МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
КРАСНОЯРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Б. С. Добронец

ИНТЕРВАЛЬНАЯ МАТЕМАТИКА

Допущено советом по прикладной математике и информатике УМО по классическому университетскому образованию для студентов высших учебных заведений, обучающихся по специальности 010200 “Прикладная математика и информатика” и направлению 510200 “Прикладная математика и информатика”

Красноярск 2004

УДК 519
ББК 22.172.3Я73
Д 56

Рецензенты:
доктор физико-математических наук, профессор М. В. Носков
доктор физико-математических наук, профессор В. И. Быков

Добронец Б. С.
Д 56 Интервальная математика: Учеб. пособие / Б. С. Добронец; Краснояр. гос. ун-т. — Красноярск, 2004. — 216 с.
ISBN 5–76–38–0529–1

В пособии рассмотрены вычислительные аспекты получения гарантированных оценок погрешности приближенных решений. Особое внимание уделено сочетанию интервальной математики и апостериорных оценок.

Пособие предназначено для аспирантов и студентов-математиков 4–5 курсов, специализирующихся в области прикладной математики и численного анализа.

ISBN 5–76–38–0529–1

Введение

В связи с развитием таких направлений науки и техники, как механика, теплотехника, математическая химия, самолетостроение, возникла потребность не только вычисления приближенных решений различных задач, но и гарантированных оценок их близости к точным решениям. Поэтому интерес к интервальному анализу и вопросам двусторонних оценок как к возможным средствам оценки погрешностей приближенных решений в последнее время нарастает. Интервальный анализ появился сравнительно недавно как метод автоматического контроля ошибок округления на ЭВМ. Впоследствии он превратился в один из разделов вычислительной математики, учитывающий также ошибки дискретизации численных методов, ошибки в начальных данных и т.п. Основная идея интервального анализа состоит в замене арифметических операций и вещественных функций над вещественными числами интервальными операциями и функциями, преобразующими интервалы, содержащие эти числа. Ценность интервальных решений заключается в том, что они содержат точные решения исходных задач.

Двусторонние методы численного анализа известны раньше интервальных, и их аппарат до последнего времени не использовал понятия интервального анализа. Для получения двусторонних оценок применяются различные приемы и методы, в частности операторные неравенства, апостериорные и априорные оценки погрешности, оценки остаточных членов и т. п. Следует отметить, что двусторонние методы обычно несколько проще в реализации, чем интервальные, хотя и обладают некоторыми недостатками. В частности, в них используются, как правило, точные входные данные, не учитывается влияние погрешностей, связанных с применением ЭВМ, некоторые приемы гарантируют включение точного решения только в асимптотике, т. е. при достаточно малых шагах сетки. Однако двусторонние методы довольно просто обобщаются с обыкновенных дифференциальных уравнений на уравнения эллиптического и параболического типа. Применение интервальн-

ного анализа к уравнениям в частных производных встречает большие трудности, а результаты его часто неудовлетворительны.

Стремление объединить положительные стороны двусторонних и интервальных методов анализа послужило одной из целей написания этой работы. Такой положительной стороной интервального анализа является возможность полного учета погрешностей, начиная с неточных данных математической модели и кончая ошибками округления на ЭВМ.

Интервальный анализ представляет собой относительно молодое и интенсивно развивающееся направление математики. Первая монография, посвященная интервальному анализу, была опубликована Р.Е. Муром в 1966 г. [89], а на русском языке — Ю.И. Шокиным в 1981 г. [67]. Затем в 1982 г. издано учебное пособие Т.И. Назаренко, Л.В. Марченко [54] по интервальным методам, а в 1986 г. — монография С.А. Калмыкова, Ю.И. Шокина, З.Х. Юлдашева [39]. Обширная и подробная библиография по интервальному анализу имеется в [3, 39, 73, 90].

Первоначально интервальные алгоритмы строились как непосредственные обобщения вещественных алгоритмов. Затем все чаще стали появляться специфические алгоритмы и дополнительные операции (такие, как пересечение).

К настоящему времени разработаны приемы интервальных вычислений [3, 39, 54, 90, 92] и несколько пакетов прикладных программ и алгоритмических макроязыков, реализующих элементы интервального анализа на машинном уровне для нескольких типов ЭВМ [1, 41]. Вместе с тем для сколько-нибудь сложных задач полное применение интервального анализа часто дает неудовлетворительные результаты из-за чрезмерных длин получаемых интервалов. Дело во внутренней установке, которую мы будем называть пессимистическим подходом. Она свойственна не только методам интервального анализа, но и некоторым априорным способам оценки погрешности [7, 15, 51, 59] и заключается в прослеживании на каждой элементарной операции всевозможных, в том числе наихудших, сочетаний погрешностей. Ясно, что при обычном ходе вычислений ошибки могут усредняться, компенсироваться и накапливаться далеко не худшим образом. В конечном итоге пессимистические оценки точности на порядок хуже, чем она есть на самом деле.

Вместе с тем статистические [14] и другие регулярные подходы [6, 7] к моделированию погрешностей дают в целом неплохое качественное представление о поведении ошибки, но не влекут гарантированных оценок для конкретных приближенных решений.

Для построения итерационных процессов используется принцип сжимающих отображений или более общий подход, основанный на теореме Шаудера о неподвижной точке [42, 44]. Итерационный метод для уточнения границ интервального решения с построением специальной матрицы перехода изложен в работе Д.М. Гея [82].

Среди прямых методов наибольшее внимание уделено интервальному обобщению метода Гаусса [3, 39]. Он наиболее эффективен для частного класса систем с M-матрицами [120]. В монографии [39] изучен также интервальный метод прогонки для систем с трехдиагональными матрицами.

Наиболее полно исследования прямых и итерационных методов решения систем линейных и нелинейных алгебраических уравнений с интервальными коэффициентами представлены в книге Г. Алефельда, Ю. Херцбергера [3].

Современное состояние двусторонних методов в исследовании применительно к линейным и нелинейным алгебраическим задачам наиболее полно изложено в монографии Н.С. Курпеля, Б.А. Шувара [44]. Там же освещены вопросы двусторонних итерационных методов решения абстрактных операторных задач и интегральных уравнений Фредгольма и Вольтерра второго рода. По существу, результаты решения интегральных уравнений являются модификацией алгоритмов решения абстрактных операторных уравнений с преодолением специфических трудностей обоснования на функциональном уровне. Поэтому мы не будем специально останавливаться на этой ветви исследований двусторонних методов решения интегральных уравнений.

При вычислении интегралов с помощью квадратурных формул вычислители начали сталкиваться с явлением дискретизации, ясно осознанным позднее при решении дифференциальных задач. Среди прочих видов погрешностей ошибки дискретизации в ряде случаев играют доминирующую роль. Поэтому на протяжении последних десятилетий при оценке погрешности приближенных решений им уделяется пристальное внимание.

Одно из первых правил для практической оценки погрешности дискретизации, позволяющее примерно оценить влияние этой погрешности, в начале прошлого века предложил К.Д. Рунге. Это правило интенсивно использовалось сначала в области квадратур, а затем в разностных методах и методе конечных элементов. Оно основано на разложении

приближенного решения в виде суммы [53]

$$u^h = u + h^k v + O(h^{k+m}), \tag{0.1}$$

где u — искомое точное решение, v — неизвестная функция, а h — малый параметр дискретизации, чаще всего шаг разностной сетки. Целое k характеризует порядок точности приближенного решения, а $m > 0$ — малость остаточного члена в сравнении с главным членом погрешности $h^k v$. Поскольку u и v не зависят от h, для параметра $h/2$ справедливо разложение

$$u^{h/2} = u + \left(\frac{h}{2}\right)^k v + O(h^{k+m}).$$

Вычтем его из (0.1), избавляясь от u:

$$u^h - u^{h/2} = v\left(\frac{h}{2}\right)^k (2^k - 1) + O(h^{k+m}).$$

Отсюда можно определить главный член погрешности:

$$u^{h/2} - u \approx \frac{u^h - u^{h/2}}{2^k - 1}. \tag{0.2}$$

Поскольку в формуле (0.2) отброшен остаточный член порядка $O(h^k)$, она не приводит к гарантированной оценке, но при достаточно малых h действительно дает представление о величине погрешности численного решения.

Еще одним важным приемом выяснения точности приближенных решений служат априорные оценки, особенно активно используемые в теории разностных схем и методе конечных элементов [6, 7, 8, 17, 11, 25, 35, 52, 59, 63, 104]. Как правило, они опираются на старшие производные точного решения и дают представление о порядке точности при измельчении сетки, что очень важно при выборе разностной схемы, конечных элементов или при оценке затрат на ЭВМ. Но попытки выяснения гарантированных констант для конкретного приближенного решения через известные данные обычно приводят к весьма грубым результатам, наряду с большими аналитическими и теоретическими трудностями.

Перейдем к двусторонним методам решения задачи Коши для обыкновенных дифференциальных уравнений и систем первого порядка. Один из первых методов решения этой задачи для одного уравнения — метод С.А. Чаплыгина — появился в 1919 г. [64]. Этот аналитический

метод позволяет итерационно строить последовательность двусторонних решений, сжимающихся к точному. Однако метод С.А. Чаплыгина использовал знакопостоянство некоторых производных от правой части и по этой причине не распространялся непосредственно на системы уравнений.

Для систем уравнений оценки уклонения точного решения от приближенного были впервые получены в работах С.М. Лозинского [48, 49] с помощью решения вспомогательных задач.

Двусторонние методы, рассмотренные в работах А. Д. Горбунова, Ю. А. Шахова [26], Е. Я. Ремеза [56, 57], Н. П. Салихова [58], основаны на построении двух численных методов интегрирования, остаточные члены которых имеют разные знаки. Поэтому полученные с помощью этих методов численные решения могут служить границами двустороннего решения.

В работе Х. Бауха [71] изложен итерационный метод построения аналитического интервального решения, основанный на оценке невязки.

В интервальном анализе следует выделить работы С. А. Калмыкова, Ю. И. Шокина, З. Х. Юлдашева [39]. В них приведены интервальные методы типа Рунге-Кутта, Адамса, основанные на получении интервальной функции, содержащей остаточный член погрешности, которая строится через известные производные правой части.

Работа Е.А. Волкова [18] посвящена оценкам погрешности решения краевых задач для обыкновенных дифференциальных уравнений второго порядка. Основной подход состоит в оценке погрешности аппроксимации с привлечением априорных оценок высших производных от точного решения.

Другой подход используется Е.А. Волковым в [19]. Для оценки погрешности имеющееся разностное решение интерполируется кубическими сплайнами. Затем уклонение сплайнового решения оценивается через максимум невязки и коэффициенты уравнения. В работе показано, что ширина полосы стремится к нулю со вторым порядком от шага сетки.

Е. Хансен [84] рассматривает интервальный аналог конечно-разностного метода для решения нелинейной краевой задачи. Используя априорные оценки старших производных точного решения, построена система интервальных уравнений, решение которой содержит точное решение исходной задачи. Ф.А. Оливейра [96] и И. Шредер [100] предложили аналогичные интервальные методы решения краевых задач для

обыкновенных дифференциальных уравнений. Ю.И. Шокин [67] описал метод повышенного порядка точности, использующий метод Галеркина с кусочно-кубическими функциями.

При исследовании уравнений в частных производных доминируют два подхода. Первый основан на оценках остаточных членов разностной схемы. Эти оценки требуют знания некоторой априорной информации о точном решении и его производных [17]. Второй подход основан на теории операторов монотонного типа [42] и теоремах сравнения. В. Аппельт [69] и И.Ф. Крюкеберг [86] такое построение провели с использованием невязки приближенного аналитического решения. Позднее в рамках этого подхода стали развиваться методы, основанные на апостериорных оценках погрешности разностных решений. Одна из первых таких работ Д.Ф. Давиденко [28] была опубликована в 1960 г. В ней разностное решение задачи Дирихле для уравнения Пуассона сначала интерполировалось полиномом, а затем по его невязке оценивалась погрешность разностного решения. Однако количество узлов сетки было ограничено единицами. Е.А. Волков [21] этот недостаток устранил. Он использовал кусочно-локальное продолжение сеточной функции, показал стремление ширины построенного коридора для погрешности к нулю со вторым порядком относительно шага сетки.

Интервальные методы решения параболических уравнений представлены в работах Е. Адамса, Х. Шпреера [68], Е. Фааса [81]. В них идет поиск полиномиальной аппроксимации разностного решения, а затем с помощью теорем сравнения оценивается уклонение приближенного решения от точного.

Список основных обозначений

Области и производные

R^n — евклидово пространство размерности n с точками $x = (x_1, ..., x_n)$.

Ω — ограниченная область в R^n, $\partial\Omega$ — ее граница, $\bar{\Omega} = \Omega \cup \partial\Omega$.

Q — открытый цилиндр, $\Omega \times (0, T)$, S —его боковая поверхность, $\overline{Q}$ — его замыкание.

$\partial^\alpha u(x) = \frac{\partial^{|\alpha|} u(x)}{\partial x_1^{\alpha_1} ... \partial x_n^{\alpha_n}}$ —классические и обобщенные производные, $\alpha = (\alpha_1, ..., \alpha_n)$ — мультииндекс с целыми $\alpha_j \geq 0$ и $|\alpha| = \alpha_1 + ... + \alpha_n$. Кроме того,

$$\partial_1 u(x) \equiv \frac{\partial u}{\partial x_1}(x), \quad \partial_2 u(x) \equiv \frac{\partial u}{\partial x_2}(x), \quad \partial_{12} u(x) \equiv \frac{\partial^2 u(x)}{\partial x_1 \partial x_2}, ...,$$

а в случае $x, t \in R$

$$\partial_x u(x,t) \equiv \frac{\partial u}{\partial x}(x,t), \quad \partial_t u(x,t) \equiv \frac{\partial u}{\partial t}(x,t), ...$$

Для обыкновенных производных в случае $x \in R$ и целого $k \geq 0$

$$\partial^k u(x) \equiv \frac{d^k u(x)}{dx^k}.$$

v — единичный вектор внешней нормали к $\partial\Omega$. Для производной в направлении внешней нормали используется обозначение $\partial_v u$.

Интервальные числа и операции

$\boldsymbol{a}$ — интервальное число $[\underline{a}, \overline{a}]$ с границами $\underline{a} \leq \bar{a}, \underline{a}, \bar{a} \in R$.

$|\boldsymbol{a}| = \max(|\underline{a}|, |\bar{a}|), \mathrm{med}\boldsymbol{a} = (\underline{a} + \bar{a})/2, \mathrm{wid}\boldsymbol{a} = \bar{a} - \underline{a}$.

$b = (b_i) = (b_1, ..., b_n)$ — вектор с элементами $b_i, i = 1, ..., n$.

$\boldsymbol{b} = (\boldsymbol{b}_i) = (\boldsymbol{b}_1, ..., \boldsymbol{b}_n)$ — интервальный вектор с элементами $\boldsymbol{b}_i, i = 1, ..., n, |\boldsymbol{b}| = (|\boldsymbol{b}_i|), \mathrm{med}\boldsymbol{b} = (\mathrm{med}\boldsymbol{b}_i), \mathrm{wid}\boldsymbol{b} = (\mathrm{wid}\boldsymbol{b}_i)$.

$A = (a_{ij})$ — матрица $n \times n$ с элементами $a_{ij}, i, j = 1, ..., n$.

$|\boldsymbol{A}| = (|\boldsymbol{a}_{ij}|$ — интервальная $n \times n$ матрица с элементами $\boldsymbol{a}_{ij}, i, j = 1, ..., n, |\boldsymbol{A}| = (|\boldsymbol{a}_{ij}|), \mathrm{med}\boldsymbol{A} = (\mathrm{med}\boldsymbol{a}_{ij}), \mathrm{wid} = (\mathrm{wid}\boldsymbol{a}_{ij})$.

$\mathcal{A}$ — подмножество матриц $R^{n \times n}$.

