dfrac{A}{2}\right)h^3$, где $B = \dfrac{f_y y''}{6}$, $A = \dfrac{f_{xx} + 2f_{xy}y' + f_{yy}(y')^2}{12}$.

Формулы Адамса. *Явной формулой Адамса* для решения уравнения $y' = f(x, y)$ называют выражение

$$y_k - y_{k-1} = h \sum_{i=0}^{m} \gamma_i \nabla^i f_{k-1};$$

неявная формула Адамса имеет вид

$$y_k - y_{k-1} = h \sum_{i=0}^{m} \bar{\gamma}_i \nabla^i f_k,$$

где

$$\nabla^i f_k = \sum_{j=0}^{i} (-1)^j C_i^j f_{k-j},$$

а коэффициенты γ_i и $\bar{\gamma}_i$ определяются следующим образом:

$$\gamma_0 = \bar{\gamma}_0 = 1,$$

$$\gamma_i = \int_0^1 \prod_{k=1}^{i} \left(1 - \frac{u}{k}\right) du\,, \quad \bar{\gamma}_i = \gamma_i - \gamma_{i-1} = -\int_0^1 \frac{u}{i} \prod_{k=1}^{i-1} \left(1 - \frac{u}{k}\right) du\,,\ i \geqslant 1.$$

8.31. Вывести явные формулы Адамса p-го порядка точности для $p = 2, 3, 4$.

Ответ: $y_{j+1} = y_j + (3\,f_j - f_{j-1})\,\frac{h}{2}\,, \quad p = 2;$

$y_{j+1} = y_j + (23\,f_j - 16\,f_{j-1} + 5\,f_{j-2})\,\frac{h}{12}\,, \quad p = 3;$

$y_{j+1} = y_j + (55\,f_j - 59\,f_{j-1} + 37\,f_{j-2} - 9\,f_{j-3})\,\frac{h}{24}\,, \quad p = 4.$

8.32. Вывести неявные формулы Адамса p-го порядка точности для $p = 2, 3, 4$.

Ответ: $y_{j+1} = y_j + (f_{j+1} + f_j)\,\frac{h}{2}\,, \quad p = 2;$

$y_{j+1} = y_j + (5\,f_{j+1} + 8\,f_j - f_{j-1})\,\frac{h}{12}\,, \quad p = 3;$

$y_{j+1} = y_j + (9\,f_{j+1} + 19\,f_j - 5\,f_{j-1} + f_{j-2})\,\frac{h}{24}\,, \quad p = 4.$

8.33. Показать, что для коэффициентов γ_i в формулах Адамса при $i \to \infty$ справедлива асимптотика

$$\gamma_i \approx \frac{\text{const}}{\ln i}\,, \quad \bar{\gamma}_i \approx \frac{\text{const}}{i \ln i}\,.$$

Уравнения второго порядка. Рассмотрим следующую задачу:

$$y'' = f(x, y, y')\,, \quad y(x_0) = a,\ \ y'(x_0) = b. \tag{8.7}$$

Вводя новую неизвестную функцию $v(x) = y'(x)$, ее можно свести к системе уравнений первого порядка

$$\begin{aligned} v' &= f(x, y, v), \quad v(x_0) = b, \\ y' &= v, \qquad\qquad y(x_0) = a, \end{aligned}$$

а для ее решения применить рассмотренные выше методы.

Однако алгоритмы, ориентированные на специальный класс задач, часто более эффективны. Далее будем предполагать, что функция f не зависит от y':

$$f(x, y, y') \equiv f(x, y)\,.$$

В этом случае (по аналогии с задачей Коши для уравнения первого порядка) *разностной схемой на равномерной сетке* $x_k = x_0 + kh$, $k \geqslant 0$ называют семейство разностных уравнений

$$\frac{1}{h^2}\sum_{i=0}^{n} a_{-i} y_{k-i} = \sum_{i=0}^{n} b_{-i} f_{k-i}\,, \quad k = n, n+1, \ldots\,. \tag{8.8}$$

с известными начальными условиями $y_0 = y(x_0)$, $y_1, \ldots, y_{n-1}$, где a_{-i}, b_{-i} не зависят от h, $a_0 \neq 0$ и $f_{k-i} = f(x_{k-i}, y_{k-i})$.

Схему для уравнения второго порядка называют *α-устойчивой*, если выполнено следующее условие: все корни характеристического уравнения принадлежат единичному кругу и на границе круга нет кратных корней, за исключением двукратного корня, равного единице.

8.34. Получить необходимые и достаточные условия аппроксимации уравнения (8.7) разностными уравнениями (8.8).

Ответ: $\sum\limits_{i=0}^{n} a_{-i} = 0$, $\sum\limits_{i=0}^{n} i a_{-i} = 0$, $\sum\limits_{i=0}^{n} i^2 a_{-i} = 2$, $\sum\limits_{i=0}^{n} b_{-i} = 1$.

8.35. Определить порядок аппроксимации на решении разностной схемы Нумерова

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = \frac{f_{k+1} + 10 f_k + f_{k-1}}{12}\,.$$

Ответ: главный член погрешности равен $\frac{h^4}{240}\, y^{(6)}(\xi), p = 4$.

8.36. Определить порядок аппроксимации на решении разностной схемы

$$\frac{y_{k+1} - y_k - y_{k-2} + y_{k-3}}{3\,h^2} = \frac{5\,f_k + 2\,f_{k-1} + 5\,f_{k-2}}{12}\,.$$

Ответ: главный член погрешности равен $\frac{17\,h^4}{720}\, y^{(6)}(\xi), p = 4$.

8.2. Краевая задача

Рассмотрим первую краевую задачу для обыкновенного дифференциального уравнения второго порядка:

$$-\left(k(x)\,u'\right)' + p(x)\,u = f(x)\,, 0 < x < 1\,, u(0) = u(1) = 0\,.$$

Предполагаем, что коэффициенты уравнения удовлетворяют условиям $0 < k_0 \leqslant k(x) \leqslant k_1$, $0 \leqslant p(x) \leqslant p_1$. На любом из концов отрезка краевое условие может быть задано в виде линейной комбинации функции и производной $a\,u + b\,u' = c$. В этом случае следует обратить внимание на способ его аппроксимации. Если это не оговаривается специально, то в задачах параграфа сетка на отрезке $[0, 1]$ выбирается равномерной: $x_i = i\,h$, $i = 0, \ldots, N$, $N\,h = 1$.

8.37. Определить локальный порядок аппроксимации в точке $x_i = i\,h$ для операторов L и L_h:

$$L\,u = (k(x)\,u')'\,,\ (L_h u_h)_i = \frac{1}{h}\left[k(x_{i+1/2})\,\frac{u_{i+1}-u_i}{h} - k(x_{i-1/2})\,\frac{u_i-u_{i-1}}{h}\right],$$

считая коэффициент $k(x)$ достаточно гладким.
Ответ: $p = 2$.

8.38. Для задачи

$$-u'' + u = f(x), \quad u(0) = u(1) = 0,$$

рассматривается разностная схема

$$-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + (\alpha u_{i+1} + \beta u_i + \gamma u_{i-1}) = f(x_i) + \frac{h^2}{12}\,f''(x_i)\,,$$

$$1 \leqslant i \leqslant N-1,\ u_0 = u_N = 0,\ Nh = 1\,.$$

При каких α, β и γ аппроксимация на решении имеет четвертый порядок?
Ответ: $\alpha = \gamma = \frac{1}{12}$, $\beta = \frac{5}{6}$.

8.39. Используя значения функции $u(x)$ в точках $x_0 = 0$ и $x_1 = h$, построить аппроксимацию второго порядка граничного условия $a\,u(0) + b\,u'(0) = c$ для уравнения

$$-u'' + p(x)\,u = f(x)\,.$$

$\lhd$ По формуле Тейлора

$$u(h) = u(0) + h\,u'(0) + \frac{h^2}{2}\,u''(0) + O(h^3)\,,$$

откуда получаем

$$u'(0) = \frac{u(h)-u(0)}{h} - \frac{h}{2}\,u''(0) + O(h^2)\,.$$

Из исходного уравнения имеем

$$-u''(0) = f(0) - p(0)\,u(0)\,,$$

следовательно,

$$a\,u(0) + b\left(\frac{u(h)-u(0)}{h} + \frac{h}{2}\,(f(0) - p(0)\,u(0))\right) = c + O(h^2)\,.$$

После замены $u(0)$ на u_0 и $u(h)$ на u_1 получим, что искомая аппроксимация имеет вид

$$\left(a - \frac{h}{2}\,b\,p(0)\right)u_0 + b\,\frac{u_1-u_0}{h} = c - \frac{h}{2}\,b\,f(0)\,. \qquad \rhd$$

8.40. Для задачи

$$-u'' + p(x)\,u = f(x),\ u'(0) = 1,\ u(1) = 0$$

построить разностную схему второго порядка аппроксимации на сетке $x_i = \left(i - \frac{1}{2}\right)h$, $i = 0, \ldots, N$, $h = \left(N - \frac{1}{2}\right)^{-1}$.

Ответ: $-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + p(x_i)\,u_i = f(x_i)$, $1 \leqslant i \leqslant N-1$, $\frac{u_1-u_0}{h} = 1$, $u_N = 0$. Отметим, что u_0 аппроксимирует $u\left(-\frac{h}{2}\right)$, u_N аппроксимирует $u(1)$.

8.41. Исследовать устойчивость разностной схемы

$$-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}=f_i,\ 1\leqslant i\leqslant N-1,\ u_0=u_N=0,\ Nh=1$$

и показать, что при $h\to 0$ число обусловленности матрицы алгебраической системы для нахождения u_i имеет порядок $O(h^{-2})$.

$\lhd$ Так как (см. 2.86) собственные значения разностной задачи

$$\frac{u_{i+1}-2\,u_i+u_{i-1}}{h^2}=-\lambda\,u_i,\ 1\leqslant i\leqslant N-1,\ u_0=u_N=0,\quad h=\frac{1}{N},$$

имеют вид

$$\lambda^{(m)}=\frac{4}{h^2}\sin^2\frac{\pi m h}{2}\,,\quad m=1,\,\ldots\,,N-1\,,$$

то можно проверить, что справедливы оценки

$$\lambda_{\min}=\lambda^{(1)}=\frac{4}{h^2}\sin^2\frac{\pi h}{2}\geqslant 4,\quad \lambda_{\max}=\lambda^{(N-1)}\leqslant\frac{4}{h^2}\ .$$

Отсюда следует порядок обусловленности системы. Выше было использовано неравенство

$$\sin|\beta|\geqslant\frac{2}{\pi}\,|\beta|\quad\text{при}\quad |\beta|\leqslant\frac{\pi}{2}\ .$$

Исходная задача записывается в виде $A\,\mathbf{u}=\mathbf{f}$, или в силу невырожденности матрицы A $\mathbf{u}=A^{-1}\mathbf{f}$. Отсюда получаем неравенство для евклидовой нормы векторов

$$\|\mathbf{u}\|_2\leqslant\|A^{-1}\|_2\,\|\mathbf{f}\|_2.$$

Подчиненная матричная норма (см. 5.5) имеет вид $\|A\|_2=\sqrt{\lambda_{\max}(A^TA)}$. В рассматриваемом случае

$$\|A^{-1}\|_2=\frac{1}{\lambda_{\min}(A)}\ ,$$

и выполнение неравенства $\lambda_{\min}(A)=\lambda^{(1)}\geqslant C$ с постоянной, не зависящей от h, по определению означает устойчивость схемы в евклидовой норме. Отсюда следует устойчивость в норме пространства $L_{2,h}$, которая отличается от евклидовой постоянным множителем h. $\rhd$

8.42. Получить на основе принципа максимума при $f\in C^{(2)}[0,1]$ оценку скорости сходимости

$$\max_{0\leqslant i\leqslant N}|u(x_i)-u_i|\leqslant\frac{h^2}{96}\max_{[0,1]}|f''(x)|$$

решения разностной схемы

$$-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}=f_i,\ 1\leqslant i\leqslant N-1,\ u_0=u_N=0,\ Nh=1,$$

к решению дифференциальной задачи

$$-u''=f,\quad u(0)=u(1)=0\,.$$

$\lhd$ Введем обозначение

$$l(u_i)=\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}$$

и покажем, что если $l(u_i) \leqslant 0$ при $i = 1, \ldots, N-1$ и $u_0 = u_N = 0$, то $u_i \geqslant 0$ при всех i.

Пусть $d = \min\limits_i u_i < 0$ и q — такое наименьшее целое, что $u_q = d$. Тогда $u_{q-1} > d$, $u_{q+1} \geqslant d$ и

$$l(u_q) = \frac{(u_{q+1} - d) + (u_{q-1} - d)}{h^2} > 0\,.$$

Полученное противоречие доказывает, что $u_i \geqslant 0$ при всех i.

Следующий шаг — доказательство неравенства

$$\max_{0 \leqslant i \leqslant N} |u_i| \leqslant \frac{1}{8}\, U, \quad \text{где} \;\; U = \max_{0 < i < N} |l(u_i)|\,.$$

Введем функцию

$$w_i = \frac{ih(1 - ih)}{2}\, U\,, \quad i = 0, \ldots, N\,,$$

удовлетворяющую условиям $w_i \geqslant 0$ и $l(w_i) = -U$. Теперь для функций $w_i \pm u_i$ справедливо

$$l(w_i \pm u_i) = -U \pm l(u_i) \leqslant 0, \quad w_0 \pm u_0 = w_N \pm u_N = 0\,.$$

Поэтому, используя доказанное выше свойство, имеем $w_i \pm u_i \geqslant 0$, откуда и следует требуемое неравенство

$$|u_i| \leqslant w_i \leqslant \max_{0 < i < N} w_i \leqslant \frac{1}{8}\, U\,.$$

Последний этап — определение величины U. Используя формулу Тейлора, запишем уравнение для погрешности $u(x_i) - u_i$

$$l(u(x_i) - u_i) = \frac{h^2}{12}\, u^{(4)}(\xi_i) = -\frac{h^2}{12}\, f''(\xi_i)\,.$$

Отсюда находим $U = \dfrac{h^2}{12} \max\limits_{[0,1]} |f''(x)|$, что и приводит к искомой оценке. $\triangleright$

Исследование устойчивости методом априорных оценок. Рассмотрим этот метод на примере дифференциальной задачи

$$-u'' + p(x)\, u = f(x),\; 0 < x < 1,\; u(0) = u(1) = 0,\; p(x) \geqslant 0\,.$$

Возьмем интеграл по отрезку $[0, 1]$ от обеих частей уравнения, предварительно умножив его на u:

$$\int\limits_0^1 (-u'')u\, dx + \int\limits_0^1 pu^2 dx = \int\limits_0^1 fu\, dx\,.$$

В результате интегрирования по частям получаем интегральное тождество

$$\int\limits_0^1 (u')^2 dx + \int\limits_0^1 pu^2 dx = \int\limits_0^1 fu\, dx\,.$$

Далее нам потребуется неравенство, связывающее интегралы от квадратов функции и ее производной. Из равенства

$$u(x_0) = \int\limits_0^{x_0} u'(x)\, dx$$

следует, что

$$|u(x_0)|^2 \leqslant \left(\int\limits_0^{x_0} 1^2 dx\right)\left(\int\limits_0^{x_0}(u')^2 dx\right) \leqslant \int\limits_0^{x_0}(u')^2 dx \leqslant \int\limits_0^1 (u')^2 dx\,.$$

Интегрируя по x_0 обе части неравенства, получаем искомое выражение

$$\int\limits_0^1 |u(x_0)|^2 dx_0 \leqslant \int\limits_0^1 (u')^2 dx \int\limits_0^1 dx_0, \quad \text{или} \quad \int\limits_0^1 u^2 dx \leqslant \int\limits_0^1 (u')^2 dx\,.$$

Окончательно имеем

$$\int\limits_0^1 u^2 dx \leqslant \int\limits_0^1 (u')^2 dx + \int\limits_0^1 p u^2 dx = \int\limits_0^1 f u\, dx \leqslant \frac{1}{2}\left(\int\limits_0^1 f^2 dx + \int\limits_0^1 u^2 dx\right),$$

откуда

$$\|u\|_{L_2} \leqslant \|f\|_{L_2}, \quad \text{где} \quad \|u\|^2_{L_2} = \int\limits_0^1 u^2 dx\,.$$

Полученная априорная оценка решения означает устойчивость задачи по правой части.

8.43. Исследовать методом априорных оценок устойчивость простейшей разностной схемы

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + p_i\, u_i = f_i, \quad 1 \leqslant i \leqslant N-1, \quad u_0 = u_N = 0,\ Nh = 1\,.$$

$\lhd$ Умножим i-е уравнение на u_i и просуммируем от 1 до $(N-1)$. Учитывая, что $u_0 = u_N = 0$, имеем

$$-\frac{1}{h^2}\sum_{i=1}^{N-1}(u_{i+1} - u_i)u_i + \frac{1}{h^2}\sum_{i=1}^{N-1}(u_i - u_{i-1})u_i =$$

$$= -\frac{1}{h^2}\sum_{i=1}^{N}(u_i - u_{i-1})u_{i-1} + \frac{1}{h^2}\sum_{i=1}^{N}(u_i - u_{i-1})u_i = \frac{1}{h^2}\sum_{i=1}^{N}(u_i - u_{i-1})^2\,.$$

Отсюда следует сумматорное (аналог интегрального) тождество

$$\frac{1}{h^2}\sum_{i=1}^{N}(u_i - u_{i-1})^2 + \sum_{i=1}^{N-1} p_i u_i^2 = \sum_{i=1}^{N-1} f_i u_i,$$

которое, учитывая обозначения $\nabla u_i = u_i - u_{i-1}$, $(u, v] = \sum\limits_{i=1}^{N} u_i v_i$, $(u, v) = \sum\limits_{i=1}^{N-1} u_i v_i$, запишем таким образом:

$$\frac{1}{h^2}\,(\nabla u, \nabla u] + (p\,u, u) = (f, u).$$

Докажем сеточный аналог неравенства для функции и ее производной. Представим значение u_k в виде суммы $u_k = \sum\limits_{i=1}^{k}(u_i - u_{i-1})$. Отсюда следует, что для любого $1 \leqslant k \leqslant N-1$ справедливо неравенство

$$u_k^2 \leqslant \sum_{i=1}^{k} 1^2 \sum_{i=1}^{k}(u_i - u_{i-1})^2 \leqslant N\sum_{i=1}^{N}(u_i - u_{i-1})^2.$$

Суммируя по переменной k, получим

$$\sum_{k=1}^{N-1} u_k^2 \leqslant N^2 \sum_{i=1}^{N}(u_i - u_{i-1})^2 \leqslant \frac{1}{h^2} \sum_{i=1}^{N}(u_i - u_{i-1})^2 + \sum_{i=1}^{N-1} p_i u_i^2 =$$
$$= \sum_{i=1}^{N-1} f_i u_i \leqslant \frac{1}{2}\left(\sum_{i=1}^{N-1} f_i^2 + \sum_{i=1}^{N-1} u_i^2\right).$$

Таким образом, $\sum\limits_{i=1}^{N-1} u_i^2 \leqslant \sum\limits_{i=1}^{N-1} f_i^2$, и априорная оценка решения разностной задачи в норме $\|u_h\|_h^2 = h(u_h, u_h)$ пространства $L_{2,h}$, согласованной с непрерывной нормой L_2, имеет вид $\|u_h\|_h \leqslant \|f_h\|_h$. $\triangleright$

8.44. Для гладких функций $u(x)$ таких, что $u(0) = u(1) = 0$, получить на основе рядов Фурье неравенство

$$\int_0^1 u^2(x)dx \leqslant \frac{1}{\pi^2}\int_0^1 (u'(x))^2 dx$$

и его дискретный аналог на равномерной сетке.

У к а з а н и е. Воспользоваться спектральной задачей

$$-u'' = \lambda u, \quad u(0) = u(1) = 0$$

и ее дискретным аналогом (см. 2.86).

Операторы с особенностями. Преобразование декартовых координат в полярные, цилиндрические или сферические может приводить к локально неограниченным дифференциальным операторам.

8.45. Построить интегро-интерполяционным методом разностную схему для задачи

$$-\frac{1}{r}\left(ru'\right)' = f(r), \quad 0 < r < R, \quad \lim_{r\to 0} r\,u' = 0, \quad u(R) = 0,$$

на сетке $\overline{D}_h = \left\{r_i = \left(i + \frac{1}{2}\right)h, \;\; 0 \leqslant i \leqslant N, \;\; N\,h = R\right\}$.

$\triangleleft$ Важными данными задачи являются условие ограниченности решения в нуле и сдвинутая на $\frac{h}{2}$ сетка. Во внутренних узлах схема имеет обычный вид

$$-\frac{1}{r_i}\,\frac{1}{h}\left[r_{i+1/2}\,\frac{u_{i+1} - u_i}{h} - r_{i-1/2}\,\frac{u_i - u_{i-1}}{h}\right] = f(r_i), \; 1 \leqslant i \leqslant N-1.$$

Построим уравнение при $i = 0$ (это соответствует значению $r = \frac{h}{2}$). Умножим уравнение на r и проинтегрируем его от ε до h. Имеем

$$\int_\varepsilon^h \left[\left(ru'\right)' + r\,f(r)\right] dr = 0.$$

Переходя к пределу при $\varepsilon \to 0$ и используя условие ограниченности, получим

$$h\,u'(h) + \int\limits_0^h r\,f(r)dr = 0\,.$$

Теперь аппроксимируем полученное выражение в точке $r = \frac{h}{2}$

$$\frac{u_1 - u_0}{h} + \frac{h}{2}\,f\left(\frac{h}{2}\right) = 0\,.$$

Аппроксимация второго порядка для условия $u(R) = 0$ имеет вид

$$\frac{u_N + u_{N-1}}{2} = 0\,. \qquad \triangleright$$

8.46. Построить интегро-интерполяционным методом разностную схему для задачи

$$-\,\frac{1}{r}\,(ru')' = f(r)\,, \quad 0 < r < R\,, \quad \lim_{r\to 0} r\,u' = 0\,,\; u(R) = 0\,,$$

на сетке $\overline{D}_h = \{r_i = i\,h,\;\; 0 \leqslant i \leqslant N,\;\; N\,h = R\}$.

Ответ: во внутренних узлах схема имеет вид, как в 8.45 (разница только в определении r_i). Правое краевое условие $u_N = 0$, а левое краевое условие таково:

$$\frac{u_1 - u_0}{h} + \frac{h}{4}\,f\left(\frac{h}{4}\right) = 0\,,$$

если интеграл $\int\limits_0^{h/2} r\,f(r)dr$ заменить выражением $\frac{h^2}{8}\,f\left(\frac{h}{4}\right)$.

8.47. Построить при $r = 0$ аппроксимацию уравнения

$$-\,\frac{1}{r}\,(ru')' = f(r),$$

считая решение $u(r)$ четной функцией, т. е. $u(r) = u(-r)$.

$\triangleleft$ Представим исходный оператор в виде двух слагаемых:

$$\frac{1}{r}\,(ru')' = u'' + \frac{1}{r}\,u'$$

и учтем, что для гладкой четной функции $u(r)$ ее производная в нуле равна нулю. В этом случае из формулы Тейлора следует:

$$u'(\varepsilon) = \varepsilon\,u''(0) + O(\varepsilon^2)\,.$$

Подставляя это выражение в слагаемое $\frac{u'}{r}$ и переходя к пределу при $\varepsilon \to 0$, получим

$$\frac{1}{r}\,(ru')'\Big|_{r=0} = 2\,u''(0)\,.$$

Запишем стандартную аппроксимацию для уравнения в нуле

$$-\,2\,\frac{u_1 - 2\,u_0 + u_{-1}}{h^2} = f(0)\,.$$

Учитывая четность, т. е. $u_{-1} = u_1$, отсюда имеем:

$$\frac{u_1 - u_0}{h} + \frac{h}{4}\,f(0) = 0\,. \qquad \triangleright$$

8.48. Построить интегро-интерполяционным методом схему для задачи

$$-\frac{1}{r^2}\left(r^2u'\right)' = f(r)\,,\ 0<r<R\,,\quad \lim_{r\to 0} r^2\,u' = 0\,,\ u(R)=0\,,$$

на сетке $\overline{D}_h = \left\{r_i = \left(i+\frac{1}{2}\right)h,\ \ 0\leqslant i\leqslant N,\ \ N\,h = R\right\}$.

8.49. Построить интегро-интерполяционным методом схему для задачи

$$-\frac{1}{r^2}\left(r^2u'\right)' = f(r)\,,\quad 0<r<R\,,\quad \lim_{r\to 0} r^2\,u' = 0\,,\ u(R)=0\,,$$

на сетке $\overline{D}_h = \{r_i = i\,h,\ \ 0\leqslant i\leqslant N,\ \ N\,h = R\}$.

8.50. Построить при $r=0$ аппроксимацию уравнения

$$-\frac{1}{r^2}\left(r^2u'\right)' = f(r),$$

считая решение $u(r)$ четной функцией, т. е. $u(r)=u(-r)$.

8.51. Построить аппроксимацию на решении второго порядка по точкам $x_N=1$ и $x_{N-1}=1-h$ краевого условия $u'(1)-3u(1)=1$ для уравнения $u''=\cos x+1$.

Ответ: $\frac{u_N-u_{N-1}}{h}+\frac{h}{2}\left(\cos(1)+1\right)-3\,u_N=1$.

8.52. Построить аппроксимацию на решении второго порядка по точкам $x_0=0$ и $x_1=h$ краевого условия $u'(0)+4u(0)=1$ для уравнения $u''-x^2\,u=1$.

Ответ: $\frac{u_1-u_0}{h}-\frac{h}{2}+4\,u_0=1$.

8.53. Построить аппроксимацию на решении второго порядка по точкам $x_N=1$ и $x_{N-1}=1-h$ краевого условия $u'(1)=0$ для уравнения $u''-3u=\exp x$.

Ответ: $\frac{u_N-u_{N-1}}{h}+\frac{h}{2}\left(3\,u_N+\exp(1)\right)=0$.

8.54. Построить аппроксимацию на решении второго порядка по точкам $x_0=0$ и $x_1=h$ краевого условия $u'(0)-u(0)=0$ для уравнения $u''-2u=\sin x-1$.

Ответ: $\frac{u_1-u_0}{h}-\frac{h}{2}\left(2\,u_0-1\right)-u_0=0$.

В задачах 8.55 — 8.58 важной является проверка ортогональности собственных функций в соответствующем скалярном произведении (см. 7.37).

8.55. Исследовать устойчивость разностной схемы $A\mathbf{u}=\mathbf{f}$

$$-\frac{u_{i+1}-2u_i+u_{i-1}}{h^2}+u_i=f_i,\quad 0<i<N,$$

$$u_0=u_1,\quad u_{N-1}=u_N,\quad (N-1)\,h=1\,.$$

Ответ: схема устойчива, так как

$$\lambda^{(n)}(A)=\frac{4}{h^2}\sin^2\frac{\pi(n-1)}{2(N-1)}+1,\quad n=1,\ldots,N-1,\quad \lambda_{\min}=1\,.$$

8.56. Исследовать устойчивость разностной схемы $A\mathbf{u} = \mathbf{f}$

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = f_i, \quad 0 < i < N,$$
$$u_0 = 0, \quad u_{N-1} = u_N, \quad (2N-1)\,h = 2\,.$$

О т в е т: схема устойчива, так как

$$\lambda^{(n)}(A) = \frac{4}{h^2}\sin^2\frac{\pi(2n-1)}{2(2N-1)}\,, \quad n = 1, \ldots, N-1, \quad \lambda_{\min} \geqslant 1\,.$$

8.57. Исследовать устойчивость разностной схемы $A\mathbf{u} = \mathbf{f}$

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = f_i, \quad 0 < i < N,$$
$$u_0 = u_1, \quad u_N = 0, \quad (2N-1)\,h = 2\,.$$

О т в е т: схема устойчива, так как

$$\lambda^{(n)}(A) = \frac{4}{h^2}\sin^2\frac{\pi(2n-1)}{2(2N-1)}\,, \quad n = 1, \ldots, N-1, \quad \lambda_{\min} \geqslant 1\,.$$

8.58. Исследовать устойчивость разностной схемы $A\mathbf{u} = \mathbf{f}$

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - (2 + \cos(2\pi ih))u_i = f_i, \quad 0 < i < N,$$
$$u_0 = u_N = 0,\ Nh = 1\,.$$

О т в е т: схема устойчива, так как для главной части оператора

$$A_0 u_i = -\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}\,, \quad 1 \leqslant i \leqslant N-1,$$

при условиях $u_0 = u_N = 0,\ Nh = 1$ имеем собственные значения

$$\lambda^{(n)}(A_0) = \frac{4}{h^2}\sin^2\frac{\pi n}{2N}\,, \quad n = 1, \ldots, N-1, \quad \lambda_{\min}(A_0) \geqslant 4\,;$$

поэтому для исходной задачи $\lambda_{\min}(A) \geqslant 4 - 3 = 1$.

Глава 9

Уравнения с частными производными

Для обыкновенных дифференциальных уравнений в настоящее время имеются удовлетворительная общая теория и алгоритмы, позволяющие в большинстве случаев эффективно численно находить решение задачи. Для уравнений в частных производных теорию численных методов приходится строить в зависимости от типа уравнения. При этом строгое обоснование сходимости и оценки погрешности чаще всего удается получить только для модельных задач. Алгоритмы для сложных нелинейных уравнений обычно строят, обобщая и комбинируя хорошо изученные методы. В этом случае исследование проводится для различных линеаризованных уравнений и с помощью численных экспериментов для задач с известными точно решениями.

В данной главе изложены численные методы решения некоторых задач математической физики. Особое внимание уделено обоснованию корректности рассмотренных алгоритмов.

9.1. Корректность разностных схем

Прежде чем приступить к формальному исследованию задач математической физики, покажем на простых примерах, что уже в линейном случае наличие аппроксимации и сколь угодно мелкой сетки недостаточно даже для получения правдоподобных результатов — решение разностной задачи и решение дифференциальной задачи могут значительно отличаться. При этом измельчение шага сетки будет только ухудшать ситуацию.