$\square S$ — наименьший по включению интервальный вектор, содержащий S.

Для произвольной функции f: $\Omega \to R$ обозначим

$$\operatorname*{int}_{\Omega} f = \left[\inf_{x\in\Omega} f(x), \sup_{x\in\Omega} f(x)\right],$$

$$f_+(x) = \max\{0, f(x)\}, f_-(x) = \max\{0, -f(x)\},$$

т.е. $f = f_+ - f_-$.

$\boldsymbol{f}_{un}(\boldsymbol{x})$ — объединенное расширение функции $f(x)$, т. е. $\boldsymbol{f}_{un}(\boldsymbol{x}) = \operatorname*{int}_{x\in\boldsymbol{x}} f(x)$.

$\boldsymbol{f}_{ne}(\boldsymbol{x})$ — естественное расширение функции $f(x)$.

$\boldsymbol{f}_{mv}(\boldsymbol{x})$ — mv-форма.

Пространства и нормы

$\boldsymbol{R}$ — множество всех интервальных чисел.

R_+ — множество всех неотрицательных чисел.

$\boldsymbol{R}^n$ — пространство всех интервальных векторов размерности n с элементами $\boldsymbol{b} = (\boldsymbol{b}_1, ..., \boldsymbol{b}_n), \boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$.

$R^{n\times n}$ — множество всех вещественных $n \times n$-матриц.

S_m^k — множество сплайнов степени m дефекта k.

$\mathcal{P}_m(\Omega)$ — множество многочленов полной степени m, определенных на Ω.

Для произвольных функций: f: $D \to R$, в том числе сеточных, на множестве $D \subset R^n$ положим

$$||f||_{\infty,D} = \sup_{x\in D} |f(x)|.$$

$L_p(\Omega)$— пространство измеримых на Ω функций с конечной нормой

$$||f||_{\infty,\Omega} = \operatorname{vrai}\sup_{\Omega} |f|, \quad p = \infty;$$

$$||f||_{p,\Omega} = \left(\int_{\Omega} |f|^p dx\right), \quad 1 \le p < \infty.$$

$L_2(\Omega)$ — гильбертово пространство, со скалярным произведением $(u, v) = \int_{\Omega} uvdx$.

$W_p^m(\Omega)(1 \le p \le \infty$, целое $m \ge 0)$ — пространство Соболева, состоящее из функций $u \in L_p(\Omega)$, имеющих обобщенные производные $\partial^\alpha u \in L_p(\Omega) \forall |\alpha| \le m$. Норма вводится равенством

$$||u||_{m,p,\Omega} = \left(\sum_{|\alpha|\le m} ||\partial^\alpha u||_{p,\Omega}^p\right)^{1/p}, \quad 1 \le p < \infty;$$

$$||u||_{m,\infty,\Omega} = \sum_{|\alpha|\le m} ||\partial^\alpha u||_{\infty,\Omega}, \quad p = \infty.$$

$\overset{\circ}{W}{}_p^m(\Omega)$ — подпространство $W_p^m(\Omega)$ являющееся замыканием в норме $||\cdot||_{m,p,\Omega}$ всех бесконечно дифференцируемых функций с носителем в Ω.

$C^m(\bar{\Omega})$ — (целое $m \ge 0$) банахово пространство функций, имеющих непрерывные на $\bar{\Omega}$ производные $\partial^\alpha u \forall |\alpha| \le m$. Норма в нем совпадает с $||\cdot||_{m,\infty,\Omega}$.

$C^{m+\gamma}(\bar{\Omega})$ — (целое $m \ge 0, 0 < \gamma < 1$) банахово пространство функций, имеющих непрерывные на Ω производные (по Гельдеру с показателем γ) $\partial^\alpha u \forall |\alpha| \le m$. Норма в нем:

$$||u||_{m+\gamma,\infty,\Omega} = ||u||_{m,\infty,\Omega} + \sum_{|\alpha|=m} \sup_{x,x'\in\Omega} \frac{|\partial^\alpha u(x) - \partial^\alpha u(x')|^\gamma}{|x - x'|}.$$

$||A||_p = \sup\limits_{||x||_p=1} ||Ax||_p$ — норма матрицы $A \in R^{n\times n}, 1 \le p \le \infty$.

$||x||_p = \left(\sum\limits_{i=1}^n x_i^p\right)^{1/p}$ — векторная норма, $1 \le p < \infty$.

$||x||_\infty = \max\limits_{1\le i\le n} |x_i|$ — равномерная норма векторов из R^n.

I — тождественный оператор.

δ_{ij} —- символ Кронекера.

Сеточные области и разностные отношения

$\omega = \{x_i | a = x_0 < x_1 < ... < x_N = b\}$— сетка (вообще говоря, неравномерная) на отрезке $[a, b]$.

$\bar{\omega}_h = \{x_i = ih, i = 0, 1, ..., n\}$— равномерная сетка на отрезке $[0, 1]$ с шагом $h = 1/n$.

Ω_h — сеточная область.

Глава 1

Вспомогательные сведения

1.1. Теоремы сравнения

Цель этого раздела — изучение свойств операторов в банаховых пространствах и доказательство теорем сравнения для решения операторных уравнений.

Пусть B — вещественное банахово пространство с нормой $\|\cdot\|$, а $R_+ = \{x \in R | x \geq 0\}$.

Замкнутое множество $K \subset B$ называется *положительным конусом*, если выполняются следующие условия:

a) из $u, v \in K$ следует, что $\alpha u + \beta v \in K \quad \forall \alpha, \beta \in R_+$;

b) если $u \neq 0$, то по крайней мере один из элементов (u или $-u$) не принадлежит K.

В соответствии с конусом K введем частичную упорядоченность следующим образом: соотношение $u \leq v$ означает, что $v - u \in K$. Тогда B становится полуупорядоченным линейным пространством [44]. Если задано отношение $\geq$, то

$$u \leq v \Longleftrightarrow u = v \text{ или } u < v.$$

В частности, для отношения $\geq$ справедливы свойства:

- из $u \geq v, v \geq w$ следует $u \geq w$ (транзитивность);
- из $u \geq v$, $v \geq u$ следует $u = v$ (антисимметричность);
- $u \geq u$ (рефлексивность);
- из $u \geq 0$ следует $\lambda u \geq 0 \quad \forall \lambda \in R_+$;
- из $u_1 \geq v_1, u_2 \geq v_2$ следует $u_1 + u_2 \geq v_1 + v_2$.

Наиболее распространенный пример конуса в линейном пространстве — множество n-мерных векторов с неотрицательными координатами, т.е. R_+^n.

Аналогично в функциональных пространствах C и L_p в качестве конуса мы будем брать множество неотрицательных функций.

Множество M называется *частично упорядоченным,* если для некоторых пар элементов $u, v \in M$ определено отношение $u \geq v$.

Отношение порядка называется *полным*, если для любых $u, v \in M$, или $u < v$, или $u > v$, или $u = v$.

Множество M называется *упорядоченным*, или цепью, если отношение порядка на этом множестве полное.

Рассмотрим несколько примеров.

1. Множество всех вещественных чисел есть упорядоченное множество.

2. Пространство R^n можно частично упорядочить следующим образом: для двух векторов $u, v \in R^n$ считаем $u \leq v$, если $u_i \leq v_i$, $\forall i = 1, 2, \ldots, n$.

3. Множество $C([a, b])$ всех вещественных непрерывных функций будет частично упорядоченным, если считать $u \leq v$, когда $u(x) \leq v(x)$, $\forall x \in [a, b]$.

Совокупность всех x, таких, что $u \leq x \leq v$, где u, v, x — элементы частично упорядоченного множества, называется *отрезком* и обозначается $[u, v]$. Такие совокупности мы также будем называть интервальными элементами и будем выделять жирным шрифтом: $\boldsymbol{u} = [\underline{u}, \overline{u}]$.

Линейный оператор $G : B \to B$ называется *положительным,* если преобразует конус K в себя, т.е. из $v \geq 0$ следует $Gv \geq 0$.

В R^n с конусом R^n_+ таким оператором является матрица размером $n \times n$ с неотрицательными элементами. В $C[0, 1]$ и $L_p[0, 1]$ пример положительного оператора дает интегральное преобразование

$$Gv = \int_0^1 k(x, t)v(t)dt$$

с неотрицательным и непрерывным на $[0, 1]^2$ ядром k.

Рассмотрим операторное уравнение второго рода

$$u = Gu + f \tag{1.1}$$

с линейным ограниченным положительным оператором $G : B \to B$ и известной правой частью $f \in B$. Предположим, что спектральный радиус

$$\rho(G) = \lim_{k \to \infty} \|G^k\|^{1/k} < 1. \tag{1.2}$$

На основании принципа сжатых отображений [44] это приводит к однозначной разрешимости задачи (1.1).

Теорема 1. *Пусть для задачи (1.1) с условием (1.2) выполнено неравенство* $f \geq 0$, *тогда* $u \geq 0$. □

Из теоремы (1) вытекают результаты, называемые теоремами сравнения [34].

Теорема 2. *Пусть для задач*

$$u_1 = Gu_1 + f_1, \quad u_2 = Gu_2 + f_2 \tag{1.3}$$

выполнено (1.2). Тогда из $f_1 \leq f_2$ *следует, что* $u_1 \leq u_2$. □

Рассмотрим оператор H, удовлетворяющий следующему неравенству $H \leq G$ или в подробной записи $Hv \leq Gv \ \forall v \in K$. В случае монотонных норм спектральные радиусы этих операторов удовлетворяют неравенствам

$$\rho(H) \leq \rho(G) < 1.$$

Теорема 3. *Пусть* u, v — *решения задач (1.1) и*

$$v = Hv + f$$

с правой частью $f \geq 0 (f \leq 0)$ *и линейными ограниченными положительными операторами* G, H *со спектральными радиусами меньше 1. Тогда из* $G \geq H$ *следует* $u \geq v \geq 0 (u \leq v \leq 0)$. □

Рассмотрим случай регулярного оператора G [42], т. е. представляемого в виде разности двух положительных линейных операторов G_+, G_- [27]:

$$G = G_+ - G_-.$$

В этом случае имеются следующие аналоги теорем сравнения.

Теорема 4. *Пусть* u — *решение задачи (1.1) с регулярным оператором* G, *а* $(v, w) \in B^2$ — *решение системы*

$$\begin{aligned} v &= G_+v - G_-w + f_1, \\ w &= G_+w - G_-v + f_2. \end{aligned} \tag{1.4}$$

И пусть для оператора $T : B^2 \to B^2$, *определяемого равенством*

$$T = \begin{pmatrix} G_+ & G_- \\ G_- & G_+ \end{pmatrix},$$

спектральный радиус

$$\rho(T) < 1. \tag{1.5}$$

Тогда из неравенства

$$f_1 \le f \le f_2 \tag{1.6}$$

вытекают оценки

$$v \le u \le w. \tag{1.7}$$

Доказательство. Положим $z = -w$ и перепишем (1.4) в виде

$$v = G_+ v + G_- z + f_1, \quad z = G_+ z + G_- v - f_2. \tag{1.8}$$

Кроме того, обозначим $y = -u$ и запишем уравнение (1.1) как систему

$$\begin{aligned} u &= G_+ u + G_- y + f, \\ y &= G_+ y - G_+ u + f. \end{aligned} \tag{1.9}$$

Из положительности операторов G_+, G_- следует положительность оператора T в B^2. Учитывая (1.5), предположения теоремы 2 выполнены. Поскольку $(f_1, -f_2) \le (f, -f)$ в B^2, то из теоремы 2 следует, что $(v, z) \le (u, y)$ или $v \le u$ и $-w \le -u$. □

Рассмотрим операторы

$$H_+ \le G_+, \quad H_- \le G_-, \quad F_+ \le G_+, \quad F_- \le G_- \tag{1.10}$$

и связанную с ними задачу

$$y = H_+ y - H_- z + f_1, \quad z = F_+ z - F_- y + f_2. \tag{1.11}$$

Предположим, что для оператора $T_1 : B^2 \to B^2$,

$$T_1 = \begin{pmatrix} H_+ & H_- \\ F_- & F_+ \end{pmatrix}$$

спектральный радиус

$$\rho(T_1) < 1. \tag{1.12}$$

Теорема 5. *Пусть для задач (1.4), (1.11) выполнены условия (1.5), (1.10), (1.12) и $f_1 \le 0 \le f_2$. Тогда*

$$y \le v \le 0 \le w \le z. \square \tag{1.13}$$

1.2. Операторы монотонного типа

Пусть R и R^* — линейные частично упорядоченные метрические пространства. Оператор T имеет область определения $D \subseteq R$ и область значений $W \subseteq R^*$.

Оператор T называется оператором *монотонного типа*, если из соотношения $Tv \leq Tw$ следует $v \leq w$.

Если задано уравнение

$$Tu = s,$$

где T — оператор монотонного типа, и известны два приближенных решения v, w, такие что:

$$Tv \leq s \leq Tw,$$

то

$$v \leq u \leq w.$$

В дальнейшем это свойство монотонных операторов будет использоваться для нахождения апостериорных оценок погрешности и построения двусторонних решений для ряда задач.

Большинство линейных и некоторых квазилинейных уравнений эллиптического и параболического типа второго порядка обладает операторами монотонного типа, поскольку функции Грина для этих уравнений $G \geq 0$.

Рассмотрим краевую задачу для квазилинейного уравнения второго порядка

$$Tu = \begin{pmatrix} -L(u)u(x) \\ u(z) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{1.14}$$

$x \in \Omega, z \in \partial\Omega$, где Ω — ограниченная, связанная область пространства R^n, $L(u)$ — квазилинейный оператор второго порядка вида

$$L(u) = \sum_{i,j=1}^{2} a_{ij}(x, u, Du)D_{ij} + b(x, u, Du), \tag{1.15}$$

$$a_{ij} = a_{ji}.$$

Мы будем предполагать, что $u \in H^2(\Omega) \cap H_0^1(\Omega)$ и a_{ij}, b — непрерывные функции, заданные на $\Omega \times R \times R^n$ [24].

Теорема 6. *Пусть выполнены условия:*

1) оператор $L(u)$ локально равномерно эллиптичен на функции u;

2) коэффициенты $a_{ij}(x,z,p)$ не зависят от z;

3) коэффициент b является неубывающей функцией аргумента z в каждой точке $(x,p) \in \Omega \times R^n$;

4) коэффициенты a_{ij}, b — непрерывно дифференцируемые функции переменных p.

Тогда оператор T — монотонного типа. □

Далее мы приведем несколько топологических теорем о неподвижной точке, которые будем использовать для построения двусторонних решений и нахождения апостериорных оценок погрешности численных решений.

Пусть непрерывный оператор T, определенный на множестве M элементов пространства R, отображает M в себя, $TM \subseteq M$. В теореме Брауэра [74] о неподвижной точке пространство R представляет собой $n-$мерное действительное пространство R^n, а M — единичный шар в нем; в этом случае оператор T имеет в M по меньшей мере одну неподвижную точку $u \in M$:

$$Tu = u.$$

Обобщение теоремы Брауэра на случай банаховых пространств — теорема Шаудера [42],[24].

Теорема 7 (Шаудера о неподвижной точке). *Пусть M — компактное выпуклое множество в банаховом пространстве R и пусть T — непрерывное отображение M в себя. Тогда отображение T имеет неподвижную точку, существует элемент $u \in M$ такой, что $Tu = u$.* □

Для того чтобы можно было применять принцип Шаудера для установления неподвижной точки оператора, необходимо иметь критерии компактности множеств в пространствах, в которых действует заданный оператор.

Множество функций $\mathcal{X} = \{x\}$, определенных в области Ω, называются *равномерно ограниченными*, если существует константа K: $|x(t)| \leq K$, $\forall t \in \Omega$.

Множество функций $\mathcal{X} = \{x\}$, определенных в области Ω, называются *равномерно непрерывными*, если $\forall \varepsilon > 0, \exists \delta > 0$ такое, что $\forall t_1, t_2 \in \Omega$: $||t_1 - t_2|| < \delta$, $\forall x \in \mathcal{X}$: $|x(t_1) - x(t_2)| < \varepsilon$.

Сформулируем критерий компактности множества в пространстве C для случая, когда Ω — компакт.

Лемма 1. *Для компактности множества $\mathcal{X}$ необходимо и достаточно, чтобы функции этого множества были равномерно ограниченными и равномерно непрерывными.*

Пусть R — полуупорядоченное банахово пространство; рассмотрим уравнение

$$Tu = u.$$

Пусть задан оператор $T(u,v)$: $T(x,x) = Tx$ и оператор $T(u,v)$ определен в выпуклом множестве $D \times D, D \subset R$, причем оператор $T(u,v)$ изотонный по u и антитонный по v, т. е. [44]

$$T(u,v) \leq T(u_1,v_2), \text{ если } u \leq u_1, v \leq v_1.$$

Заметим, что любой отрезок $\mathbf{v} = [\underline{v}, \overline{v}], v, w \in R$ является выпуклым и замкнутым.

Теорема 8. *Пусть определены последовательности $u_n, v_n (u_0 \leq v_0)$ по формулам*

$$u_n = T(u_{n-1}, v_{n-1}), \tag{1.16}$$
$$v_n = T(v_{n-1}, u_{n-1}), \tag{1.17}$$

причем для элементов u_0, v_0, u_1, v_1, справедливо соотношение

$$u_0 \leq u_1 \leq v_1 \leq v_0.$$

и множество $[u_n, v_n]$ компактно. Тогда уравнение

$$Tx = x$$

имеет на отрезке $[u_n, v_n]$ по крайней мере одно решение. □

1.3. Специальные аппроксимации численных решений

В разделе рассматриваются методы построения специальных аппроксимаций численных решений, необходимых для вычисления невязок.

Предположим, что нам известно численное решение u^h некоторой задачи

$$Lu = f, \ x \in \Omega,$$

на сетке $\Omega_h = \{x_i | x_i \in \Omega, i = 1, \ldots, N\}$. Наша задача построить некоторую функцию $s \in H^l(\Omega) \cap H_0^1(\Omega)$, удовлетворяющую следующим условиям

$$|s(x) - u^h(x)| \leq Kh^\mu, \ x \in \Omega_h.$$

Кроме того, мы будем стремиться к тому, чтобы невязка была как можно меньше в некоторой норме

$$||Ls - f|| \to min,$$

где h — характерный размер сетки.

Одномерные задачи

Рассмотрим случай, когда $\Omega = [0, 1]$ и

$$Lu \equiv -\frac{d}{dx}(p\frac{d}{dx}u) + qu = f, \tag{1.18}$$

$$u(0) = u(1) = 0. \tag{1.19}$$

Относительно коэффициентов предположим, что

$$p(x) > c_1 > 0, \ q(x) > 0, \ x \in (0, 1), \tag{1.20}$$

$$q, f \in C^r[0, 1], \ p \in C^{r+1}[0, 1] \tag{1.21}$$

для некоторого $r \geq 0$.

Введем на отрезке $[0, 1]$ сетку

$$\Omega_h = \{x_i | x_i = ih, i = 0, 1, \ldots, N, h = 1/N\}, N \geq 2.$$

Рассмотрим разностную схему для численного решения задачи (1.18), (1.19)

$$-p_i d_h^2 u_i^h - p'(x_i) d_h^1 u_i^h + q_i u_i^h = f_i, i = 1, \ldots, N-1, \tag{1.22}$$

$$u_0^h = u_N^h = 0, \tag{1.23}$$

где

$$d_h^1 u_i = \frac{u_{i+1} - u_{i-1}}{2h},$$

$$d_h^2 u_i = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}.$$

Далее будем считать

$$d_h^0 u_i = u_i.$$

Предположим, что нам необходимо построить функцию $s \in W_2^2[0,1] \cap \overset{o}{W}{}_2^1[0,1]$. Наиболее простой способ нахождения s — воспользоваться эрмитовыми сплайнами третьей степени. Для этого в каждом узле сетки необходимо задать значения $s(x_i), s'(x_i)$. На каждом отрезке $[x_i, x_{i+1}]$ сплайн s можно представить в виде[36]

$$s(x) = \phi_1 s(x_i) + \phi_2 s(x_{i+1}) + \xi_1 s'(x_i) + \xi_2 s(x_{i+1}),$$

где

$$\phi_1(t) = (1-t)^2(1+2t), \phi_2(t) = t^2(3-2t),$$
$$\xi_1(t) = t(1-t^2), \xi_2(t) = -t^2(1-t), t = (x-x_i)/h.$$

Если сплайн s интерполирует на сетке Ω_h некоторую функцию u и в узлах x_i выполняются условия

$$s(x_i) = u(x_i), s'(x_i) = u'(x_i). \tag{1.24}$$

то справедлива следующая теорема [36].

Теорема 9. *Для эрмитовых сплайнов третьей степени справедливы оценки*

$$||s^\nu - u^\nu||_{\infty,\Omega} \le K_\nu h^{4-\nu} ||u^4||_{\infty,\Omega}, \quad \nu = 0,1,2,3.\square$$

Рассмотрим случай интерполяции, когда значения производных в узлах сетки нам неизвестны. Тогда мы можем воспользоваться разностными производными и положить

$$s_d(x) = u(x),\ s'_d(x) = \partial_x u(x), x \in \Omega_h. \tag{1.25}$$

Разностный оператор ∂_x определим следующим образом:

$$\partial_x^h v(x) = \frac{1}{6h} \begin{cases} -11v(x) + 18v(x+h) - 9v(x+2h) + 2v(x+3h), & x = 0; \\ v(x-2h) - 6v(x-h) + 3v(x) + 2v(x+h), & x = 1-h; \\ -2v(x-3h) + 9v(x-2h) - 18v(x-h) + 11v(x), & x = 1; \\ -2v(x-h) - 3v(x) + 6v(x+h) - v(x+2h). \end{cases}$$

Этот разностный оператор аппроксимирует первую производную с точностью $O(h^3)$. Следуя работе [36], мы можем оценить погрешность построенного сплайна s_d. Заметим, что на отрезке $[x_i, x_{i+1}]$ [36]

$$|s_d(x) - s(x)| \le t(1-t)h \max_{j=i,i+1} |u'(x_j) - \partial_x u(x_j)|, \tag{1.26}$$

$$|s'_d(x) - s'(x)| \le ((1-t)|1-3t| + t|2-3t|) \times \max_{j=i,i+1} |u'(x_j) - \partial_x u(x_j)|, \tag{1.27}$$

$$|s''_d(x) - s''(x)| \le |6t-4||u'(x_i) - \partial_x u(x_i)|/h + |2-6t||u'(x_{i+1}) - \partial_x u(x_{i+1})|/h. \tag{1.28}$$

Следовательно, для s_d справедлива оценка

$$||s_d^\nu - u^\nu||_{\infty,\Omega} \le K_\nu h^{4-\nu}||u^4||_{\infty,\Omega}, \quad \nu = 0,1,2. \tag{1.29}$$

Обозначим

$$u_{xx}(x) = (u(x-h) - 2u(x) + u(x+h))/h^2.$$

Лемма 2. *Для эрмитовых сплайнов, интерполирующих функцию и на сетке Ω_h, справедливо следующее неравенство:*

$$|s''(x)| \le C||u_{xx}||_{\infty,\Omega_h},$$

где C — константа, не зависящая от h.

Доказательство. Заметим, что на каждом интервале $[x_i, x_{i+1}]$ s'' можно представить в виде

$$s''(x) = a_i(x_{i+1} - x)/h + b_i(x - x_i)/h.$$

Непосредственной подстановкой нетрудно убедиться, что

$$a_i = \frac{4}{3}u_{xx}(x_i) - \frac{2}{3}u_{xx}(x_{i+1}) + \frac{1}{3}u_{xx}(x_{i+2}),$$

$$b_i = -\frac{2}{3}u_{xx}(x_i) + \frac{7}{3}u_{xx}(x_{i+1}) - \frac{2}{3}u_{xx}(x_{i+2}).$$

Поскольку s'' — линейная функция, то

$$|s''(x)| \le \max\{a_i, b_i\},\ x \in [x_i, x_{i+1}]$$

и

$$|s''(x)| \le C\max\{|u_{xx}(x_i)|, |u_{xx}(x_{i+1})|, |u_{xx}(x_{i+2})|\},\ x \in [x_i, x_{i+1}].$$

Из последнего неравенства и вытекает утверждение леммы. □

Лемма 3. *Существует константа K, не зависящая от h, такая, что*

$$||d_h^i u^h||_{\infty,\Omega_h} \le K||f||_{\infty,\Omega_h}, i = 0,1,2. \tag{1.30}$$

Доказательство. Случай $i = 0$ рассмотрен в [59]:

$$||d_h^0 u^h||_{\infty,\Omega_h} \le K||f||_{\infty,\Omega_h}. \tag{1.31}$$

В работе [60] приводится следующее неравенство для сеточных функций с условием (1.19):

$$||d_h^1 u^h||_{\infty,\Omega_h} \le \epsilon ||d_h^2 u^h||_{\infty,\Omega_h} + \frac{1}{\epsilon}||d_h^0 u^h||_{\infty,\Omega_h}, \tag{1.32}$$

где ϵ — произвольная положительная постоянная. Из разностной схемы (1.22) вытекает неравенство:

$$||d_h^2 u^h||_{\infty,\Omega_h} \le C_1||d_h^1 u^h||_{\infty,\Omega_h} + C_2||d_h^0 u^h||_{\infty,\Omega_h} + ||f||_{\infty,\Omega_h}. \tag{1.33}$$

Решая совместно (1.31), (3.1), (1.33), получаем (1.30) при $i = 1, 2$. Лемма доказана.□

Обозначим $e_i = u(x_i) - u_i^h,\ e_i' = u'(x_i) - d_h^1 u_i^h,\ e_i'' = u''(x_i) - d_h^2 u_i^h$. Тогда из леммы 3 непосредственно вытекает оценка

$$||e^i||_{\infty,\Omega_h} \le Kh^2(||u^{(3)}||_{\infty,\Omega} + ||u^{(4)}||_{\infty,\Omega}), i = 0, 1, 2. \tag{1.34}$$

Для построения эрмитового сплайна s, интерполирующего численное решение задачи (1.18), (1.19) положим:

$$s(x) = u^h(x),\ s'(x) = \partial_x u^h(x), x \in \Omega_h.$$

Рассмотрим невязку построенного эрмитового сплайна s

$$\phi(x, s) = Ls - f.$$

Поскольку $Lu - f = 0$, то

$$L(s - u) = \phi(x, s).$$

Оценим $||L(s-u)||_{\infty,\Omega}$. Для этого представим

$$L(u - s) = L(u - s_I) + L(+s_I - s_d) + L(s_d - s),$$

где s_I — эрмитовый сплайн, интерполирующий u, со значениями (1.24), s_d — эрмитовый сплайн, интерполирующий u, со значениями (1.25). Тогда имеем:

$$|L(u - s_I)| \le Kh^2||u^4||_{\infty,\Omega},$$

$$|L(-s_I + s_d)| \le Kh^2||u^4||_{\infty,\Omega}.$$

Покажем, что

$$|s_d^\nu(x) - s^\nu(x)| = O(h^2).$$

Рассмотрим случай $\nu = 2$. Заметим, что в силу леммы 2

$$|s''_d(x) - s''(x)| \le K||d_h^2(u - u^h)||_{\infty,\Omega}.$$

Далее, в силу (1.34)

$$|d_h^2(u - u^h)(x)| \le Kh^2(||u^{(3)}||_{\infty,\Omega} + ||u^4||_{\infty,\Omega}), x \in \Omega_h.$$

Случаи $\nu = 0, 1$ показываются аналогично. Следовательно, доказана теорема.

Теорема 10. *Существует константа K, не зависящая от h, такая, что*

$$||\phi(\cdot, s)||_{\infty,\Omega} \le Kh^2||u||_{W_2^4(\Omega)}.\square$$

Рассмотрим еще один подход для построения одномерных аппроксимаций. Пусть нам необходимо построить $s \in C^q$. На каждом отрезке $[x_i, x_{i+1}]$ представим s как полином степени n, т.е. необходимо определить $n + 1$ коэффициент. Для достижения необходимой гладкости мы можем воспользоваться разностными производными и положить

$$s^i(x) = \partial_x^i u^h(x), i = 0, 1, 2, \ldots, q; x \in \Omega_h, \tag{1.35}$$

где $\partial_x^i u(x), i = 0, 1, 2, \ldots$ — специальные операторы, аппроксимирующие $u(x)$, $u^{(i)}(x)$, $i = 1, 2, \ldots$, которые мы определим ниже. Для построения сплайна s на каждом отрезке $[x_i, x_{i+1}]$ эти условия дают $2q + 2$ ограничения. В случае $n > 2q + 1$ нам необходимо задать еще $n - 2q - 1$ ограничение. Для этой цели можно потребовать выполнения следующих равенств:

$$Ls(\xi_i) - f(\xi_i) = 0, i = 1, 2, \ldots, n - 2q - 1,$$

где $\{\xi_i\} \in [x_i, x_{i+1}]$ — множество узлов некоторой вспомогательной сетки. Можно несколько усложнить задачу, потребовав

$$\sum_{i=1}^{Q}(Ls(\xi_i) - f(\xi_i))^2 \to \min.$$

Рассмотренные выше задачи сводятся в общем случае к решению систем линейных алгебраических уравнений.

Остановимся на построении операторов $\partial_x^i u(x), i = 0, 1, 2, \ldots$ Для этой цели зафиксируем некоторую точку x_0 и определим $p \in \mathcal{P}^n$ как полином $p = \sum_{i=0}^n a_i(x - x^0)^i$ такой, что

$$\sum_i^N \alpha_i|p(v_i) - u^h(v_i)|^2 + \sum_i^N \beta_i|Lp(w_i) - f(w_i)|^2 \to min, \tag{1.36}$$

где $\{v_i\}$ и $\{w_i\}$ — узлы вспомогательных сеток таких, что x_0 — точка концентрации узлов и

$$\alpha_i = a_1/(g_1(v_i, x_0) + a_2),$$

$$\beta_i = b_1/(g_2(w_i, x_0) + b_2).$$

Причем a_1, a_2 можно положить $O(h^2)$, $b_1 = O(h^2), b_2 \ll 1$ и

$$g_1(x, y) = ||x - y||^{n+1}, g_2(x, y) = ||x - y||^{n-1}.$$

Подобный выбор α_i, β_i обусловлен погрешностями $p(x) - u(x)$ и $L(p(x) - u(x))$ в случае выполнения равенств $p(x_0) = u(x_0)$ и $d^i(p(x_0)/dx^i = d^i(u(x_0)/dx^i, i = 1, 2, \ldots, n$. Задача (1.36) сводится к решению некоторой системы линейных алгебраических уравнений

$$Ba = d.$$

Структура этой системы для общего случая будет подробно выписана в следующем разделе.

Таким образом, в точке x_0 мы можем вычислить $p(x_0), d^i p(x_0)/dx^i$ и положить $\partial_x^i = a_i$, $i = 0, 1, 2, \ldots$.

Двумерные задачи

Рассмотрим случай сглаживания численного решения двумерного эллиптического уравнения, полученного методом конечных элементов (МКЭ). Нашей задачей будет построение функции $s \in H^2(\Omega) \cap H_0^1(\Omega)$, являющейся конечно-элементным восполнением численного решения. Рассмотрим для определенности следующую модельную задачу:

$$\begin{aligned} Lu &= f, \ x \in \Omega, \\ u(x) &= 0, \ x \in \partial\overline{\Omega}, \end{aligned} \tag{1.37}$$

$$Lu = -\sum_{i=1}^{2} \frac{\partial}{\partial x_i}\left(a_i \frac{\partial}{\partial x_i} u\right) + qu.$$

Мы предположим, что коэффициенты $a_i \in C^1(\Omega)$, $q, f \in C(\Omega)$ и что $a_i \geq c > 0, q \geq 0, x \in \Omega$. Предположим, что решение задачи (1.37) существует и единственно.