Пример 1. Пусть в полуплоскости $t \geqslant 0$ решается задача Коши для уравнения $u_t + au_x = 0$ при начальном условии $u(x, 0) = u_0(x)$. Зададимся сеткой с узлами в точках $(mh, n\tau)$ и заменим исходную дифференциальную задачу разностной

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\,\frac{u_{m+1}^n - u_m^n}{h} = 0,\ u_m^0 = u_0(mh).$$

Тогда значения u_m^n при $n > 0$ определяются последовательно из соотношения

$$u_m^{n+1} = \left(1 + \frac{a\tau}{h}\right) u_m^n - \left(\frac{a\tau}{h}\right) u_{m+1}^n.$$

Пусть при измельчении сетки справедливо $\frac{\tau}{h} = r = \text{const}$. В этом случае значения решения сеточной задачи в точке (x_0, t_0) не зависят от начальных условий вне отрезка $[x_0, x_0 + r^{-1}t_0]$. Точное решение дифференциальной задачи имеет вид $u(x, t) = u_0(x - at)$. Поэтому в классе начальных условий, обладающих некоторым ограниченным числом производных, областью зависимости для дифференциальной задачи является

точка $x_0 - at_0$. Взяв в качестве начальной функции финитный «всплеск» в точке $x_0 - at_0$, можно показать, что необходимым условием сходимости решений в точке (x_0, t_0) для произвольной начальной функции $u_0(x)$ является условие $x_0 - at_0 \in [x_0, x_0 + r^{-1}t_0]$. Это эквивалентно одновременному выполнению неравенств $a \leqslant 0$ и $\left|\frac{a\tau}{h}\right| \leqslant 1$. Таким образом, при $a > 0$ схема непригодна для расчетов.

Строгая формулировка этого утверждения для общего случая называется *теоремой Куранта об областях зависимости*. Для рассматриваемой схемы необходимое условие сходимости совпадает с условием устойчивости и обеспечивает сходимость для гладких начальных данных.

Пример 2. Пусть в области $[0,1] \times [0,T]$ решается начально-краевая задача для уравнения теплопроводности

$$u_t = u_{xx}, \quad u(0,t) = u(1,t) = 0,$$
$$u(x,0) = C_k \sin(\pi k x), \quad k \geqslant 1.$$

Зададимся сеткой с узлами в точках $(mh, n\tau)$ и заменим исходную задачу разностной

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}, \quad 0 < m < M, \quad h = \frac{1}{M},$$
$$u_0^n = u_M^n = 0 \ \ \forall n \geqslant 0, \quad u_m^0 = C_k \sin(\pi k m h), \quad 1 \leqslant k \leqslant M-1.$$

Применяя метод разделения переменных, найдем точные решения дифференциальной и разностной задач

$$u(x,t) = C_k \mu^t(k) \sin(\pi k x), \quad \mu(k) = \mathrm{e}^{-(\pi k)^2},$$

и

$$u_m^n = C_k \mu_h^n(k) \sin(\pi k m h), \quad \mu_h(k) = 1 - \tau \frac{4}{h^2} \sin^2\left(\pi k \frac{h}{2}\right).$$

Справедливость неравенств $0 < \mu(k) < 1$ в дифференциальной задаче определяет экспоненциальное убывание решения с течением времени. Решение разностной задачи также экспоненциально убывает при любых начальных данных, если только $|\mu_h(k)| < 1$. Определим отсюда соотношение для τ и h. Так как

$$\max_{1 \leqslant k \leqslant M-1} |\mu_h(k)| < \max\left\{1, \left|1 - \tau \frac{4}{h^2}\right|\right\},$$

то имеем оценку $\tau \leqslant \frac{h^2}{2}$. Если $\frac{\tau}{h^2} \geqslant \frac{1}{2} + \gamma$, $\gamma = \mathrm{const} > 0$, то величина $\mu_h(k)$ при больших k отрицательна и по модулю больше единицы, что полностью изменяет поведение решения разностной задачи при больших значениях n.

Рассмотренные примеры показывают, что для уравнения в частных производных при замене дифференциальной задачи его разностной аппроксимацией возникают следующие вопросы (аналогичные имевшим место ранее при рассмотрении методов решения других задач):

1) сходится ли точное решение разностной задачи к решению дифференциальной;

2) насколько сильно изменяется решение разностной задачи, если при вычислениях допускаются некоторые погрешности?

Требуемый для соответствующих исследований математический аппарат изложен в гл. refch7. Однако операторная форма записи в определении устойчивости недостаточно детальна при анализе нестационарных уравнений, что затрудняет формализацию процедуры получения необходимых оценок.

Для нестационарного уравнения теплопроводности, уравнения колебаний струны и уравнения Шрёдингера устойчивость схем удобно проверять, если в оператор разностной задачи явно включить эволюцию по времени, записав схему в *каноническом (по Самарскому) виде*. Наиболее употребительными являются двухслойные схемы, связывающие значения решения на следующем и текущем временны́х слоях, и трехслойные, которые требуют для построения решения в следующий момент времени значения с текущего и предыдущего временны́х слоев.

Для задач гиперболического типа допустимо заменять проверку условия устойчивости применением спектрального признака (СПУ). Такой подход позволяет отсеивать большинство непригодных для расчета схем при значительном упрощении техники исследования.

9.2. Гиперболические уравнения

Построение и исследование разностных схем для уравнений в частных производных гиперболического типа традиционно проводят в открытой полуплоскости

$$D = \{(x,t) : \infty > x > -\infty,\ t > 0\}$$

на примере линейного уравнения переноса

$$Lu \equiv \frac{\partial u}{\partial t} + a(x,t)\ \frac{\partial u}{\partial x} = f(x,t)$$

с заданными функциями $a(x,t)$, $f(x,t)$ и начальным условием $u(x,0) = u_0(x)$ при $t = 0$.

Если это не оговаривается специально, то в задачах 9.1–9.33 сетка выбирается равномерной по обеим переменным

$$x_m = m\,h,\quad m = 0, \pm 1, \ldots;\quad t_n = n\,\tau,\quad n = 0, 1, \ldots,$$

а для сеточной функции u в точке (x_m, t_n) используется обозначение u_m^n.

9.1. Определить порядок аппроксимации разностной схемы

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\ \frac{u_m^n - u_{m-1}^n}{h} = 0$$

для уравнения $u_t + au_x = 0$, $a = \text{const} > 0$. При каком соотношении τ и h решение дифференциального уравнения в узлах сетки совпадает с решением разностной схемы?

Ответ: $O(\tau + h)$, $\frac{\tau}{h} = a$.

9.2. Определить порядок аппроксимации разностной схемы

$$\frac{u_m^{n+1}-u_m^n}{\tau}+a\,\frac{u_{m+1}^n-u_{m-1}^n}{2h}=0$$

для уравнения $u_t+a\,u_x=0$.
Ответ: $O\left(\tau+h^2\right)$.

9.3. Определить порядок аппроксимации разностной схемы

$$\frac{u_m^{n+1}-\dfrac{u_{m+1}^n+u_{m-1}^n}{2}}{\tau}+a\,\frac{u_{m+1}^n-u_{m-1}^n}{2h}=0$$

для уравнения $u_t+a\,u_x=0$.
Ответ: $O\left(\tau+h^2+\frac{h^2}{\tau}\right)=O\left(\tau+\frac{h^2}{\tau}\right)$.

9.4. Для однородного уравнения $u_t+a\,u_x=0$, $a=\text{const}$, построить схемы первого и второго порядков аппроксимации на решении (если это возможно), используя шаблон из точек (x_m,t_n), (x_m,t_{n+1}), (x_{m+1},t_n) и условие $\tau=rh$ $(r=\text{const})$.

Указание. При $a<0$ и $r=\frac{1}{|a|}$ существует схема с порядком аппроксимации $O(h^2)$.

9.5. Для уравнения $u_t-u_x=f$ построить разностную схему

$$a^0u_m^{n+1}+a_{-1}u_{m-1}^n+a_0u_m^n+a_1u_{m+1}^n=\varphi_m^n$$

максимального порядка аппроксимации при условии $\tau=rh$, $r=\text{const}$.

Ответ: имеется однопараметрическое семейство схем первого порядка, коэффициенты которого удовлетворяют системе уравнений

$$a^0r\,h=1,\quad a^0+a_0+a_1+a_{-1}=0,\quad a^0r+a_1-a_{-1}=0.$$

Схемы второго порядка не существуют.

9.6. Пусть $\tau=r\,h$. Определим оператор $P_h=I+\frac{rh}{2}\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}\right)$, где I — тождественный оператор. Для уравнения $P_h(u_t-u_x)=P_h(f)$ построить разностную схему

$$a^0u_m^{n+1}+a_{-1}u_{m-1}^n+a_0u_m^n+a_1u_{m+1}^n=\varphi_m^n$$

максимального порядка аппроксимации на решении.

Ответ: схема второго порядка аппроксимации на решении

$$a^0=\frac{1}{rh}\,,\quad a_0=-\,\frac{1}{rh}+\frac{r}{h}\,,\quad a_{-1}=\frac{1-r}{2h}\,,\quad a_1=-\,\frac{1+r}{2h}\,,$$

или

$$\frac{u_m^{n+1}-u_m^n}{\tau}-\frac{u_{m+1}^n-u_{m-1}^n}{2h}-\frac{r}{2h}\,(u_{m+1}^n-2u_m^n+u_{m-1}^n)=\left[f+\frac{rh}{2}\,(f_t+f_x)\right]_m^n.$$

9.7. Для уравнения $u_t + u_x = f$ построить разностную схему

$$a^0 u_m^{n+1} + a^1 u_{m+1}^{n+1} + a_0 u_m^n + a_1 u_{m+1}^n = \varphi_m^n\,,$$

имеющую на решении второй порядок аппроксимации при условии $\tau = h$.

9.8. Пусть для задачи $u_t - u_x = f,\ u(x,0) = \varphi(x)$ используется схема

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} - \frac{u_{m+1}^n - u_{m-1}^n}{2h} = f_m^n, \quad u_m^0 = \varphi(mh).$$

Как определить значения функции u_m^1, чтобы не ухудшить порядок аппроксимации на решении?

Ответ: $u_m^1 = \varphi(mh) + \tau\,[\varphi_x(mh) + f(mh,0)]$.

9.9. Для уравнения $u_t + a\,u_x = 0$ рассматривается схема с пересчетом

$$\frac{u_{m+1/2}^{n+1/2} - \dfrac{u_{m+1}^n + u_m^n}{2}}{0{,}5\tau} + a\ \frac{u_{m+1}^n - u_m^n}{h} = 0\,,$$

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\,\frac{u_{m+1/2}^{n+1/2} - u_{m-1/2}^{n+1/2}}{h} = 0\,.$$

Определить ее порядок аппроксимации на решении.

Ответ: исключая $u_m^{n+1/2}$ при дробных m, получаем схему

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\ \frac{u_{m+1}^n - u_{m-1}^n}{2h} - a^2\,\frac{\tau}{2}\ \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = 0\,,$$

имеющую второй порядок аппроксимации ($\tau = rh$, $r = \text{const}$).

9.10. Для уравнения $u_t + a\,u_x = 0$ рассматривается схема с пересчетом

$$\frac{u_m^{n+1/2} - u_m^n}{\tau} + a\ \frac{u_{m+1}^n - u_m^n}{h} = 0\,,$$

$$\frac{u_m^{n+1} - \dfrac{u_m^n + u_m^{n+1/2}}{2}}{0{,}5\tau} + a\ \frac{u_m^{n+1/2} - u_{m-1}^{n+1/2}}{h} = 0.$$

Определить ее порядок аппроксимации на решении.

Ответ: схема имеет порядок аппроксимации $O(\tau^2 + h^2)$.

9.11. Для уравнения $u_t + u_x = 0$ рассматривается семейство схем с параметром θ

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \theta\ \frac{u_m^{n+1} - u_{m-1}^{n+1}}{h} + (1-\theta)\ \frac{u_m^n - u_{m-1}^n}{h} = 0.$$

При каких значениях θ схема имеет на решении порядок аппроксимации $O(\tau^2 + h^2)$?

Ответ: $\theta = \frac{1}{2} - \frac{h}{2\tau}$.

9.12. Для уравнения $u_t + u_x = 0$ построить схему с порядком аппроксимации на решении $O(\tau^2 + h^4)$ на шаблоне из десяти точек: $(x_{m\pm2}, t_k)$, $(x_{m\pm1}, t_k)$, (x_m, t_k), $k = n, n+1$.

У к а з а н и е. Взять разностную схему

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \frac{1}{2}\left[\frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2h} + \frac{u_{m+1}^n - u_{m-1}^n}{2h}\right] = 0,$$

имеющую порядок аппроксимации $O(\tau^2 + h^2)$ при разложении в ряд Тейлора в точке $(x_m, t_{n+1/2})$. Исключить ее главный член погрешности по h, аппроксимируя с четвертым порядком производную по переменной x на заданном шаблоне.

9.13. Для уравнения переноса $u_t + a(x,t)\, u_x = f(x,t)$ построить двухслойную схему порядка аппроксимации на решении: 1) $O(\tau^2 + h)$; 2) $O(\tau + h^2)$; 3) $O(\tau^2 + h^2)$; 4) $O(\tau + h)$, с минимальным, по возможности, числом узлов l в шаблоне.

У к а з а н и е. Рассмотреть шаблоны из l узлов: 1)$l = 4$; 2)$l = 4$; 3)$l = 4$; 4)$l = 3$.

Спектральный признак устойчивости. Разностные схемы для однородного уравнения переноса с постоянным коэффициентом a можно записать так:

$$L_h\, u_m^n \equiv \sum_{k,l} b_{lk}\, u_{m+l}^{n+k} = 0.$$

Рассмотрим их частные решения вида

$$u_m^n = (\lambda(\varphi))^n\, \mathrm{e}^{\mathrm{i}\, m\, \varphi}.$$

Спектральный признак устойчивости (СПУ) разностной схемы формулируется следующим образом: если при заданном законе стремления τ и h к нулю существует постоянная $0 \leqslant c < \infty$ такая, что для всех φ справедливо неравенство $|\lambda(\varphi)| \leqslant \mathrm{e}^{c\tau}$, то спектральный признак выполнен, и схема может быть применена для численного решения соответствующей задачи Коши для уравнения $L\, u = f$.

Можно показать, что если СПУ не выполняется, то для решения задачи не существует априорной оценки вида $\|u^n\| \leqslant M$ с константой M, не зависящей от параметров сетки, в норме $\|\cdot\|$, которая не зависит от временного слоя.

В упражнениях 9.14–9.23 требуется с помощью спектрального признака исследовать устойчивость разностных схем для случая постоянного коэффициента a в операторе L_h.

9.14. Исследовать устойчивость схемы

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\, \frac{u_m^n - u_{m-1}^n}{h} = 0.$$

◁ Подставим в схему частное решение $u_m^n = \lambda^n \mathrm{e}^{\mathrm{i}\,m\,\varphi}$. В результате имеем

$$\frac{\lambda^{n+1}\mathrm{e}^{\mathrm{i}\,m\,\varphi} - \lambda^n \mathrm{e}^{\mathrm{i}\,m\,\varphi}}{\tau} + a\ \frac{\lambda^n \mathrm{e}^{\mathrm{i}\,m\,\varphi} - \lambda^n \mathrm{e}^{\mathrm{i}\,(m-1)\,\varphi}}{h} = 0.$$

Сокращая на $\lambda^n\,\mathrm{e}^{\mathrm{i}\,m\,\varphi}$, получаем

$$\frac{\lambda - 1}{\tau} + a\ \frac{1 - \mathrm{e}^{-\mathrm{i}\,\varphi}}{h} = 0,$$

откуда следует, что

$$\lambda(\varphi) = 1 - \frac{a\tau}{h} + \frac{a\tau}{h}\ \mathrm{e}^{-\mathrm{i}\,\varphi}.$$

Пусть $a > 0$. Тогда при $0 < \frac{a\tau}{h} \leqslant 1$ имеем $|\lambda(\varphi)| \leqslant 1 - \frac{a\tau}{h} + \frac{a\tau}{h} = 1$, т. е. схема устойчива при выполнении указанных выше условий. При $\frac{a\tau}{h} = 1 + \gamma > 1$, $\gamma = \mathrm{const}$ получаем $\lambda(\pi) = -1 - 2\gamma < -1$, т. е. в этом случае схема неустойчива. Таким образом, разностная схема условно устойчива. При $a < 0$ схема неустойчива.

Аналогичные рассуждения справедливы при $a < 0$ для схемы

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\ \frac{u_{m+1}^n - u_m^n}{h} = 0.$$

Данная схема для $a < 0$ является устойчивой при $0 < \frac{|a|\tau}{h} \leqslant 1$ и неустойчивой при $\frac{|a|\tau}{h} = 1 + \gamma > 1$, $\gamma = \mathrm{const}$ или при $a > 0$. ▷

9.15. Исследовать устойчивость схемы $\frac{u_m^{n+1} - u_m^n}{\tau} + a\ \frac{u_{m+1}^n - u_{m-1}^n}{2h} = 0$.

◁ Поступая аналогично 9.14, получаем

$$\lambda(\varphi) = 1 - \frac{a\tau}{2h}\left(\mathrm{e}^{\mathrm{i}\varphi} - \mathrm{e}^{-\mathrm{i}\varphi}\right) = 1 - \mathrm{i}\,\frac{a\tau}{h}\sin\varphi,$$

откуда следует, что

$$\max_{\varphi} |\lambda(\varphi)| = \left|\lambda\left(\frac{\pi}{2}\right)\right| = \sqrt{1 + \frac{a^2\tau^2}{h^2}}.$$

Пусть $\tau = Ah^2$, тогда $\left|\lambda\left(\frac{\pi}{2}\right)\right| \leqslant \mathrm{e}^{\frac{a^2 A}{2}\tau}$, т. е. схема устойчива при $\tau = O(h^2)$.

Исследование устойчивости с помощью спектрального признака позволяет находить искомые (т. е. устойчивые) законы стремления τ и h к нулю. ▷

9.16. Исследовать устойчивость схемы

$$\frac{u_m^{n+1} - u_m^n}{\tau} + a\ \frac{u_{m+1}^n - u_{m-1}^n}{2h} - \frac{h^2}{2\tau}\ \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2} = 0.$$

Ответ: $\lambda(\varphi) = \frac{1}{2}\left(1 - \frac{a\tau}{h}\right)\mathrm{e}^{\mathrm{i}\varphi} + \frac{1}{2}\left(1 + \frac{a\tau}{h}\right)\mathrm{e}^{-\mathrm{i}\varphi} = \cos\varphi - \frac{a\tau}{h}\,\mathrm{i}\sin\varphi$. Схема устойчива ($|\lambda(\varphi)| \leqslant 1$) при выполнении условия $\frac{|a|\tau}{h} \leqslant 1$.

9.17. Исследовать устойчивость схемы

$$\frac{u_m^{n+1}-u_m^n}{\tau}+a\,\frac{u_{m+1}^n-u_{m-1}^n}{2h}-\frac{a^2\tau}{2}\,\frac{u_{m+1}^n-2u_m^n+u_{m-1}^n}{h^2}=0.$$

Ответ: $\lambda(\varphi)=1-\mathrm{i}\left(\frac{a\tau}{h}\right)\sin\varphi+\left(\frac{a^2\tau}{h^2}\right)(\cos\varphi-1)$. Схема устойчива при выполнении условия $\frac{|a|\tau}{h}\leqslant 1$.

9.18. Исследовать устойчивость схемы $\frac{u_m^{n+1}-u_m^n}{\tau}+a\,\frac{u_m^{n+1}-u_{m-1}^{n+1}}{h}=0$.

Ответ: $\lambda(\varphi)=\left(1+\frac{a\tau}{h}-\frac{a\tau}{h}\,\mathrm{e}^{-\mathrm{i}\varphi}\right)^{-1}$. Введем следующие обозначения: $\delta=\sup\limits_{0\leqslant\varphi\leqslant 2\pi}|\lambda(\varphi)|$ и $\gamma=\frac{a\tau}{h}$. Тогда:

при $a>0$ или при $\gamma\leqslant -1$ выполняется неравенство $\delta\leqslant 1$, т. е. схема устойчива;

при $-1<\gamma<0$ имеем $\delta=\frac{1}{|2\gamma+1|}>1$, т. е. схема неустойчива.

9.19. Исследовать устойчивость схемы $\frac{u_m^{n+1}-u_m^n}{\tau}+a\,\frac{u_{m+1}^{n+1}-u_{m-1}^{n+1}}{2h}=0$.

Ответ: $\lambda(\varphi)=\left(1+\frac{a\tau}{h}\,\mathrm{i}\sin\varphi\right)^{-1}$. Так как $|\lambda(\varphi)|\leqslant 1$, то схема устойчива при любых τ и h.

9.20. Исследовать устойчивость схемы

$$\frac{u_m^{n+1}-\frac{u_{m+1}^n+u_{m-1}^n}{2}}{\tau}+a\,\frac{u_{m+1}^n-u_{m-1}^n}{2h}=0.$$

Ответ: $\lambda(\varphi)=\cos\varphi-\frac{a\tau}{h}\,\mathrm{i}\sin\varphi$. Схема устойчива ($|\lambda(\varphi)|\leqslant 1$) при выполнении условия $\frac{|a|\tau}{h}\leqslant 1$ (ср. с 9.16).

9.21. Исследовать устойчивость схемы

$$\frac{u_m^{n+1}-\frac{u_{m+1}^n+u_{m-1}^n}{2}}{\tau}+a\,\frac{u_{m+1}^{n+1}-u_{m-1}^{n+1}}{2h}=0.$$

Ответ: схема безусловно устойчива.

9.22. Для уравнения $u_t+u_x=0$ рассматривается семейство схем с параметром θ

$$\frac{u_m^{n+1}-u_m^n}{\tau}+\theta\,\frac{u_{m+1}^n-u_m^n}{h}+(1-\theta)\,\frac{u_m^n-u_{m-1}^n}{h}=0.$$

При каких $\theta\in[0,1]$ схема устойчива?

Ответ: $0\leqslant\theta\leqslant\frac{1}{2}$.

9.23. Для уравнения $u_t + u_x = 0$ рассматривается семейство схем с параметром θ

$$\frac{u_m^{n+1} - u_m^n}{\tau} + \theta\ \frac{u_m^{n+1} - u_{m-1}^{n+1}}{h} + (1-\theta)\ \frac{u_m^n - u_{m-1}^n}{h} = 0.$$

При каких значениях $\theta \in [0,1]$ схема безусловно устойчива?

Ответ: $\frac{1}{2} \leqslant \theta \leqslant 1$.

Дифференциальное приближение. Пусть для дифференциальной задачи $L\,u = f$ построена разностная схема $L_h\,v_h = f_h$ и найдено ее решение v_h. Предположим, что это решение является следом на сетке некоторой гладкой функции v, т. е. $v_h = (v)_h$.

Дифференциальное уравнение, решением которого является функция v, называют *дифференциальным приближением разностной схемы.*

Как правило, дифференциальное приближение содержит бесконечное число слагаемых, зависящих от производных функции v с коэффициентами, пропорциональными шагам сетки. Так как интересна асимптотическая зависимость относительно сеточных параметров, то обычно ограничиваются одним или двумя старшими членами асимптотики (т. е. наиболее медленно убывающими слагаемыми). В этом случае говорят о *первом* дифференциальном приближении. Дифференциальное приближение при этом стараются записать в такой форме, чтобы в дополнительные (по сравнению с исходным дифференциальным уравнением) слагаемые не входили частные производные по временной переменной. Это удобно для анализа различий между решениями u и v_h. В частности, первое дифференциальное приближение полезно при исследовании корректности (устойчивости) разностной схемы.

9.24. Получить дифференциальное приближение разностной схемы

$$\frac{v_m^{n+1} - v_m^n}{\tau} + a\ \frac{v_m^n - v_{m-1}^n}{h} = 0 \tag{9.1}$$

с точностью до членов порядка $O(\tau^3 + h^3)$.

$\lhd$ Используя гладкость функции v, получим разложения в ряды Тейлора в точке $(x_m, t_n,)$ значений v_m^{n+1} и v_{m-1}^n с точностью $O(\tau^4 + h^4)$ и подставим их в (9.1). Имеем

$$\frac{1}{\tau}\left(\left[v + \tau\,\frac{\partial v}{\partial t} + \frac{\tau^2}{2}\,\frac{\partial^2 v}{\partial t^2} + \frac{\tau^3}{6}\,\frac{\partial^3 v}{\partial t^3} + O(\tau^4)\right] - v\right) +$$
$$+ \frac{a}{h}\left(v - \left[v - h\,\frac{\partial v}{\partial x} + \frac{h^2}{2}\,\frac{\partial^2 v}{\partial x^2} - \frac{h^3}{6}\,\frac{\partial^3 v}{\partial x^3} + O(h^4)\right]\right) = 0.$$

Это соотношение удобно преобразовать к виду

$$\frac{\partial v}{\partial t} + a\,\frac{\partial v}{\partial x} = -\frac{\tau}{2}\,\frac{\partial^2 v}{\partial t^2} + \frac{ah}{2}\,\frac{\partial^2 v}{\partial x^2} - \frac{\tau^2}{6}\,\frac{\partial^3 v}{\partial t^3} - \frac{ah^2}{6}\,\frac{\partial^3 v}{\partial x^3} + O(\tau^3 + h^3). \tag{9.2}$$

В левой части равенства (9.2) находится оператор уравнения, которое аппроксимирует разностная схема (9.1), а в правой — погрешность аппроксимации, которая в общем случае отлична от нуля.

Производные по времени, входящие в погрешность аппроксимации, заменим с требуемой точностью производными по пространственной переменной. Для этого выразим производную $\dfrac{\partial^2 v}{\partial t^2}$ через производную по x. Формально дифференцируя (9.2) по времени, получаем

$$\frac{\partial^2 v}{\partial t^2} + a\,\frac{\partial^2 v}{\partial t \partial x} = -\,\frac{\tau}{2}\,\frac{\partial^3 v}{\partial t^3} + \frac{ah}{2}\,\frac{\partial^3 v}{\partial t \partial x^2} - \frac{\tau^2}{6}\,\frac{\partial^4 v}{\partial t^4} - \frac{ah^2}{6}\,\frac{\partial^4 v}{\partial t \partial x^3} + O(\tau^3 + h^3),$$

а дифференцируя (9.2) по x и умножая на $-a$, находим

$$-a\,\frac{\partial^2 v}{\partial t \partial x} - a^2\,\frac{\partial^2 v}{\partial x^2} = \frac{a\tau}{2}\,\frac{\partial^3 v}{\partial t^2 \partial x} - \frac{a^2 h}{2}\,\frac{\partial^3 v}{\partial x^3} + \frac{a\tau^2}{6}\,\frac{\partial^4 v}{\partial t^3 \partial x} + \frac{a^2 h^2}{6}\,\frac{\partial^4 v}{\partial x^4} + O(\tau^3 + h^3).$$

Складывая два последних равенства, имеем

$$\frac{\partial^2 v}{\partial t^2} = a^2\,\frac{\partial^2 v}{\partial x^2} + \tau\left(-\,\frac{1}{2}\,\frac{\partial^3 v}{\partial t^3} + \frac{a}{2}\,\frac{\partial^3 v}{\partial t^2 \partial x} + O(\tau)\right) + h\left(\frac{a}{2}\,\frac{\partial^3 v}{\partial t \partial x^2} - \frac{a^2}{2}\,\frac{\partial^3 v}{\partial x^3} + O(h)\right). \tag{9.3}$$

Из уравнения (9.2) аналогично можно получить следующие выражения для производных $\dfrac{\partial^3 v}{\partial t^3}$, $\dfrac{\partial^3 v}{\partial t^2 \partial x}$, $\dfrac{\partial^3 v}{\partial t \partial x^2}$:

$$\frac{\partial^3 v}{\partial t^3} = -a^3\,\frac{\partial^3 v}{\partial x^3} + O(\tau + h)\,, \quad \frac{\partial^3 v}{\partial t^2 \partial x} = a^2\,\frac{\partial^3 v}{\partial x^3} + O(\tau + h)\,,$$

$$\frac{\partial^3 v}{\partial t \partial x^2} = -a\,\frac{\partial^3 v}{\partial x^3} + O(\tau + h)\,. \tag{9.4}$$

Заменяя по формулам (9.3) и (9.4) в правой части уравнения (9.2) производные по временной переменной производными по пространственной переменной, получаем

$$\frac{\partial v}{\partial t} + a\,\frac{\partial v}{\partial x} = \frac{ah}{2}\,(1-\gamma)\,\frac{\partial^2 v}{\partial x^2} - \frac{ah^2}{6}\,(2\gamma^2 - 3\gamma + 1)\,\frac{\partial^3 v}{\partial x^3} + O(\tau^3 + h^3),$$

где $\gamma = \dfrac{a\tau}{h}$. Это уравнение и является дифференциальным приближением разностной схемы (9.1) с точностью до членов порядка $O(\tau^3 + h^3)$.

Важно, что для замены временных производных пространственными используется уравнение (9.2), а не исходное уравнение $u_t + a\,u_x = 0$. Это связано с тем, что искомое решение $u(x,t)$ в общем случае не совпадает с решением $v(x,t)$ дифференциального приближения. $\triangleright$

9.25. Получить с точностью до членов порядка $O(\tau^3 + h^3)$ дифференциальное приближение разностной схемы

$$\frac{v_m^{n+1} - v_m^n}{\tau} + a\,\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0.$$

Ответ: $\dfrac{\partial v}{\partial t} + a\,\dfrac{\partial v}{\partial x} = -\,\dfrac{\tau\,a^2}{2}\,\dfrac{\partial^2 v}{\partial x^2} - \left(\dfrac{ah^2}{6} + \dfrac{a^3\tau^2}{3}\right)\dfrac{\partial^3 v}{\partial x^3} + O(\tau^3 + h^3)$.

9.26. Получить с точностью до членов порядка $O(\tau^3+h^3)$ дифференциальное приближение разностной схемы

$$\frac{v_m^{n+1}-v_m^{n-1}}{2\tau}+a\,\frac{v_{m+1}^n-v_{m-1}^n}{2h}=0.$$

Ответ: $\dfrac{\partial v}{\partial t}+a\,\dfrac{\partial v}{\partial x}=-\left(\dfrac{ah^2}{6}-\dfrac{a^3\tau^2}{6}\right)\dfrac{\partial^3 v}{\partial x^3}+O(\tau^3+h^3)$.