Пусть $\mathcal{T}$ — разбиение области Ω, составленное из элементов T и $\overline{\Omega} = \cup_i T_i$, Пересечение элементов $T_i, T_j, i \neq j$: $T_i \cap T_j = \emptyset$, или общая сторона, или общая вершина. h — параметр $\mathcal{T}$.

Если $\mathcal{T}_i$ — триангуляция, то пространство конечных элементов S_l^n определим, вводя кусочно-полиномиальный базис на $\mathcal{T}$:

$$S^n = \{s(x) | s \in H^l(\Omega) \cap H_0^1(\Omega), s|_T \in \mathcal{P}^n, T \in \mathcal{T}\}, \qquad (1.38)$$

где $\mathcal{P}^n$ — множество полиномов степени n. Если $\mathcal{T}$ — прямоугольное разбиение, то подпространство S_l^n определим с помощью сплайнов Эрмита степени n [102, 76].

Функцию s можно представить как полином пятой степени [102]. Для этого на каждом треугольнике необходимо найти 21 коэффициент, из которых 18 определяются значениями в вершинах u, $\partial_1 u$, $\partial_2 u$, $\partial_1^2 u$, $\partial_{1,2} u$, $\partial_2^2 u$. Полиномы пятой степени на общей стороне двух треугольников совпадают, так как три условия в каждой вершине — значения функции u и ее производных u_t, u_{tt} вдоль стороны — однозначно задают его коэффициенты.

Остается отыскать три дополнительных ограничения, чтобы производная по нормали s_n была непрерывна между треугольниками. Один способ — добавить значения u_n в середине каждой стороны треугольника. Так как s_n — полином четвертой степени от t вдоль стороны, он единственным образом задается этим параметром вместе со значениями в вершинах (рис. 1.1).

Существует способ построить макроэлементы на объединении нескольких треугольников. Наиболее известен элемент Клафа-Точера: объединение кубических полиномов на трех подтреугольниках. Для его построения необходимо задавать значения $u, \partial_1 u, \partial_2 u$, и $\partial_n u$ в середине каждой стороны треугольника, всего 12 параметров. Так как каждый из трех кубических элементов имеет 10 степеней свободы, а макроэлемент — только 12 узловых параметров, нужно ввести 18 ограничений. Этого достаточно для достижения гладкости C^1 внутри треугольника (рис. 1.2).

Приведем общие теоремы аппроксимации конечными элементами [102].

Теорема 11. *Пусть $s \in S_l^{k-1}$ и базис удовлетворяет условию однородности порядка q. Тогда для $i \le q$*

$$\max_{x \in T_i, |\alpha| <= i} |D^\alpha u(x) - D^\alpha s(x)| \le C_i h^{k-i} \max_{x \in T_i, |\beta| <= k} |D^\beta u(x)|. \square$$

Теорема 12. *Пусть $s \in S_l^{k-1}$ и базис удовлетворяет условию однородности порядка q. Тогда для $i \le q$*

$$|D^\alpha u(x) - D^\alpha s(x)|_{i,\Omega} \le C_i h^{k-i} |u(x)|_{k,\Omega}. \square$$

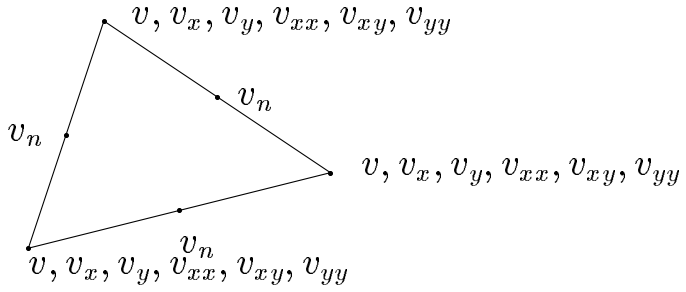

Рис. 1.1. Элементы пятой степени

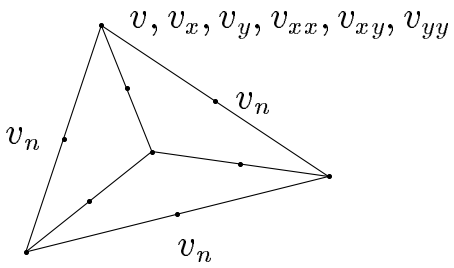

Рис. 1.2. Элементы Клафа-Точера

Рассмотрим проблему аппроксимации численного решения u^h конечными элементами $s \in S_2^k, k \geq 3$, так что

$$||s||_{L_2(\partial\overline{\Omega})} \leq Kh^{\sigma 1},$$

$$||Ls - f||_{L_2(\Omega)} \leq Kh^{\sigma 2}.$$

Для построения конечных элементов высоких степеней мы нуждаемся в знании определенного набора значений $s(x), \partial_{i,j}^{i+j} s(x), i, j = 0, 1, 2, \ldots$ в некоторых точках $x \in \overline{\Omega}$. Пусть $x_0 = (x_{0,1}, x_{0,2})$ одна из таких точек.

Для этих целей рассмотрим локальные вспомогательные сетки: $Z_{r,\delta,d} = \{z_{i,j}\}$, $z_{i,j} = (z_{i,j,1}, z_{i,j,2})$:

$$z_{i,j,1} = sign(i)abs(i\delta)^r/d + x_{0,1}, z_{i,j,2} = sign(j)abs(j\delta)^r/d + x_{0,2}, i, j = 0, \pm 1, \pm 2, \ldots,$$

где r, δ, d — параметры и $r \geq 1, \delta > 0, d > 0$ (рис. 1.3).

Определим p как обобщенный полином $p = \sum_{l=0}^{n_p} a_l \psi_l(x - x_0)$ так, что

$$\sum_{i=1}^{n_1} \alpha_i |p(v_i^1) - u^h(v_i^1)|^2 + \sum_{i=1}^{n_2} \beta_i |Lp(v_i^2) - f(v_i^2)|^2 \to min, \qquad (1.39)$$

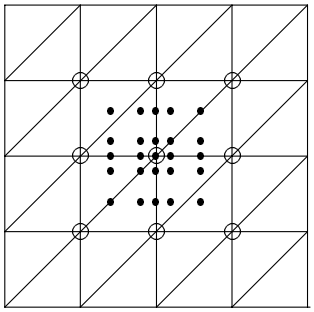

Рис. 1.3. Пример вспомогательных сеток

где ψ_i — линейно независимые функции. Пусть для определенности

$$\psi_0(x) = 1, \psi_1(x) = x_1, \psi_2(x) = x_2, \psi_3(x) = x_1^2, \ldots,$$

$$\psi_{n_p}(x) = x_2^n, n_p = (n+2)(n+1)/2,$$

$\{v_i^1\}_{i=1}^{n_1}$ and $\{v_i^2\}_{i=1}^{n_2}$ — узлы вспомогательных сеток, расположенных в непосредственной близости от точки x_0. Здесь $v_i^1 \in \Omega_h, v_i^2 \in Z_{r,\delta,d}$, $n_1 + n_2 \geq n_p + 1$, $n_1 \geq 2n + 1$,

$$\alpha_i = \alpha_{-1}/(\rho(v_i^1, x_0)^{n+1} + \alpha_{-2}),$$

$$\beta_i = \beta_{-1}/(\rho(v_i^2, x_0)^{n-1} + \beta_{-2}),$$

где $\rho(x, y)$ — расстояние между точками x, y; α_{-i}, β_{-i} выражаются в терминах точности численного решения u^h:

$$\alpha_{-1}, \alpha_{-2}, \beta_{-1} \approx ||u - u^h||_{L_\infty(\Omega_h)},$$

$$\beta_{-2} \approx h||u - u^h||_{L_\infty(\Omega_h)}.$$

Задача (1.39) сводится к решению системы линейных алгебраических уравнений

$$Ba = d,$$

где

$$B = \{b_{ij}\}_{i,j=0}^{n_p},$$

$$a = (a_0, a_1, a_2, \ldots, a_{n_p}),$$

$$b_{ij} = \sum_{l=1}^{n_1} \alpha_l \psi_i(v_l^1 - x_0)\psi_j(v_l^1 - x_0) + \sum_{l=1}^{n_2} \beta_l L\psi_i(v_l^2 - x_0)L\psi_j(v_l^2 - x_0),$$

$$d_i = \sum_{l=1}^{n_1} \alpha_l u^h(v_l^1)\psi_i(v_l^1 - x_0) + \sum_{l=1}^{n_2} \beta_l f(v_l^2) L\psi_i(v_l^2 - x_0).$$

Таким образом, мы можем положить $s(x_0) = p(x_0), \partial_{i,j}^{i+j} s(x_0) = \partial_{i,j}^{i+j} p(x_0)$.

1.4. Некоторые свойства вариационно-разностных решений

Рассмотрим следующее линейное уравнение:

$$-\Delta u + qu = f,\ x \in \Omega, \tag{1.40}$$

$$u = 0,\ x \in \partial\Omega. \tag{1.41}$$

Введем билинейную форму

$$\mathcal{L}(u,v) = \int_\Omega \sum_{i=1}^{2} \partial_i u \partial_j v d\Omega, \quad \forall u, v \in \overset{\circ}{W}{}_2^1(\Omega).$$

Рассмотрим систему линейных алгебраических уравнений для нахождения вариационного решения задачи (1.40), (1.41)

$$\sum_{j=1}^{N} \mathcal{L}(v_i, v_j)\alpha_j = (f, v_i). \tag{1.42}$$

Обозначим через $\vec{f}$ вектор правых частей системы $f_i = (f, v_i)$. Для любой функции $u^h \in W_2^1(\Omega)$ формально определим

$$u^h{}_{ij}(x_l) = -h^{-2} \int_\Omega \partial_i u^h \partial_j v d\Omega,$$

$$||u^h||_2^2 = \sum_{l=1}^{N} \sum_{i,j=1}^{2} (u_{ij}^h(x_l))^2.$$

В случае равномерной триангуляции выражения v_{ij} совпадают с соответствующими конечно-разностными аппроксимациями вторых производных.

Запишем (1.42) в следующем виде:

$$-(u^h{}_{11} + u^h{}_{22}) + (qu^h, v_l)/h^2 = f_l/h^2, l = 1, \ldots, N.$$

Возведем обе части равенства в квадрат и просуммируем:

$$\sum_{l=1}^{N} (u^h{}_{11}(x_l))^2 + (u^h{}_{22}(x_l))^2 +$$

$$(qu^h, v_l)^2/h^4 + 2u^h{}_{11}u^h{}_{22} - 2u^h{}_{11}(qu^h, v_l)/h^2$$
$$-2u^h{}_{22}(qu^h, v_l)/h^2 = \sum_{l=1}^{N} f_l^2/h^4.$$

Используя формулы интегрирования по частям [59], преобразуем выражение

$$\sum_{l=1}^{N} u^h{}_{11}(x_l)u^h{}_{22}(x_l) = -\sum_{l=1}^{N} u^h{}_1(x_l)(u^h{}_{22}(x_l))_1 =$$
$$= \sum_{l=1}^{N} u^h{}_{12}(x_l))^2.$$

Далее, используя ε — неравенство, оценим

$$\sum_{l=1}^{N} u^h{}_{11}(qu^h, v_l)/h^2 \le \sum_{l=1}^{N} \varepsilon/2(u^h{}_{11})^2 + \sum_{l=1}^{N}(qu^h, v_l)^2/(2\varepsilon h^4).$$

Слагаемое $\sum_{l=1}^{N}(qu^h, v_l)^2$ можно оценить

$$\sum_{l=1}^{N}(qu^h, v_l)^2 \le C||\vec{f}||^2,$$

где C — некоторая константа, не зависящая от h. Окончательно получаем оценку

$$\sum_{l=1}^{N} u^h{}^2_{11} + u^h{}^2_{22} + u^h{}^2_{12} \le C||\vec{f}||^2/h^4,$$

или

$$||u^h||_2 \le C||\vec{f}||/h^2, \tag{1.43}$$

где C — некоторая константа, не зависящая от h.

Далее, если $\varepsilon = u - u^h$, то, следуя общей теории разностных схем [59], получаем оценку

$$||\varepsilon||_2 C \le h^2\psi, \tag{1.44}$$

где ψ — некоторая ограниченная функция, зависящая от u, f, q и производных.

1.5. Аппроксимация кубическими элементами

Рассмотрим случай равномерной триангуляции. Пусть T — прямоугольный треугольник с катетами, равными h, z_1, z_2, z_3 — его вершины, z_0 — точка пересечения медиан.

Будем строить кубические элементы из C^0 таким образом. На каждом треугольнике мы потребуем выполнения соотношений:

$$s(z_l) = u^h(z_l), \partial_i s(z_l) = \partial_i^h u^h(z_l), l = 1,2,3; i = 1,2;$$

потребуем также выполнения равенства в центре треугольника

$$Ls(z_0) = f(z_0).$$

Таким образом, для построения s на T получаем систему из 10 линейных алгебраических уравнений, что полностью определяет s.

Представим s в виде

$$s(x,y) = \sum_{\alpha=0}^{3} \sum_{i+j=\alpha} a_{ij} x^i y^j.$$

Лемма 4. *Для коэффициентов $a_{ij}, i+j = 2,3$ полинома s справедливы следующие оценки:*

$$|a_{ij}| \le K_{ij} h^{-i-j} (\sum_{l=1}^{3} (u_{11}^h(z_l) + u_{12}^h(z_l) + u_{22}^h(z_l)) + f(z_0)).$$

Для доказательства достаточно выписать представление a_{ij} через $u^h(z_l)$, $\partial_i^h u^h(z_l)$, $f(z_0)$. Затем подставить вместо $\partial_i^h u^h(z_l)$ их конечно-разностные выражения и перегруппировать слагаемые. Тогда, например,

$$a_{11} = u_{12}^h(z_2) + 3/4 f(z_0),$$
$$a_{21} = h^{-1}(u_{11}^h(z_3) - 4u_{11}^h(z_2) - u_{22}^h(z_1) + u_{22}^h(z_1) + 2u_{22}^h(z_2))/4.$$

□

Несложно убедиться, что в силу доказанной леммы выполнено соотношение:

$$|\partial_{ij} s(x,y)| \le K \sum_{i+j=2}^{3} \sum_{\alpha=1}^{3} |u_{ij}^h(z_\alpha)|. \qquad (1.45)$$

Глава 2

Элементы интервального анализа

2.1. Интервальные числа

Под *интервальным числом* $\boldsymbol{a}$ мы будем понимать вещественный отрезок $[\underline{a}, \overline{a}]$, где $\underline{a} \leq \overline{a}$. Множество интервальных чисел мы будем обозначать через $\mathbf{R}$. При $\underline{a} = \overline{a} = \boldsymbol{a}$ интервальное число будем отождествлять с вещественным числом a, следовательно $R \subset \boldsymbol{R}$. В дальнейшем мы будем называть интервальные числа просто интервалами. *Шириной* $\boldsymbol{a}$ — это величина

$$\mathrm{wid}(\boldsymbol{a}) = \overline{a} - \underline{a},$$

середина — полусумма

$$\mathrm{med}(\boldsymbol{a}) = (\underline{a} + \overline{a})/2.$$

Если S — непустое ограниченное множество в R^n, то его *интервальной оболочкой* $\square S$ определим наименьший по включению интервальный вектор, содержащий S.