9.27. Получить с точностью до членов порядка $O(\tau^3+h^3)$ дифференциальное приближение разностной схемы

$$\frac{v_m^{n+1}-v_m^n}{\tau}+a\,\frac{v_{m+1}^n-v_{m-1}^n}{2h}=\frac{h^2}{2\tau}\,\frac{v_{m+1}^n-2v_m^n+v_{m-1}^n}{h^2}\,.$$

Ответ: $\dfrac{\partial v}{\partial t}+a\,\dfrac{\partial v}{\partial x}=\dfrac{a\,h}{2}\left(\dfrac{1}{\gamma}-\gamma\right)\dfrac{\partial^2 v}{\partial x^2}+\dfrac{a\,h^2}{3}\left(1-\gamma^2\right)\dfrac{\partial^3 v}{\partial x^3}+O(\tau^3+h^3),\ \gamma=\tau\,\dfrac{a}{h}$.

9.28. Получить с точностью до членов порядка $O(\tau^3+h^3)$ дифференциальное приближение разностной схемы

$$\frac{v_m^{n+1}-v_m^n}{\tau}+a\,\frac{v_{m+1}^{n+1}-v_{m-1}^{n+1}}{2h}=0.$$

Ответ: $\dfrac{\partial v}{\partial t}+a\,\dfrac{\partial v}{\partial x}=\dfrac{\tau\,a^2}{2}\,\dfrac{\partial^2 v}{\partial x^2}-\left(\dfrac{ah^2}{6}+\dfrac{a^3\tau^2}{3}\right)\dfrac{\partial^3 v}{\partial x^3}+O(\tau^3+h^3)$.

9.29. Получить с точностью до членов порядка $O(\tau^3+h^3)$ дифференциальное приближение разностной схемы

$$\frac{v_m^{n+1}-v_m^n}{\tau}+a\,\frac{v_{m+1}^{n+1}-v_m^{n+1}}{h}=0.$$

Ответ: $\dfrac{\partial v}{\partial t}+a\,\dfrac{\partial v}{\partial x}=\dfrac{a\,h}{2}\,(\gamma-1)\,\dfrac{\partial^2 v}{\partial x^2}+\dfrac{a\,h^2}{6}\left(-1-2\,\gamma^2+3\,\gamma\right)\dfrac{\partial^3 v}{\partial x^3}+O(\tau^3+h^3)$, $\gamma=\tau\,\dfrac{a}{h}$.

9.30. Получить с точностью до членов порядка $O(\tau^3+h^3)$ дифференциальное приближение разностной схемы

$$\frac{v_m^{n+1}-v_m^n}{\tau}+\frac{a}{2}\left(\frac{v_{m+1}^{n+1}-v_m^{n+1}}{h}+\frac{v_m^n-v_{m-1}^n}{h}\right)=0$$

с точностью до членов порядка $O(\tau^3+h^3)$.

Указание. Воспользовавшись соотношениями

$$\frac{v_{m+1}^{n+1}-v_m^{n+1}}{h}=\frac{1}{2}\left(\frac{v_{m+1}^{n+1}-v_m^{n+1}}{h}+\frac{v_{m+1}^n-v_m^n}{h}\right)+\frac{1}{2}\left(\frac{v_{m+1}^{n+1}-v_m^{n+1}}{h}-\frac{v_{m+1}^n-v_m^n}{h}\right),$$

$$\frac{v_m^n-v_{m-1}^n}{h}=\frac{1}{2}\left(\frac{v_m^{n+1}-v_{m-1}^{n+1}}{h}+\frac{v_m^n-v_{m-1}^n}{h}\right)-\frac{1}{2}\left(\frac{v_m^{n+1}-v_{m-1}^{n+1}}{h}-\frac{v_m^n-v_{m-1}^n}{h}\right),$$

сначала привести схему к виду

$$\frac{v_m^{n+1}-v_m^n}{\tau}+\frac{a}{2}\left(\frac{v_{m+1}^{n+1}-v_{m-1}^{n+1}}{2h}+\frac{v_{m+1}^n-v_{m-1}^n}{2h}\right)+$$
$$+\frac{a\,h}{4}\left(\frac{v_{m+1}^{n+1}-2\,v_m^{n+1}+v_{m-1}^{n+1}}{h^2}+\frac{v_{m+1}^n-2\,v_m^n+v_{m-1}^n}{h^2}\right)=0.$$

Ответ: $\dfrac{\partial v}{\partial t}+a\,\dfrac{\partial v}{\partial x}=\left(\dfrac{a^2h\,\tau}{4}-\dfrac{ah^2}{6}-\dfrac{a^3\tau^2}{12}\right)\dfrac{\partial^3 v}{\partial x^3}+O(\tau^3+h^3)$.

Уравнение колебаний струны. Рассмотрим первую краевую задачу для однородного уравнения колебаний струны

$$\frac{\partial^2 u}{\partial t^2}=\frac{\partial^2 u}{\partial x^2}\,,\quad 0<t\leqslant T,\quad u(0,t)=u(1,t)=0,$$
$$u(x,0)=u_0(x),\quad \frac{\partial u}{\partial t}\,(x,0)=v_0(x),\quad 0<x<1.$$

9.31. На пятиточечном шаблоне «крест» построить разностную схему второго порядка сходимости.

Ответ: при $\tau\leqslant h$ схема

$$\frac{u_m^{n+1}-2u_m^n+u_m^{n-1}}{\tau^2}=\Lambda u_m^n,$$

где

$$\Lambda u_m=\frac{u_{m+1}-2u_m+u_{m-1}}{h^2}\,,\quad m=1,\ldots,M-1,$$
$$u_0^{n+1}=u_M^{n+1}=0, n=0,\ldots,N-1,\quad u_m^0=u_0(mh),$$
$$\frac{u_m^1-u_m^0}{\tau}=v_0(mh)+\frac{\tau}{2}\,\frac{u_{m+1}^0-2u_m^0+u_{m-1}^0}{h^2}\,,\quad m=1,\ldots,M-1,$$

имеет второй порядок сходимости.

9.32. Найти порядок аппроксимации и методом разделения переменных условия устойчивости для семейства схем с весами

$$\frac{u_m^{n+1}-2u_m^n+u_m^{n-1}}{\tau^2}=\theta\Lambda u_m^{n+1}+(1-2\theta)\Lambda u_m^n+\theta\Lambda u_m^{n-1},$$
$$u_0^{n+1}=u_M^{n+1}=0,\quad u_m^0=u_0(mh),\quad u_m^1=\bar v_0(mh),$$

где θ — весовой параметр, оператор Λ определен в 9.31, а функция $\bar v_0(mh)$ выбрана с нужным порядком аппроксимации.

$\lhd$ Если θ не зависит от шагов сетки, то порядок аппроксимации $O(\tau^2+h^2)$. Если $\theta=\theta_0-\dfrac{h^2}{12\tau^2}$, то порядок аппроксимации на решении $O(\tau^2+h^4)$. Параметр θ_0 не зависит от шагов сетки и выбирается из условия устойчивости схемы.

Найдем условие устойчивости. Рассмотрим частные решения разностной задачи следующего вида: $u_m^n = \mu_h^n(k)\sin(\pi k m h)$. Здесь и далее для простоты исследования условие $|\mu_h| \leqslant \mathrm{e}^{c\tau}$ заменяем условием $|\mu_h| \leqslant 1$. Для μ_h получаем уравнение

$$\mu^2 - 2(1-\alpha)\mu + 1 = 0, \quad \alpha = \frac{1}{2}\,\frac{\tau^2\lambda_h(k)}{1+\theta\tau^2\lambda_h(k)}\,.$$

Условие $|\mu_{1,2}| \leqslant 1$, $\mu_1 \neq \mu_2$, выполняется, если дискриминант меньше нуля, т. е. при $0 < \alpha < 2$. Отсюда имеем $\theta > \frac{1}{4} - \frac{1}{\tau^2\lambda_h(k)}$. Так как $\lambda_h(k) = \frac{4}{h^2}\sin^2\left(\frac{\pi h k}{2}\right) < \frac{4}{h^2}$, то условие устойчивости имеет вид $\theta \geqslant \frac{1}{4} - \frac{h^2}{4\tau^2}$.

Для явной схемы ($\theta = 0$) полученное неравенство приводит к условию $\tau \leqslant h$. $\triangleright$

9.33. Найти условия устойчивости двухпараметрического семейства схем

$$\frac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{\tau^2} = \theta_1\Lambda u_m^{n+1} + (1-\theta_1-\theta_2)\Lambda u_m^n + \theta_2\Lambda u_m^{n-1},$$

$$u_0^{n+1} = u_M^{n+1} = 0, \quad u_m^0 = u_0(mh), \quad u_m^1 = \bar{v}_0(mh),$$

где θ_1, θ_2 — параметры, а функция $\bar{v}_0(mh)$ выбрана с нужным порядком аппроксимации.

$\triangleleft$ Определив самосопряженный оператор A как оператор $(-\Lambda)$, действующий на пространстве сеточных функций, равных нулю на границе, получаем, что трехслойная схема записывается в канонической форме (см. (9.15) в разделе 9.4) с самосопряженными операторами

$$B = (\theta_1 - \theta_2)\tau A, \quad R = \frac{1}{\tau^2}\,I + \frac{\theta_1+\theta_2}{2}\,A.$$

Условия устойчивости трехслойных схем имеют вид

$$B \geqslant 0, \quad R = R^* > \frac{1}{4}\,A, \quad A = A^* > 0.$$

Условие $B \geqslant 0$ приводит к неравенству $\theta_1 \geqslant \theta_2$, а условие $R > \frac{1}{4}\,A$ можно записать в следующем виде:

$$\frac{1}{\tau^2}\,I + \left(\frac{\theta_1+\theta_2}{2} - \frac{1}{4}\right)A > 0.$$

Данное неравенство по определению означает, что

$$\frac{1}{\tau^2}\,\|u\| + \left(\frac{\theta_1+\theta_2}{2} - \frac{1}{4}\right)(Au, u) > 0 \quad \forall u \neq 0,$$

и так как наибольшее собственное число оператора A равно $\frac{4}{h^2}\sin^2\frac{\pi h(M-1)}{2} < \frac{4}{h^2}$, то искомая оценка имеет вид

$$\frac{\theta_1+\theta_2}{2} \geqslant \frac{1}{4}\left(1 - \frac{h^2}{\tau^2}\right).$$ $\triangleright$

О т в е т: из теории устойчивости трехслойных разностных схем (см. раздел 9.4) следует, что достаточными условиями устойчивости схемы являются следующие неравенства:

$$\theta_1 \geqslant \theta_2, \quad \frac{\theta_1+\theta_2}{2} \geqslant \frac{1}{4}\left(1-\frac{h^2}{\tau^2}\right).$$

Методом разделения переменных можно показать, что найденные достаточные условия асимптотически совпадают с необходимыми.

В случае $\theta_1 = \theta_2 = \theta$ (см. 9.32) схема имеет второй порядок аппроксимации, а условие устойчивости сводится к неравенству $\theta \geqslant \frac{1}{4} - \frac{h^2}{4\tau^2}$.

9.3. Эллиптические уравнения

Построение и исследование разностных схем для уравнений в частных производных эллиптического типа в простейшем случае проводят в области прямоугольной формы

$$D = \{(x,y) : X > x > 0, Y > y > 0\}$$

на примере уравнения с заданными переменными коэффициентами $a_i(x,y) \geqslant a_0 > 0,\ i = 1,2,$

$$Lu \equiv \frac{\partial}{\partial x}\left(a_1(x,y)\,\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a_2(x,y)\,\frac{\partial u}{\partial y}\right) = f(x,y)$$

с однородными краевыми условиями первого рода

$$u(0,y) = u(X,y) = 0 \quad \text{при} \quad Y \geqslant y \geqslant 0\,,$$
$$u(x,0) = u(x,Y) = 0 \quad \text{при} \quad X \geqslant x \geqslant 0\,.$$

Наиболее употребительным является случай уравнения Пуассона $(a_i(x,y) \equiv 1,\ i = 1,2)$

$$\Delta\, u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x,y)\,.$$

В общем случае на любой части границы краевое условие может быть задано в виде линейной комбинации функции и производной первого порядка. Тогда необходимо обратить внимание на способ его аппроксимации.

Типичным примером эллиптического оператора четвертого порядка является бигармонический оператор

$$\Delta^2 u \equiv \frac{\partial^4 u}{\partial x^4} + 2\,\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4}\ ,$$

для которого краевое условие может содержать линейную комбинацию производных неизвестной функции до третьего порядка включительно.

Особенность постановки эллиптических задач — наличие только краевых условий. Поэтому аппроксимация и устойчивость исследуются как в случае линейных краевых задач для обыкновенных дифференциальных уравнений.

Для удобства независимые переменные также будем использовать в виде $x_1 = x$, $x_2 = y$. Введем обозначение $\Lambda_\alpha u(x_1, x_2)$, $\alpha = 1, 2$, для разностного аналога оператора второй производной $L_\alpha u = \dfrac{\partial^2 u}{\partial x_\alpha^2}$ по переменной x_α, например,

$$\Lambda_1 u(x_1, x_2) = \frac{u(x_1 + h_1, x_2) - 2\,u(x_1, x_2) + u(x_1 - h_1, x_2)}{h_1^2} \ .$$

Аналогичный смысл имеет выражение

$$\Lambda_1 u_{i,j} = \frac{u_{i+1,j} - 2\,u_{i,j} + u_{i-1,j}}{h_1^2} \ ,$$

где $u_{i,j}$ соответствует $u(ih_1, jh_2)$.

9.34. Оценить погрешность аппроксимации оператора Лапласа Δ оператором $\Delta^h = \Lambda_1 + \Lambda_2$ (стандартная аппроксимация на шаблоне «крест»).

Ответ: $\Delta^h u - \Delta u = \left(\frac{h_1^2}{12}\,L_1^2 + \frac{h_2^2}{12}\,L_2^2\right) u + O\left(h_1^4 + h_2^4\right) \equiv O\left(h_1^2 + h_2^2\right)$.

9.35. Построить аппроксимацию оператора Лапласа и оценить ее погрешность на шаблоне «косой крест» при $h_1 = h_2 = h$

$$\Delta^h u = \frac{1}{h^2}\,[a_{0,0}u(x_1, x_2) + a_{1,1}u(x_1 + h, x_2 + h) + a_{1,-1}u(x_1 + h, x_2 - h) +$$

$$+ a_{-1,1}u(x_1 - h, x_2 + h) + a_{-1,-1}u(x_1 - h, x_2 - h)]\,,$$

где $a_{k,j}$ не зависят от h.

$\lhd$ «Косой крест» — это обычный «крест» с шагом $\sqrt{2}h$ в системе координат, полученной поворотом исходной системы на $\frac{\pi}{4}$. $\rhd$

Ответ: $\Delta^h = \Lambda_1 + \Lambda_2 + \frac{h^2}{2}\,\Lambda_1\Lambda_2$, или $a_{0,0} = -2$ и $a_{k,j} = \frac{1}{2}$ в остальных случаях. Погрешность аппроксимации равна $O(h^2)$.

9.36. Построить аппроксимацию оператора Лапласа и оценить ее погрешность на треугольной решетке (область разбита на непересекающиеся правильные треугольники со стороной h).

$\lhd$ Вторую производную функции $u(x, y)$ в любом требуемом направлении можно выразить тремя вторыми производными: $\dfrac{\partial^2 u}{\partial x^2}$, $\dfrac{\partial^2 u}{\partial y^2}$, $\dfrac{\partial^2 u}{\partial x \partial y}$.

Введем новые координаты x_n и y_n (индекс n означает «новые»):

$$x = x_n \cos\theta - y_n \sin\theta, \quad y = x_n \sin\theta + y_n \cos\theta.$$

Геометрически это означает, что начальная координатная система поворачивается против часовой стрелки на угол θ относительно начальных осей. Поэтому вторая производная $u(x, y)$ в направлении θ вычисляется так:

$$\frac{\partial^2 u}{\partial x_n^2} = \frac{\partial}{\partial x_n}\left(\frac{\partial u}{\partial x}\,\frac{\partial x}{\partial x_n} + \frac{\partial u}{\partial y}\,\frac{\partial y}{\partial x_n}\right) = \frac{\partial}{\partial x_n}\left(\cos\theta\,\frac{\partial u}{\partial x} + \sin\theta\,\frac{\partial u}{\partial y}\right) =$$

$$= \left(\cos\theta\,\frac{\partial}{\partial x} + \sin\theta\,\frac{\partial}{\partial y}\right)^2 u = \cos^2\theta\,\frac{\partial^2 u}{\partial x^2} + \sin^2\theta\,\frac{\partial^2 u}{\partial y^2} + 2\sin\theta\cos\theta\,\frac{\partial^2 u}{\partial x \partial y}\,.$$

Обозначим основные направления линий из узла сетки x_0 через a, b, c. Значения угла θ для этих трех направлений соответственно таковы: $0, \frac{\pi}{3}$ и $\frac{2\pi}{3}$. Теперь оператор Лапласа можно выразить через частные производные второго порядка по данным направлениям. Имеем

$$\frac{\partial^2 u}{\partial a^2} = \frac{\partial^2 u}{\partial x^2},$$
$$\frac{\partial^2 u}{\partial b^2} = \frac{1}{4}\,\frac{\partial^2 u}{\partial x^2} + \frac{3}{4}\,\frac{\partial^2 u}{\partial y^2} + \frac{\sqrt{3}}{2}\,\frac{\partial^2 u}{\partial x \partial y},$$
$$\frac{\partial^2 u}{\partial c^2} = \frac{1}{4}\,\frac{\partial^2 u}{\partial x^2} + \frac{3}{4}\,\frac{\partial^2 u}{\partial y^2} - \frac{\sqrt{3}}{2}\,\frac{\partial^2 u}{\partial x \partial y}.$$

Сложив эти равенства, получим

$$\frac{\partial^2 u}{\partial a^2} + \frac{\partial^2 u}{\partial b^2} + \frac{\partial^2 u}{\partial c^2} = \frac{3}{2}\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right).$$

Заменим вторые производные по направлениям a, b и c обычными аппроксимациями второго порядка

$$\left(\frac{\partial^2 u}{\partial a^2}\right)_0 = \frac{1}{h^2}\,(u_4 - 2u_0 + u_1) + O(h^2),$$
$$\left(\frac{\partial^2 u}{\partial b^2}\right)_0 = \frac{1}{h^2}\,(u_5 - 2u_0 + u_2) + O(h^2),$$
$$\left(\frac{\partial^2 u}{\partial c^2}\right)_0 = \frac{1}{h^2}\,(u_6 - 2u_0 + u_3) + O(h^2),$$

где u_i — значения в соседних с x_0 узлах решетки, отстоящих на расстояние h (нумерация ведется против часовой стрелки). В результате имеем искомую аппроксимацию оператора Лапласа

$$\Delta^h u_0 = \frac{2}{3h^2}\,(u_1 + u_2 + \cdots + u_6 - 6\,u_0). \qquad \triangleright$$

9.37. Используя значения функции u в центре A_0 и в вершинах A_k правильного n-угольника со стороной h, получить аппроксимацию оператора Лапласа $\Delta^h u$ в центре многоугольника. Оценить ее порядок для различных n.

Ответ: $\Delta^h u(A_0) = \frac{16}{h^2}\sin^2\frac{\pi}{n}\left[-u(A_0) + \frac{1}{n}\sum\limits_{k=1}^{n} u(A_k)\right]$.

9.38. Описать все девятиточечные разностные аппроксимации оператора Лапласа Δ^h, имеющие вид

$$\frac{1}{h^2}\,[a_{0,0}u(x_1, x_2) + a_{1,0}u(x_1 + h, x_2) + a_{-1,0}u(x_1 - h, x_2) +$$
$$+ a_{0,1}u(x_1, x_2 + h) + a_{0,-1}u(x_1, x_2 - h) + a_{1,1}u(x_1 + h, x_2 + h) +$$
$$+ a_{1,-1}u(x_1 + h, x_2 - h) + a_{-1,1}u(x_1 - h, x_2 + h) + a_{-1,-1}u(x_1 - h, x_2 - h)],$$

где $a_{k,j}$ не зависят от h, и обладающие вторым порядком аппроксимации, т. е.

$$\Delta^h u(x_1, x_2) - \Delta u(x_1, x_2) = O\left(h^2\right) \quad \text{при} \quad u \in C^{(4)}\,.$$

Ответ: $a_{0,0} = 4c - 4$, $a_{1,1} = a_{1,-1} = a_{-1,1} = a_{-1,-1} = c$, $a_{1,0} = a_{-1,0} = a_{0,1} = a_{0,-1} = 1 - 2c$, или, что то же самое, $\Delta^h = \Lambda_1 + \Lambda_2 + c\,h^2 \Lambda_1 \Lambda_2$, где c — произвольная постоянная.

9.39. Какие из разностных операторов в 9.38 отрицательно определенные?

Ответ: все операторы при $c < \frac{1}{2}$.

9.40. Построить тринадцатиточечную разностную аппроксимацию бигармонического оператора Δ^2, использующую узлы (x_1, x_2), $(x_1 \pm h, x_2)$, $(x_1, x_2 \pm h)$, $(x_1 \pm 2h, x_2)$, $(x_1, x_2 \pm 2h)$, $(x_1 \pm h, x_2 \pm h)$, и оценить погрешность аппроксимации на функциях $u \in C^{(6)}$.

Ответ: $\left(\Delta^h\right)^2 u = \left(\Lambda_1^2 + 2\,\Lambda_1 \Lambda_2 + \Lambda_2^2\right) u = \left(\Lambda_1 + \Lambda_2\right)^2 u$; погрешность аппроксимации равна $O\left(h^2\right)$.

9.41. Если u — гармоническая функция в ограниченной области D с границей Γ, то $\int\limits_\Gamma \frac{\partial u}{\partial n}\, d\Gamma = 0$, где $\frac{\partial}{\partial n}$ — производная по направлению внешней нормали к границе Γ. Сформулировать и доказать аналог этого равенства для решений разностного уравнения

$$\Delta^h u_{i,j} \equiv \left(\Lambda_1 + \Lambda_2\right) u_{i,j} = 0\,, \quad 1 \leqslant i < N_1,\ \ 1 \leqslant j < N_2\,,$$

в прямоугольной области, покрытой равномерной сеткой с шагом h ($N_1 h = X$, $N_2 h = Y$).

Ответ:
$$h \sum_{j=1}^{N_2 - 1} \left(\frac{u_{N_1,j} - u_{N_1 - 1,j}}{h} - \frac{u_{1,j} - u_{0,j}}{h} \right) +$$
$$+ h \sum_{i=1}^{N_1 - 1} \left(\frac{u_{i,N_2} - u_{i,N_2 - 1}}{h} - \frac{u_{i,1} - u_{i,0}}{h} \right) = 0\,.$$

9.42. Записать разностную схему во внутренних узлах сетки для уравнения Пуассона с аппроксимацией на решении $O\left(h^4\right)$.

Указание. $\left(\Lambda_1 + \Lambda_2 + \frac{h^2}{6}\,\Lambda_1 \Lambda_2\right) u = f + \frac{h^2}{12}\left(\Lambda_1 + \Lambda_2\right)\, f\, + O\left(h^4\right)$.

9.43. Записать разностную схему во внутренних узлах сетки для уравнения Пуассона с аппроксимацией на решении $O\left(h^6\right)$.

Указание. $\left(\Lambda_1 + \Lambda_2 + \frac{h^2}{6}\,\Lambda_1 \Lambda_2\right) u = f + \frac{h^2}{12}(\Lambda_1 + \Lambda_2) f - \frac{h^4}{240}(\Lambda_1^2 + \Lambda_2^2) f +$
$+ \frac{h^4}{90}\,\Lambda_1 \Lambda_2 f + O(h^6)$.

9.44. Для уравнения $\Delta u = f$ построить аппроксимацию на решении с порядком $O(h^2)$ граничного условия $\dfrac{\partial u}{\partial x_1} - \alpha u = 0$ при $x_1 = 0$, используя минимальное количество узлов вдоль оси x_1.

Ответ: $\dfrac{u_{1,j} - u_{0,j}}{h_1} - \dfrac{h_1}{2}(f_{0,j} - \Lambda_2 u_{0,j}) - \alpha u_{0,j} = 0$.

9.45. Для уравнения $\Delta u = f$ построить аппроксимацию на решении с порядком $O(h^4)$ граничного условия $\dfrac{\partial u}{\partial x_1} - \alpha u = 0$ при $x_1 = 0$, используя минимальное количество узлов вдоль оси x_1.

9.46. Пусть в прямоугольной области, покрытой равномерной сеткой с шагом h, определен разностный аналог оператора Лапласа

$$\Delta^h u_{i,j} \equiv (\Lambda_1 + \Lambda_2)\, u_{i,j}\,, \quad 1 \leqslant i < N_1,\ \ 1 \leqslant j < N_2\,.$$

Показать, что если справедливо неравенство $\Delta^h u_{i,j} \leqslant 0$ при всех $1 \leqslant i < N_1,\ 1 \leqslant j < N_2$, то функция $u_{i,j}$ достигает наименьшего значения хотя бы в одной точке границы, т. е. при $i = 0$ или $i = N_1$, либо при $j = 0$ или $j = N_2$.

$\lhd$ Будем считать, что функция $u_{i,j}$ отлична от константы, так как в этом случае утверждение является тривиальным.

Предположим теперь противное, т. е. что минимальное значение достигается во внутреннем узле сетки (вообще таких узлов может быть несколько). Пусть его номер (m, n); в этом случае справедливо неравенство $\Delta^h u_{m,n} \leqslant 0$:

$$u_{m,n} \geqslant \frac{1}{4}\left(u_{m+1,n} + u_{m-1,n} + u_{m,n+1} + u_{m,n-1}\right)\,.$$

Знак $>$ не имеет места, так как $u_{m,n}$ — минимальное значение функции на сетке. Знак равенства означает, что в окрестности узла (m, n) значения функции $u_{i,j}$ совпадают

$$u_{m+1,n} = u_{m-1,n} = u_{m,n+1} = u_{m,n-1} = u_{m,n}\,.$$

Продолжая рассуждения для этих узлов, затем для их соседей, получим в силу связности сетки, что при выполнении неравенства $\Delta^h u_{i,j} \leqslant 0$ функция обязана быть константой. Это противоречит исходной посылке, значит наименьшее значение обязано достигаться на границе, где указанное в условии неравенство места не имеет. $\rhd$

9.47. Доказать, что если в обозначениях 9.46 справедливо неравенство $\Delta^h u_{i,j} \geqslant 0$ при всех $1 \leqslant i < N_1$, $1 \leqslant j < N_2$ то функция $u_{i,j}$ достигает наибольшего значения хотя бы в одной точке границы.

9.48. Пусть в прямоугольной области, покрытой равномерной сеткой с шагом h, разностный аналог оператора Лапласа

$$\Delta^h u_{i,j} \equiv (\Lambda_1 + \Lambda_2)\, u_{i,j}\,, \quad 1 \leqslant i < N_1,\ \ 1 \leqslant j < N_2\,,$$

определен на сеточных функциях $u_{i,j} \equiv u_h$, обращающихся в нуль на границе, т. е. при $i = 0, N_1$ и при $j = 0, N_2$. Доказать, что оператор

$(-\Delta^h)$ является симметричным, положительно определенным, и для него справедливы оценки

$$c_1\,(u_h, u_h) \leqslant \left(-\Delta^h u_h, u_h\right) \leqslant c_2\,(u_h, u_h)\ ,\ \ (u_h, v_h) = \sum_{i,j} u_{i,j} v_{i,j} h^2,$$

в которых постоянная $c_1 > 0$ не зависит от сеточного параметра h, а постоянная c_2 может быть выбрана равной $\dfrac{8}{h^2}$.

9.49. Показать, что для решения методом Гаусса разностного уравнения Пуассона $\Delta^h u_h = f_h$ с однородными условиями Дирихле на границе (см. 9.48) при естественной нумерации ненулевых неизвестных (либо по строкам, либо по столбцам, например, $u_h = (u_{1,1},\, u_{2,1},\, \ldots\,,\, u_{N_1-1,1},\, u_{1,2},\, \ldots\,,\, u_{N_1-1,N_2-1})^T$) требуется порядка $O\,(h^{-4})$ арифметических действий.

9.50. Упорядочить неизвестные в 9.49 так, чтобы количество арифметических действий при решении методом Гаусса имело порядок $O\,(h^{-3})$.

9.51. Пусть в единичном квадрате D задана регулярная («северо-восточная») триангуляция с шагом h и в качестве базисных функций используются кусочно-линейные над треугольниками функции. Записать систему уравнений метода Ритца (конечных элементов) для задачи

$$-\,\Delta\, u = f(x, y) \text{ в } \ D\,,\ u = 0 \text{ на } \ \Gamma\,.$$

$\lhd$ На множестве непрерывно дифференцируемых функций, обращающихся в нуль на границе Γ, введем норму

$$\|u\|_U = \left\{\int\limits_D \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right]\, dxdy\right\}^{1/2} = \left(\int\limits_D (\nabla u)^2\, dxdy\right)^{1/2}.$$

Замыкание указанного множества функций в этой норме является гильбертовым пространством; обозначим его через U. Рассмотрим задачу нахождения минимума функционала

$$\min_{v\in U} J(v) = \min_{v\in U} \left\{\int\limits_D (\nabla v)^2\, dxdy - 2\int\limits_D f v\, dxdy\right\}. \tag{9.5}$$

Если классическое решение u исходной задачи существует, то оно доставляет минимум функционалу (9.5). Обратное, вообще говоря, неверно: функция, доставляющая минимум функционалу (9.5) на U, не обязательно должна иметь непрерывные вторые производные. Таким образом, нахождение решения исходной задачи можно заменить более общей задачей нахождения минимума квадратичного функционала (9.5) на U.