Арифметические операции над интервальными числами введем следующим образом. Пусть $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{R}$, тогда положим

$$\boldsymbol{a} * \boldsymbol{b} = \{x * y | x \in \boldsymbol{a},\ y \in \boldsymbol{b}\},$$

где знак $(*)$ — одна из операций $+, -, \cdot, /$. При делении интервал $\boldsymbol{b} = [\underline{b}, \overline{b}]$ не должен содержать ноль. Введенные выше операции эквивалентны следующим:

$$\boldsymbol{a} + \boldsymbol{b} = [\underline{a}, \overline{a}] + [\underline{b}, \overline{b}] = [\underline{a} + \underline{b}, \overline{a} + \overline{b}],$$

$$\boldsymbol{a} - \boldsymbol{b} = [\underline{a}, \overline{a}] - [\underline{b}, \overline{b}] = [\underline{a} - \overline{b}, \overline{a} - \underline{b}],$$

$$\boldsymbol{a} \cdot \boldsymbol{b} = [\underline{a}, \overline{a}] \cdot [\underline{b}, \overline{b}] = [\min(\underline{ab}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}), \quad \max(\underline{ab}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b})],$$

$$\boldsymbol{a}/\boldsymbol{b} = [\underline{a}, \overline{a}]/[\underline{b}, \overline{b}] = [\underline{a}, \overline{a}] \cdot [1/\overline{b}, 1/\underline{b}], \quad 0 \notin [\underline{b}, \overline{b}].$$

Если $\boldsymbol{a}$ и $\boldsymbol{b}$ вырождаются в вещественные числа, то эти равенства совпадают с обычными арифметическими операциями. Интервальные операции сложения и умножения остаются коммутативными и ассоциативными, т.е. $\boldsymbol{a}$,$\boldsymbol{b}$,$\boldsymbol{c} \in R$

$$\boldsymbol{a} + \boldsymbol{b} = \boldsymbol{b} + \boldsymbol{a}, \ \boldsymbol{a} + (\boldsymbol{b} + \boldsymbol{c}) = (\boldsymbol{a} + \boldsymbol{b}) + \boldsymbol{c},$$

$$\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{b} \cdot \boldsymbol{a}, \ \boldsymbol{a} \cdot (\boldsymbol{b} \cdot \boldsymbol{c}) = (\boldsymbol{a} \cdot \boldsymbol{b}) \cdot \boldsymbol{c}.$$

Если $r(x)$ — непрерывная унарная операция на R, то

$$r(X) = [\min_{x \in X} r(x), \max_{x \in X} r(x)]$$

определяет соответствующую ей операцию на $\boldsymbol{R}$. Примерами таких унарных операций могут служить

$$\exp(X), \ln(X), \sin(X), \ldots$$

Теперь укажем отличия интервальной арифметики от обычной. Вместо дистрибутивности умножения относительно сложения для $a, b, c \in R$, т.е. $a(b + c) = ab + ac$, выполняется субдистрибутивность

$$\boldsymbol{a}(\boldsymbol{b} + \boldsymbol{c}) \subset \boldsymbol{a}\boldsymbol{b} + \boldsymbol{a}\boldsymbol{c}.$$

Заметим, что $\boldsymbol{R}$ не является полем: элементы $\boldsymbol{R}$ не имеют обратных элементов относительно сложения и умножения. В частности, $\boldsymbol{a}$ - $\boldsymbol{a} \neq 0$ и $\boldsymbol{a}/\boldsymbol{a} \neq 1$. Вместо этого действуют два других правила сокращения:

1. Из $\boldsymbol{a} + \boldsymbol{b} = \boldsymbol{a} + \boldsymbol{c}$ следует $\boldsymbol{b} = \boldsymbol{c}$;
2. Из $\boldsymbol{a}b = \boldsymbol{a}\boldsymbol{c}$ и $0 \notin \underline{a}$ следует $\boldsymbol{b} = \boldsymbol{c}$.

Интервальные арифметические операции обладают свойством *монотонности по включению:* из условий $\boldsymbol{a} \subset \boldsymbol{c}, \boldsymbol{b} \subset \boldsymbol{d}$ следуют включения

$$\boldsymbol{a} + \boldsymbol{b} \subset \boldsymbol{c} + \boldsymbol{d}, \quad \boldsymbol{a} - \boldsymbol{b} \subset \boldsymbol{c} - \boldsymbol{d},$$

$$\boldsymbol{a}b \subset \boldsymbol{c}\boldsymbol{d}, \quad \boldsymbol{a}/\boldsymbol{b} \subset \boldsymbol{c}/\boldsymbol{d} \quad \text{если} \quad 0 \notin \boldsymbol{d}.$$

Унарные операции обладают сходными свойствами:

$$X \subseteq Y \Rightarrow r(X) \subseteq r(X),$$

$$x \in Y \Rightarrow r(x) \in r(X).$$

Мы расширим отношения порядка $* \in \{<, \leq, >, \geq\}$ на интервальные переменные:

$$\boldsymbol{x} * \boldsymbol{y} \Leftrightarrow \tilde{x} * \tilde{y} \text{ для всех } \tilde{x} \in \boldsymbol{x}, \tilde{y} \in \boldsymbol{y}.$$

Расстояние ρ между двумя интервалами $\boldsymbol{a}$, $\boldsymbol{b}$ определяется следующим образом

$$\rho(\boldsymbol{a}, \boldsymbol{b}) = \max\{|\underline{a} - \underline{b}|, |\overline{a} - \overline{b}|\}.$$

Для вырожденных интервалов введенное расстояние сводится к обычному расстоянию между вещественными числами:

$$\rho(a, b) = |a - b|.$$

Рассмотренная выше метрика на $\boldsymbol{R}$ является частным случаем хаусдорфовой.

Если U и V — непустые компактные множества вещественных чисел, то *хаусдорфово* расстояние определяется как

$$\rho(U, V) = \max\{\sup_{v \in V} \inf_{u \in U} \rho(u, v), \sup_{u \in U} \inf_{v \in V} \rho(u, v)\}.$$

Вводя на множестве метрику, мы делаем его топологическим пространством. При этом понятия сходимости и непрерывности могут быть использованы обычным образом. Последовательности интервалов сходятся, если сходятся последовательности границ интервалов. Метрическое пространство $\boldsymbol{R}$ с метрикой ρ является замкнутым метрическим пространством, и введенные арифметические операции непрерывны.

2.2. Гистограммная арифметика

Пусть $X_1, X_2, \ldots, X_N$ — независимые случайные величины и Y определяется функциональной зависимостью

$$Y = f(X_1, X_2, \ldots, X_N). \tag{2.1}$$

Предположим, что при вычислении функции f используются арифметические операции $\{+, -, \cdot, /\}$ и операции возведения в степень $\{\uparrow\}$, нахождения максимума и минимума $\{\max, \min\}$. Случайные величины $X_l, l = 1, 2, \ldots, N$ принимают значения в интервалах $[\underline{a}_l, \overline{a}_l]$, и плотность их распределения задается кусочно-постоянной функцией $P_l(x)$ следующего вида: точки $\{\alpha_m^l, m = 0, 1, \ldots, M_l\}$ образуют интервалы постоянных значений функции P_l. Такие случайные величины будем называть *гистограммными числами* или *гистограммами.*

Требуется вычислить плотность вероятности P_Y величины Y с заданной точностью в классе кусочно-постоянных функций — гистограмм.

Для этой цели мы заменим обычные операции на операции над гистограммными переменными.

Известны аналитические формулы для определения плотности вероятности результатов арифметических действий над случайными величинами. Например, для нахождения плотности вероятности $P_{X_1+X_2}$ суммы двух случайных величин $X_1 + X_2$ используется соотношение

$$P_{X_1+X_2} = \int_{-\infty}^{\infty} P_1(x-v)P_2(v)dv = \int_{-\infty}^{\infty} P_1(v)P_2(x-v)dv, \qquad (2.2)$$

для нахождения плотности вероятности P_{X_1/X_2} частного двух случайных величин X_1/X_2

$$P_{X_1/X_2} = \int_0^{\infty} vP_1(xv)P_2(v)dv - \int_{-\infty}^{0} P_1(v)P_2(x-v)dv. \qquad (2.3)$$

На основе формул типа (2.2), (2.3) для специально подобранных разбиений интервалов значений исходных случайных величин в некоторых случаях удается восстановить точно функцию плотности вероятности. Однако в случаях произвольных разбиений эти формулы не удобны для численных расчетов.

Основные принципы разработки гистограммных операций продемонстрируем на примере операции сложения. Пусть $Z = X_1 + X_2$, тогда плотность вероятности Z отлична от нуля на интервале $[\alpha_0^1 + \alpha_0^2, \alpha_{M_1}^1 + \alpha_{M_2}^2]$. Обозначим $d_k,\ k = 0, 1, \ldots, K$ — точки деления этого интервала на K отрезков. Тогда вероятность попадания величины Z в интервал $[d_k, d_{k+1}]$ определяется по формуле

$$P_{z,k} = (\int_{\Omega_k} \int P_1(x)P_2(y)dxdy)/(d_{k+1} - d_k),$$

где $\Omega_k = \{(x,y) | d_k \le x + y \le d_{k+1}\}$.

Обозначим Π_{ij} — декартово произведение отрезка $[\alpha_i^1, \alpha_{i+1}^1]$, на $[\alpha_i^2, \alpha_{i+1}^2]$. Тогда

$$P_{z,k} = (\sum_{i=0}^{M_1-1} \sum_{j=0}^{M_2-1} \int_{\Pi_{ij}\cap\Omega_k} \int P_{1i}P_{2j}dxdy)/(d_{k+1} - d_k) = \qquad (2.4)$$

$$(\sum_{i=0}^{M_1-1} \sum_{j=0}^{M_2-1} P_{1i}P_{2j}|\Pi_{ij} \cap \Omega_k|)/(d_{k+1} - d_k).$$

Здесь $|\Pi_{ij} \cap \Omega_k|$ — площадь пересечения прямоугольника Π_{ij} с Ω_k.

Очевидно, что точность восстановления функции плотности вероятности случайной величины Z зависит от дробности деления интервала ее значений. В практических расчетах для упрощения операций по резервированию памяти все промежуточные операции можно выполнять с фиксированным, но достаточно большим количеством точек деления. Для сокращения количества вычислений в формуле (2.4) отношение $|\Pi_{ij} \cap \Omega_k|/|\Pi_{ij}|$ можно заменить приближенной оценкой $(B_{ij} - A_{ij})/(C_2 - C_1)$, где

$$A_{ij} = \inf_{(x,y)\in\Pi_{ij}} (x + y) = \alpha_0^1 + \alpha_0^2,$$

$$B_{ij} = \sup_{(x,y)\in\Pi_{ij}} (x + y) = \alpha_{M_1}^1 + \alpha_{M_2}^2,$$

$$[C_1, C_2] = [A_{ij}, B_{ij}] \cap [d_k, D_{k+1}].$$

Проведем оценку точности восстановления функции плотности вероятности для операции сложения. Предположим, что точки $\{\alpha_m^l, m = 0, 1, \ldots, M_l\}$ образуют равномерное разбиение отрезков $[\underline{a}_l, \overline{a}_l]$, а точки $\{d_k, k = 0, \ldots, K\}$ осуществляют равномерное деление отрезка $[\underline{a}_0, \overline{a}_0]$, являющегося интервалом допустимых значений случайной величины $Z = X_1 + X_2$. Пусть $\nu = (\overline{a}_0 - \underline{a}_0)/K$ — длина каждого из интервалов $[d_k, d_{k+1}], k = 0, \ldots, K - 1$.

В соответствии с принятым подходом к определению значения плотности $P_{z,k}$ на интервале $[d_k, d_{k+1}]$ имеем

$$P_{Z,k} = \frac{1}{\nu} \int_{d_k}^{d_{k+1}} P_Z(v) dv,$$

где $P_Z(v)$ — точное значение функции плотности вероятности случайной величины Z. Поэтому

$$\min_{v\in[d_k,d_{k+1}]} P_Z(v) \le P_{Z,k} \le \max_{v\in[d_k,d_{k+1}]} P_Z(v).$$

В силу этих неравенств при $x \in [d_k, d_{k+1}]$ имеем

$$|P_Z(x) - P_{Z,k}| \le \max_{x,y\in[d_k,d_{k+1}]} |P_Z(x) - P_Z(y)|.$$

Воспользовавшись формулой (2.2), получаем

$$|P_Z(x) - P_Z(y)| \le \int_{-\infty}^{\infty} P_1(v)|P_2(x - v) - P_2(y - v)|dv \le M_2\omega(P_2)\nu,$$

где $\omega(P_2)\max_{|v|\le\nu,x\in[\underline{a}_2,\overline{a}_2]} P_2(x+v)-P_2(x)$. Аналогично можно показать, что

$$|P_Z(x)-P_Z(y)|\le M_1\omega(P_1)\nu.$$

Обозначим $C(+,P_1,P_2)=\min\{M_1\omega(P_1),M_2\omega(P_2)\}$, тогда для любого интервала $[d_k,d_{k+1}]$

$$|P_Z(x)-P_{Z,k}|\le C(+,P_1,P_2)\nu.$$

Поэтому для того, чтобы построенная кусочно-постоянная аппроксимация функции $P_Z(x)$ отличалась не более чем на заданную величину ε, количество точек равномерного деления интервала $[\underline{a}_0,\overline{a}_0]$ должно удовлетворять неравенству

$$K\ge C(+,P_1,P_2)(\overline{a}_0-\underline{a}_0)/\varepsilon.$$

Аналогичные оценки можно получить и для других арифметических операций. Если плотности вероятности случайных величин X_1 и X_2 уже являются результатом промежуточных действий или кусочно-постоянной аппроксимацией гладких плотностей распределений с конечным носителем, то для оценки точности гистограммных операций поступаем следующим образом. Вводим отклонения этих величин от точных распределений $d_i=P_i-P_i^0$, где P_i^0 — точное значение плотности вероятности, а P_i — кусочно-постоянная аппроксимация. Величины d_i удовлетворяют соотношению $|d_i|\le C_i\nu_i$, где ν_i — длина интервала разбиения величины X_i; C_i — соответствующая постоянная.

Точное значение плотности вероятности случайной величины X_1+X_2, согласно формуле (2.2), равно

$$P^0_{X_1+X_2}=\int_{-\infty}^{\infty}P_1^0(v)P_2^0(x-v)dv=$$

$$=\int_{-\infty}^{\infty}(P_1^0(v)+d_1(v))(P_2^0(x-v)+d_2(x-v))dv=$$

$$=\int_{-\infty}^{\infty}P_1(v)P_2(x-v)dv+\int_{-\infty}^{\infty}d_1(v)P_2(x-v)dv+$$

$$+\int_{-\infty}^{\infty}P_1(v)d_2(x-v)dv+\int_{-\infty}^{\infty}d_1(v)d_2(x-v)dv.$$

Обозначим через $\overline{P}_{X_1+X_2}$ кусочно-постоянную аппроксимацию выражения $\int_{-\infty}^{\infty}P_1(v)P_2(x-v)dv$. Тогда

$$|P^0_{X_1+X_2}-\overline{P}_{X_1+X_2}|\le|P^0_{X_1+X_2}-\int_{-\infty}^{\infty}P_1(v)P_2(x-v)dv|+$$

$$+|\int_{-\infty}^{\infty} P_1(v)P_2(x-v)dv - \overline{P}_{X_1+X_2}|.$$

Второе слагаемое в этой формуле оценивается приведенным выше методом. Для первого слагаемого справедливы соотношения

$$|P^0_{X_1+X_2} - \int_{-\infty}^{\infty} P_1(v)P_2(x-v)dv|$$

$$\leq C_1\nu_1 + C_2\nu_2 + C_1\nu_1 C_2\nu_2 \min\{|\text{ supp }\ P_1|, |\text{ supp }\ P_2|\},$$

где $|\text{ supp }\ P_i|$ — мера носителя функции P_i.

Используя алгоритм расчета функции (2.1) и приведенную выше технику оценки точности, получаем оценку отклонения кусочно-постоянной аппроксимации, найденной с помощью гистограммной арифметики, от точного значения плотности вероятности. Эта оценка будет зависеть от количества точек деления интервалов значений промежуточных величин. Используя эту зависимость, можно вычислить точность промежуточных операций.

В качестве примера рассмотрим решение следующей простейшей системы линейных алгебраических уравнений:

$$\begin{aligned} a_{1,1}x_1 \ + \ a_{1,2}x_2 &= b_1, \\ a_{2,2}x_2 &= b_2, \end{aligned} \tag{2.5}$$

где $a_{1,1}, a_{2,2} = [1,2]$, $a_{1,2} = [0,1]$, $b_1, b_2 = [1,2]$. Гистограммы x_1, x_2 приведены на рис.2.1.

Как видно из приведенного рисунка, использование гистограммной арифметики позволяет определить наиболее вероятные участки попадания неизвестных.

2.3. Интервальные расширения

В этом разделе мы будем рассматривать непрерывные вещественные функции. Примем при этом, что все функции, с которыми мы будем иметь дело, можно вычислить используя конечное число арифметических операций и операндов. Такие функции мы в дальнейшем будем называть *рациональными*. Одна и та же рациональная функция f может иметь несколько аналитических представлений.

Интервально-значная функция $\boldsymbol{f}$ называется *монотонной по включению*, если для векторов $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{R}^n$ из соотношения $\boldsymbol{a} \subset \boldsymbol{b}$ вытекает,

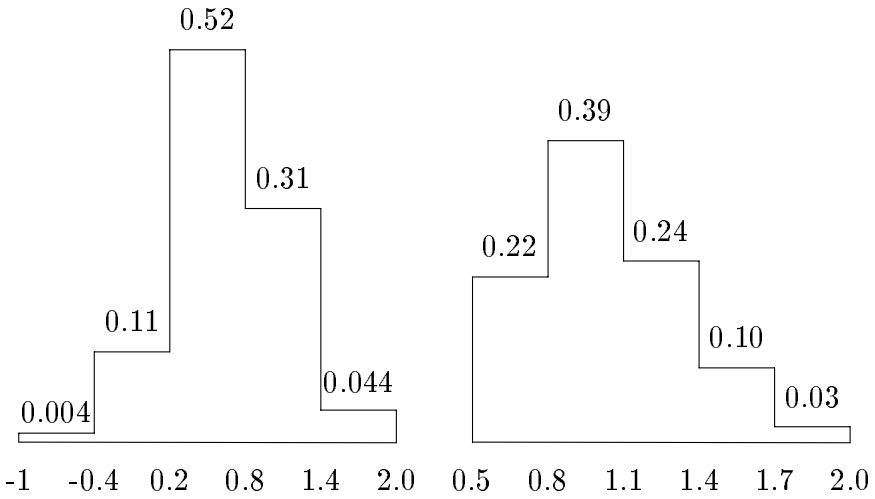

Рис. 2.1.

что

$$\boldsymbol{f}(a) \subset \boldsymbol{f}(b).$$

Заметим, что согласно методу математической индукции из монотонности по включению для интервальных операций это свойство справедливо для любого рационального выражения $\boldsymbol{f}(\boldsymbol{x})$, содержащего переменные $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ и интервальные константы $\boldsymbol{c}_1, ..., \boldsymbol{c}_m$.

Пусть $f(x)$ — вещественная функция, непрерывная в области $\boldsymbol{a} \subset \boldsymbol{R}^n$. *Объединенным расширением* этой функции мы будем называть интервальную функцию $\boldsymbol{f}_{un}(\boldsymbol{x})$ переменных $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \subset \boldsymbol{a}$, задаваемую равенством

$$\boldsymbol{f}_{un}(\boldsymbol{x}) = \bigcup_{x \in \boldsymbol{x}} f(x). \tag{2.6}$$

Непрерывность функции f является гарантией того, что при каждом $\boldsymbol{x}$ правая часть в (2.6) будет конечным отрезком.

Как видно из (2.6), объединенное расширение монотонно по включению. Кроме того оно минимально из всех возможных интервальных расширений.

Назовем *интервальным расширением* вещественной функции $f(x)$, $x \in D \subset R^n$, интервальную функцию $\boldsymbol{f}$, $\boldsymbol{x} \in R^n$ такую, что

$$f(x) = \boldsymbol{f}(x), \quad \forall x \in D.$$

Тогда для любого монотонного по включению интервального расширения $\boldsymbol{f}$ имеет место включение

$$\boldsymbol{f}_{un}(\boldsymbol{x}) \subset \boldsymbol{f}(\boldsymbol{x}), \quad \boldsymbol{x} \subset D.$$

Рассмотрим вещественную рациональную функцию $f(x), x \in R^n$. Для нее интервальное расширение можно получить естественным путем, если заменить аргументы x_i и арифметические операции соответственно интервальными числами и операциями. Как уже говорилось, полученное таким образом *естественное расширение* $\mathbf{f}_{ne}(\mathbf{x})$ будет монотонно по включению. Поэтому

$$\mathbf{f}_{un}(\mathbf{x}) \subset \mathbf{f}_{ne}(\mathbf{x})$$

для любых $\mathbf{x} \in \mathbf{R}^n$, для которых определена правая часть.

Отметим, что естественное интервальное расширение существенно зависит от способа записи рационального выражения.

Существуют способы представления рациональных выражений когда естественное интервальное расширение совпадает с объединенным [90].

Теорема 13. *Пусть $f(x_1, .., x_n)$ — рациональное выражение, в котором каждая переменная встречается не более одного раза и только в первой степени, $\boldsymbol{f}_{ne}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ — его естественное расширение. Тогда*

$$\boldsymbol{f}_{un}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \boldsymbol{f}_{ne}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$$

для любого набора $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, такого, что $\boldsymbol{f}_{ne}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ имеет смысл. □

В случае, когда не удается достигнуть ситуации, когда каждая переменная встречается только один раз, по крайней мере надо стремиться к уменьшению числа вхождений каждой переменной. Это обычно приводит к уменьшению ширины интервального расширения.

Суперпозиции рациональных и элементарных функций охватывают подавляющую часть нужных нам применений. Мы будем называть *естественным интервальным расширением* таких выражений интервальные функции, в которых вещественные аргументы заменены интервальными числами, а вещественные арифметические операции — интервальными операциями.

Пусть $\mathcal{F}$ — множество вещественных функций, для которых мы можем строить объединенные интервальные расширения. В это множество мы включим все элементарные функции $\{\sin, \cos, \ln, \exp, \ldots\}$, их рациональные комбинации и суперпозиции.

Рассмотрим обобщение теоремы 13. Для $f \in \mathcal{F}$ предположим, что мы можем представить $f(x_1, \ldots, x_n)$ как рациональное выражение от некоторых функций $g_i(x_k, \ldots, x_l)$.

Пусть функции g_i обладают следующими свойствами:

1) для каждой функции g_i построено объединенное расширение;

2) наборы переменных для разных функций g_i попарно не пересекаются.

Тогда, сделав замену переменных $z_i = g_i(x_k, \ldots, x_l)$, мы попадаем в область действия теоремы 13 и можем построить для f объединенное расширение.

Пример 1. Рассмотрим функцию двух переменных $f(x, y) = xy + x + y + 1$. Естественное интервальное расширение этой функции не будет совпадать с объединенным, поскольку переменные x, y встречаются более одного раза. В качестве функций g_i можно взять $g_1(x) = x + 1$, $g_2(y) = y + 1$. Тогда $f(x, y) = g_1(x)g_2(y)$ и, соответственно, $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{g}_1(\boldsymbol{x})\boldsymbol{g}_2(\boldsymbol{y})$.

Заметим, что представление функции f со свойствами 1), 2) в большинстве случаев невозможно. Однако мы будем стремиться выполнить эти свойства как можно полнее. Естественно предположить, что таких представлений будет несколько. Обозначим через $\boldsymbol{f}_j$ различные интервальные расширения, порожденные этими представлениями. Тогда в качестве наилучшего из возможных интервальных расширений мы можем взять

$$\boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \cap_j \boldsymbol{f}_j(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n).$$

Этот простой прием на практике позволяет существенно снижать ширину интервальных расширений.

Заметим, что естественные интервальные расширения, как правило, имеют значительно более широкие интервалы в сравнении с объединенными расширениями. Рассмотрим прием уменьшения ширины интервальных вычислений, описанный в [89]. Идею поясним на примере функций одной переменной.

Пусть f: $\boldsymbol{a} \to R$ непрерывная функция, а $\boldsymbol{f}_{ne}(\boldsymbol{x})$ — ее естественное интервальное расширение. Разобьем $\boldsymbol{a}$ на p интервалов равной длины

$$\boldsymbol{a} = \bigcup_{i=1}^{p} \boldsymbol{a}_i, \quad \mathrm{wid}(\boldsymbol{a}_i) = \mathrm{wid}(\boldsymbol{a})/p.$$

Тогда интервал $\boldsymbol{b} = \bigcup\limits_{i=1}^{p} \boldsymbol{f}_{ne}(\boldsymbol{a}_i)$ по-прежнему будет содержать объединенное расширение $\boldsymbol{f}_{un}(\boldsymbol{a})$, но, как правило, будет иметь меньшую ширину, чем $\boldsymbol{f}_{ne}(\boldsymbol{a}_i)$:

$$\boldsymbol{f}_{un}(\boldsymbol{a}) \subset \mathbf{b} \subset \boldsymbol{f}_{ne}(\boldsymbol{a}). \tag{2.7}$$

Более того, при $p \to \infty$

$$\mathrm{wid}(\boldsymbol{b}) - \mathrm{wid}(\boldsymbol{f}_{un}(\boldsymbol{a})) = O(1/p). \tag{2.8}$$

В принципе этот прием можно распространить и на большие размерности, в том числе останутся справедливыми свойства (2.7) и (2.8). Но для n переменных вектор $(\boldsymbol{a}_1, ..., \boldsymbol{a}_n)$ будет подразделяться уже на p^n равных частей. Поэтому при практических вычислениях этот прием используется только для небольших размерностей.

Если известны интервальные расширения первых производных, то можно построить интервальные расширения следующим образом.

Пусть, например, f: $R^n \to R$ непрерывно дифференцируемая функция на $\boldsymbol{a} \in R^n$ и $\boldsymbol{g}_i$ — интервальные расширения первых производных $g_i = \partial f / \partial x_j$. Пусть $c = \mathrm{med}(\boldsymbol{x})$, где $\mathbf{x} \in \mathbf{R}^n$. Тогда для всех $x \in \boldsymbol{x}$

$$f(x) = f(c) + \sum_{j=1}^{n} g_j(\xi)(x_j - c_j), \quad \xi \in \boldsymbol{x}.$$

Заменяя производные их интервальными расширениями, получим соотношение

$$\boldsymbol{f}(\boldsymbol{x}) \subset \boldsymbol{f}(c) + \sum_{j=1}^{n} \boldsymbol{g}_j(\boldsymbol{x})(\boldsymbol{x}_j - c_j),$$

его правая часть определяет интервальную функцию

$$\boldsymbol{f}_{mv}(\boldsymbol{x}) = \boldsymbol{f}(c) + \sum_{j=1}^{n} \boldsymbol{g}_j(\boldsymbol{x})(\boldsymbol{x}_j - c_j),$$

которая обычно называется mv-формой (mean value form [90]). Справедлив следующий результат [75].

Теорема 14. *Пусть производные* $\boldsymbol{g}_j$ *для всех* $j = 1, ..., n$ *удовлетворяют условию Липшица на* $\boldsymbol{a}$, *и при* $\mathrm{wid}(\boldsymbol{x}) \to 0$

$$\mathrm{wid}(\boldsymbol{g}_j(\boldsymbol{x}) - \mathrm{wid}(\boldsymbol{g}_{j,un}(\boldsymbol{x})) = O(\mathrm{wid}(\boldsymbol{x})) \quad \forall j = 1, ..., n \quad \text{и} \quad \forall \mathbf{x} \subset \boldsymbol{a}. \tag{2.9}$$

Тогда

$$\mathrm{wid}(\boldsymbol{f}_{mv}(\boldsymbol{x}) - \mathrm{wid}(\boldsymbol{f}_{un}(\boldsymbol{x})) = O(\mathrm{wid}^2(\boldsymbol{x})) \quad \forall \boldsymbol{x} \subset \boldsymbol{a}. \tag{2.10}$$

Если вместо (2.9) справедлива более грубая оценка

$$\mathrm{wid}(\boldsymbol{g}_j(\mathbf{x})) \le c \quad \forall j = 1, ..., n, \quad \forall \boldsymbol{x} \subset \boldsymbol{a}, \tag{2.11}$$

то вместо (2.10) выполняется соотношение

$$\mathrm{wid}(\boldsymbol{f}_{mv}(\boldsymbol{x}) - \mathrm{wid}(\boldsymbol{f}_{un}(\boldsymbol{x})) = O(\mathrm{wid}(\boldsymbol{x})) \quad \forall \boldsymbol{x} \subset \boldsymbol{a}. \square$$

Таким образом, имея интервальные расширения производных, можно получать mv-форму функции, обладающей большей асимптотической точностью, чем исходные расширения производных.

Ниже мы рассмотрим подход, позволяющий строить близкие к оптимальным интервальные расширения для некоторых классов функций.

2.4. Интервальные расширения полиномов многих переменных

В этом разделе изложен алгоритм нахождения интервальных расширений полиномов многих переменных, приведенный в работе [33].

Пусть $f(x),\ x \in R^n$ — вещественная рациональная функция. Для нее интервальное расширение можно получить естественным путем, если заменить переменные x_i и арифметические операции — интервальными переменными и операциями. Полученное таким образом интервальное расширение будет монотонным по включению.

Рассмотрим метод нахождения интервальных расширений для функций вида

$$f(x) = \sum_{i,j=0}^{2} a_{ij} x_i x_j,$$

где $x_0 = 1$ и $a_{ij} \in \boldsymbol{a}_{ij}, x_i \in \boldsymbol{x}_i$. Далее обозначим $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Необходимость нахождения оптимальных расширений для полиномиальных функций возникает в различных приложениях, например, при решении систем нелинейных уравнений методом простой итерации, при нахождении двусторонних решений дифференциальных уравнений с правыми частями, имеющими указанный вид. Наиболее часто такие задачи встречаются в химической кинетике.

Заметим, что построение интервального расширения $\mathbf{f}$ можно свести к нахождению $\min_{\boldsymbol{x}} f, \max_{\boldsymbol{x}} f$ и представления

$$\boldsymbol{f} = [\min_{\boldsymbol{x}} f, \max_{\boldsymbol{x}} f].$$

Найдем $\max_{\boldsymbol{x}} f$. Основная идея метода заключается в том, чтобы попытаться найти переменные x_i, для которых

$$0 \notin \partial_{x_i} \boldsymbol{f}(\boldsymbol{x}). \tag{2.12}$$

Тогда, в зависимости от знака $\partial_{x_i} \boldsymbol{f}(\boldsymbol{x})$, значение $\max_{\boldsymbol{x}} f$ будет достигаться для граничных значений x_i. В силу приведенной выше теоремы

13, естественное интервальное расширение для f будет совпадать с объединенным.

Пусть $\mathcal{I}$: $\{i|wid(x_i) \neq 0\}$ — множество индексов, таких, что $i \in \mathcal{I} \Rightarrow wid(x_i) \neq 0$. Циклично проверяем переменные $x_i, i \in \mathcal{I}$. Для тех переменных x_i, по которым удалось установить условия (2.12), производим замену $\boldsymbol{x}_i$ на $\underline{x}_i$ или $\overline{x}_i$ в соответствии со знаком $\partial_{x_i}\boldsymbol{f}(\boldsymbol{x})$ и исключаем их индекс из $\mathcal{I}$. Цикл продолжается до тех пор, пока условие (2.12) выполняется хотя бы один раз на множестве $\mathcal{I}$ и идет сужение интервалов. Тем самым мы окончательно формируем $\boldsymbol{x}^* \subseteq x$.