Аппроксимируем U конечномерным подпространством U_h, которое построим следующим способом. Пусть заданы узлы

$$\overline{D}_h = \{(x,\, y):\ x = mh,\ y = nh;\ 0 \leqslant m,\, n \leqslant N\}.$$

Разобьем $\overline{D}$ на квадратные ячейки со стороной h и вершинами в узлах $\overline{D}_h$. Каждую ячейку $D_{mn} = \{(x,\, y):\ mh \leqslant x \leqslant (m+1)h,\ nh \leqslant y \leqslant (n+1)h\}$,

разобьем диагональю, проходящей через вершины $(m,\, n)$, $(m+1,\, n+1)$. Таким образом, вся область $\overline{D} = D \cup \Gamma$ будет разбита на прямоугольные треугольники с катетами, равными h. Эти треугольники назовем *элементарными*, а разбиение области $\overline{D}$ на треугольники — *триангуляцией области* $\overline{D}$. В качестве подпространства U_h пространства U возьмем пространство непрерывных в $\overline{D}$ функций, линейных на каждом элементарном треугольнике и обращающихся в нуль на Γ. Функции $\varphi_{mn}(x, y)$, которые принимают значения, равные единице в узле $(m,\, n)$ и нулю в других узлах, образуют базис в U_h.

Для построения конечноэлементной схемы воспользуемся методом Ритца. В качестве приближенного решения задачи (9.5) будем рассматривать функцию u_h, которая минимизирует функционал (9.5) на подпространстве U_h, т. е.

$$\min_{v \in U_h} J(v) = J(u_h).$$

Представим u_h в виде

$$u_h = \sum_{i,\, j=1}^{N-1} u_{ij} \varphi_{ij}(x, y),$$

где u_{ij} — коэффициенты, подлежащие определению. Отметим, что u_{ij}, в силу выбора функций φ_{ij}, является значением u_h в точке $(i,\, j)$. Запишем уравнения для определения этих коэффициентов. В точке минимума u_h функционала $J(v)$ должны выполняться равенства

$$\frac{\partial J(u_h)}{\partial u_{mn}} = 0; \quad m,\, n = 1, \ldots,\, N-1.$$

Вычислим левую часть этого соотношения:

$$\frac{\partial J}{\partial u_{mn}} = \frac{\partial}{\partial u_{mn}} \int\limits_D \left[\left(\sum_{i,j=1}^{N-1} u_{ij} \frac{\partial \varphi_{ij}}{\partial x} \right)^2 + \left(\sum_{i,j=1}^{N-1} u_{ij} \frac{\partial \varphi_{ij}}{\partial y} \right)^2 - 2f \sum_{i,j=1}^{N-1} u_{ij} \varphi_{ij} \right] dxdy =$$

$$= 2 \int\limits_D \left[\sum_{i,j=1}^{N-1} u_{ij} \left(\frac{\partial \varphi_{ij}}{\partial x} \frac{\partial \varphi_{mn}}{\partial x} + \frac{\partial \varphi_{ij}}{\partial y} \frac{\partial \varphi_{mn}}{\partial y} \right) - f \varphi_{mn} \right] dxdy.$$

Следовательно, система уравнений относительно u_{ij} имеет вид

$$\sum_{i,\, j=1}^{N-1} u_{ij} \int\limits_D \left(\frac{\partial \varphi_{ij}}{\partial x} \frac{\partial \varphi_{mn}}{\partial x} + \frac{\partial \varphi_{ij}}{\partial y} \frac{\partial \varphi_{mn}}{\partial y} \right) dxdy = \int\limits_D f \varphi_{mn}\, dxdy, \qquad (9.6)$$

$$m,\, n = 1, \ldots,\, N-1.$$

Функция φ_{mn} отлична от нуля лишь в тех элементарных треугольниках, которые имеют узел $(m,\, n)$ своей вершиной. Поэтому в каждом из уравнений (9.6) интегрирование ведется не по всей области D, а только по пересечению таких треугольников с D.

Множество точек, где $\varphi_{mn} \neq 0$, образует шестиугольник (рис. 1). Обозначим этот шестиугольник через S_{mn}, а входящие в него треугольники через $T_1, \ldots, T_6$. Положим

$$I^x_{mn}(i,\, j) = \int\limits_D \frac{\partial \varphi_{mn}}{\partial x}\, \frac{\partial \varphi_{ij}}{\partial x}\, dxdy;$$

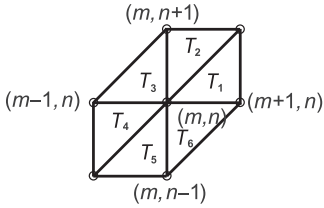


Рис. 1

так как $\dfrac{\partial \varphi_{mn}}{\partial x} \equiv 0$ в T_2 и T_5, то

$$I^x_{mn}(i,j)=\int\limits_D \frac{\partial \varphi_{mn}}{\partial x}\, \frac{\partial \varphi_{ij}}{\partial x}\, dxdy \;\; = \int\limits_{T_1 \cup T_6} \frac{\partial \varphi_{mn}}{\partial x}\, \frac{\partial \varphi_{ij}}{\partial x}\, dxdy + \int\limits_{T_3 \cup T_4} \frac{\partial \varphi_{mn}}{\partial x}\, \frac{\partial \varphi_{ij}}{\partial x}\, dxdy.$$

Отсюда следует, что $I^x_{mn} \neq 0$ лишь при $j=n$ и $i=m-1, i=m, i=m+1$. Вычисляя интегралы, получаем

$$I^x_{mn}(m,\, n)= \int\limits_{T_1 \cup T_6} \left(\frac{\partial \varphi_{mn}}{\partial x}\right)^2 dxdy + \int\limits_{T_3 \cup T_4} \left(\frac{\partial \varphi_{mn}}{\partial x}\right)^2 dxdy = 2,$$

$$I^x_{mn}(m+1,\, n)=I^x_{mn}(m-1,\, n)= \int\limits_{T_3 \cup T_4} \frac{\partial \varphi_{mn}}{\partial x}\, \frac{\partial \varphi_{m+1,n}}{\partial x}\, dxdy = -1.$$

Аналогично для

$$I^y_{mn}(i,\, j)=\int\limits_D \frac{\partial \varphi_{mn}}{\partial y}\, \frac{\partial \varphi_{ij}}{\partial y}\, dxdy$$

имеем

$$I^y_{mn}(m,\, n)=2, \quad I^y_{mn}(m,\, n+1)=I^y_{mn}(m,\, n-1)=-1;$$

в остальных случаях $I^y_{mn}(i,\, j)=0$.

Таким образом, уравнение, соответствующее узлу $(m,\, n)$, при всех $m,\, n = 1, \ldots, N-1$, записывается в виде

$$4u_{mn} - u_{m+1,\,n} - u_{m-1,\,n} - u_{m,\,n+1} - u_{m,\,n-1} = \int\limits_{S_{mn}} f\varphi_{mn}\, dxdy\,. \qquad \triangleright$$

9.52. Пусть единичный квадрат D разбит на элементарные квадраты со стороной h и в качестве базисных используются билинейные функции. Записать систему уравнений метода Ритца (конечных элементов) для задачи

$$-\Delta\, u = f(x,y) \text{ в } \; D\,, \; u = 0 \text{ на } \; \Gamma\,.$$

О т в е т: уравнение, соответствующее узлу $(m,\, n)$, при всех $m,\, n = 1, 2, \ldots, N-1$; $Nh = 1$, записывается в виде

$$\tfrac{8}{3}\, u_{m,n} - \tfrac{1}{3}\, (u_{m,n+1} + u_{m,n-1} + u_{m-1,n} + u_{m+1,n} +$$

$$+u_{m+1,n+1} + u_{m+1,n-1} + u_{m-1,n+1} + u_{m-1,n-1}) = \int\limits_{S_{mn}} f\varphi_{mn} dxdy,$$

где S_{mn} — носитель базисной функции φ_{mn}, т. е. квадрат со стороной $2\,h$ и центром в узле $(m,\, n)$.

9.53. Построить аппроксимацию $O(h^2)$ в точке (x, y) для смешанной производной $\dfrac{\partial^2 u}{\partial x \partial y}$ на семиточечном шаблоне

$$(x, y), (x \pm h, y), (x, y \pm h), (x - h, y + h), (x + h, y - h).$$

9.54. Построить аппроксимацию $O(h^2)$ в точке (x, y) для смешанной производной $\dfrac{\partial^2 u}{\partial x \partial y}$ на семиточечном шаблоне

$$(x, y), (x \pm h, y), (x, y \pm h), (x + h, y + h), (x - h, y - h).$$

9.4. Параболические уравнения

Построение и исследование разностных схем для уравнений в частных производных параболического типа традиционно проводят в открытой полуполосе

$$D = \{(x, t) : 1 > x > 0,\, t > 0\}$$

на примере простейшего уравнения теплопроводности

$$Lu \equiv \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f(x, t)$$

с начальным $u(x, 0) = u_0(x)$ при $t = 0$ и краевыми условиями $u(0, t) = u(1, t) = 0$ при $\forall t \geqslant 0$.

Предполагается, что начальная функция $u_0(x)$ удовлетворяет краевым условиям.

В общем случае на любом из концов отрезка краевое условие может быть задано в виде линейной комбинации функции и производной первого порядка. Тогда необходимо обратить внимание на способ его аппроксимации.

Характерная особенность параболической задачи — смешанный тип данных: краевые условия по x и начальные по t. Поэтому исследование аппроксимации такое же, как в гиперболических и эллиптических уравнениях, а исследование устойчивости проводят специальным образом.

Сетку, если это особо не оговаривается, считают равномерной по обеим переменным

$$x_m = m\,h, \quad m = 0, 1, \ldots, M, \quad M\,h = 1; \quad t_n = n\tau, \quad n = 0, 1, \ldots,$$

для сеточной функции u в точке (x_m, t_n) используют обозначение u_m^n, а краевые условия берут однородными: $u_0^n = u_M^n = 0\ \forall n$.

9.55. При каком соотношении τ и h схема

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}$$

имеет на решении порядок аппроксимации $O(\tau^2 + h^4)$?

Ответ: при $\dfrac{\tau}{h^2} = \dfrac{1}{6}$.

9.56. При каких θ разностная схема

$$\frac{u_m^{n+1}-u_m^n}{\tau}=\theta\,\frac{u_{m-1}^{n+1}-2u_m^{n+1}+u_{m+1}^{n+1}}{h^2}+(1-\theta)\,\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2}$$

имеет на решении порядок аппроксимации $O\left(\tau^2+h^4\right)$?

Ответ: при $\theta=\frac{1}{2}-\frac{h^2}{12\tau}$.

9.57. Определить порядок аппроксимации на решении схемы

$$\frac{1}{12}\,\frac{u_{m+1}^{n+1}-u_{m+1}^n}{\tau}+\frac{5}{6}\,\frac{u_m^{n+1}-u_m^n}{\tau}+\frac{1}{12}\,\frac{u_{m-1}^{n+1}-u_{m-1}^n}{\tau}=$$
$$=\frac{1}{2}\left(\frac{u_{m-1}^{n+1}-2u_m^{n+1}+u_{m+1}^{n+1}}{h^2}+\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2}\right).$$

Ответ: $O\left(\tau^2+h^4\right)$.

9.58. Для уравнения теплопроводности построить схему наивысшего порядка аппроксимации на шаблоне из точек:

1) $(x_{m-1},t_{n-1}),(x_{m+1},t_{n+1}),(x_m,t_k),k=n-1,n,n+1$;
2) $(x_{m\pm1},t_k),(x_m,t_k),k=n,n+1$;
3) $(x_{m\pm1},t_n),(x_m,t_k),k=n-1,n,n+1$;
4) $(x_{m-1},t_{n+1}),(x_{m+1},t_{n-1}),(x_m,t_k),k=n-1,n,n+1$.

Анализ устойчивости схем в равномерной метрике. Определим норму сеточной функции u_m^n на n-м временно́м слое следующим образом:

$$\|u^n\|=\max_{0\leqslant m\leqslant M}|u_m^n|.$$

Схема для простейшего уравнения теплопроводности называется *устойчивой в равномерной метрике* на отрезке $[0,T]$, $T=N\tau$, если имеет место неравенство

$$\max_{1\leqslant n\leqslant N}\|u^n\|\leqslant\|u^0\|+c\max_{0\leqslant n\leqslant N}\|f^n\|,$$

где c не зависит от шагов сетки τ и h, но может линейно зависеть от T.

При исследовании устойчивости схем с краевыми условиями первого рода важную роль играют сеточные функции

$$y_m^{(k)}=\sin(\pi\,k\,m\,h),\quad m=0,\ldots,M,\quad k=1,\ldots,M-1,$$

являющиеся решениями задачи на собственные значения

$$\frac{y_{m+1}-2\,y_m+y_{m-1}}{h^2}=-\lambda\,y_m,\quad 0<m<M,\ y_0=y_M=0,\quad h=\frac{1}{M}.$$

С их помощью легко строятся частные решения однородного уравнения вида

$$u_m^n=\mu_h^n(k)\,y_m^{(k)}=\mu_h^n(k)\,\sin(\pi\,k\,m\,h),\quad m=0,\ldots,M,\quad k=1,\ldots,M-1,$$

удовлетворяющие однородным краевым условиям.

9.59. Найти порядок аппроксимации явной схемы

$$\frac{u_m^{n+1}-u_m^n}{\tau}=\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2}+f_m^n,\quad 1\leqslant m\leqslant M-1,$$

$$u_m^0=u_0(mh),\quad u_0^n=u_M^n=0\ \forall n\geqslant 0\,,$$

и исследовать устойчивость при $\tau\leqslant\dfrac{h^2}{2}$.

$\lhd$ Схема имеет порядок аппроксимации $O(\tau+h^2)$. Введем обозначение $\rho=\dfrac{\tau}{h^2}$ и перепишем схему в удобном для анализа виде

$$u_m^{n+1}=(1-2\rho)\,u_m^n+\rho\,(u_{m+1}^n+u_{m-1}^n)+\tau f_m^n\,.$$

Максимальные значения обеих частей равенства по m совпадают, поэтому при $\rho\leqslant\frac{1}{2}$ имеем

$$\begin{aligned}\|u^{n+1}\|&\leqslant(1-2\rho)\|u^n\|+2\rho\|u^n\|+\tau\|f^n\|=\|u^n\|+\tau\|f^n\|\leqslant\\&\leqslant\|u^{n-1}\|+\tau(\|f^n\|+\|f^{n-1}\|)\leqslant\ldots\leqslant\|u^0\|+\sum_{k=0}^{n}\tau\|f^k\|\leqslant\\&\leqslant\|u^0\|+(n+1)\,\tau\max_k\|f^k\|.\end{aligned}$$

Следовательно, схема удовлетворяет определению устойчивости с постоянной $c=T$ (так как $(n+1)\tau=t_{n+1}$) при условии $\dfrac{\tau}{h^2}\leqslant\dfrac{1}{2}$. $\rhd$

9.60. Исследовать аппроксимацию и устойчивость полностью неявной схемы

$$\frac{u_m^{n+1}-u_m^n}{\tau}=\frac{u_{m-1}^{n+1}-2u_m^{n+1}+u_{m+1}^{n+1}}{h^2}+f_m^{n+1},\ 1\leqslant m\leqslant M-1,$$

$$u_m^0=u_0(mh),\quad u_0^n=u_M^n=0\ \ \forall n\geqslant 0\,.$$

$\lhd$ Порядок аппроксимации схемы $O\,(\tau+h^2)$. Удобная для анализа форма записи имеет вид

$$u_m^{n+1}+\rho\left(-u_{m-1}^{n+1}+2u_m^{n+1}-u_{m+1}^{n+1}\right)=u_m^n+\tau f_m^{n+1},\quad \rho=\frac{\tau}{h^2}.$$

Выбирая из всех значений u_m^{n+1}, по модулю равных $\|u^{n+1}\|$, такое, у которого индекс m принимает наименьшее значение, имеем

$$|u_m^{n+1}|>|u_{m-1}^{n+1}|\ \text{ и }\ |u_m^{n+1}|\geqslant|u_{m+1}^{n+1}|.$$

Отсюда $|2u_m^{n+1}|>|u_{m-1}^{n+1}|+|u_{m+1}^{n+1}|$, и знак выражения $2u_m^{n+1}-u_{m-1}^{n+1}-u_{m+1}^{n+1}$ совпадает со знаком u_m^{n+1}, т. е. справедлива оценка снизу

$$\|u^{n+1}\|=|u_m^{n+1}|\leqslant|u_m^{n+1}+\rho\,(2u_m^{n+1}-u_{m-1}^{n+1}-u_{m+1}^{n+1})|=|u_m^n+\tau f_m^{n+1}|.$$

Таким образом, при любых шагах сетки τ и h справедливо неравенство $\|u^{n+1}\|\leqslant\|u^n\|+\tau\|f^{n+1}\|$. Дальнейший вывод оценки безусловной устойчивости аналогичен решению 9.59. $\rhd$

9.61. Первая краевая задача для однородного уравнения теплопроводности $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$ аппроксимируется явной двухслойной схемой

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}\ ,\ 1 \leqslant m \leqslant M-1,$$

$$u_m^0 = u_0(mh), \quad u_0^n = u_M^n = 0\ \forall n \geqslant 0\,.$$

Определить порядок сходимости решения разностной схемы к решению дифференциальной задачи при различных $\rho = \frac{\tau}{h^2}$.

Ответ: сходимость имеет место только при выполнении условия устойчивости $\rho \leqslant \frac{1}{2}$, при этом порядок сходимости $O\left(\tau + h^2\right)$ для $\rho \neq \frac{1}{6}$ и $O(\tau^2 + h^4)$ для $\rho = \frac{1}{6}$.

9.62. Доказать, что явная схема

$$\frac{u_m^{n+1} - u_m^n}{\tau} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}\ ,\ 1 \leqslant m \leqslant M-1,$$

$$u_m^0 = u_0(mh), \quad u_0^n = u_M^n = 0 \ \ \forall n \geqslant 0$$

неустойчива, если $\overline{\lim\limits_{\tau,h\to 0}}\ \frac{1}{\tau}\left(\frac{\tau}{h^2} - \frac{1}{2}\right) = \infty$.

Указание. Проверить, что при выполнении этого условия среди частных решений вида $u_m^n = \mu_h^n(k)\,\sin(\pi mhk)$ найдется решение с номером k таким, что $|\mu_h(k)|^{t/\tau} \to \infty$ при $\tau \to 0$.

9.63. Исследовать устойчивость схемы по начальным данным

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}\ ,\ 1 \leqslant m \leqslant M-1,$$

$$u_0^n = u_M^n = 0\ \forall n \geqslant 0\,.$$

Указание. С помощью частных решений вида $u_m^n = \mu_h^n(k)\,\sin(\pi mhk)$ показать, что схема неустойчива.

9.64. Сравнить численные решения однородного уравнения теплопроводности с разрывной начальной функцией, полученные по явной и полностью неявной разностным схемам.

Анализ устойчивости схем в интегральной метрике. Положим

$$\|u^n\|_{L_{2,h}} = \left(h \sum_{m=1}^{M-1} (u_m^n)^2\right)^{1/2}$$

и назовем однородную разностную схему *устойчивой по начальным данным в метрике* $L_{2,h}$ на отрезке $[0,T], T = N\tau$, если справедливо неравенство

$$\max_{1\leqslant n\leqslant N} \|u^n\|_{L_{2,h}} \leqslant c\,\|u^0\|_{L_{2,h}}\,,$$

где c не зависит от шагов сетки τ и h.

9.65. При каких $\theta \in [0,1]$ схема

$$\frac{u_m^{n+1}-u_m^n}{\tau} = \theta\,\frac{u_{m-1}^{n+1}-2u_m^{n+1}+u_{m+1}^{n+1}}{h^2} + (1-\theta)\,\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2},$$

$$1 \leqslant m \leqslant M-1,\ u_m^0 = u_0(mh), \quad u_0^n = u_M^n = 0\ \forall n \geqslant 0,$$

является устойчивой?

$\lhd$ Введем оператор

$$\Lambda u_m = \frac{u_{m+1}-2u_m+u_{m-1}}{h^2}\ .$$

Учитывая однородность краевых условий, под Λ будем также понимать матрицу, которая ставит в соответствие вектору $u^n = (u_1^n,\dots,u_{M-1}^n)^T$ вектор $\Lambda u^n = (\Lambda u_1^n,\dots,\Lambda u_{M-1}^n)^T$. Тогда рассматриваемую схему можно записать в виде

$$u^{n+1} = S\,u^n = S^{n+1}u^0,$$

где $S = (I-\tau\theta\Lambda)^{-1}(I+\tau(1-\theta)\Lambda)$, а I — единичная матрица.

Рассмотрим задачу на собственные значения на отрезке $[0,l]$

$$\Lambda y_m = -\lambda y_m, \quad 1 \leqslant m \leqslant M-1,\ Mh = l, \quad y_0 = y_M = 0.$$

Ее решение можно записать в форме

$$y_m^{(k)} = \sqrt{\frac{2}{l}}\sin\left(\frac{\pi mk}{M}\right), \quad \lambda^{(k)} = \frac{4}{h^2}\sin^2\frac{\pi kh}{2l}, \quad 1 \leqslant k \leqslant M-1\,.$$

Так как матрица $(I+\beta\Lambda)^{\pm1}$ имеет ту же систему собственных векторов, что и Λ, то собственные значения матрицы S можно выразить через $\lambda^{(k)}$. Действительно, пусть $S\,y = \mu\,y$. Возьмем в качестве y вектор $y^{(k)}$; тогда для соответствующего $\mu^{(k)}$ получим явное выражение

$$\mu^{(k)}(S) = \frac{1-\tau(1-\theta)\lambda^{(k)}}{1+\tau\theta\lambda^{(k)}}\ .$$

Отсюда следует, что матрица S является симметричной, так как имеет представление $S = QDQ^{-1}$, где столбцами ортогональной матрицы Q являются векторы $y^{(k)}$, а D — диагональная матрица, состоящая из соответствующих $\mu^{(k)}$. Матричная норма

$$\|A\| = \sup_{\mathbf{x}\neq0}\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}\ ,$$

подчиненная векторной евклидовой норме, равна $\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$. Причем для симметричной матрицы выражение упрощается: $\|A\|_2 = \max|\lambda(A)|$. Метрика $L_{2,h}$ отличается от евклидовой только множителем h, поэтому справедливо выражение

$$\|S\|_{L_{2,h}} = \max_k|\mu^{(k)}(S)|\,.$$

Выясним теперь, в каком случае $\|S\|_{L_{2,h}} \leqslant 1$. Имеем

$$-1 \leqslant \frac{1-\tau(1-\theta)\lambda^{(k)}}{1+\tau\theta\lambda^{(k)}} \leqslant 1,$$

для $k = 1, 2, \ldots, M-1$. Так как $\tau, \lambda^{(k)} > 0,\ \theta \geqslant 0$, то знаменатель дроби всегда положителен, поэтому

$$-(1 + \tau\theta\lambda^{(k)}) \leqslant 1 - \tau(1-\theta)\lambda^{(k)} \leqslant 1 + \tau\theta\lambda^{(k)}\,.$$

Правое неравенство выполняется всегда, значит, содержательным является левое неравенство. Перепишем его в виде

$$\tau(1-2\theta)\lambda^{(k)} \leqslant 2\,.$$

При $\frac{1}{2} \leqslant \theta \leqslant 1$ это неравенство выполняется при любом τ, а при $0 \leqslant \theta < \frac{1}{2}$ имеет место ограничение

$$\tau \leqslant \frac{2}{(1-2\theta)\max\limits_k \lambda^{(k)}} \leqslant \frac{h^2}{2-4\theta}\ .$$

Из полученной выше формулы $u^n = S^n u^0$ следует, что

$$\|u^n\|_{L_{2,h}} \leqslant \|S\|^n_{L_{2,h}} \|u^0\|_{L_{2,h}}\,.$$

Поэтому для всех $\theta \in [0,1]$ имеем устойчивость в метрике $L_{2,h}$ с постоянной $c = 1$. $\triangleright$

Ответ: для $\frac{1}{2} \leqslant \theta \leqslant 1$ схема устойчива при любых τ и h, а для $0 \leqslant \theta < \frac{1}{2}$ устойчива при выполнении условия $\tau \leqslant \frac{h^2}{2-4\theta}$. Здесь в основу решения положен следующий принцип: все собственные значения оператора перехода должны по модулю не превышать единицы. Это ограничение можно ослабить до величины $1+\gamma\tau$ с постоянной γ, не зависящей от τ и h (аналогично спектральному признаку для гиперболических уравнений). В этом случае постоянная c в определении устойчивости принимает значение $c = \mathrm{e}^{\gamma T}$.

9.66. Уравнение теплопроводности $\dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2}$ аппроксимируется схемой Дюфорта—Франкела (схема «ромб»):

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} = \frac{u_{m+1}^n - u_m^{n-1} - u_m^{n+1} + u_{m-1}^n}{h^2}\,,$$

$$1 \leqslant m \leqslant M-1, \quad u_0^n = u_M^n = 0,\ \forall n \geqslant 1.$$

Выяснить условия ее устойчивости и показать, что если $h \to 0$, $\tau \to 0$ так, что $\dfrac{\tau}{h} = c \neq 0$, то эта схема аппроксимирует гиперболическое уравнение

$$\frac{\partial u}{\partial t} + c^2\ \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}\,.$$

$\triangleleft$ Преобразуем схему к виду

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} + \frac{\tau^2}{h^2}\ \frac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{\tau^2} = \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}\,.$$

Отсюда следует ответ на вопрос об аппроксимации.

Для анализа устойчивости воспользуемся частными решениями $u_m^n = \mu_h^n(k)\sin(\pi\, m\, k\, h)$, для которых достаточно показать справедливость неравенств $|\mu_h(k)| \leqslant 1$. Из преобразованной схемы имеем

$$\frac{\mu^2-1}{2\tau} + \frac{\tau^2}{h^2}\,\frac{\mu^2-2\mu+1}{\tau^2} = -\lambda^{(k)}\mu\,, \quad \lambda^{(k)} = \frac{4}{h^2}\sin^2\frac{\pi k}{2M}\,,\ Mh = 1,$$

или, введя обозначение $\rho = \dfrac{2\tau}{h^2}$, получим

$$p(\mu) \equiv (\rho+1)\mu^2 - 2\mu\left(\rho - \tau\lambda^{(k)}\right) + (\rho-1) = 0.$$

Если дискриминант этого уравнения неположительный, то

$$|\mu_1|^2 = |\mu_2|^2 = \frac{\rho-1}{\rho+1} < 1.$$

В случае положительного дискриминанта в точках $\mu = \pm 1$ парабола $p(\mu)$ принимает положительные значения, а координата $\mu_* = \dfrac{\rho - \tau\lambda^{(k)}}{\rho+1}$ ее вершины располагается внутри интервала $(-1,1)$. Это гарантирует оценку $|\mu_{1,2}| < 1$ для вещественных корней. Таким образом, в обоих случаях получаем, что схема устойчива для всех τ и h. $\triangleright$

9.67. Исследовать устойчивость по начальным данным схемы

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} = \frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2}\,,$$
$$1 \leqslant m \leqslant M-1, \quad u_0^n = u_M^n = 0\ \forall n \geqslant 1.$$

9.68. Исследовать устойчивость по начальным данным схемы

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} = \frac{u_{m-1}^{n-1} - 2u_m^{n-1} + u_{m+1}^{n-1}}{h^2}\,,$$
$$1 \leqslant m \leqslant M-1, \quad u_0^n = u_M^n = 0\ \forall n \geqslant 1.$$

9.69. Исследовать устойчивость по начальным данным схемы

$$\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} = \frac{u_{m-1}^{n} - 2u_m^{n+1} + u_{m+1}^{n}}{h^2}\,,$$
$$1 \leqslant m \leqslant M-1, \quad u_0^n = u_M^n = 0\ \forall n \geqslant 1.$$

Операторная устойчивость двухслойных разностных схем. Далее, когда это не вызывает неоднозначности, будем опускать индекс, соответствующий пространственной координате.

В общем случае двухслойная разностная схема записывается следующим образом:

$$B_1 u^{n+1} + B_0 u^n = \varphi^n \tag{9.7}$$

с известным начальным вектором u^0 из некоторого конечномерного пространства U со скалярным произведением $(\cdot,\cdot)$ и порожденной им нормой $\|\cdot\|$. Учитывая тождество

$$u^{n+1} \equiv u^n + \tau\,\frac{u^{n+1}-u^n}{\tau}\,,$$

перепишем схему в *канонической форме*

$$B\,\frac{u^{n+1}-u^n}{\tau} + Au^n = \varphi^n, \tag{9.8}$$

где $A = B_0 + B_1$ и $B = \tau B_1$.

Матричный оператор A (в общем случае комплекснозначный) обычно задает аппроксимацию дифференциального оператора по пространству, а оператор B задает аппроксимацию по времени. Такой вид записи позволяет проверять устойчивость различных схем по общей методике, формулируя условия в терминах свойств операторов A и B.

Достаточные условия устойчивости для двухслойных схем имеют вид

$$B > 0, \quad A = A^* > 0, \quad B \geqslant \frac{\tau}{2}\,A. \tag{9.9}$$

Обратим внимание на возможность использования естественной формы записи двухслойных схем (9.7). В этом случае достаточные условия устойчивости (9.9) эквивалентны условиям

$$B_1 > 0, \quad (B_0 + B_1) = (B_0 + B_1)^* > 0, \quad B_1 \geqslant B_0. \tag{9.10}$$

Для задач с несамосопряженным оператором A, т. е. $A \neq A^*$, устойчивость схем проверить значительно сложнее. В качестве примера приведем достаточные условия устойчивости схемы с весами. Пусть дана двухслойная схема с весами

$$\frac{u^{n+1}-u^n}{\tau} + A(\theta u^{n+1} + (1-\theta)u^n) = \varphi^n, \tag{9.11}$$

где A — несамосопряженный оператор, θ — весовой параметр. Схема устойчива при

$$\left(\theta - \tfrac{1}{2}\right)\tau\|Au\|^2 + (Au,u) \geqslant 0.$$

Отсюда следует, что если $A > 0$, то для $\theta \geqslant \frac{1}{2}$ схема устойчива при всех τ. Если, кроме того, дано, что

$$\|Au\|^2 \leqslant C(Au,u) \tag{9.12}$$

с некоторой константой C, то схема устойчива при

$$\theta \geqslant \tfrac{1}{2} - \tfrac{1}{\tau C}\,. \tag{9.13}$$

Наконец, если $A = A^* > 0$, то в качестве C можно взять $C = \|A\|$, и схема (9.11) устойчива при

$$\theta \geqslant \tfrac{1}{2} - \tfrac{1}{\tau\|A\|}\,. \tag{9.14}$$

9.70. Исследовать устойчивость по начальным данным схемы

$$\frac{1}{12}\,\frac{u_{m+1}^{n+1}-u_{m+1}^{n}}{\tau}+\frac{5}{6}\,\frac{u_{m}^{n+1}-u_{m}^{n}}{\tau}+\frac{1}{12}\,\frac{u_{m-1}^{n+1}-u_{m-1}^{n}}{\tau}=$$

$$=\frac{1}{2}\left(\Lambda u_m^{n+1}+\Lambda u_m^{n}\right),\quad \Lambda u_m^n=\frac{u_{m-1}^n-2u_m^n+u_{m+1}^n}{h^2},$$

$$1\leqslant m\leqslant M-1,\quad u_0^n=u_M^n=0\ \forall n\geqslant 0.$$

$\lhd$ Используя соотношения

$$u_{m+1}+10u_m+u_{m-1}=(u_{m+1}-2u_m+u_{m-1})+12u_m=h^2\Lambda u_m+12u_m,$$

$$u^{n+1}=u^n+\tau\,\frac{u^{n+1}-u^n}{\tau},$$

схему можно преобразовать к виду

$$\frac{u_m^{n+1}-u_m^n}{\tau}-\sigma\tau\Lambda\,\frac{u_m^{n+1}-u_m^n}{\tau}=\Lambda u_m^n,$$

или

$$(I-\sigma\tau\Lambda)\,\frac{u_m^{n+1}-u_m^n}{\tau}-\Lambda u_m^n=0,$$

где $\sigma=\frac{1}{2}-\frac{h^2}{12\tau}$. $\rhd$

Ответ: схема устойчива при любых h и τ.

9.71. Исследовать устойчивость по начальным данным и найти порядок аппроксимации схемы

$$2\gamma u_m^{n+1}=\left(\gamma-\frac{1}{2}\right)(u_{m-1}^{n+1}+u_{m+1}^{n+1})+\frac{1}{2}\,(u_{m-1}^n+u_{m+1}^n),\quad \gamma=\frac{\tau}{h^2},$$

$$1\leqslant m\leqslant M-1,\quad u_0^n=u_M^n=0\ \forall n\geqslant 0.$$

Ответ: $(I-\sigma\tau\Lambda)\,\frac{u_m^{n+1}-u_m^n}{\tau}=\Lambda u_m^{n+1}$, $\sigma=1-\frac{1}{2\gamma}$. Схема устойчива при $\gamma\geqslant\frac{1}{2}$ и имеет порядок аппроксимации $O(\tau+h^2)$.

Операторная устойчивость трехслойных разностных схем. В общем случае трехслойная разностная схема записывается следующим образом:

$$B_1u^{n+1}+B_0u^n+B_{-1}u^{n-1}=\varphi^n. \tag{9.15}$$

Канонической формой называют представление *трехслойной схемы* в виде

$$B\,\frac{u^{n+1}-u^{n-1}}{2\tau}+\tau^2R\,\frac{u^{n+1}-2u^n+u^{n-1}}{\tau^2}+Au^n=\varphi^n. \tag{9.16}$$

Имеют место следующие достаточные условия устойчивости для трехслойных схем:

$$B\geqslant 0,\quad R=R^*>\frac{A}{4},\quad A=A^*>0. \tag{9.17}$$

Сравнивая (9.16) и (9.15), находим

$$B_1 = R + \frac{1}{2\tau}\,B, \quad B_0 = A - 2R, \quad B_{-1} = R - \frac{1}{2\tau}\,B,$$

$$B = \tau(B_1 - B_{-1}), \quad R = \frac{1}{2}\,(B_1 + B_{-1}), \quad A = B_1 + B_0 + B_{-1}.$$

Отсюда и из (9.17) получаем достаточные условия устойчивости для схемы (9.15)

$$\begin{gathered} B_1 + B_{-1} = (B_1 + B_{-1})^*, \quad B_0 = B_0^*, \quad B_1 \geqslant B_{-1}, \\ B_1 - B_0 + B_{-1} > 0, \quad B_1 + B_0 + B_{-1} > 0. \end{gathered} \tag{9.18}$$

Приведем достаточные условия устойчивости трехслойной схемы с весами. Для схемы с весами (θ_1 и θ_2 — весовые параметры)

$$\frac{u^{n+1} - u^{n-1}}{2\tau} + A(\theta_1 u^{n+1} + (1 - \theta_1 - \theta_2)u^n + \theta_2 u^{n-1}) = \varphi^n \tag{9.19}$$

с несамосопряженным оператором $A > 0$, для которого выполнена оценка $\|Au\|^2 \leqslant C(Au, u)$, достаточные условия устойчивости имеют вид

$$\theta_1 \geqslant \theta_2 - \frac{1}{\tau C}\,, \quad \theta_1 + \theta_2 > \frac{1}{2}\,. \tag{9.20}$$

Трехслойную схему всегда можно записать в виде некоторой двухслойной схемы, но уже для системы уравнений, что соответствует принятому в теории дифференциальных уравнений сведению задачи второго порядка к системе уравнений первого порядка. Такое представление неединственно; приведем одну из возможных форм записи. Схема (9.16) эквивалентна

$$\mathcal{B}\,\frac{Y^{n+1} - Y^n}{\tau} + \mathcal{A}Y^n = \Phi^n,$$

где

$$Y = \left(\frac{u^n + u^{n-1}}{2}, u^n - u^{n-1}\right)^T, \quad \Phi = (\varphi, 0)^T,$$

$$\mathcal{A} = \begin{pmatrix} A & 0 \\ 0 & R - \dfrac{A}{4} \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} B + \dfrac{\tau A}{2} & \tau\left(R - \dfrac{A}{4}\right) \\ -\tau\left(R - \dfrac{A}{4}\right) & \tau\left(\dfrac{R}{2} - \dfrac{A}{8}\right) \end{pmatrix}.$$

Такая замена позволяет формально применять результаты теории двухслойных разностных схем для исследования трехслойных схем.

Рассмотренный операторный подход помогает свести доказательство устойчивости схемы к проверке сформулированных достаточных условий для конкретных операторов B, R, A, определяемых задачей. В ряде случаев удобно использовать запись схем (9.8) и (9.16) в естественном виде (9.7) и (9.15), после чего проверять достаточные условия устойчивости (9.10) и (9.18) соответственно.

9.72. Исследовать устойчивость по начальным данным и найти порядок аппроксимации схемы

$$\frac{1}{12}\,\frac{1,5u_{m+1}^{n+1} - 2u_{m+1}^n + 0,5u_{m+1}^{n-1}}{\tau} + \frac{5}{6}\,\frac{1,5u_m^{n+1} - 2u_m^n + 0,5u_m^{n-1}}{\tau} +$$

$$+ \frac{1}{12}\,\frac{1,5u_{m-1}^{n+1} - 2u_{m-1}^n + 0,5u_{m-1}^{n-1}}{\tau} = \Lambda u_m^{n+1},$$

$$1 \leqslant m \leqslant M - 1, \quad u_0^n = u_M^n = 0\ \forall n \geqslant 1.$$

У к а з а н и е. Воспользоваться соотношением

$$1,5u_m^{n+1} - 2u_m^n + 0,5u_m^{n-1} = \tau\,\frac{u_m^{n+1} - u_m^{n-1}}{2\tau} + \tau^2\,\frac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{\tau^2}$$

и переписать схему в трехслойной канонической форме с операторами

$$B = I + \Big(1 - \frac{h^2}{12\tau}\Big)\tau A, \quad R = \frac{1}{\tau}\,I + \Big(\frac{1}{2} - \frac{h^2}{12\tau}\Big)A, \quad A = -\Lambda.$$

О т в е т: схема абсолютно устойчива, погрешность аппроксимации на решении равна $O(\tau^2 + h^4)$.

9.73. Исследовать устойчивость схемы по начальным данным в зависимости от параметра θ

$$(1-\theta)\,\frac{u_m^{n+1} - u_m^n}{\tau} + \theta\,\frac{u_m^n - u_m^{n-1}}{\tau} = \Lambda u_m^{n+1},$$
$$1 \leqslant m \leqslant M-1, \quad u_0^n = u_M^n = 0\ \forall n \geqslant 1.$$

$\lhd$ Эквивалентная схема

$$(I - \tau\Lambda)\,\frac{u_m^{n+1} - u_m^n}{2\tau} + \Big(\Big(\frac{1}{2} - \theta\Big)\tau I - \frac{\tau^2}{2}\,\Lambda\Big)\,\frac{u_m^{n+1} - 2u_m^n + u_m^{n-1}}{\tau^2} = \Lambda u_m^n$$

соответствует трехслойной канонической схеме с операторами

$$B = I + \tau A, \quad R = \frac{1}{2}\,A + \frac{1-2\theta}{2\tau}\,I, \quad A = -\Lambda.$$

Отсюда следует, что достаточное условие устойчивости имеет вид

$$R - \frac{1}{4}\,A = \frac{1}{4}\,A + \frac{1-2\theta}{2\tau}\,I \geqslant \Big(\frac{1}{4}\,\lambda + \frac{1-2\theta}{2\tau}\Big)I > 0,\ \lambda = \frac{4}{h^2}\sin^2\Big(\frac{\pi h}{2}\Big),$$

так как $A \geqslant \lambda I$. Оно выполняется при $\theta < \frac{1}{2} + \frac{\tau\lambda}{4}$. $\rhd$

О т в е т: схема устойчива при $\theta < \frac{1}{2} + \frac{\tau\lambda}{4}$.

Факторизованные схемы. Рассмотрим двухслойную разностную схему

$$B\,\frac{u^{n+1} - u^n}{\tau} + Au^n = \varphi^n.$$

Схему называют *факторизованной*, если оператор B представим в виде произведения $B = B_1 \cdots B_p$. Для факторизованных схем обращение оператора B сводится к последовательному решению задач

$$B_\alpha u^{n+\alpha/p} = F^\alpha, \quad \alpha = 1, \ldots, p,$$

где F^α — известные функции; сами матрицы B_α выбирают, например, треугольными или трехдиагональными.

При решении многомерных задач математической физики требуемое представление может быть получено заменой многомерной задачи с оператором B последовательностью одномерных задач с операторами B_α.

В этом случае для промежуточных функций $u^{n+\alpha/p}$ должны быть заданы соответствующие краевые условия на границе области (однородные, если исходные условия однородные). Такой подход позволяет совместить абсолютную устойчивость неявных схем и простоту реализации одномерных задач. Напомним, что явные схемы ($B = I$) имеют жесткое ограничение на шаг по времени $\tau = O(h^2)$, что делает их слишком трудоемкими для реальных расчетов.

Для изучения устойчивости факторизованной схемы с операторами $B_\alpha = I + \tau R_\alpha$ можно применять следующий критерий: если схема с $B = I + \tau \sum\limits_{\alpha=1}^{p} R_\alpha$ устойчива и операторы R_α положительные, самосопряженные и попарно перестановочные, то факторизованная схема с $B = B_1 \dots B_p$ также устойчива.

В задачах 9.74–9.78 требуется провести полное исследование разностных схем для уравнения

$$\frac{\partial u}{\partial t} = \Delta u,$$

где Δ — оператор Лапласа в пространстве двух или трех измерений, при нулевых граничных условиях первого рода. Используемые обозначения разностных операторов введены в разделе 9.3.

9.74. Исследовать устойчивость и найти порядок аппроксимации схемы Дугласа—Рекфорда

$$\frac{u^{n+1/2} - u^n}{\tau} = \Lambda_1 u^{n+1/2} + \Lambda_2 u^n, \quad \frac{u^{n+1} - u^{n+1/2}}{\tau} = \Lambda_2 (u^{n+1} - u^n).$$

Указание. Привести схему к виду

$$(I - \tau\Lambda_1)\,\frac{u^{n+1/2} - u^n}{\tau} = \Lambda u^n, \quad \Lambda = \Lambda_1 + \Lambda_2,$$

$$(I - \tau\Lambda_2)\,\frac{u^{n+1} - u^n}{\tau} = \frac{u^{n+1/2} - u^n}{\tau}\,.$$

Исключить $(u^{n+1/2} - u^n)$ и получить факторизованную схему

$$(I - \tau\Lambda_1)(I - \tau\Lambda_2)\,\frac{u^{n+1} - u^n}{\tau} = \Lambda u.$$

Ответ: порядок аппроксимации схемы равен $O(\tau + h^2)$. Схема устойчива.

9.75. Исследовать устойчивость и найти порядок аппроксимации (θ_1 и θ_2 — весовые параметры) разностной схемы

$$\frac{u^{n+1/2} - u^n}{\tau} = \theta_1 \Lambda_1 u^{n+1/2} + (1 - \theta_1)\Lambda_1 u^n + \Lambda_2 u^n,$$

$$\frac{u^{n+1} - u^{n+1/2}}{\tau} = \theta_2 \Lambda_2 (u^{n+1} - u^n), \quad \theta_1\theta_2 \geqslant 0.$$

$\lhd$ Преобразуем схему к виду

$$(I-\theta_1\tau\Lambda_1)(I-\theta_2\tau\Lambda_2)\,\frac{u^{n+1}-u^n}{\tau}=\Lambda u^n,\quad \Lambda=\Lambda_1+\Lambda_2.$$

Обозначим $A_\alpha=-\Lambda_\alpha$, $A=A_1+A_2$, где A_1 и A_2 — положительно определенные, самосопряженные и перестановочные операторы такие, что $A_1A_2>0$. Проверим условие устойчивости двухслойной схемы $B-\frac{\tau}{2}A\geqslant 0$. Имеем

$$B=(I+\tau\theta_1A_1)(I+\tau\theta_2A_2)=I+\tau\theta_1A_1+\tau\theta_2A_2+\tau^2\theta_1\theta_2A_1A_2,$$

$$B-\frac{\tau}{2}A=I+\left(\theta_1-\frac12\right)\tau A_1+\left(\theta_2-\frac12\right)\tau A_2+\tau^2\theta_1\theta_2A_1A_2\geqslant$$

$$\geqslant I+\left(\theta_1-\frac12\right)\tau A_1+\left(\theta_2-\frac12\right)\tau A_2,$$

так как $\theta_1\theta_2\geqslant 0$. Учитывая, что $I\geqslant A_\alpha/\|A_\alpha\|$, и вводя требование

$$\frac12 I+\left(\theta_\alpha-\frac12\right)\tau A_\alpha\geqslant\left(\frac{1}{2\|A_\alpha\|}+\left(\theta_\alpha-\frac12\right)\tau\right)A_\alpha\geqslant 0,$$

получаем $\theta_\alpha\geqslant\frac12-\frac{1}{2\tau\|A_\alpha\|}$. $\rhd$

Ответ: порядок аппроксимации схемы равен $O(\tau^2+h^2)+O\left(\left|\theta_1-\frac12\right|\tau\right)+O\left(\left|\theta_2-\frac12\right|\tau\right)$, т. е. $O(\tau^2+h^2)$ при $\theta_1=\theta_2=\frac12$. Схема устойчива при $\theta_1\theta_2\geqslant 0$ и

$$\theta_\alpha\geqslant\frac12-\frac{1}{2\tau\|A_\alpha\|},\quad \alpha=1,2.$$

Схема безусловно устойчива при $\theta_\alpha\geqslant\frac12$.

9.76. Исследовать устойчивость и найти порядок аппроксимации (θ_1,θ_2 и θ_3 — весовые параметры) разностной схемы

$$\frac{u^{n+1/3}-u^n}{\tau}=\theta_1\Lambda_1u^{n+1/3}+(1-\theta_1)\Lambda_1u^n+(\Lambda_2+\Lambda_3)u^n,$$

$$\frac{u^{n+2/3}-u^{n+1/3}}{\tau}=\theta_2\Lambda_2(u^{n+2/3}-u^n),$$

$$\frac{u^{n+1}-u^{n+2/3}}{\tau}=\theta_3\Lambda_3(u^{n+1}-u^n)$$

(частный случай $\theta_1=\theta_2=\theta_3=\frac12$ называется *схемой Дугласа*).

Указание. Обозначим $v_\alpha=\frac{u^{n+\alpha/3}-u^n}{\tau}$, $\alpha=1,2,3$. Тогда схема принимает вид

$$(I-\theta_1\tau\Lambda_1)v_1=\Lambda u^n,\quad (I-\theta_2\tau\Lambda_2)v_2=v_1,\quad (I-\theta_3\tau\Lambda_3)v_3=v_2.$$

Последовательно исключая v_2 и v_1 и заменяя $v_3=\frac{u^{n+1}-u^n}{\tau}$, получить факторизованную схему

$$(I-\theta_1\tau\Lambda_1)(I-\theta_2\tau\Lambda_2)(I-\theta_3\tau\Lambda_3)\,\frac{u^{n+1}-u^n}{\tau}=\Lambda u^n,$$

$$\Lambda=\Lambda_1+\Lambda_2+\Lambda_3.$$

О т в е т: порядок аппроксимации схемы равен $O(\tau^2+h^2)$ при $\theta_1=\theta_2=\theta_3=\frac{1}{2}$ и $O(\tau+h^2)$ при $\theta_\alpha\neq\frac{1}{2}$, $\alpha=1,2,3$. Схема устойчива при $\theta_\alpha\geqslant\frac{1}{2}$, $\alpha=1,2,3$.

9.77. Исследовать устойчивость и найти порядок аппроксимации (θ_1 и θ_2 — весовые параметры) разностной схемы

$$\frac{u^{n+1/2}-u^n}{\tau}=\theta_1\Lambda_1u^{n+1/2}+(1-\theta_2)\Lambda_2u^n,$$
$$\frac{u^{n+1}-u^{n+1/2}}{\tau}=\theta_2\Lambda_2u^{n+1}+(1-\theta_1)\Lambda_1u^{n+1/2}.$$

Показать, что при $\theta_\alpha=\frac{1}{2}-\frac{h_\alpha^2}{12\tau}$, $\alpha=1,2$, порядок аппроксимации схемы равен $O(\tau^2+h^4)$.

$\lhd$ Запишем уравнения в виде

$$(I-\theta_1\tau\Lambda_1)u^{n+1/2}=(I+(1-\theta_2)\tau\Lambda_2)u^n,$$
$$(I+(1-\theta_1)\tau\Lambda_1)u^{n+1/2}=(I-\theta_2\tau\Lambda_2)u^{n+1}.$$

Умножая первое из уравнений на $(1-\theta_1)$, второе — на θ_1 и складывая полученные выражения, получим $u^{n+1/2}=\theta_1(I-\theta_2\tau\Lambda_2)u^{n+1}+$ $+(1-\theta_1)(I+(1-\theta_2)\tau\Lambda_2)u^n$. Подставляя это выражение в первое уравнение, имеем

$$(I-\theta_1\tau\Lambda_1)(I-\theta_2\tau\Lambda_2)u^{n+1}=(I+(1-\theta_1)\tau\Lambda_1)(I+(1-\theta_2)\tau\Lambda_2)u^n.$$

Таким образом, факторизованная схема принимает вид

$$(I-\theta_1\tau\Lambda_1)(I-\theta_2\tau\Lambda_2)\,\frac{u^{n+1}-u^n}{\tau}=\Lambda u^n+(1-\theta_1-\theta_2)\tau\Lambda_1\Lambda_2u^n,$$

где $\Lambda=\Lambda_1+\Lambda_2$. $\rhd$

О т в е т: порядок аппроксимации схемы при $\theta_1=\theta_2=\frac{1}{2}$ равен $O(\tau^2+h^2)$; при $\theta_\alpha=\frac{1}{2}-\frac{h_\alpha^2}{12\tau}$, $\alpha=1,2$, порядок равен $O(\tau^2+h^4)$.

Схема устойчива при $\theta_\alpha\geqslant\frac{1}{2}$ и $\theta_\alpha=\frac{1}{2}-\frac{h_\alpha^2}{12\tau}$, $\alpha=1,2$.

9.78. Исследовать устойчивость разностной схемы

$$\frac{u^{n+1/3}-u^n}{\tau}=\frac{1}{3}\left(\Lambda_1u^{n+1/3}+(\Lambda_2+\Lambda_3)u^n\right),$$
$$\frac{u^{n+2/3}-u^{n+1/3}}{\tau}=\frac{1}{3}\left(\Lambda_1u^{n+1/3}+\Lambda_2u^{n+2/3}+\Lambda_3u^{n+1/3}\right),$$
$$\frac{u^{n+1}-u^{n+2/3}}{\tau}=\frac{1}{3}\left((\Lambda_1+\Lambda_2)u^{n+2/3}+\Lambda_3u^{n+1}\right).$$

О т в е т: факторизованная схема имеет вид $B_1B_2B_3u^{n+1}=C_1C_2C_3u^n$, где $B_\alpha=I-\frac{\tau}{3}\Lambda_\alpha$, $C_\alpha=B_\alpha+\frac{\tau}{3}\Lambda_\alpha$, $\alpha=1,2,3$, $\Lambda=\Lambda_1+\Lambda_2+\Lambda_3$. Схема не является безусловно устойчивой.

9.5. Уравнение Шрёдингера

Рассмотрим первую краевую задачу для линейного однородного пространственно одномерного нестационарного уравнения Шрёдингера

$$\mathrm{i}\,\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}\,, \quad \mathrm{i} = \sqrt{-1}, \quad 0 < x < 1, \quad 0 < t \leqslant T,$$

$$u(x,0) \equiv u^0(x) = \sum_{k=1}^{M-1} C_k \sin(\pi k x), \quad u(0,t) = u(1,t) = 0.$$

Применяя метод разделения переменных, можно показать, что решение дифференциальной задачи имеет вид

$$u(x,t) = \sum_{k=1}^{M-1} C_k \mu^t(k) \sin(\pi k x), \quad \mu(k) = \mathrm{e}^{\mathrm{i}(\pi k)^2}.$$

Так как для коэффициентов $\mu(k)$ выполняется равенство $|\mu(k)| = 1$, то норма

$$\|u(t)\|^2 \equiv \int\limits_0^1 u\overline{u}dx = \frac{1}{2}\sum_{k=1}^{M-1} |C_k|^2 = \mathrm{const}$$

не изменяется с течением времени.

При построении разностных схем разумно учитывать данное свойство дифференциальной задачи и требовать выполнения подобной оценки $|\mu_h(k)| \leqslant 1$ для разностной задачи, которая гарантирует устойчивость разностной схемы. Однако, как будет показано далее, явные схемы для уравнения Шрёдингера данному условию не удовлетворяют. Поэтому здесь допустимо применение устойчивых разностных схем с растущими решениями. При этом, как следует из определения устойчивости, необходимо выполнение неравенства $\max\limits_k |\mu_h(k)| \leqslant 1 + c\tau \leqslant \mathrm{e}^{c\tau}$. Такому условию удовлетворяют, например, явные схемы при $\tau = O(h^4)$.

Двухслойная схема с весами. Зададимся сеткой с узлами в точках $(mh, n\tau)$ и заменим исходную задачу следующей разностной схемой (θ — весовой параметр):

$$\mathrm{i}\,\frac{u_m^{n+1} - u_m^n}{\tau} = \theta\Lambda u_m^{n+1} + (1-\theta)\Lambda u_m^n, \quad 1 \leqslant m \leqslant M-1,\ \forall n \geqslant 0,$$

$$u_0^n = u_M^n = 0,\ u_m^0 = u^0(mh),$$

где

$$\Lambda u_m = \frac{u_{m+1} - 2u_m + u_{m-1}}{h^2}\,,\ \theta = \theta_0 + \mathrm{i}\theta_1,\ h = \frac{1}{M}.$$

9.79. Найти порядок аппроксимации схемы с весами.

Указание. Погрешность аппроксимации представима в виде

$$\left(\mathrm{i}\,\frac{h^2}{12} + \left(\theta - \frac{1}{2}\right)\tau\right)\frac{\partial}{\partial t}\,u_{xx}(x_m, t_n + 0{,}5\tau) + O\left(\tau^2 + h^4\right).$$

Отсюда, в частности, следует, что схема имеет максимальный порядок аппроксимации при $\theta = \frac{1}{2} - \mathrm{i}\,\frac{h^2}{12\tau}$.

9.80. Применяя метод разделения переменных, записать явный вид решения u_m^n схемы с весами и показать, что необходимым условием невозрастания решения разностной задачи при всех начальных данных является неравенство $\theta_0 \geqslant \frac{1}{2}$.

Указание. Так как

$$u_m^n = \sum_{k=1}^{M-1} C_k \left(\mu_h(k)\right)^n \sin(\pi k m h),$$
$$|\mu_h(k)|^2 = 1 - \frac{(2\theta_0 - 1)\tau^2 \lambda_h^2(k)}{\left(1 + \theta_1 \tau \lambda_h(k)\right)^2 + \theta_0^2 \tau^2 \lambda_h^2(k)},$$

где $\lambda_h(k) = \frac{4}{h^2} \sin^2\left(\frac{\pi k h}{2}\right)$ — собственные числа разностного оператора $(-\Lambda)$, то решение не возрастает при $|\mu_h(k)| \leqslant 1$.

9.81. Применяя метод разделения переменных, записать явный вид решения u_m^n схемы с весами и показать, что для $\theta_0 < \frac{1}{2}$ схема является устойчивой при $\tau = O(h^4)$.

$\lhd$ Из 9.80 следует, что $|\mu_h(k)|^2 = 1 + K$, где

$$K = \frac{(1 - 2\theta_0)\tau^2 \lambda_h^2(k)}{\left(1 + \theta_1 \tau \lambda_h(k)\right)^2 + \theta_0^2 \tau^2 \lambda_h^2(k)}.$$

Так как $K \leqslant (1 - 2\theta_0)\tau \lambda_{\max}^2 \tau$, $\lambda_{\max} = \lambda_h(M-1)$, то условие $K \leqslant c_0 \tau$, означающее устойчивость, имеет место при

$$\tau \leqslant \frac{c_0}{(1 - 2\theta_0)\lambda_{\max}^2} \leqslant \frac{c_0 h^4}{(1 - 2\theta_0)16} = O(h^4).$$

Отсюда следует, что явная схема ($\theta = 0$) устойчива, если $\tau \leqslant \frac{c_0 h^4}{16} = O(h^4)$. При $\theta = \frac{1}{2} - \frac{c_1 h^4}{\tau}$ с произвольной комплексной постоянной c_1 условие $K \leqslant c_0 \tau$ также выполняется. $\rhd$

Таким образом, наличие в уравнении мнимой единицы i принципиально меняет его свойства. В отличие от схем для уравнения теплопроводности ($\theta \geqslant \frac{1}{2}$ — безусловно устойчивые схемы, $\theta < \frac{1}{2}$ — устойчивость при $\tau = O(h^2)$, но в обоих случаях нет растущих решений) для уравнения Шрёдингера все схемы с $\theta_0 \geqslant \frac{1}{2}$ безусловно устойчивы, нет растущих решений, а схемы с $\theta_0 < \frac{1}{2}$ устойчивы при $\tau = O(h^4)$, но допускают растущие решения.

Трехслойная схема с весами

9.82. Для уравнения Шрёдингера найти порядок аппроксимации схемы с весами (θ — действительный весовой параметр)

$$\mathrm{i}\,\frac{u_m^{n+1}-u_m^{n-1}}{2\tau}=\theta\Lambda u_m^{n+1}+(1-2\theta)\Lambda u_m^n+\theta\Lambda u_m^{n-1},\ u_0^n=u_M^n=0,\ n\geqslant 1\,,$$

где

$$\Lambda u_m=\frac{u_{m+1}-2u_m+u_{m-1}}{h^2}\,,\quad h=\frac{1}{M},\ 1\leqslant m\leqslant M-1,$$

и предложить аппроксимацию начальных условий того же порядка.
О т в е т: при всех θ схема имеет порядок аппроксимации $O(\tau^2+h^2)$. Требуемая аппроксимация начальных условий:

$$u_m^0=u^0(x_m),\quad \mathrm{i}\,\frac{u_m^1-u_m^0}{\tau}=u_{xx}^0(x_m)+\mathrm{i}\,\frac{\tau}{2}\,u_{xxxx}^0(x_m),\quad x_m=mh.$$

9.83. Применяя метод разделения переменных, показать, что необходимым условием невозрастания решения трехслойной схемы с весами при любых начальных данных является выполнение неравенства $\theta\geqslant\frac{1}{4}-\frac{h^4}{64\tau^2}$.

$\lhd$ Подставляя в разностную схему частное решение вида $u_m^n=\mu_h^n\sin(\pi kmh)$, получаем для $\mu_h=\mu_h(k)$ квадратное уравнение

$$(\mathrm{i}+2\tau\lambda_h(k)\theta)\mu_h^2-2\tau\lambda_h(k)(2\theta-1)\mu_h+2\tau\lambda_h(k)\theta-\mathrm{i}=0,\ \lambda_h(k)=\frac{4}{h^2}\sin^2\left(\pi k\,\frac{h}{2}\right).$$

Найдем его дискриминант

$$D=4(1-4\theta)(\tau\lambda_h(k))^2-4.$$

Так как модуль произведения корней равен единице, то частные решения не возрастают с увеличением n при $D<0$. Отсюда с учетом $\lambda_h(k)<\frac{4}{h^2}$ получаем ответ: $\theta\geqslant\frac{1}{4}-\frac{h^4}{64\tau^2}$. $\rhd$

9.6. Задача Стокса

Рассмотрим в прямоугольной области $D=\{(x,y):0<x<l_1,0<y<l_2\}$ с границей Γ задачу Стокса в классической формулировке:

$$-\Delta\mathbf{u}+\operatorname{grad}p=\mathbf{f},\ \operatorname{div}\mathbf{u}=0,\ \mathbf{u}|_\Gamma=\mathbf{0}. \tag{9.21}$$

Здесь задана правая часть — вектор-функция $\mathbf{f}=(f^{(1)}(x,y),f^{(2)}(x,y))^T$, а неизвестными являются вектор-функция $\mathbf{u}=(v(x,y),w(x,y))^T$ и скалярная функция $p=p(x,y)$, определенная с точностью до константы. Для однозначности p, как правило, предполагают выполненным условие нормировки $\int\limits_D p(x,y)dxdy=0$. Особенностью постановки задачи является

набор краевых условий: если для $\mathbf{u}$ заданы условия первого рода, то для p никаких дополнительных условий не требуется.