Выберем переменные x_i, по которым выполнены два условия:

а) f не содержит квадрат данной переменной;

б) x_i^* – имеет ненулевую ширину.

Преобразуем функцию f. При переменной такого вида можно привести подобные члены так, чтобы она встречалась не более одного раза и только в первой степени. Тогда она не будет давать вклада в ошибку при вычислении интервального расширения на интервале x^*. Может оказаться, что при всех выделенных таким образом переменных одновременно привести подобные члены не удается. Тогда приводим подобные члены при тех переменных, при которых это возможно сделать одновременно, и получаем функцию $\hat{f}$.

Далее вычислим $\bar{f}$ – оценку максимума функции $\hat{f}$ на x^*. Эту оценку удобно сделать с помощью метода Волкова [20], так как в нем используются оценки с помощью парабол по каждой переменной. В нашем случае параболы точно отражают поведение функции. В методе используется информация о вторых производных, которые для рассматриваемой функции f являются фиксированными интервалами.

Зададимся точностью ε, нахождения максимума функции $\hat{f}$. Пусть f_0 – значение функции $\hat{f}$ в середине интервала x^*. Если

$$|f_0 - \bar{f}| < \varepsilon,$$

то счет окончен. В противном случае мы можем воспользоваться алгоритмом деления большей стороны интервала пополам или небольшой его модернизацией: деление на три части, т.е. представление $x^* = x_1 \cup x_2 \cup x_3$. Далее для каждого x_i проделываем описанные выше процедуры и определяем, в каком x_i находится точка максимума функции. Алгоритм заканчивает свою работу в одном из следующих случаев:

1) множество $\mathcal{I}$ пусто;

2) ширина большей грани x^* меньше ε —

$$\max_{i\in\mathcal{I}} wid(x_i^*) < \varepsilon.$$

3) $\bar{f}$ содержит каждую переменную $x_i, i \in \mathcal{I}$ только один раз и только в первой степени.

Рассмотрим несколько примеров.

Пример 2. Пусть дана функция $f = 2x^2 - 2xy + y + 2x$. Требуется найти ее интервальное расширение на интервале $\boldsymbol{x} = [0,1]\times[0,1]$. Найдем естественные интервальные расширения производных $f'_x = [0,6]$, $f'_y = [-1,1]$. Таким образом, для переменной x_1 на интервале $\boldsymbol{x}$ есть монотонность. Следовательно, $x^* = 1 \times [0,1]$. Вычисляя f'_y, получаем $f'_y = -1$. Новый интервал $x^* = 1 \times 0$ имеет нулевую ширину, точка $(1,0)$ — точка максимума функции f. Аналогично находится точка минимума. Таким образом, $\boldsymbol{f}(\boldsymbol{x}) = [0,4]$.

Пример 3. Пусть дана функция $f = 2xy + z^2 + xz - 2x + 2z$. Требуется найти ее интервальное расширение на интервале $\boldsymbol{x} = [-1,1] \times [0,1] \times [0,1]$. Как и в примере 2, находим интервальные расширения частных производных $\boldsymbol{f}'_x = [-2,1]$, $\boldsymbol{f}'_y = [-2,2]$, $\boldsymbol{f}'_z = [1,5]$. Тогда $\boldsymbol{x}^* = [-1,1]\times[0,1]\times 1$. Повторно вычисляем $\boldsymbol{f}'_x = [-1,1]$, $\boldsymbol{f}'_y = [-2,2]$. Сужение интервалов дальше не происходит.

Переменная x удовлетворяет условиям алгоритма: не содержит квадрата и $wid(\boldsymbol{x}^*) = 2$. Приводим подобные члены по x. Получаем функцию $f = x(2y + z - 2) + z^2 + 2z$, естественное интервальное расширение этой функции на интервале $\boldsymbol{x}^* = [-1,1]\times[0,1]\times 1$ совпадает с объединенным, поскольку все интервальные переменные встречаются только один раз и только в первой степени (z не интервальная переменная).

Заметим, что алгоритм построения интервального расширения можно распространить и на более общие случаи полиномов степени m от n переменных. Покажем как это можно сделать на примере полинома третьей степени от трех переменных

$$f(x) = \sum_{i,j,k=0}^{3} a_{ijk} x_i x_j x_k.$$

В п.1 алгоритма объединенное интервальное расширение уже не будет совпадать с их естественным интервальным расширением. Вследствие этого наличие монотонности по некоторым переменным может оказаться невыявленным. Для уточнения значений производных заметим, что

они — полиномы второй степени и необходимо установить либо отрицательность значения максимума частной производной на интервале $\boldsymbol{x}$, либо положительность значения ее минимума на интервале $\mathbf{x}$. Ставя задачу таким образом, мы сводим ее к уже упомянутому случаю — нахождению интервального расширения для полиномов второй степени.

Таким образом, можно рекурсивно понижать степень полиномов и добиваться значительного сужения ширины интервальных расширений.

Описанный выше подход можно распространить на случай произвольных вещественных функций, для которых известны интервальные расширения $\boldsymbol{g}_i$. Предположим, что известны индексы i, при которых

$$0 \notin \mathbf{g_i}(\mathbf{x}). \tag{2.13}$$

Это означает, что по этим переменным функция монотонна и, следовательно, нижняя и верхняя границы функции достигаются при граничных значениях интервала. Исходя из этого мы можем положить $\underline{f}(\boldsymbol{x}) = \underline{f}(\boldsymbol{z}_1), \overline{f}(\boldsymbol{x}) = \overline{f}(\boldsymbol{z}_2)$, где

$$\boldsymbol{z}_{1,i} = \begin{cases} \overline{x}_i, & \text{если } \boldsymbol{g}_\mathrm{i} > 0; \\ \underline{x}_i, & \text{если } \boldsymbol{g}_\mathrm{i} < 0; \\ \boldsymbol{x}_i, & \text{если } \boldsymbol{g}_\mathrm{i} \ni 0, \end{cases}$$

$$\boldsymbol{z}_{2,i} = \begin{cases} \overline{x}_i, & \text{если } \boldsymbol{g}_\mathrm{i} < 0; \\ \underline{x}_i, & \text{если } \boldsymbol{g}_\mathrm{i} > 0; \\ \boldsymbol{x}_i, & \text{если } \boldsymbol{g}_\mathrm{i} \ni 0. \end{cases}$$

Таким образом, вычисление каждой из границ интервального расширения сводится к вычислению интервальных расширений той же функции, но при более узких значениях интервального аргумента. Далее к каждому такому вычислению интервального расширения мы можем вновь применить одну из перечисленных выше процедур.

Пример 4. Рассмотрим следующую функцию трех переменных, заданную на $[0,1]^3$, $a = b = c = 0.5$:

$$f(x_1, x_2, x_3) = ax_1x_2x_3 + 2x_1x_2 + 2bx_2x_3 + cx_1 - (1-b)x_2 - bx_3.$$

Несложно подсчитать, что

$$\partial_1 \boldsymbol{f}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = [c, 2 + a + c] > 0,$$

$$\partial_2 \boldsymbol{f}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = [-1 - b, 1 + a + b] \ni 0,$$

$$\partial_3 \boldsymbol{f}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = [-b, a + b] \ni 0.$$

Поскольку $f(x_1, x_2, x_3)$ монотонно возрастает по x_1, то для нахождения $\overline{f}$ мы можем положить $x_1 = \overline{x}_1 = 1$. Относительно поведения функции по второму и третьему аргументам мы ничего определенного сказать не можем. В этом случае

$$\partial_2 f(\overline{x}_1, \mathbf{x_2}, \mathbf{x_3}) = [1 - b, 1 + a + b] > 0,$$

$$\partial_3 f(\overline{x}_1, \mathbf{x_2}, \mathbf{x_3}) = [-b, a + b] \ni 0.$$

Далее при $x_1 = \overline{x}_1 = 1$ в силу положительности частной производной по второму аргументу $f(x_1, x_2, x_3)$ возрастает по x_2, следовательно, можно положить $x_2 = \overline{x}_2 = 1$. При фиксированных значениях $x_1 = 1, x_2 = 1$

$$\partial_3 f(\overline{x}_1, \overline{x}_2, \mathbf{x_3}) = a + b > 0.$$

Таким образом, нам удалось последовательно построить верхнюю границу интервального расширения в виде

$$\overline{f}(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}) = f(\overline{x}_1, \overline{x}_2, \overline{x}_3).$$

Аналогично можно построить и нижнюю границу

$$\underline{f}(\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}) = f(\underline{x}_1, \underline{x}_2, \underline{x}_3).$$

Несложно видеть, что построенное интервальное расширение совпадает с объединенным интервальным расширением.

2.5. Методы минимизации функций

Метод золотого сечения

Пусть Ω — некоторое множество из R, $f(x)$ — функция, определенная на Ω и принимающая во всех точках Ω конечные значения. Будем рассматривать задачу минимизации функции $f(x)$ на Ω.

Определение 1. *Точку $x_* \in \Omega$ называют точкой минимума функции на Ω, если $f(x_*) \leq f(x), \forall x \in \Omega$; величину $f(x_*)$ называют наименьшим значением $f(x)$ на Ω. Множество всех точек минимума на Ω будем обозначать X_*.*

Определение 2. *Функцию $f(x)$ назовем унимодальной на отрезке $[a, b]$, если она непрерывна на $[a, b]$, и существуют числа α, β, $a \leq \alpha \leq \beta \leq b$ такие, что 1) $f(x)$ строго монотонно убывает при $a \leq x \leq \alpha$ (если $a < \alpha$); 2) $f(x)$ строго монотонно возрастает при $\beta \leq x \leq b$ (если $\beta < b$), 3) $f(x) = \inf_{x \in [a,b]} f(x)$ при $\alpha \leq x \leq \beta$.*

Случаи, когда один или два из отрезков $[a,\alpha]$, $[\alpha,\beta]$, $[\beta,b]$ вырождаются в точку, здесь не исключаются. В частности, если $\alpha=\beta$, то $f(x)$ назовем строго унимодальной на отрезке $[a,b]$.

Пример 5. Функция $f(x)=|x|+|x-1|-1$ унимодальна на любом отрезке $[a,b]$; функция $\sin^2(\pi/x)$ строго унимодальна на $[2/3,2]$, но не будет унимодальной на $[1/2,2]$.

Перейдем к описанию метода минимизации унимодальной функции на отрезке.

Как известно, золотым сечением отрезка называется деление отрезка на две неравные части так, чтобы отношение длины всего отрезка к длине большей части равнялось отношению длины большей части к длине меньшей части отрезка. Нетрудно проверить, что золотое сечение отрезка $[a,b]$ производится двумя точками $x_1=a+(3-\sqrt{5})(b-a)/2=a+0,381966011...(b-a)$ и $x_2=a+(\sqrt{5}-1)(b-a)/2=a+0,618033989...(b-a)$, расположенными симметрично относительно середины отрезка, причем $a<x_1<x_2<b$,

$$(b-a)/(b-x_1)=(b-x_1)/(x_1-a)=(b-a)/(x_2-a)=(x_2-a)/(b-x_2)=$$

$$=(\sqrt{5}+1)/2=1,618033989...$$

Замечательно здесь то, что точка x_1 в свою очередь производит золотое сечение отрезка $[a,x_2]$, так как $x_2-x_1<x_1-a=b-x_2$ и $(x_2-a)/(x_1-a)=(x_1-a)/(x_2-x_1)$. Аналогично точка x_2 производит золотое сечение отрезка $[x_1,b]$. Опираясь на это свойство золотого сечения, можно предложить следующий метод минимизации унимодальной функции $f(x)$ на отрезке $[a,b]$.

Положим $a_1=a, b_1=b$. На отрезке $[a_1,b_1]$ возьмем точки $[x_1,x_2]$, производящие золотое сечение, и вычислим значения $f(x_1)$, $f(x_2)$. Далее, если $f(x_1)\le f(x_2)$, то примем $a_2=a_1$, $b_2=x_2$, $\overline{x}_2=x_1$; если же $f(x_1)>f(x_2)$, то $a_2=x_1$, $b_2=b_1$, $\overline{x}_2=x_2$. Так как функция $f(x)$ унимодальна на $[a,b]$, то отрезок $[a_2,b_2]$ имеет хотя бы одну общую точку со множеством X_*, точек минимума $f(x)$ на $[a,b]$. Кроме того, $b_2-a_2=(\sqrt{5}-1)(b-a)/2$, и весьма важно то, что внутри $[a_2,b_2]$ содержится точка $\overline{x}_2$ с вычисленным значением $f(\overline{x}_2)=\min\{f(x_1);f(x_2)\}$, которая производит золотое сечение отрезка $[a_2,b_2]$.

Пусть уже определены точки $x_1,x_2,...,x_{n-1}$, вычислены значения $f(x_1)$, $f(x_2)$,...,$f(x_{n-1})$, найден отрезок $[a_{n-1},b_{n-1}]$ такой, что $[a_{n-1},b_{n-1}]\cap X_*\ne\emptyset$, и известна точка $\overline{x}_{n-1}$, производящая золотое сечение отрезка

$[a_{n-1}, b_{n-1}]$ и такая, что $f(\overline{x}_{n-1}) = \min_{1 \le i n-1} f(x_i), n \ge 2$. Тогда в качестве следующей точки возьмем точку $x_n = a_{n-1} + b_{n-1} - \overline{x}_{n-1}$, также производящую золотое сечение отрезка $[a_{n-1}, b_{n-1}]$, вычислим значение $f(x_n)$. Пусть для определенности $a_{n-1} < x_n < \overline{x}_{n-1} < b_{n-1}$ (случай $\overline{x}_{n-1} < x_n$ рассматривается аналогично). Если $f(x_n) \le f(\overline{x}_{n-1})$, то полагаем $a_n = a_{n-1}$, $b_n = \overline{x}_{n-1}$, $\overline{x}_n = x_n$, если же $f(x_n) > f(\overline{x}_{n-1})$, то $a_n = x_n$, $b_n = b_{n-1}$, $\overline{x}_n = \overline{x}_{n-1}$, Новый отрезок $[a_n, b_n]$ таков, что $[a_n, b_n] \cap X_* \ne \emptyset$, $b_n - a_n = ((\sqrt{5} - 1)/2)^{n-1}(b - a)$, точка $\overline{x}_n$ производит золотое сечение $[a_n, b_n]$ и $f(\overline{x}_n) = \min_{1 \le i \le n} f(x_i)$.

Если число вычислений значений $f(x)$ заранее не ограничено, то описанный процесс можно продолжать, например, до тех пор, пока не выполнится неравенство $b_n - a_n < \varepsilon$, где ε — заданная точность. Если же число вычислений значений функции $f(x)$ заранее жестко задано и равно n, то процесс на этом заканчивается.

Обсудим возможности численной реализации метода золотого сечения на ЭВМ. Заметим, что число $\sqrt{5}$ на ЭВМ неизбежно будет задаваться приближенно, поэтому первая точка x_1 будет найдена с некоторой погрешностью. Эта погрешность с возрастанием n растет очень быстро [12]. Поэтому использование направленных округлений при вычислении $[a_n, b_n]$ будет гарантировать выполнение условия $[a_n, b_n] \cap X_* \ne \emptyset$.

Имеется достаточно простая модификация метода золотого сечения, позволяющая избежать слишком быстрого возрастания погрешностей при определении точек x_n, а именно на каждом отрезке $[a_n, b_n]$ при выборе следующей точки x_n нужно остерегаться пользоваться формулой $x_{n+1} = a_n + b_n - \overline{x}_n$ и вместо этого лучше непосредственно произвести золотое сечение отрезка $[a_n, b_n]$.

Метод ломаных

Описанные выше методы часто приходится применять без априорного знания о том, что минимизируемая функция является унимодальной. Однако в этом случае погрешности в определении минимального значения и точек минимума функции могут быть значительными. Например, применение этих методов к минимизации непрерывных на отрезке функций приведет, вообще говоря, лишь в окрестность точки локального минимума, в которой значение функции может очень отличаться от искомого минимального значения на отрезке. Поэтому представляется важным разработка методов поиска глобального минимума, позволяю-

щих строить минимизирующие последовательности и получить гарантированные решения задач минимизации для функций, не обязательно унимодальных.