Приведем скалярную форму записи уравнений (9.21)

$$\begin{aligned} -\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}\right)+\frac{\partial p}{\partial x}&=f^{(1)}(x,y),\\ -\left(\frac{\partial^2 w}{\partial x^2}+\frac{\partial^2 w}{\partial y^2}\right)+\frac{\partial p}{\partial y}&=f^{(2)}(x,y),\quad \frac{\partial v}{\partial x}+\frac{\partial w}{\partial y}=0. \end{aligned} \tag{9.22}$$

Схемы метода конечных элементов. Введем следующие пространства:

$$P=\left\{p\,|\,p\in L_2(D),\ \int\limits_D p dxdy=0\right\},\quad \mathbf{U}=\left(\overset{\circ}{W}{}_2^1(D)\right)^2,$$

$$\overset{\circ}{W}{}_2^1(D)=\left\{u\,|\,u,\frac{\partial u}{\partial x},\frac{\partial u}{\partial y}\in L_2(D),\ u|_\Gamma=0\right\}$$

и определим обобщенное решение задачи (9.21) как функции $\mathbf{u}\in\mathbf{U}$ и $p\in P$, удовлетворяющие системе интегральных тождеств

$$\begin{aligned} &(\operatorname{grad}\mathbf{u},\operatorname{grad}\mathbf{v})-(p,\operatorname{div}\mathbf{v})=(\mathbf{f},\mathbf{v}) && \forall\mathbf{v}\in\mathbf{U},\\ &-(q,\operatorname{div}\mathbf{u})=0 && \forall q\in P. \end{aligned} \tag{9.23}$$

Здесь круглые скобки означают скалярные произведения для обычных функций f, g и для векторных функций $\mathbf{u}=(u_1,u_2)^T$, $\mathbf{v}=(v_1,v_2)^T$:

$$(f,g)=\int\limits_D fgdxdy,$$

$$(\operatorname{grad}\mathbf{u},\operatorname{grad}\mathbf{v})=\left(\frac{\partial u_1}{\partial x},\frac{\partial v_1}{\partial x}\right)+\left(\frac{\partial u_1}{\partial y},\frac{\partial v_1}{\partial y}\right)+\left(\frac{\partial u_2}{\partial x},\frac{\partial v_2}{\partial x}\right)+\left(\frac{\partial u_2}{\partial y},\frac{\partial v_2}{\partial y}\right).$$

Для дискретизации задачи (9.23) необходимо ввести конечномерные подпространства $\mathbf{U}_h\subset\mathbf{U}$ и $P_h\subset P$, задаваемые, как правило, в виде линейных оболочек наборов базисных функций:

$$\mathbf{U}_h=\left(\operatorname{span}\{\varphi_1,\varphi_2,\ldots,\varphi_{N_u/2}\}\right)^2,\quad P_h=\operatorname{span}\{\psi_1,\psi_2,\ldots,\psi_{N_p}\},$$

так что

$$\mathbf{u}_h=\sum_{i=1}^{N_u/2}\mathbf{u}_i\varphi_i,\quad p_h=\sum_{i=1}^{N_p}p_i\psi_i,$$

где N_p — общее количество неизвестных, описывающих дискретное давление (размерность P_h), N_u — общее количество неизвестных, описывающих дискретную скорость (размерность $\mathbf{U}_h$). Это приводит к конечномерной задаче: найти $\mathbf{u}_h\in\mathbf{U}_h$ и $p_h\in P_h$ такие, что

$$\begin{aligned} &(\operatorname{grad}\mathbf{u}_h,\operatorname{grad}\mathbf{v})-(p_h,\operatorname{div}\mathbf{v})=(\mathbf{f},\mathbf{v}) && \forall\mathbf{v}\in\mathbf{U}_h,\\ &-(q,\operatorname{div}\mathbf{u}_h)=0 && \forall q\in P_h. \end{aligned} \tag{9.24}$$

При этом независимо от конкретного выбора $\mathbf{U}_h$ и P_h система уравнений (9.24) может быть записана в матричной блочной (2×2) форме

$$\begin{pmatrix} A & B\\ B^T & 0 \end{pmatrix}\begin{pmatrix}\mathbf{u}\\ p\end{pmatrix}=\begin{pmatrix}\mathbf{f}\\ 0\end{pmatrix}, \tag{9.25}$$

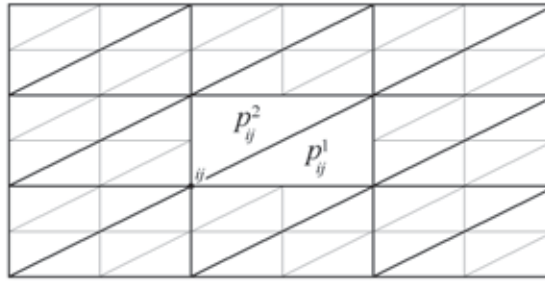


Рис. 2

где $\mathbf{u}$ — дискретный аналог вектора скорости, p — дискретный аналог функции давления, порожденные разложениями искомых неизвестных по базисам пространств $\mathbf{U}_h$ и P_h соответственно, т. е. коэффициенты в $\mathbf{u}_h$ и p_h.

Важным аспектом конечноэлементных аппроксимаций задачи Стокса является неравенство Ладыженской—Бабушки—Брецци (или LBB-условие). Оно может быть представлено в следующей форме: существует постоянная $\delta > 0$, не зависящая от параметра дискретизации области h, такая, что

$$\sup_{\mathbf{u}_h \in \mathbf{U}_h} \frac{(p_h, \operatorname{div} \mathbf{u}_h)}{(\operatorname{grad} \mathbf{u}_h, \operatorname{grad} \mathbf{u}_h)^{1/2}} \geqslant \delta \, \|p_h\|_P \quad \forall p_h \in P_h \,. \tag{9.26}$$

При выполнении этого условия гарантируется корректность дискретного аналога задачи Стокса (9.24), т. е. равномерная ограниченность нормы обратного оператора из (9.25) по параметру дискретизации h. Невыполнение LBB-условия означает: $\delta \to 0$ при $h \to 0$ или $\delta = 0$ при $\forall h$.

Запишем (9.26) в матричных обозначениях

$$\max_{\mathbf{u} \in R^{N_u}} \frac{(p, B^T \mathbf{u})}{(\mathbf{u}, A\mathbf{u})^{1/2}} \geqslant \delta \, (p, Cp)^{1/2} \quad \forall p \in \mathbf{R}^{N_p}. \tag{9.27}$$

В (9.27) C — матрица масс для давления, т. е. матрица Грама с элементами $c_{ij} = (\psi_i, \psi_j)$ для базисных функций из P_h. Определим для матрицы системы (9.25) дополнение по Шуру $S = B^T A^{-1} B$ и рассмотрим спектральную задачу $Sp = \lambda Cp$. Выполнение LBB-условия (9.27) порождает оценку $\lambda_{\min}(C^{-1}S) \geqslant \delta^2 > 0$ при всех h.

9.84. Для разбиения прямоугольной области D (рис. 2) построить конечноэлементную схему для задачи Стокса, используя в качестве базисных линейные функции над маленькими треугольниками для компонент $\mathbf{u}_h$ и постоянные функции над большими треугольниками для p_h.

$\lhd$ Пусть $h_1 N_1 = l_1$, $h_2 N_2 = l_2$. Будем считать N_i, $i = 1, 2$, четными. Тогда множество прямых вида $x = 2ih_1$, $i = 0, 1, \ldots, \frac{N_1}{2}$;

$y = 2jh_2,\ \ j = 0, 1, \ldots, \frac{N_2}{2}$, разбивает $\overline{D} = D \bigcup \Gamma$ на элементарные (большие) прямоугольники. Разобьем каждый элементарный прямоугольник на два треугольника, проводя диагональ, параллельную прямой $y = \frac{h_2}{h_1} x$. Таким образом, мы получим регулярное («северо-восточное») разбиение T_h прямоугольной области $\overline{D}$ на треугольники. Далее разобьем каждый треугольник из T_h на четыре треугольника средними линиями; это даст триангуляцию $T_{h/2}$ (см. рис. 2). Определим пространства

$$\mathbf{U}_h = \{\mathbf{u}_h \,|\, \mathbf{u}_h \in S_1(\triangle), \triangle \in T_{h/2}; \quad \mathbf{u}_h \in C(D); \quad \mathbf{u}_h = \mathbf{0} \ \text{ на } \ \Gamma\},$$

$$P_h = \left\{ p_h \,|\, p_h \in S_0(\triangle), \triangle \in T_h; \quad \int\limits_D p_h dx dy = 0 \right\}.$$

Здесь $C(D)$ — пространство непрерывных на $\overline{D}$ функций, $S_r(\triangle)$ — пространство многочленов степени не выше r, определенных на множестве $\triangle \subset \mathbf{R}^2$, или векторное (размерности 2) пространство функций, каждая компонента которых принадлежит $S_r(\triangle)$.

Функцией формы первого порядка $\varphi_k(x, y)$ для треугольного конечного элемента $\triangle$ называют линейную $(a_0 + a_1 x + a_2 y)$ функцию, удовлетворяющую соотношениям $\varphi_k(x_i, y_i) = \delta_k^i$, где $k, i = 1, 2, 3$. Количество различных функций форм для фиксированного элемента, как правило, соответствует количеству его вершин.

Каждая базисная функция $\varphi_{ij}(x, y)$ определена на всей области D и строится формальным объединением всех функций форм, принимающих в фиксированной точке (ih_1, jh_2) значение единица. Вне конечных элементов $\triangle$, из которых брались эти функции формы, базисные функция продолжаются нулем. Таким образом, базисная функция является пирамидой, в основании которой лежит объединение элементарных треугольников из $T_{h/2}$.

Функции формы $\varphi_k \in S_1(\triangle),\ k = 1, 2, 3$, каждого ij-треугольника $\triangle \in T_{h/2}$ с вершинами (ih_1, jh_2), $((i+1)h_1, jh_2)$, $((i+1)h_1, (j+1)h_2)$ в локальных координатах с началом в точке $(i\,h_1, j\,h_2)$ соответственно имеют вид

$$\varphi_1(x, y) = 1 - \frac{1}{h_1}\, x,\ \varphi_2(x, y) = \frac{1}{h_1}\, x - \frac{1}{h_2}\, y,\ \varphi_3(x, y) = \frac{1}{h_2}\, y.$$

Базисные функции $\varphi_{ij}(x, y)$ в $\mathbf{U}_h$ совпадают с базисными функциями в 9.51, построенными при решении уравнения Пуассона. Это дает возможность записать явно матричную форму (9.25). Пусть $\mathbf{u} = (v, w)^T$ и каждая компонента вектора, в свою очередь, есть вектор с ij-элементами. Тогда после соответствующей нормировки (деления на $h_1 h_2$) получим

$$\left(-\Delta^h v\right)_{ij} = -\left(\frac{v_{i+1,j} - 2v_{ij} + v_{i-1,j}}{h_1^2} + \frac{v_{i,j+1} - 2v_{ij} + v_{i,j-1}}{h_2^2} \right),$$

$$i = 1, \ldots, N_1 - 1, \quad j = 1, \ldots, N_2 - 1.$$

Выражение для $\left(-\Delta^h w\right)_{ij}$ получается заменой v на w в предыдущем выражении. Отметим, что в (9.25) матрица $A = \operatorname{diag}(-\Delta^h, -\Delta^h)$.

Далее, учитывая, что функция формы $\varphi_0 \in S_0(\triangle)$ каждого треугольника $\triangle \in T_h$ имеет вид

$$\varphi_0(x,y) = \begin{cases} 1 & \text{при} \quad (x,y) \in \triangle, \\ 0 & \text{при} \quad (x,y) \notin \triangle \end{cases}$$

и совпадает с соответствующей базисной функцией, имеем

$$(B_x\, p)_{ij} =$$

$$= \frac{1}{2h_1} \begin{cases} -p^1_{ij} + p^1_{i-2,j} + p^2_{i-2,j-2} - p^2_{i,j-2} & \text{при} \quad i = 2l, j = 2k, \\ 0 & \text{при} \quad i = 2l+1, j = 2k, \\ -2p^2_{i,j-1} + 2p^1_{i-2,j-1} & \text{при } i = 2l, j = 2k+1, \\ -2p^1_{i-1,j-1} + 2p^2_{i-1,j-1} & \text{при } i = 2l+1, j = 2k+1; \end{cases}$$

$$(B_y\, p)_{ij} =$$

$$= \frac{1}{2h_2} \begin{cases} -p^2_{ij} - p^1_{i-2,j} + p^1_{i-2,j-2} + p^2_{i,j-2} & \text{при} \quad i = 2l, j = 2k, \\ -2p^1_{i-1,j} + 2p^2_{i-1,j-2} & \text{при} \quad i = 2l+1, j = 2k, \\ 0 & \text{при} \quad i = 2l, j = 2k+1, \\ -2p^2_{i-1,j-1} + 2p^1_{i-1,j-1} & \text{при } i = 2l+1, j = 2k+1; \end{cases}$$

$$i = 1, \ldots, N_1 - 1, \quad j = 1, \ldots, N_2 - 1.$$

Кроме того,

$$(B^T \mathbf{u})^1_{ij} = \frac{-v_{ij} + v_{i+2,j} - 2v_{i+1,j+1} + 2v_{i+2,j+1}}{4h_1} +$$

$$+ \frac{-2w_{i+1,j} - w_{i+2,j} + 2w_{i+1,j+1} + w_{i+2,j+2}}{4h_2},$$

$$(B^T \mathbf{u})^2_{ij} = \frac{-2v_{i,j+1} - v_{i,j+2} + 2v_{i+1,j+1} + v_{i+2,j+2}}{4h_1} +$$

$$+ \frac{-w_{ij} + w_{i,j+2} - 2w_{i+1,j+1} + 2w_{i+1,j+2}}{4h_2},$$

$$i = 0, 2, \ldots, N_1 - 2, \quad j = 0, 2, \ldots, N_2 - 2.$$

Здесь p^1_{ij} и p^2_{ij} соответствуют значениям p в правом нижнем и левом верхнем «больших» треугольниках; точка с координатами ij находится в левом нижнем углу объединяющего их прямоугольника.

Матричная форма (9.25) для данной схемы окончательно принимает вид

$$\left(-\Delta^h\, v\right)_{ij} + (B_x\, p)_{ij} = f^{(1)}_{ij},$$

$$\left(-\Delta^h\, w\right)_{ij} + (B_y\, p)_{ij} = f^{(2)}_{ij},$$

$$(B^T \mathbf{u})^1_{ij} = (B^T \mathbf{u})^2_{ij} = 0,$$

где $f^{(k)}_{ij}$, $k = 1, 2,$ — нормированная ij-компонента проекции заданной функции $f^{(k)}$ на базис $\mathbf{U}_h$. Напомним, что в полученной системе $v_{ij} = w_{ij} = 0$ при $i = 0$, $i = N_1$ и всех j, а также при $j = 0$, $j = N_2$ и всех i.

Построенная схема удовлетворяет LBB-условию, но доказательство этого факта технически сложно. $\triangleright$

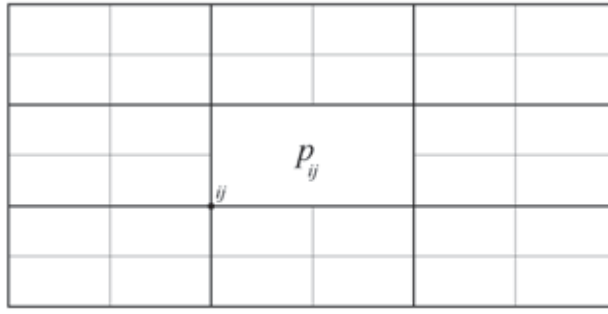


Рис. 3

9.85. Для разбиения прямоугольной области D (рис. 3) построить конечноэлементную схему для задачи Стокса, используя в качестве базисных билинейные функции над маленькими прямоугольниками для компонент $\mathbf{u}_h$ и постоянные функции над большими прямоугольниками для p_h.

$\lhd$ Пусть $h_1N_1 = l_1$, $h_2N_2 = l_2$. Будем считать N_i, $i = 1,2$, четными. Тогда множество прямых вида $x = 2ih_1$, $i = 0,1,\ldots,\frac{N_1}{2}$; $y = 2jh_2$, $j = 0,1,\ldots,\frac{N_2}{2}$, разбивает $\overline{D}$ на элементарные (большие) прямоугольники. Обозначим это разбиение через Q_h. Затем каждый полученный элемент дополнительно разобьем на четыре подобных, соединив середины противоположных сторон прямыми. Построенное таким образом разбиение области обозначим через $Q_{h/2}$ (см. рис. 3).

Рассмотрим аппроксимацию поля скоростей кусочно-билинейными функциями по отношению к разбиению $Q_{h/2}$, непрерывными на $\overline{D}$ и обращающимися в нуль на Γ, т. е.

$$\mathbf{U}_h = \left\{\mathbf{u}_h \,|\, \mathbf{u}_h \in \hat{S}_1(\square), \square \in Q_{h/2}; \quad \mathbf{u}_h \in C(D); \quad \mathbf{u}_h = \mathbf{0} \ \text{ на } \ \Gamma\right\}.$$

Для аппроксимации давления будем использовать кусочно-постоянные функции, определенные на больших прямоугольниках разбиения Q_h и имеющие нулевые средние на D, т. е.

$$P_h = \left\{p_h \,|\, p_h \in \hat{S}_0(\square), \square \in Q_h; \quad \int\limits_D p_h dxdy = 0\right\}.$$

В данном случае $\hat{S}_r(\square)$ имеет смысл пространства полиномов не выше r по каждой координате.

Функции формы $\varphi_k \in \hat{S}_1(\square)$, $k = 1,2,3,4$, каждого ij-прямоугольника $\square \in Q_{h/2}$ с вершинами $(i\,h_1, j\,h_2)$, $((i+1)\,h_1, j\,h_2)$, $(i\,h_1, (j+1)\,h_2)$, $((i+1)\,h_1, (j+1)\,h_2)$ в локальных координатах с началом в точке

$(i\,h_1, j\,h_2)$ имеют соответственно вид

$$\varphi_1(x,y) = \frac{(h_1-x)(h_2-y)}{h_1h_2}\,, \qquad \varphi_2(x,y) = \frac{x(h_2-y)}{h_1h_2}\,,$$
$$\varphi_3(x,y) = \frac{(h_1-x)y}{h_1h_2}\,, \qquad \varphi_4(x,y) = \frac{xy}{h_1h_2}\,.$$

Это дает возможность записать явно матричную форму (9.25). Пусть $\mathbf{u} = (v,w)^T$ и каждая компонента вектора, в свою очередь, есть вектор с ij-элементами. Тогда после соответствующей нормировки (деления на h_1h_2) находим

$$(-\Delta^h v)_{ij} = \frac{1}{h_1h_2}\left[\frac{4(h_1^2+h_2^2)}{3h_1h_2}\,v_{ij} - \frac{h_1^2+h_2^2}{6h_1h_2}\,v_{i+1,j+1} - \frac{-h_1^2+2h_2^2}{3h_1h_2}\,v_{i+1,j} - \right.$$
$$- \frac{h_1^2+h_2^2}{6h_1h_2}\,v_{i+1,j-1} - \frac{2h_1^2-h_2^2}{3h_1h_2}\,v_{i,j+1} - \frac{2h_1^2-h_2^2}{3h_1h_2}\,v_{i,j-1} -$$
$$\left. - \frac{h_1^2+h_2^2}{6h_1h_2}\,v_{i-1,j+1} - \frac{-h_1^2+2h_2^2}{3h_1h_2}\,v_{i-1,j} - \frac{h_1^2+h_2^2}{6h_1h_2}\,v_{i-1,j-1}\right],$$
$$i = 1,\ldots,N_1-1, \quad j = 1,\ldots,N_2-1.$$

Выражение для $(-\Delta^h w)_{ij}$ получается заменой v на w в предыдущем выражении. Отметим, что в (9.25) матрица $A = \operatorname{diag}(-\Delta^h, -\Delta^h)$. Далее, учитывая, что функция формы $\varphi_0 \in \hat{S}_0(\square)$ каждого прямоугольника $\square \in Q_h$ имеет вид

$$\varphi_0(x,y) = \begin{cases} 1 & \text{при} \quad (x,y) \in \square, \\ 0 & \text{при} \quad (x,y) \notin \square \end{cases}$$

и совпадает с соответствующей базисной функцией, получим

$$(B_x\,p)_{ij} = \frac{1}{2h_1}\begin{cases} -p_{ij} + p_{i-2,j} + p_{i-2,j-2} - p_{i,j-2} & \text{при} \quad i = 2l, j = 2k, \\ -2p_{i,j-1} + 2p_{i-2,j-1} & \text{при} \quad i = 2l, j = 2k+1, \\ 0 & \text{при} \quad i = 2l+1, \end{cases}$$
$$(B_y\,p)_{ij} = \frac{1}{2h_2}\begin{cases} -p_{ij} - p_{i-2,j} + p_{i-2,j-2} + p_{i,j-2} & \text{при} \quad i = 2l, j = 2k, \\ -2p_{i-1,j} + 2p_{i-1,j-2} & \text{при} \quad i = 2l+1, j = 2k, \\ 0 & \text{при} \quad j = 2k+1; \end{cases}$$
$$i = 1,\ldots,N_1-1, \quad j = 1,\ldots,N_2-1.$$

Кроме того,

$$(B^T\mathbf{u})_{ij} = \frac{v_{i+2,j+2} + 2v_{i+2,j+1} + v_{i+2,j} - v_{i,j+2} - 2v_{i,j+1} - v_{ij}}{8h_1} +$$
$$+ \frac{w_{i+2,j+2} + 2w_{i+1,j+2} + w_{i,j+2} - w_{i+2,j} - 2w_{i+1,j} - w_{ij}}{8h_2}\,,$$

$$i = 0,2,\ldots,N_1-2, \quad j = 0,2,\ldots,N_2-2.$$

Здесь p_{ij} соответствует значению p в «большом» прямоугольнике; точка с координатами ij находится в левом нижнем углу прямоугольника.

Матричная форма (9.25) для данной схемы окончательно принимает вид

$$\begin{gathered}\left(-\Delta^h v\right)_{ij} + (B_x p)_{ij} = f_{ij}^{(1)},\\ \left(-\Delta^h w\right)_{ij} + (B_y p)_{ij} = f_{ij}^{(2)},\\ (B^T \mathbf{u})_{ij} = 0,\end{gathered}$$

где $f_{ij}^{(k)}$, $k = 1, 2,$ — нормированная ij-компонента проекции заданной функции $f^{(k)}$ на базис $\mathbf{U}_h$. Напомним, что в полученной системе $v_{ij} = w_{ij} = 0$ при $i = 0$, $i = N_1$ и всех j, а также при $j = 0$, $j = N_2$ и всех i. Построенная схема удовлетворяет LBB-условию, но доказательство этого факта технически сложно. $\triangleright$

Дискретный аналог оператора Лапласа (диагональный блок матрицы A) в данной схеме является нестандартным, и для него имеет место соотношение

$$\begin{gathered}-\Delta^h \sin(m\pi i h_1)\sin(n\pi j h_2) = \lambda_{mn}\sin(m\pi i h_1)\sin(n\pi j h_2),\\ i = 1, \ldots, N_1 - 1, \quad j = 1, \ldots, N_2 - 1,\end{gathered}$$

где

$$\begin{gathered}\lambda_{mn} = \frac{2}{3h_1h_2}\left[\frac{2(h_1^2+h_2^2)}{h_1h_2} - \frac{h_1^2+h_2^2}{h_1h_2}\cos(\pi m h_1)\cos(\pi n h_2) - \right.\\ \left. - \frac{-h_1^2+2h_2^2}{h_1h_2}\cos(\pi m h_1) - \frac{2h_1^2-h_2^2}{h_1h_2}\cos(\pi n h_2)\right],\\ m = 1, \ldots, N_1 - 1, \quad n = 1, \ldots, N_2 - 1.\end{gathered}$$

Приведенное выражение можно использовать для решения систем вида $\left(-\Delta^h v\right)_{ij} = f_{ij}$ методом разложения в двойной ряд.

9.86. Показать, что если для разбиения прямоугольной области D (см. рис. 2) построить конечноэлементную схему для задачи Стокса, выбирая в качестве базисных линейные функции для компонент $\mathbf{u}_h$ и постоянные функции для p_h над одинаковыми треугольниками, то она не будет удовлетворять LBB-условию.

$\triangleleft$ В дискретной задаче Стокса имеется уравнение

$$B^T \mathbf{u} = 0\,. \tag{9.28}$$

Покажем сначала, что из этого уравнения следует тождество $\mathbf{u} \equiv \mathbf{0}$. Введем обозначения: $v_{i,j} = v(ih_1, jh_2)$, $w_{i,j} = w(ih_1, jh_2)$ и рассмотрим прямоугольник в нижнем левом углу. В силу нулевых краевых условий, имеем

$$v_{0,0} = v_{1,0} = v_{0,1} = w_{0,0} = w_{1,0} = w_{0,1} = 0\,,$$

т. е. неизвестными являются только значения $v_{1,1}$ и $w_{1,1}$. Для левого верхнего треугольника T_l в этом прямоугольнике справедливо представление

$$\mathbf{u}_h = \left(v_{1,1}\varphi_l(x,y), w_{1,1}\varphi_l(x,y)\right)^T,$$

где $\varphi_l(x,y) = \frac{x}{h_1}$ — функция формы T_l, относящаяся к вершине (h_1, h_2). Соотношение

$$\int\limits_{T_l} \operatorname{div} \mathbf{u}_h \, dxdy = 0\,,$$

т. е. ортогональность дивергенции произвольной постоянной для фиксированного треугольника T_l, порождает равенство $v_{1,1} = 0$, так как $\varphi_l(x,y)$ не зависит от y.

Теперь рассмотрим правый нижний треугольник T_r в этом прямоугольнике. Для T_r имеем

$$\mathbf{u}_h = (v_{1,1}\varphi_r(x,y), w_{1,1}\varphi_r(x,y))^T\,,$$

где $\varphi_r(x,y) = \frac{y}{h_2}$ — функция формы треугольника T_r, относящаяся к вершине (h_1, h_2). Соотношение

$$\int\limits_{T_r} \operatorname{div} \mathbf{u}_h \, dxdy = 0$$

порождает равенство $w_{1,1} = 0$, так как $\varphi_r(x,y)$ не зависит от x.

Далее рассмотрим соседний прямоугольник, расположенный справа. Аналогичные рассуждения, базирующиеся на значениях

$$v_{1,1} = v_{1,0} = v_{2,0} = w_{1,1} = w_{1,0} = w_{2,0} = 0\,,$$

приводят к равенствам $v_{2,1} = 0$ и $w_{2,1} = 0$. Перебирая прямоугольники направо до границы, получим

$$v_{i,1} = w_{i,1} = 0\,, \quad 0 \leqslant i \leqslant N_1\,.$$

Исчерпав первую линию прямоугольников, перейдем к следующей: опять рассмотрим примыкающий к левой границе прямоугольник и т. д. Такой последовательный перебор приводит к $\mathbf{u} \equiv \mathbf{0}$, т. е. исходная линейная система

$$\begin{gathered} A\,\mathbf{u} + B\,p = \mathbf{f}, \\ B^T\mathbf{u} = 0 \end{gathered}$$

принимает вид

$$B\,p = \mathbf{f}\,. \tag{9.29}$$

Рассмотрим размерность этой системы. Количество уравнений системы совпадает с числом степеней свободы вектора $\mathbf{u}$, т. е. равно $N_u = 2(N_1 - 1)(N_2 - 1)$. Количество неизвестных N_p определяется числом треугольников (так как $p_h(x,y)$ — кусочно-постоянная на каждом треугольнике функция). В силу специфики носителей базисных функций для компонент скорости, верхний левый и правый нижний треугольники в построении схемы не участвуют, поэтому $N_p = 2N_1N_2 - 2$. Таким образом, получено, что в системе уравнений (9.29) число неизвестных N_p больше числа уравнений N_u. Это означает, что ядро матрицы B нетривиально, т. е. $\delta = 0\ \forall h$, следовательно, LBB-условие не выполнено. $\triangleright$

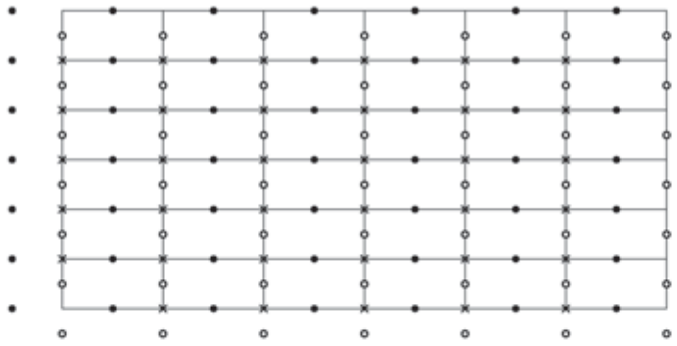


Рис. 4

9.87. Показать, что если для разбиения прямоугольной области D (см. рис. 3) построить конечноэлементную схему для задачи Стокса, используя в качестве базисных билинейные функции для компонент $\mathbf{u}_h$ и постоянные функции для p_h над одинаковыми прямоугольниками, то она не будет удовлетворять LBB-условию.

Схемы метода конечных разностей. Схемы, построенные методом конечных разностей, часто называют схемами на *смещенных (staggered) сетках Лебедева, или MAC-схемами.* Их спецификой является использование различных сеточных областей определения для компонент решения v, w и p. Пусть D_1, D_2 и D_3 — некоторые дискретные аналоги области D (со своими сеточными границами Γ_i, $i = 1, 2, 3$). Введем следующие обозначения: $\mathbf{U}_h$ — линейное пространство вектор-функций, определенных на $\overline{D}_1 \times \overline{D}_2$ и обращающихся в нуль на соответствующих сеточных границах; P_h — пространство функций, определенных на D_3 и ортогональных единице. При фиксированных D_i, $i = 1, 2, 3$, целью является построение сеточных уравнений, решение которых $\mathbf{u}_h = (v, w)^T$ и p_h принадлежит пространствам $\mathbf{U}_h$ и P_h соответственно.

На рисунках 4–6 множества $\overline{D}_1$, $\overline{D}_2$ и D_3 обозначены символами $\bullet$, $\circ$ и $\times$ соответственно.

9.88. Пусть сеточные области (см. рис. 4) определены следующим образом:

$$\overline{D}_1 = \{x_{ij} = \left((i-\tfrac{1}{2})\,h_1, jh_2\right):\quad i=0,\ldots,N_1,\quad j=0,\ldots,N_2\},$$

$$\overline{D}_2 = \{x_{ij} = \left(ih_1, \left(j-\tfrac{1}{2}\right)h_2\right): i=0,\ldots,N_1,\quad j=0,\ldots,N_2\},$$

$$D_3 = \{x_{ij} = (ih_1, jh_2):\qquad i=0,\ldots,N_1-1, j=0,\ldots,N_2-1,\ i^2+j^2\neq 0\}.$$

Построить разностную схему для задачи Стокса.

О т в е т: запишем соответствующие пространства в индексных обозначениях:

$$\mathbf{U}_h = V_{1,h} \times V_{2,h},$$

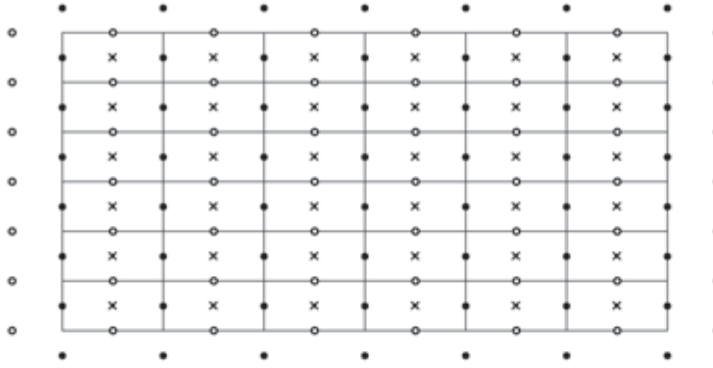


Рис. 5

где

$$V_{1,h} = \{v_{ij} = v(x_{ij}) : x_{ij} \in \overline{D}_1, \quad v_{0,j} = v_{N_1,j} = v_{i,0} = v_{i,N_2} = 0\},$$
$$V_{2,h} = \{w_{ij} = w(x_{ij}) : x_{ij} \in \overline{D}_2,\, w_{0,j} = w_{N_1,j} = w_{i,0} = w_{i,N_2} = 0\},$$
$$P_h = \{p_{ij} = p(x_{ij}) : x_{ij} \in D_3, \sum_{ij} h_1 h_2\, p_{ij} = 0\}.$$

Сеточные уравнения при этом принимают вид

$$\frac{v_{i+1,j} - 2v_{ij} + v_{i-1,j}}{h_1^2} + \frac{v_{i,j+1} - 2v_{ij} + v_{i,j-1}}{h_2^2} - \frac{p_{ij} - p_{i-1,j}}{h_1} = -f_{ij}^{(1)},$$
$$\frac{w_{i+1,j} - 2w_{ij} + w_{i-1,j}}{h_1^2} + \frac{w_{i,j+1} - 2w_{ij} + w_{i,j-1}}{h_2^2} - \frac{p_{ij} - p_{i,j-1}}{h_2} = -f_{ij}^{(2)},$$
$$\frac{v_{i+1,j} - v_{ij}}{h_1} + \frac{w_{i,j+1} - w_{ij}}{h_2} = 0.$$

В этой системе первое, второе и третье уравнения заданы на множествах D_1, D_2 и D_3 соответственно (здесь и далее $D_i = \overline{D}_i \setminus \Gamma_i$).

9.89. Пусть сеточные области (см. рис. 5) определены следующим образом:

$$\overline{D}_1 = \{x_{ij} = \left(ih_1, \left(j - \tfrac{1}{2}\right) h_2\right) : \qquad i = 0, \ldots, N_1, \qquad j = 0, \ldots, N_2 + 1\},$$
$$\overline{D}_2 = \{x_{ij} = \left(\left(i - \tfrac{1}{2}\right) h_1, jh_2\right) : \qquad i = 0, \ldots, N_1 + 1,\, j = 0, \ldots, N_2\},$$
$$D_3 = \{x_{ij} = \left(\left(i + \tfrac{1}{2}\right) h_1, \left(j + \tfrac{1}{2}\right) h_2\right) : \; i = 0, \ldots, N_1 - 1,\, j = 0, \ldots, N_2 - 1\}.$$

Построить разностную схему для задачи Стокса.

О т в е т: определим пространства $\mathbf{U}_h = V_{1,h} \times V_{2,h}$, где

$$V_{1,h} = \{v_{ij} = v(x_{ij}) : x_{ij} \in \overline{D}_1, \quad v_{0,j} = v_{N_1,j} = 0, v_{i,0} = -v_{i,1}, v_{i,N_2+1} = -v_{i,N_2}\},$$
$$V_{2,h} = \{w_{ij} = w(x_{ij}) : x_{ij} \in \overline{D}_2, \; w_{0,j} = -w_{1,j}, w_{N_1+1,j} = -w_{N_1,j}, w_{i,0} = w_{i,N_2} = 0\},$$
$$P_h = \{p_{ij} = p(x_{ij}) : x_{ij} \in D_3, \sum_{ij} h_1 h_2\, p_{ij} = 0\},$$

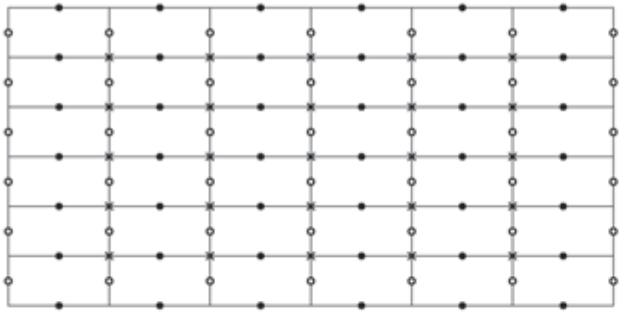


Рис. 6

и запишем уравнения

$$\frac{v_{i+1,j}-2v_{ij}+v_{i-1,j}}{h_1^2}+\frac{v_{i,j+1}-2v_{ij}+v_{i,j-1}}{h_2^2}-\frac{p_{i,j-1}-p_{i-1,j-1}}{h_1}=-f_{ij}^{(1)},$$

$$\frac{w_{i+1,j}-2w_{ij}+w_{i-1,j}}{h_1^2}+\frac{w_{i,j+1}-2w_{ij}+w_{i,j-1}}{h_2^2}-\frac{p_{i-1,j}-p_{i-1,j-1}}{h_2}=-f_{ij}^{(2)},$$

$$\frac{v_{i+1,j+1}-v_{i,j+1}}{h_1}+\frac{w_{i+1,j+1}-w_{i+1,j}}{h_2}=0.$$

Здесь первое, второе и третье уравнения заданы на множествах D_1, D_2 и D_3 соответственно.

9.90. Пусть сеточные области (см. рис. 6) определены следующим образом:

$$\overline{D}_1=\{x_{ij}=\left(\left(i+\tfrac{1}{2}\right)h_1,jh_2\right):\ i=0,\ldots,N_1-1,\ j=0,\ldots,N_2\},$$

$$\overline{D}_2=\{x_{ij}=\left(ih_1,\left(j+\tfrac{1}{2}\right)h_2\right):\ i=0,\ldots,N_1,\qquad j=0,\ldots,N_2-1\},$$

$$D_3=\{x_{ij}=(ih_1,jh_2):\qquad\qquad i=1,\ldots,N_1-1,\ j=1,\ldots,N_2-1\}.$$

Построить разностную схему для задачи Стокса.

О т в е т: построим пространства $\mathbf{U}_h=V_{1,h}\times V_{2,h}$, где

$$V_{1,h}=\{v_{ij}=v(x_{ij}):x_{ij}\in\overline{D}_1,\ \ v_{0,j}=v_{N_1-1,j}=v_{i,0}=v_{i,N_2}=0\},$$

$$V_{2,h}=\{w_{ij}=w(x_{ij}):x_{ij}\in\overline{D}_2,w_{0,j}=w_{N_1,j}=w_{i,0}=w_{i,N_2-1}=0\},$$

$$P_h=\{p_{ij}=p(x_{ij}):x_{ij}\in D_3,\sum_{ij}h_1h_2\,p_{ij}=0\},$$

и сеточные уравнения

$$\frac{v_{i+1,j}-2v_{ij}+v_{i-1,j}}{h_1^2}+\frac{v_{i,j+1}-2v_{ij}+v_{i,j-1}}{h_2^2}-\frac{p_{i+1,j}-p_{ij}}{h_1}=-f_{ij}^{(1)},$$

$$\frac{w_{i+1,j}-2w_{ij}+w_{i-1,j}}{h_1^2}+\frac{w_{i,j+1}-2w_{ij}+w_{i,j-1}}{h_2^2}-\frac{p_{i,j+1}-p_{ij}}{h_2}=-f_{ij}^{(2)},$$

$$\frac{v_{ij}-v_{i-1,j}}{h_1}+\frac{w_{ij}-w_{i,j-1}}{h_2}=0.$$

Здесь первое, второе и третье уравнения заданы на множествах D_1, D_2 и D_3 соответственно.

9.91. Пусть сеточные области определены следующим образом:

$$\overline{D}_1 = \{x_{ij} = (ih_1, jh_2): \quad i = 0, \ldots, N_1,\ j = 0, \ldots, N_2\},$$
$$\overline{D}_2 = \{x_{ij} = (ih_1, jh_2): \quad i = 0, \ldots, N_1,\ j = 0, \ldots, N_2\},$$
$$D_3 = \{x_{ij} = (ih_1, jh_2): \quad i = 0, \ldots, N_1 - 1,\ j = 0, \ldots, N_2 - 1\}.$$

Построить разностную схему для задачи Стокса, используя для аппроксимации операторов div и grad разности вперед и назад соответственно.

О т в е т: формулы для разностных уравнений совпадают с ответом 9.88.

9.92. Пусть сеточные области определены следующим образом:

$$\overline{D}_1 = \{x_{ij} = (ih_1, jh_2): \qquad i = 0, \ldots, N_1,\ j = 0, \ldots, N_2\},$$
$$\overline{D}_2 = \{x_{ij} = (ih_1, jh_2): \qquad i = 0, \ldots, N_1,\ j = 0, \ldots, N_2\},$$
$$D_3 = \{x_{ij} = \left(\left(i + \tfrac{1}{2}\right) h_1, \left(j + \tfrac{1}{2}\right) h_2\right): \ i = 0, \ldots, N_1 - 1,\ j = 0, \ldots, N_2 - 1\}.$$

Построить разностную схему второго порядка аппроксимации для задачи Стокса, используя для операторов div^h и grad^h выражения, определенные на симметричных относительно x_{ij} шаблонах из ближайших четырех узлов. Показать, что ядро оператора grad^h не содержит нетривиальных функций, кроме

$$p_{ij}^{(1)} = \frac{1 - (-1)^{i+j}}{2} \quad \text{и} \quad p_{ij}^{(2)} = \frac{1 - (-1)^{i+j+1}}{2}.$$

Глава 10

Интегральные уравнения

Обозначим через G_b интегральный оператор следующего вида:

$$G_b y(x) = \int_a^b K(x,s)y(s)\,ds\,,$$

где заданная функция $K(x,s)$ называется *ядром*, а верхний предел интегрирования b в общем случае может быть переменным. Типичными примерами интегральных уравнений являются уравнения Фредгольма и Вольтерры, каждое из которых может быть первого или второго рода. В уравнениях Фредгольма b является постоянной заданной величиной

$$G_b y(x) = f(x) \quad \text{и} \quad y(x) - \lambda G_b y(x) = f(x),$$

причем в уравнения второго рода входит дополнительный числовой параметр λ, который может быть задан или подлежит определению в зависимости от постановки задачи. В уравнениях Вольтерры верхний предел интегрирования совпадает с текущим значением независимой переменной: $b = x$, поэтому уравнения Вольтерры первого и второго рода соответственно имеют вид

$$G_x y(x) = f(x) \quad \text{и} \quad y(x) - \lambda G_x y(x) = f(x).$$

Функция $f(x)$ задана, а $y(x)$ подлежит определению в уравнениях всех типов.

Наиболее распространенной модельной постановкой является следующая задача для уравнения Фредгольма второго рода: на множествах $[a,b]$ и $[a,b]\times[a,b]$ заданы квадратично интегрируемые функции $f(x)$ и $K(x,s)$; для заданного значения параметра λ требуется определить квадратично интегрируемую функцию $y(x)$, удовлетворяющую уравнению

$$y(x) - \lambda \int_a^b K(x,s)y(s)\,ds = f(x)\,. \tag{10.1}$$

Предполагается, что (10.1) имеет единственное решение. Это справедливо, например, если выполнено условие

$$|\lambda| < \left(\int_a^b \int_a^b |K(x,s)|^2 dx ds \right)^{-1/2}.$$

10.1. Метод замены интеграла

Самым простым подходом к решению модельной задачи (10.1) является замена интеграла в уравнении какой-либо квадратурной формулой. В результате получается система линейных алгебраических уравнений относительно значений неизвестной функции, которая решается рассмотренными ранее прямыми или итерационными методами.

Для определения приближенного значения интеграла $I(\varphi) = \int\limits_a^b \varphi(s)ds$ воспользуемся квадратурной формулой $S_N(\varphi) = \sum\limits_{j=1}^{N} c_j\varphi(s_j)$. Тогда дискретный аналог (10.1) в узлах $x_i = s_i$, $1 \leqslant i \leqslant N$, принимает вид

$$y_i - \lambda \sum_{j=1}^{N} c_j K(x_i, s_j) y_j = f(x_i), \quad 1 \leqslant i \leqslant N\,, \tag{10.2}$$

что представляет собой систему линейных алгебраических уравнений $A\,\mathbf{y} = \mathbf{f}$ относительно неизвестных $y_1, y_2, \ldots, y_N$, являющихся приближениями к точным значениям $y(x_1), y(x_2), \ldots, y(x_N)$.

Предметом рассмотрения при таком подходе являются оценки погрешности приближенного решения (10.1), возникающей в результате замены интеграла квадратурной формулой, и методы решения полученной системы (10.2).

В основе метода замены интеграла квадратурной формулой лежит следующее утверждение.

Теорема. *Пусть однородное уравнение имеет только нулевое решение, ядро и решение уравнения (10.1) непрерывно дифференцируемы до m-го порядка включительно, погрешность квадратурной формулы на гладких функциях имеет асимптотику не ниже, чем $I(\varphi) - S_N(\varphi) = O(N^{-m})$. Тогда справедлива оценка погрешности*

$$\max_{[a,b]} |y(x) - u(x)| \leqslant C\, N^{-m}\,,$$

где $u(x)$ — решение уравнения

$$u(x) - \lambda \sum_{j=1}^{N} c_j K(x, s_j) u(s_j) = f(x),$$

а постоянная C не зависит от N — количества узлов квадратурной формулы.

10.1. Найти приближенное решение интегрального уравнения

$$y(x) + \int\limits_0^1 x(\mathrm{e}^{xs} - 1) y(s)\, ds = \mathrm{e}^x - x$$

методом замены интеграла квадратурной формулой Симпсона и оценить его погрешность.

$\lhd$ Для отыскания приближенного решения $y(x)$ в точках $x = 0, \frac{1}{2}, 1$ запишем систему уравнений

$$y_1 = 1\,,\ \frac{1}{3}\,(\mathrm{e}^{0,25} - 1)y_2 + \frac{1}{12}\,(\mathrm{e}^{0,5} - 1)y_3 = \mathrm{e}^{0,5} - \frac{1}{2}\ ,$$

$$\frac{2}{3}\,(\mathrm{e}^{0,5} - 1)y_2 + \frac{1}{6}\,(\mathrm{e} + 5)y_3 = \mathrm{e} - 1,$$

или (с точностью до четырех знаков после запятой)

$$y_1 = 1\,,\ 1,0947y_2 + 0,0541y_3 = 1,1487\,,$$
$$0,4325y_2 + 1,2864y_3 = 1,7183\,.$$

Решение этой системы: $y_1 = 1$, $y_2 \approx 0,9999$, $y_3 \approx 0,9996$. Несложно проверить, что точное решение уравнения $y(x) \equiv 1$, поэтому абсолютная погрешность не превышает $0,0004$. $\triangleright$

10.2. Найти приближенное решение интегрального уравнения

$$y(x) = \frac{5}{6}\,x + \frac{1}{2}\int\limits_0^1 xsy(s)\,ds$$

методом замены интеграла квадратурной формулой Симпсона и оценить его погрешность.

У к а з а н и е. Точное решение $y(x) = x$.

10.3. Найти приближенное решение интегрального уравнения

$$y(x) = \mathrm{e}^{-x} + \frac{1}{2}\int\limits_0^1 x\mathrm{e}^s y(s)\,ds$$

методом замены интеграла квадратурной формулой Симпсона и оценить его погрешность.

О т в е т: с тремя верными десятичными знаками $y(x) = 1,003x + \mathrm{e}^{-x}$, точное решение $y(x) = x + \mathrm{e}^{-x}$.

10.4. Найти в узлах $a = x_1 < x_2 < \cdots < x_N = b$ приближенное решение уравнения Вольтерры второго рода

$$y(x) - \lambda\int\limits_a^x K(x,s)\,y(s)\,ds = f(x)$$

методом замены интеграла составной квадратурной формулой трапеций.

$\triangleleft$ Для вычисления элементарного интеграла $I_i = \int\limits_{x_{i-1}}^{x_i} \varphi(s)\,ds$ по отрезку длины $h_i = x_i - x_{i-1}$ квадратурная формула трапеций имеет вид

$$I_i \approx h_i\,\frac{\varphi(x_{i-1}) + \varphi(x_i)}{2}\,.$$

Поэтому для определения приближенного решения имеем систему уравнений

$$y_1 = f(x_1),\ y_i - \lambda\sum_{j=2}^{i} h_j\,\frac{K_{i,j-1}y_{j-1} + K_{i,j}y_j}{2} = f(x_i), \quad i = 2,3,\ldots,N,$$

где $K_{i,j} = K(x_i, s_j)$. Решение системы можно получить рекуррентно: $y_1 = f(x_1)$; для $i = 2,3,\ldots,N$ его находят по формуле

$$y_i = \frac{2f(x_i) + \lambda h_2 K_{i,1}y_1 + \lambda\sum\limits_{j=2}^{i-1}(h_j + h_{j+1})\,K_{i,j}y_j}{2 - \lambda h_i K_{i,i}}\,,$$

если знаменатель не обращается в нуль. Результат суммирования в этой формуле равен нулю, если верхний индекс суммирования меньше нижнего (для $i = 2$). $\triangleright$

10.5. Найти в точках $x = 0, \frac{1}{2}, 1$ приближенное решение интегрального уравнения

$$y(x) = \mathrm{e}^{-x} + \int_0^x \mathrm{e}^{-(x-s)} y(s)\, ds$$

методом замены интеграла составной квадратурной формулой трапеций.
У к а з а н и е. Точное решение $y(x) \equiv 1$.

10.6. Найти в точках $x = 0, \frac{1}{2}, 1$ приближенное решение интегрального уравнения

$$y(x) = \mathrm{e}^{x} + \int_0^x \mathrm{e}^{x-s} y(s)\, ds$$

методом замены интеграла составной квадратурной формулой трапеций.
У к а з а н и е. Точное решение $y(x) = \mathrm{e}^{2x}$.

10.7. Пусть $K(x, s) \equiv K = \mathrm{const} > 0$. Записать в явном виде элементы матрицы в системе (10.2), если в качестве квадратурной формулы используется составная формула прямоугольников с узлом в центральной точке.

$\triangleleft$ Пусть $h = (b - a)/N$; тогда составная формула прямоугольников для вычисления интеграла $\int_a^b \varphi(s) ds$ такова:

$$S_1^N(\varphi) = h \sum_{j=1}^{N} \varphi\left(a + h\left(j - \frac{1}{2}\right)\right),$$

а систему (10.2) можно записать в виде

$$\sum_{j=1}^{N} a_{ij} y_j = f(x_i), \quad 1 \leqslant i \leqslant N,$$

где $a_{ij} = \delta_i^j - \lambda\, K\, h$, δ_i^j — символ Кронекера. $\triangleright$

10.8. Исследовать сходимость метода простой итерации для решения системы линейных уравнений, полученной в 10.7.

$\triangleleft$ Запишем метод простой итерации в виде $\mathbf{y}^{k+1} = B\mathbf{y}^k + \mathbf{f}$. Из решения 10.7 имеем явный вид элементов матрицы B : $b_{ij} = \lambda K h$. Матрица размера $N \times N$ с постоянными элементами $\lambda K h$ имеет ядро размерности $N - 1$ и одно собственное значение, равное $N\lambda K h$. В этом легко убедиться, если рассмотреть действие матрицы на векторы $(-1, 1, 0 \ldots, 0)^T$, $(-1, 0, 1, 0 \ldots, 0)^T, \ldots,$ $(-1, 0, 0 \ldots, 0, 1)^T$ — базис ядра и $(1, 1, \ldots, 1)^T$ — базис образа. Отсюда, по теореме о необходимом и достаточном условии

сходимости метода простой итерации, получаем, что сходимость с произвольного начального приближения имеется при выполнении условия $|\lambda|\,K\,(b-a)<1$. $\triangleright$

10.9. Пусть $K(x,s)\equiv K=\text{const}>0$, $|\lambda|\,K\,(b-a)<\alpha<1$. Оценить погрешность решения при замене интеграла составной квадратурной формулой прямоугольников.

$\triangleleft$ Обозначим через $R_1^N(\varphi)$ остаточный член составной квадратурной формулы прямоугольников

$$\int\limits_a^b \varphi(s)ds = S_1^N(\varphi)+R_1^N(\varphi).$$

Тогда система уравнений для точных значений решения (10.1) в узлах x_i имеет вид

$$y(x_i)-\lambda K\,\frac{b-a}{N}\sum_{j=1}^{N} y(x_j)=f(x_i)+\lambda K R_1^N(y)\big|_{x=x_i},\quad 1\leqslant i\leqslant N\,.$$

Введя для компонент погрешности обозначение $z_i=y(x_i)-y_i$ и обозначение $r_i=R_1^N(y)\big|_{x=x_i}$ для остаточного члена, получим систему $A\mathbf{z}=\lambda K\mathbf{r}$. Отсюда в силу диагонального преобладания элементов матрицы A, так как $|\lambda|\,K\,(b-a)<\alpha<1$, имеем $\|\mathbf{z}\|_\infty\leqslant|\lambda|K\|A^{-1}\|_\infty\|\mathbf{r}\|_\infty$. Для остаточного члена составной квадратурной формулы прямоугольников справедлива оценка (см. раздел 4.1)

$$|R_1^N(y)|\leqslant\|y''\|\,\frac{(b-a)^3}{24N^2}\,,$$

откуда следует неравенство

$$\|\mathbf{r}\|_\infty\leqslant\|y''\|\,\frac{(b-a)^3}{24N^2}\,.$$

В обоих этих неравенствах использовалась равномерная норма.

Оценим величину $\|A^{-1}\|_\infty$. Так как матрица обладает свойством диагонального преобладания, то, на основании 5.65, имеем $\|A^{-1}\|_\infty\leqslant$ $\leqslant\frac{1}{1-\alpha}\,\frac{1}{1-|\lambda|Kh}\,,h=(b-a)/N$. Отсюда следует окончательная оценка

$$\|\mathbf{z}\|_\infty\leqslant\frac{1}{1-\alpha}\;\frac{|\lambda|K}{1-|\lambda|Kh}\,\|y''(x)\|\,\frac{(b-a)^3}{24N^2}=O(h^2)\,.$$

Таким образом, если система линейных уравнений «не слишком плоха» и решение достаточно гладкое, то порядок погрешности в узлах x_i совпадает с порядком остаточного члена составной квадратурной формулы.

Для получения приближенных значений $y(x)$ при $x\neq x_i$ можно воспользоваться кусочно-линейной интерполяцией, которая в этом случае не ухудшит порядок ошибки, равный двум (см. 3.140). Если используемая

квадратурная формула является более точной, то для получения приближенного решения интегрального уравнения рекомендуется использовать формулу

$$y(x) \approx f(x) + \lambda \sum_{j=1}^{N} c_j K(x_i, s_j) y_j$$

или воспользоваться интерполяцией более высокого порядка. $\triangleright$

10.10. Показать, что для решения системы уравнений, полученной в 10.7, метод Гаусса является устойчивым, т. е. все элементы матриц, возникающих в процессе треугольной LR-факторизации, равномерно ограничены.

10.2. Метод замены ядра

Если в уравнении (10.1) ядро $K(x, s)$ вырождено, т. е.

$$K(x, s) = \sum_{i=1}^{p} A_i(x) B_i(s),$$

где $\{A_i(x)\}_{i=1}^{p}$ и $\{B_i(x)\}_{i=1}^{p}$ — системы линейно независимых на отрезке $[a, b]$ функций, то (10.1) можно записать в виде

$$y(x) - \lambda \sum_{i=1}^{p} A_i(x) \int_a^b B_i(s) y(s) ds = f(x)\,.$$

Решение уравнения такой структуры удобно представить формулой

$$y(x) = f(x) + \lambda \sum_{i=1}^{p} D_i A_i(x)\,,$$

где D_i — некоторые постоянные, подлежащие определению. В результате подстановки в уравнение формулы для решения $y(x)$ и сокращения на λ получим

$$\sum_{i=1}^{p} D_i A_i(x) - \lambda \sum_{i=1}^{p} A_i(x) \sum_{j=1}^{p} D_j \int_a^b A_j(s) B_i(s) ds = \sum_{i=1}^{p} A_i(x) \int_a^b f(s) B_i(s) ds\,.$$

Обозначим

$$f_i = \int_a^b f(s) B_i(s) ds, \quad a_{ij} = \int_a^b A_j(s) B_i(s) ds$$

и на основании линейной независимости функций $\{A_i(x)\}_{i=1}^{p}$ перепишем полученное равенство в виде

$$D_i - \lambda \sum_{j=1}^{p} a_{ij} D_j = f_i, \quad 1 \leqslant i \leqslant p\,. \tag{10.3}$$

Если определитель системы (10.3) отличен от нуля, то система имеет единственное решение $D_1, \ldots, D_p$ и решение интегрального уравнения

$y(x)$ будет явно найдено в указанном виде. Если при заданном λ определитель (10.3) равен нулю, то λ является характеристическим числом уравнения (ядра $K(x,s)$). В этом случае, находя все линейно независимые решения соответствующей однородной системы, в явном виде можно записать собственные функции ядра $K(x,s)$, соответствующие этому характеристическому числу λ (собственным значением называется величина $\frac{1}{\lambda}$).

Метод приближенного решения интегральных уравнений Фредгольма с помощью замены ядра $K(x,s)$ близким к нему вырожденным ядром $H(x,s)$ основан на следующем результате.

Теорема. *Если*

$$y(x) - \lambda \int_a^b K(x,s)y(s)\,ds = f(x)\,,$$

$$z(x) - \lambda \int_a^b H(x,s)z(s)\,ds = \varphi(x)$$

— два интегральных уравнения, $R(x,s,\lambda)$ — резольвента второго из этих уравнений, существуют такие константы δ, ε, M, что имеют место неравенства

$$\int_a^b |K(x,s) - H(x,s)|ds \leqslant \delta,\ |f(x) - \varphi(x)| \leqslant \varepsilon,\ \int_a^b |R(x,s,\lambda)|ds \leqslant M$$

и выполнено условие

$$|\lambda|\,\delta\,(1 + |\lambda|\,M) < 1,$$

то первое интегральное уравнение имеет единственное решение $y(x)$ и справедлива оценка

$$|y(x) - z(x)| \leqslant \frac{F\,|\lambda|\,\delta\,(1 + |\lambda|\,M)^2}{1 - |\lambda|\,\delta\,(1 + |\lambda|\,M)} + \varepsilon\,(1 + |\lambda|\,M)\,,$$

где $F = \max\limits_{a \leqslant x \leqslant b} |f(x)|$.

Способы построения вырожденных ядер, близких к данному ядру $K(x,s)$, могут быть самые различные. Например, $K(x,s)$ можно приближать частичными суммами степенного или двойного тригонометрического ряда, если оно разлагается в соответствующий равномерно сходящийся в прямоугольнике $a \leqslant x, s \leqslant b$ ряд, или приближать ядро алгебраическими или тригонометрическими интерполяционными многочленами.

10.11. Найти приближенное решение интегрального уравнения

$$y(x) + \int_0^1 x(\mathrm{e}^{xs} - 1)y(s)\,ds = \mathrm{e}^x - x$$

с помощью замены ядра вырожденным $H(x,s) = x^2 s + \frac{x^3 s^2}{2} + \frac{x^4 s^3}{6}$, взятым в виде первых трех членов разложения $K(x,s)$ в ряд Тейлора.

$\lhd$ Вместо исходного рассмотрим интегральное уравнение

$$z(x) + \int\limits_0^1 H(x,s)z(s)\,ds = \mathrm{e}^x - x\,,$$

решение которого будем искать в виде

$$z(x) = \mathrm{e}^x - x + D_1x^2 + D_2x^3 + D_3x^4\,.$$

Для определения постоянных D_1, D_2, D_3 получаем систему

$$\tfrac{5}{4}\,D_1 + \tfrac{1}{5}\,D_2 + \tfrac{1}{6}\,D_3 = -\,\tfrac{2}{3}\ ,\ \tfrac{1}{5}\,D_1 + \tfrac{13}{6}\,D_2 + \tfrac{1}{7}\,D_3 = \tfrac{9}{4} - \mathrm{e}\,,$$

$$\tfrac{1}{6}\,D_1 + \tfrac{1}{7}\,D_2 + \tfrac{49}{8}\,D_3 = 2\,\mathrm{e} - \tfrac{29}{5}\ ,$$

решая которую, находим $D_1 \approx -\,0,5010, D_2 \approx -\,0,1671, D_3 \approx -\,0,0422$. Точное решение уравнения $y(x) \equiv 1$. Несложно проверить, что абсолютная погрешность не превышает $0,008$. $\rhd$

10.12. Найти приближенное решение интегрального уравнения

$$y(x) = -\,\tfrac{1}{6}\,x - \tfrac{1}{2} + \int\limits_0^1 (1 + 2xs)y(s)\,ds$$

с вырожденным ядром.

У к а з а н и е. Точное решение $y(x) = x + \frac{1}{2}$.

10.3. Проекционные методы

Метод наименьших квадратов. Будем искать приближенное решение $z(x)$ интегрального уравнения (10.1) в виде

$$z(x) = \sum_{j=1}^{N} c_j\varphi_j(x)\,,$$

где $\{\varphi_j(x)\}_{j=1}^N$ — известные линейно независимые функции, которые часто называют *координатными*. Определим невязку уравнения

$$R\,z(x) = z(x) - \lambda\int\limits_a^b K(x,s)z(s)\,ds - f(x)\,. \tag{10.4}$$

Неизвестные постоянные коэффициенты $c_1, \ldots, c_N$ находят из условия минимума интеграла $J = \int\limits_a^b [Rz(x)]^2dx$, т. е. из условий

$$\frac{\partial J}{\partial c_i} = 0, \quad 1 \leqslant i \leqslant N\,.$$

Используя выражение для невязки (10.4), получаем для отыскания коэффициентов систему линейных алгебраических уравнений $A\,\mathbf{c} = \mathbf{f}$

$$a_{ij} = \int\limits_a^b \left[\varphi_j(x) - \lambda \int\limits_a^b K(x,s)\varphi_j(s)\,ds\right]\left[\varphi_i(x) - \lambda \int\limits_a^b K(x,s)\varphi_i(s)\,ds\right] dx\,,$$

$$f_i = \int\limits_a^b f(x)\left[\varphi_i(x) - \lambda \int\limits_a^b K(x,s)\varphi_i(s)\,ds\right] dx\,.$$

Особенность построенной матрицы A — ее симметричность и положительная определенность в случае, если λ не является характеристическим числом интегрального оператора. Это приводит к тому, что алгебраическая система имеет единственное решение и $z(x)$ стремится к $y(x)$ с ростом N.

10.13. Найти методом наименьших квадратов приближенное решение интегрального уравнения

$$y(x) = x + \int\limits_{-1}^{1} xsy(s)\,ds\,,$$

используя в качестве координатных функций $\varphi_1(x) = 1$, $\varphi_2(x) = x$.

$\lhd$ Для отыскания приближенного решения $z(x) = c_1 + c_2 x$ получаем систему

$$\frac{8}{3}\,c_1 - \frac{2}{9}\,c_2 = -\frac{2}{3}\ ,\ -\frac{2}{9}\,c_1 + \frac{2}{27}\,c_2 = \frac{2}{9}\ ,$$

решая которую, находим $c_1 = 0$, $c_2 = 3$, т. е. приближенное решение интегрального уравнения совпадает с точным: $z(x) = y(x) = 3\,x$. $\rhd$

Метод Петрова—Галеркина. Пусть имеются две различные системы линейно независимых функций. Будем искать приближенное решение $z(x)$ интегрального уравнения (10.1) в виде

$$z(x) = f(x) + \sum_{j=1}^{N} c_j\varphi_j(x)\,,$$

где $\{\varphi_j(x)\}_{j=1}^N$ — первая система линейно независимых функций. Коэффициенты $c_1, \ldots, c_N$ находят из условий ортогональности невязки (10.4) каждой функции из второй системы линейно независимых функций $\{\psi_i(x)\}_{i=1}^N$:

$$\int\limits_a^b Rz(x)\,\psi_i(x)\,dx = 0, \quad 1 \leqslant i \leqslant N\,.$$

Эти условия представляют собой систему линейных алгебраических уравнений $A\,\mathbf{c} = \mathbf{f}$ для отыскания неизвестных постоянных коэффициентов,

где

$$a_{ij} = \int\limits_a^b \varphi_j(x)\,\psi_i(x)dx - \lambda \int\limits_a^b \psi_i(x) \int\limits_a^b K(x,s)\varphi_j(s)\,ds\,dx\,,$$

$$f_i = \lambda \int\limits_a^b \psi_i(x) \int\limits_a^b K(x,s) f(s)\,ds\,dx\,.$$

10.14. Найти методом Петрова—Галеркина приближенное решение интегрального уравнения

$$y(x) = 1 + \int\limits_{-1}^{1} (xs + x^2) y(s)\,ds\,,$$

используя в качестве первой системы функций $\varphi_1(x) = x$, $\varphi_2(x) = x^2$, а в качестве второй системы $\psi_1(x) = 1$, $\psi_2(x) = x$.

$\lhd$ Приближенное решение $z(x) = 1 + c_1 x + c_2 x^2$ находим из системы

$$0 \cdot c_1 + \frac{2}{9}\,c_2 = \frac{4}{3}\ ,\ \frac{2}{9}\,c_1 + 0 \cdot c_2 = 0\,.$$

Ее решение имеет вид $c_1 = 0,\ c_2 = 6$. Приближенное решение интегрального уравнения совпадает с точным: $z(x) = y(x) = 1 + 6\,x^2$. $\rhd$

Метод Бубнова—Галеркина. Этот метод является частным случаем метода Петрова—Галеркина, когда обе системы функций совпадают: $\varphi_i(x) = \psi_i(x),\ 1 \leqslant i \leqslant N$. Иногда для приближенного решения удобно использовать представление

$$z(x) = \sum_{j=1}^{N} c_j \varphi_j(x)$$

(без учета правой части $f(x)$), например, при определении характеристических значений интегрального оператора.

10.15. Найти методом Бубнова—Галеркина приближенное решение интегрального уравнения

$$y(x) = 1 + \int\limits_{-1}^{1} (xs + x^2) y(s)\,ds,$$

используя в качестве системы функций $\varphi_1(x) = x$, $\varphi_2(x) = x^2$.

$\lhd$ Приближенное решение $z(x) = 1 + c_1 x + c_2 x^2$ получаем из системы

$$\frac{2}{9}\,c_1 + 0 \cdot c_2 = 0\,,\ 0 \cdot c_1 + \frac{2}{15}\,c_2 = \frac{4}{5}\ ,$$

решая которую, находим $c_1 = 0,\ c_2 = 6$. Приближенное решение интегрального уравнения совпадает с точным: $z(x) = y(x) = 1 + 6\,x^2$. $\rhd$

10.16. Найти методом Бубнова—Галеркина два младших характеристических числа и соответствующие им собственные функции однородного интегрального уравнения

$$y(x) = \lambda \int_0^1 K(x,s) y(s)\, ds\,,$$

где

$$K(x,s) = \begin{cases} x(1-s) & \text{при} \quad 0 \leqslant x \leqslant s \leqslant 1\,, \\ s(1-x) & \text{при} \quad 0 \leqslant s \leqslant x \leqslant 1\,, \end{cases}$$

если координатные функции имеют вид: $\varphi_1(x) = 1$, $\varphi_2(x) = x(1-x)$, $\varphi_3(x) = x(1-x)(1-2x)$.

$\lhd$ Для приближенных собственных функций вида $z(x) = c_1 + c_2 x(1-x) + c_3 x(1-x)(1-2x)$ с пока неизвестными коэффициентами имеем уравнения

$$\int_0^1 \varphi_i(x)\, Rz(x)\, dx = 0, \quad 1 \leqslant i \leqslant 3\,.$$

Вычисляя интегралы, получаем систему

$$c_1 \left(1 - \frac{\lambda}{12}\right) + \frac{c_2}{6} \left(1 - \frac{\lambda}{10}\right) = 0\,,$$

$$\frac{c_1}{6} \left(1 - \frac{\lambda}{10}\right) + \frac{c_2}{30} \left(1 - \frac{17\lambda}{168}\right) = 0\,, \ \frac{c_3}{210} \left(1 - \frac{\lambda}{40}\right) = 0\,.$$

Приравнивая определитель этой системы к нулю, приближенно (с четырьмя знаками после запятой) находим

$$\lambda_1 \approx 9,8751, \quad \lambda_2 \approx 40, \quad \lambda_3 \approx 170,1249.$$

Отсюда, учитывая условие нормировки $\int_0^1 [z(x)]^2\, dx = 1$, определим

$$z_1(x) = -0,0684 + 5,817x(1-x), \quad z_2(x) = 14,49x(1-x)(1-2x).$$

Точное решение задачи имеет вид: $\lambda_k = (k\pi)^2$, $y_k(x) = \sqrt{2}\sin(k\pi x)$, $k = 1, 2, \ldots$. $\rhd$

10.17. Найти методом Бубнова—Галеркина два младших характеристических числа и соответствующие им собственные функции однородного интегрального уравнения

$$y(x) = \lambda \int_0^\pi K(x,s) y(s)\, ds\,,$$

с вырожденным ядром $K(x,s) = \cos^2 x \cos 2s + \cos 3x \cos^3 s$.

Указание. Точное решение: $\lambda_1 = \frac{4}{\pi}$, $y_1(x) = c_1 \cos^2 x$; $\lambda_2 = \frac{8}{\pi}$, $y_2(x) = c_2 \cos 3x$.

10.18. Найти методом Бубнова—Галеркина приближенное решение интегрального уравнения

$$y(x) = 1 + \int\limits_{-1}^{1} (xs + x^2) y(s)\, ds\,,$$

используя в качестве координатной системы функций первые три многочлена Лежандра.

$\lhd$ Подставим приближенное решение

$$z(x) = c_1 + c_2 x + c_3\, \frac{3x^2 - 1}{2}$$

в исходное уравнение. Интегрируя правую часть полученного равенства, имеем

$$c_1 + c_2 x + c_3\, \frac{3x^2 - 1}{2} = 1 + 2xc_1 + \frac{2}{3}\, xc_2\,.$$

Последовательно умножаем равенство на функции 1, x, $\frac{3x^2 - 1}{2}$ и интегрируем по отрезку $[-1, 1]$. В результате получаем систему линейных уравнений

$$2c_1 = 2 + \frac{4}{3}\, c_1\,,\ \frac{2}{3}\, c_2 = \frac{4}{9}\, c_2\,,\ \frac{5}{2}\, c_3 = \frac{8}{15}\, c_1\,,$$

откуда находим $c_1 = 3$, $c_2 = 0$, $c_3 = 4$, т. е. $z(x) = y(x) = 1 + 6\, x^2$. $\rhd$

Метод моментов. Пусть задана ортонормированная система координатных функций $\{\varphi_i(x)\}_{i=1}^{\infty}$; рассмотрим вспомогательные функции

$$u_i(x) = \int\limits_{a}^{b} K(x, s)\varphi_i(s)\, ds\,, \quad i = 1, 2, \ldots\,,$$

и построим вырожденное ядро $H_N(x, s) = \sum\limits_{i=1}^{N} u_i(s)\varphi_i(x)$. Определим приближенное решение методом моментов как точное решение $z(x)$ интегрального уравнения

$$z(x) - \lambda \int\limits_{a}^{b} H_N(x, s) z(s)\, ds = f(x).$$

10.19. Найти методом моментов приближенное решение интегрального уравнения

$$y(x) = 1 + \int\limits_{-1}^{1} (xs + x^2) y(s)\, ds,$$

используя в качестве координатных функций первые два многочлена Лежандра.

Метод коллокации. Будем искать приближенное решение $z(x)$ интегрального уравнения (10.1) в виде

$$z(x) = \sum_{j=1}^{N} c_j \varphi_j(x)\,,$$

где $\{\varphi_j(x)\}_{j=1}^{N}$ — известные линейно независимые функции. Искомые коэффициенты определяются из требования обращения в нуль невязки $Rz(x)$ в заданных точках (точках коллокации) отрезка $[a, b]$.

10.20. Найти методом коллокации с точками $-1, 0, 1$ приближенное решение интегрального уравнения

$$y(x) = 1 + \tfrac{4}{3}\,x + \int\limits_{-1}^{1} (xs^2 - x) y(s)\,ds,$$

используя в качестве координатной системы функций первые три многочлена Лежандра.

$\lhd$ Для отыскания приближенного решения в виде

$$z(x) = c_1 + c_2 x + c_3\,\frac{3x^2 - 1}{2}$$

имеем невязку интегрального уравнения

$$Rz(x) = c_1 \left(1 + \tfrac{4}{3}\,x\right) + c_2 x + c_3 \left(\frac{3x^2 - 1}{2} + \tfrac{7}{3}\,x\right) - 1 - \tfrac{4}{3}\,x\,.$$

Если потребовать, чтобы она обращалась в нуль в точках $-1, 0, 1$, то получим систему линейных уравнений для отыскания неизвестных коэффициентов

$$\tfrac{1}{3}\,c_1 + c_2 + \tfrac{4}{3}\,c_3 = \tfrac{1}{3}\ ,\ c_1 = 1\,,\ \tfrac{7}{3}\,c_1 + c_2 + \tfrac{10}{3}\,c_3 = \tfrac{7}{3}\ ,$$

откуда находим $c_1 = 1$, $c_2 = 0$, $c_3 = 0$, т. е. $z(x) = y(x) \equiv 1$. $\rhd$

10.4. Некорректные задачи

Типичным примером некорректной задачи является интегральное уравнение Фредгольма первого рода $G_b y(x) = f(x)$, т. е.

$$\int\limits_{a}^{b} K(x, s) y(s)\,ds = f(x)\,. \tag{10.5}$$

Под некорректностью здесь понимается либо несуществование решения $y(x)$, либо его неединственность, которая, как правило, приводит к неустойчивости решения относительно возмущения правой части. Эти эффекты проявляются (или не проявляются) в зависимости от свойств ядра и правой части уравнения.

Напомним некоторые полезные свойства ядра $K(x,s)$ интегрального оператора G_b. Пусть ядро $K(x,s)$ вещественное, симметричное и непрерывное. Тогда существует полная, ортонормированная в метрике пространства $L_2(a,b)$, система собственных функций $\{z_i(x)\}_{i=1}^{\infty}$ оператора G_b

$$G_b z_i(x) = \lambda_i z_i(x), \quad (z_i, z_j) = \delta_i^j.$$

При этом само ядро определяется сходящимся рядом

$$K(x,s) = \sum_{i=1}^{\infty} \lambda_i z_i(x) z_i(s) \quad \text{и} \quad \|K(x,s)\|^2 = \sum_{i=1}^{\infty} |\lambda_i|^2,$$

где все характеристические числа λ_i вещественные. В упражнениях 10.21–10.27 по умолчанию речь идет о ядрах с такими свойствами.

10.21. Пусть все характеристические числа G_b отличны от нуля, $f(x)$ — достаточно гладкая вещественная функция. Показать, что решение интегрального уравнения (10.5) существует и оно единственно.

$\lhd$ Будем искать решение (10.5) в виде формального ряда

$$y(x) = \sum_{i=1}^{\infty} y_i z_i(x). \tag{10.6}$$

Так как правая часть гладкая, имеем сходящийся ряд

$$f(x) = \sum_{i=1}^{\infty} f_i z_i(x), \quad f_i = \int_a^b f(x) z_i(x)\, dx\,.$$

Подставляя в исходное уравнение предполагаемое решение (10.6) и разложения по собственным функциям для ядра и правой части, в силу ортонормированности системы $\{z_i(x)\}_{i=1}^{\infty}$, получаем

$$y_i = \frac{f_i}{\lambda_i}\,, \quad i = 1, 2, \ldots\,.$$

Таким образом, сходимость функционального ряда (10.6) определяется сходимостью числового ряда

$$S = \sum_{i=1}^{\infty} |y_i|^2 = \sum_{i=1}^{\infty} \left| \frac{f_i}{\lambda_i} \right|^2 .$$

Будем считать, что неявно наложенное условие гладкости на правую часть $f(x)$ приводит к неравенству $S < \infty$. Тогда решение поставленной задачи существует и оно единственное. $\rhd$

10.22. Пусть характеристические числа G_b обладают следующим свойством: $\lambda_i \neq 0$ при $1 \leqslant i \leqslant p$ и $\lambda_i = 0$ при $i > p$. Показать, что при сколь угодно гладкой правой части $f(x)$ решение интегрального уравнения (10.5) может не существовать, а если оно существует, то не является единственным.

$\lhd$ При указанных в условии свойствах характеристических чисел ядро интегрального оператора имеет вид

$$K(x,s)=\sum_{i=1}^{p}\lambda_i z_i(x)z_i(s)\,,$$

т. е. является вырожденным. Поэтому на основании раздела 10.2 решение уравнения (10.5) имеет вид конечной суммы

$$y(x)=\sum_{i=1}^{p}y_i z_i(x)$$

с некоторыми коэффициентами y_i. Пусть правая часть уравнения (10.5) имеет вид $f(x)=f_k z_k(x),\ f_k\neq 0$, при некотором $k>p$. Тогда равенство $G_b y(x)=f(x)$, т. е.

$$\int\limits_a^b\left(\sum_{j=1}^{p}\lambda_j z_j(x)z_j(s)\right)\left(\sum_{i=1}^{p}y_i z_i(s)\right)ds=f_k z_k(x),$$

противоречит свойству ортонормированности системы собственных функций $\{z_i(x)\}_{i=1}^{\infty}$ интегрального оператора G_b. Поэтому решение задачи $y(x)$ существует только для правых частей, представимых в виде

$$f(x)=\sum_{i=1}^{p}f_i z_i(x).$$

При этом, в силу 10.21, неизвестные коэффициенты в решении определяются равенствами $y_i=\dfrac{f_i}{\lambda_i},\ i=1,2,\ldots,p$.

Теперь, чтобы установить неединственность решения для таких правых частей, достаточно проверить, что произвольная функция

$$y(x)=\sum_{i=1}^{p}\frac{f_i}{\lambda_i}\,z_i(x)+\sum_{i=p+1}^{\infty}c_i z_i(x)$$

с коэффициентами, удовлетворяющими условию $\sum\limits_{i=p+1}^{\infty}|c_i|^2<\infty$, также является решением исходного интегрального уравнения. $\rhd$

Для решения некорректных задач используют различные способы *регуляризации*. Например, в качестве приближенного решения y уравнения Фредгольма первого рода (10.5) рассматривают некоторое решение y_α корректного уравнения Фредгольма второго рода (10.1), зависящее от параметра регуляризации $\alpha>0$.

Простейшая регуляризация. Этот способ заключается в том, что исходное уравнение $G_b y=f$ заменяют возмущенным уравнением $\alpha y_\alpha+G_b y_\alpha=f_\delta$, где f_δ возникает в результате незнания точной правой части f: $\|f_\delta-f\|\leqslant\delta$.

Метод регуляризации Тихонова. Приближенное решение y_α находится из условия минимума функционала

$$J(z) = \|G_b z - f_\delta\|^2 + \alpha \|z\|^2,$$

что приводит к уравнению Эйлера

$$\alpha y_\alpha + G_b^* G_b y_\alpha = G_b^* f_\delta \,,$$

где $G_b^*\varphi(x) = \int\limits_a^b K(s,x)\varphi(s)\,ds$ — сопряженный интегральный оператор.

Основным моментом в методах регуляризации является выбор параметра α и его согласование с величиной погрешности входных данных δ: необходимо, чтобы при $\delta \to 0$ выполнялось $y_\alpha \to y$.

10.23. Пусть для решения y уравнения (10.5) с оператором $G_b = G_b^* > 0$ справедливо неравенство

$$\|G_b^{-1} y\| \leqslant M = \text{const}.$$

Показать, что метод простейшей регуляризации сходится ($y_\alpha \to y$) при $\alpha = \sqrt{\dfrac{\delta}{M}}$, $\delta \to 0$.

$\lhd$ Запишем решения точного и регуляризованного уравнений в виде разложений по собственным функциям интегрального оператора (см. 10.21):

$$y = \sum_{i=1}^{\infty} \frac{(f, z_i)}{\lambda_i}\, z_i(x), \quad y_\alpha = \sum_{i=1}^{\infty} \frac{(f_\delta, z_i)}{\alpha + \lambda_i}\, z_i(x).$$

Обозначим через u решение регуляризованного уравнения с точной правой частью

$$\alpha u + G_b u = f, \quad u = \sum_{i=1}^{\infty} \frac{(f, z_i)}{\alpha + \lambda_i}\, z_i(x).$$

Это решение потребуется для получения оценки по неравенству треугольника

$$\|y_\alpha - y\| \leqslant \|y_\alpha - u\| + \|u - y\| \,.$$

В силу ортонормированности системы собственных функций $\{z_i(x)\}_{i=1}^{\infty}$, для оценки слагаемых в правой части можно использовать равенство Парсеваля

$$\|y_\alpha - u\|^2 = \sum_{i=1}^{\infty} \frac{(f_\delta - f, z_i)^2}{(\alpha + \lambda_i)^2} \leqslant \frac{\delta^2}{\alpha^2} \,,$$

$$\|u - y\|^2 = \alpha^2 \sum_{i=1}^{\infty} \frac{(f, z_i)^2}{(\alpha + \lambda_i)^2 \lambda_i^2} \leqslant \alpha^2 \sum_{i=1}^{\infty} \frac{(f, z_i)^2}{\lambda_i^4} \leqslant \alpha^2 M^2.$$

В последней оценке использовано неравенство из условия задачи. Из полученных выражений следует

$$\|y_\alpha - y\| \leqslant \frac{\delta}{\alpha} + \alpha\, M.$$

Минимум правой части выражения достигается при $\alpha = \sqrt{\frac{\delta}{M}}$, что приводит к неравенству $\|y_\alpha - y\| \leqslant 2\sqrt{\delta\, M}$. Эта оценка означает, что $y_\alpha \to y$ при $\delta \to 0$. $\triangleright$

10.24. Записать для (10.5) регуляризованное интегральное уравнение в методе Тихонова.

Ответ: уравнение имеет вид $\alpha y_\alpha + Q_b y_\alpha = \varphi$, где $\varphi(x) = \int\limits_a^b K(s,x) f(s)\, ds$; ядро $H(x,s)$ в новом интегральном операторе $Q_b = G_b^* G_b$ определяется через исходное $K(x,s)$ по формуле

$$H(x,s) = \int\limits_a^b K(t,x) K(t,s)\, dt$$

и является симметричным.

10.25. Пусть в задаче определения характеристических чисел для уравнения $y(x) = \lambda G_b y(x)$ ядро $K(x,s)$ интегрального оператора симметрично и положительно определено и по заданной начальной функции $\varphi_0(x)$ строится последовательность функций $\varphi_i(x)$

$$\varphi_{i+1}(x) = G_b \varphi_i(x), \quad i = 0, 1, \ldots .$$

Показать, что для минимального характеристического числа $\lambda_{\min}$ можно получить приближения

$$\lambda_{\min} \approx \frac{\|\varphi_i\|}{\|\varphi_{i+1}\|}, \quad \lambda_{\min} \approx \|\varphi_i\|^{-1/i}.$$

10.26. Пусть в задаче определения характеристических чисел уравнения $y(x) = \lambda G_b y(x)$ с симметричным ядром найдены числа S_i (i-е следы ядра $K(x,s)$). Показать, что для минимального характеристического числа $\lambda_{\min}$ при больших i имеется приближение

$$|\lambda_{\min}| \approx \sqrt{\frac{S_{2i}}{S_{2i+2}}}.$$

Указание. След четного порядка для симметричного ядра определяется по формуле

$$S_{2i} = \int\limits_a^b \int\limits_a^b K_i^2(x,s)\, dx ds,$$

где $K_1 = K(x,s)$, $K_i(x,s) = \int\limits_a^b K(x,t)\, K_{i-1}(t,s)\, dt$, $i = 2, 3, \ldots$.

10.27. Пусть уравнение $G_b y(x) = f(x)$ имеет единственное решение и ядро $K(x,s)$ симметрично и положительно определено. Показать, что метод простой итерации

$$\frac{y^{k+1}(x) - y^k(x)}{\tau} + \int_a^b K(x,s)\, y^k(s)\, ds = f(x)$$

сходится при $0 < \tau < \dfrac{2}{\lambda_{\max}}$, где $\lambda_{\max}$ — максимальное характеристическое число.

10.28. Вычислить методом из (10.25) наименьшее характеристическое число ядра $K(x,s) = xs$, $0 \leqslant x, s \leqslant 1$, начиная с $\varphi_0(x) \equiv 1$.

Ответ: $\lambda_{\min} = 3$.

Литература

1. *Арушанян И. О., Чижонков Е. В.* Материалы семинарских занятий по курсу «Методы вычислений» / под ред. О. Б. Арушаняна: Учеб. пособие. — 2-е изд. — М. : Изд-во ЦПИ при механико-математическом факультете МГУ, 1999 (1-е изд., 1998).
2. *Бабенко К. И.* Основы численного анализа. — М. : Наука, 1986.
3. *Бахвалов Н. С.* Численные методы. — М. : Наука, 1975.
4. *Бахвалов Н. С., Лапин А. В., Чижонков Е. В.* Численные методы в задачах и упражнениях. Учеб. пособие. / Под ред. В. А. Садовничего — 2-е издание; перераб. и доп. — М. : БИНОМ. Лаборатория знаний, 2010.
5. *Бахвалов Н. С., Жидков Н. П., Кобельков Г. М.* Численные методы. — 6-е издание — М. : БИНОМ. Лаборатория знаний, 2008.
6. *Верлань А. Ф., Сизиков В. С.* Интегральные уравнения: методы, алгоритмы, программы. Справочное пособие. — Киев : Наукова думка, 1986.
7. *Гельфонд А. О.* Исчисление конечных разностей. — М. : Наука, 1967.
8. *Годунов С. К., Рябенький В. С.* Разностные схемы. — М. : Наука, 1977.
9. *Деммель Дж.* Вычислительная линейная алгебра. — М. : Мир, 2001.
10. *Дробышевич В. И., Дымников В. П., Ривин Г. С.* Задачи по вычислительной математике. — М. : Наука, 1980.
11. *Икрамов Х. Д.* Несимметричная проблема собственных значений. — М. : Наука, 1991.
12. *Корнев А. А., Чижонков Е. В.* Упражнения по численным методам. Части I, II / под ред. Н. С. Бахвалова — М. : Изд-во ЦПИ при механико-математическом факультете МГУ, 2002, 2003.
13. *Кунц К. С.* Численный анализ. — Киев : ТЕХНІКА, 1964.
14. *Марчук Г. И.* Методы вычислительной математики. — М. : Наука, 1980.
15. *Парлетт Б.* Симметричная проблема собственных значений. — М. : Мир, 1983.
16. *Попов А. В.* Практикум на ЭВМ. Разностные методы решения квазилинейных уравнений первого порядка. — М. : Изд-во ЦПИ при механико-математическом ф-те МГУ, 1999.
17. *Самарский А. А.* Теория разностных схем. — М. : Наука, 1983.
18. *Самарский А. А., Вабищевич П. Н., Самарская Е. А.* Задачи и упражнения по численным методам. — М. : Эдиториал УРСС, 2000.

19. *Самарский А. А., Гулин А. В.* Численные методы. — М. : Наука, 1989.

20. *Самарский А. А., Карамзин Ю. Н.* Разностные уравнения. — М. : Знание, 1978.

21. *Самарский А. А., Николаев Е. С.* Методы решения сеточных уравнений. — М. : Наука, 1978.

22. *Трауб Дж.* Итерационные методы решения уравнений. — М. : Мир, 1985.

23. *Тыртышников Е. Е.* Методы численного анализа. — М. : Издательский центр «Академия», 2007.

24. *Хемминг Р. В.* Численные методы. — М. : Наука, 1968.

25. *Хорн Р., Джонсон Ч.* Матричный анализ. — М. : Мир, 1989.

26. *Шокин Ю. И.* Методы дифференциального приближения. — Новосибирск : Наука, 1979.

Минимальные системные требования определяются соответствующими требованиями программы Adobe Reader версии не ниже 11-й для платформ Windows, Mac OS, Android, iOS, Windows Phone и BlackBerry; экран 10"

Учебное электронное издание

Серия: «Классический университетский учебник»

Бахвалов Николай Сергеевич, **Корнев** Андрей Алексеевич,
Чижонков Евгений Владимирович

ЧИСЛЕННЫЕ МЕТОДЫ.
РЕШЕНИЯ ЗАДАЧ И УПРАЖНЕНИЯ

Учебное пособие для вузов

Ведущий редактор *М. С. Стригунова*. Художественный редактор *В. Е. Шкерин*
Оригинал-макет подготовлен *О. Г. Лапко* в пакете LaTeX 2_ε

Подписано к использованию 11.08.16.
Формат 145×225 мм

Подготовлено при участии
ООО «Лаборатория Базовых Знаний»
129110, Москва, ул. Гиляровского, д. 54, стр. 1

Издательство «Лаборатория знаний»
125167, Москва, проезд Аэропорта, д. 3
Телефон: (499) 157-5272, e-mail: info@pilotLZ.ru, http://www.pilotLZ.ru

Московский государственный университет имени М.В. Ломоносова

КЛАССИЧЕСКИЙ УНИВЕРСИТЕТСКИЙ УЧЕБНИК

Основная парадигма вычислительной математики гласит: «Цель расчетов – понимание, а не числа». Это означает, что ни гигантские хранилища информации, ни вычислительная мощь современных суперкомпьютеров не в состоянии заменить интеллектуальные возможности математика-исследователя. Развитие цивилизации ставит перед обществом проблемы, решение которых вынуждает не только использовать все уже накопленные знания, но и интенсивно раздвигать научные горизонты. Изучение нелинейностей окружающего мира приводит к естественной математизации других, в том числе гуманитарных наук, что проявляется в построении и анализе математических моделей численными и аналитическими методами.

Данное пособие написано на основе многолетнего опыта преподавания численных методов в МГУ им. М. В. Ломоносова. Содержание пособия тесно связано с классическим учебником Н. С. Бахвалова, Н. П. Жидкова, Г. М. Кобелькова «Численные методы» и книгой Н. С. Бахвалова, А. В. Лапина, Е. В. Чижонкова «Численные методы в задачах и упражнениях».