Здесь мы рассмотрим один из таких методов для класса функций, удовлетворяющих условию Липшица на отрезке. Напомним

Определение 3. *Говорят, что функция $f(x)$ удовлетворяет условию Липшица на отрезке $[a,b]$, если существует постоянная $\boldsymbol{L} > 0$ такая, что*

$$f(x) - f(y) \in \boldsymbol{L}(x-y),\ x,y \in [a,b]. \tag{2.14}$$

Постоянную $\boldsymbol{L}$ называют константой Липшица функции $f(x)$ на $[a,b]$.

Условие (2.14) имеет простой геометрический смысл: оно означает, что угловой коэффициент (тангенс угла наклона) хорды, соединяющей две точки $(x,f(x))$, $(y,f(y))$ графика функции, содержится в $\mathbf{L}$ для всех точек $x,y,\in [a,b]$. Из (2.14) следует, что функция $f(x)$ непрерывна на отрезке $[a,b]$, так что [12] множество X_* точек минимума $f(x)$ на $[a,b]$ непусто [12].

Теорема 15. *Пусть функция $f(x)$ непрерывна на отрезке $[a,b]$ и на каждом, отрезке $\mathbf{a}_i = [a_i, a_{i+1}]$, $[a,b] = \cup_i \mathbf{a}_i$, удовлетворяет условию (2.14) с константой $\mathbf{L}_i$. Тогда $f(x)$ удовлетворяет условию (2.14) на всем отрезке с константой $\mathbf{L} = \cup_i \mathbf{L}_i$.*

Теорема 16. *Пусть функция $f(x)$ дифференцируема на отрезке $[a,b]$ и ее производная $f'(x)$ на этом отрезке ограничена. Тогда $f(x)$ удовлетворяет условию (2.14) с константой $\boldsymbol{L} = f'([a,b])$.*

Пусть функция $f(x)$ удовлетворяет условию (2.14) на отрезке $[a,b]$. Зафиксируем какую-либо точку $v \in [a,b]$ и определим функцию

$$l(x,v) = \begin{cases} f(v) + \overline{L}x - v, & x \in [a,v], \\ f(v) + \underline{L}x - v, & x \in [v,b]. \end{cases}$$

Очевидно, функция $l(x,v)$ кусочно-линейна на $[a,b]$, и график ее — ломаная линия, составленная из отрезков двух прямых, имеющих угловые коэффициенты $\overline{L}$ и $\underline{L}$ и пересекающихся в точке $(v,f(v))$. Кроме того, в силу условия (2.14)

$$f(x) \geq l(x,v), \forall x \in [a,b]. \tag{2.15}$$

Это значит, что график функции $f(x)$ лежит выше ломаной $l(x,v)$ при всех и$x \in [a,b]$ и имеет с ней общую точку $(v, f(v))$.

Свойство (2.15) ломаной $l(x,v)$ можно использовать для построения следующего метода, который назовем методом ломаных. Этот метод начинается с выбора произвольной точки $x_0 \in [a,b]$ и составления функции $l_0(x,x_0) = l(x,x_0)$. Следующая точка x_1 определяется из условий $l_0(x_1) = \min_{x\in[a,b]} l_0(x)$. Далее берется новая функция $l_1(x) = \max\{l(x,u_1), l_0(x)\}$ и очередная точка x_2 находится из условий $l_1(x_2) = \min_{x\in[a,b]} l_1(x)$, $x_2 \in [a,b]$ и т.д.

Пусть точки $x_0, x-1, ..., x_n$ уже известны. Тогда составляется функция

$$l_n(x) = \max\{l(x,x_n), l_{n-1}\} = \max_{0\le\le n} l(x,x_n)$$

и следующая точка x_{n+1} определяется условиями

$$l_n(x_{n+1}) = \min_{x\in[a,b]} l_n(x), x_{n+1} \in [a,b].$$

Если минимум $l_n(x)$ на $[a,b]$ достигается в нескольких точках, то в качестве x_{n+1} можно взять любую из них. Метод ломаных описан.

Очевидно, $l_n(x)$ является кусочно-линейной функцией и график ее представляет собой непрерывную ломаную линию, состоящую из отрезков прямых с угловыми наклонами $\overline{L}$ или $\underline{L}$.

Таким образом, на каждом шаге метода ломаных задача минимизации функции $f(x)$ заменяется более простой задачей минимизации кусочно-линейной функции $l_n(x)$, которая приближает $f(x)$ снизу, причем последовательность функций $\{l_n(x)\}$ монотонно возрастает.

Теорема 17. *Пусть $f(x)$ — произвольная функция, удовлетворяющая на отрезке $[a,b]$ условию (2.14). Тогда последовательность x_n, полученная с помощью описанного метода ломаных, такова, что*

1) $\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} l_n(x_{n+1}) = f_* = \inf_{x\in[a,b]} f(x)$, *причем справедлива оценка*

$$0 \le f(x_n) - f_* \le f(x_n) - l_n(x_n), n = 1,2,...; \qquad (2.16)$$

2) $\{x_n\}$ *сходится ко множеству* X_* *точек минимума* $f(x)$ *на* $[a,b]$.

Таким образом, с помощью метода ломаных можно получить решение задач минимизации первого и второго типов для функций, удовлетворяющих условию (2.14). Проста и удобна для практического использования формула (2.16), дающая оценку неизвестной погрешности,

через известные величины, вычисляемые в процессе реализации метода ломаных. Этот метод не требует унимодальности минимизируемой функции, более того, функция может иметь сколько угодно точек локального экстремума на рассматриваемом отрезке. На каждом шаге метода ломаных нужно минимизировать кусочно-линейную функцию, что может быть сделано простым перебором известных вершин ломаной $l_n(x)$, причем здесь перебор существенно упрощается благодаря тому, что ломаная $l_n(x)$ отличается от ломаной $l_{n-1}(x)$ не более чем двумя новыми вершинами. К достоинству метода относится и то, что он сходится при любом выборе начальной точки x_0.

К недостаткам метода ломаных следует отнести то, что с увеличением числа шагов n растет требуемый объем памяти ЭВМ для хранения координат вершин ломаной $l_n(x)$. В следующем параграфе будет рассмотрен другой метод, по своей идее близкий к методу ломаных, но предъявляющий менее жесткие требования к объему памяти и более удобный для реализации на ЭВМ.

Метод касательных

Определение 4. *Функция $f(x)$, определенная на отрезке $[a,b]$, называется выпуклой на этом отрезке, если*

$$f(\alpha u + (1-\alpha)v) \le \alpha f(u) + (1-\alpha)f(v)$$

при всех $u, v \in [a,b]$, $\forall \alpha \in [0,1]$.

Пусть функция $f(x)$ выпукла и дифференцируема на отрезке $[a,b]$, такая функция удовлетворяет условию Липшица и унимодальна на $[a,b]$. Поэтому для минимизации $f(x)$ на $[a,b]$ применимы почти все описанные выше методы, в частности метод ломаных. Однако если значения функции $f(x)$ и ее производной $f'(x)$ вычисляются достаточно просто, то можно предложить другой, вообще говоря, более эффективный вариант метода ломаных, когда в качестве звеньев ломаных берутся отрезки касательных к графику $f(x)$ в соответствующих точках.

Зафиксируем какую-либо точку $v \in [a,b]$ и определим функцию

$$l(x,v) = f(v) + f'(v)(x-v),\ x \in [a,b].$$

Согласно [12]

$$l(x,v) \le f(x), \forall x \in [a,b]. \tag{2.17}$$

В качестве начального приближения возьмем любую точку $x_0 \in [a,b]$ (например, $x_0 = a$), составим функцию $l_0(x) = l(x, x_0)$ и определим точку $x_1 \in [a,b]$ из условия $l_0(x_1) = \min_{x\in[a,b]} l_0(x)$ (ясно, что при $f'(x_0) \neq 0$ будет $x_1 = a$ или $x_1 = b$). Далее, берем новую функцию $l_1(x) = \max\{l(x,x_1), l_0(x)\}$ и следующую точку $x_2 \in [a,b]$ найдем из условия $l_1(x_2) = \min_{x\in[a,b]} l_1(x)$ и т.д. Если точки $x_0, x-1, ..., x_n$ уже известны, то составляем функцию

$$l_n(x) = \max\{l(x,x_n), l_{n-1}\} = \max_{0\leq\leq n} l(x,x_n)$$

и следующую точку x_{n+1} определим из условий

$$l_n(x_{n+1}) = \min_{x\in[a,b]} l_n(x), x_{n+1} \in [a,b].$$

Если при каком-либо $n \geq 0$ окажется $f'(x_n + 0) \geq 0$, $f'(x_n - 0) \leq 0$ (если $a < x_n < b$, то это равносильно условию $f'(x_n) = 0$, то задача минимизации уже решена и итерации на этом заканчиваются.

Нетрудно видеть, что $l_n(x)$ — непрерывная кусочно-линейная функция и ее график представляет собой ломаную, состоящую из отрезков касательных к графику функции $f(x)$ в точках $x_0, x-1, ..., x_n$. Поэтому описанный метод естественно назвать методом касательных.

Теорема 18. *Пусть функция $f(x)$ на отрезке $[a,b]$ выпукла и дифференцируема, а последовательность $\{x_n\}$ получена описанным выше методом касательных, причем $x_n \notin X_*$, $n = 0, 1, ...$ Тогда*

1) $\lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} l_n(x_{n+1}) = f_* = \inf_{x\in[a,b]} f(x)$, *и справедлива оценка*

$$0 \leq f(x_n) - f_* \leq f(x_n) - l_n(x_n), n = 1, 2, ...;$$

2) $\{x_n\}$ *сходится ко множеству X_* точек минимума $f(x)$ на $[a,b]$ или, точнее, $\{x_n\}$ имеет не более двух предельных точек, совпадающих с $\underline{x}_* = \inf X_*$ или $\overline{x}_* = \sup X_*$.*

Метод касательных обладает всеми достоинствами метода ломаных. Недостаток этого метода: он применим лишь в случае, когда минимизируемая функция выпукла и значения функции и ее производных вычисляются достаточно просто.

Можно предложить более удобную для использования на ЭВМ вычислительную схему метода касательных, которая не требует хранения в машинной памяти информации обо всей ломаной $l_n(x)$. А именно возьмем $a_1 = a, b_1 = b$, вычислим $f'(a_1) = f'(a_1 + 0), f'(b_1) = f'(b_1 - 0)$. Если

$f'(a_1) \geq 0$ или $f'(b_1) \leq 0$, то по теореме 11.5 [12] $a \in X_*$ или $b \in X_*$ — задача решена. Поэтому пусть $f'(a_1) < 0$, $f'(b_1) > 0$, то согласно теореме 11.6 [12] означает $X_* \subset [a, b]$. Пусть отрезок $[a_{n-1}, b_{n-1}]$, $n \leq 2$, уже построен, причем $f'(a_{n-1}) < 0$, $f'(b_{n-1}) > 0$, $X_* \subset [a_{n-1}, b_{n-1}]$. Обозначим через x_n точку пересечения касательных $l(x, a_{n-1})$, $l(x, b_{n-1})$

Ясно, что $a_{n-1} \leq x_n \leq b_{n-1}$. Вычислим $f'(x_n)$. Если $f'(x_n) = 0$, то задача решена, итерации на этом заканчиваются. Если $f'(x_n) \neq 0$, то положим

$$a_n = \begin{cases} a_{n-1} & \text{if} f'(x_n) > 0, \\ x_n & \text{if} f'(x_n) < 0, \end{cases} \quad b_n = \begin{cases} x_n & \text{if} f'(x_n) > 0, \\ b_{n-1} & \text{if} f'(x_n) < 0. \end{cases} \tag{2.18}$$

По построению $f'(a_n) < 0$, $f'(b_n) > 0$, и $X_* \subset [a_n, b_n]$. Индуктивное описание метода закончено. Из геометрических построений нетрудно усмотреть, что этот метод совпадает с описанным выше методом касательных, в котором за начальную точку берется $x_0 = a$. В то же время приведенная схема метода более проста и удобна для реализации на ЭВМ; на каждом шаге метода здесь достаточно хранить в памяти ЭВМ величины a_n, b_n, $f(a_n)$, $f(b_n)$, $f'(a_n)$, $f'(b_n)$. Нетрудно выписать явное выражение для точки x_{n+1}, определяемой условием $l(x, a_n) = l(x, b_n)$ пересечения касательных в точках a_n, b_n при $f'(a_n) < 0$, $f'(b_n) > 0$:

$$x_{n+1} = \frac{f(a_n) - f(b_n) + b_n f'(b_n) - a_n f'(a_n)}{f'(b_n) - f'(a_n)}, \; n \geq 1. \tag{2.19}$$

Поскольку ломаная из отрезков касательных аппроксимирует функцию лучше, чем ломаные, то следует ожидать, что метод касательных для выпуклых функций сходится быстрее метода ломаных.

Оценка минимума строго выпуклой функции

Рассмотрим алгоритм оценки минимального значения многомерной выпуклой функции. В нем сначала строят n-мерный прямоугольник, содержащий точку минимума, а затем при использовании оценок производных находят гарантированную оценку минимального значения.

Пусть U — выпуклая ограниченная область в $R^n, n \geq 1, f$ — строго выпуклая, непрерывно дифференцируемая функция на U. Предположим, что известны константы L_i, удовлетворяющие неравенствам

$$\left| \frac{\partial f}{\partial x_i} \right| \leq L_i \quad \text{on} \quad \bar{U}, \quad i = 1, 2, ..., n. \tag{2.20}$$

Обозначим через x^* точку минимума, т. е.

$$f(x^*) = \min_{x \in \bar{U}} f(x).$$

Пусть x^0 — точка, близкая к x^* и полученная каким-либо из методов минимизации функций (см., например, [12]).

Возьмем $\varepsilon > 0$ и поставим целью построение прямоугольника, удовлетворяющего двум условиям:

– он содержит точку минимума x^*;

– внутри него значения функции $f(x)$ отличаются от $F(x^*)$ на величину не более ε.

Сначала приведем алгоритм в одномерном случае. Исходя из (2.20), достаточно построить интервал, содержащий точку x^* и имеющий длину не более $d_0 = \varepsilon / L_1$. Найдем отрезок произвольной длины, содержащий x^*. Этот поиск опишем рекуррентно. В качестве начального значения возьмем x^0. Рассмотрим k-й шаг. Имеется некоторая точка x^k. Она разбивает $\bar{u}$ на две части. В силу выпуклости функции f точка минимума лежит на полуинтервале, определяемом неравенством $-\mathrm{sign}(f'(x^k))(x^k - x) > 0$.

Построим точку

$$x^{k+1} = x^k - \mathrm{sign}(f'(x^k))d_k, \tag{2.21}$$

где $d_k = (1.5)^k d_0$. Если

$$\mathrm{sign}(f'(x^k)) = \mathrm{sign}(f'(x^{k+1})), \tag{2.22}$$

то переходим к шагу $k+1$. В противном случае в силу выпуклости функции f

$$x^* \in [a, b], \tag{2.23}$$

где $a = \min(x^k, x^{(k+1)}), b = \max(x^k, x^{k+1})$. Ясно, что при достаточной близости x^0 к x^* через конечное число шагов условие (2.23) выполняется. После этого необходимо ширину интервала довести до требуемой величины. Если $|x^{k+1} - x^k| > d_0$, то поделим отрезок $[a, b]$ пополам и, исходя из знака величины $f'((a+b)/2)$, выберем половину, содержащую x^*. Повторяя эту процедуру, через несколько шагов мы получим отрезок, содержащий точку x^* с длиной не более d_0. Тогда условие 1 справедливо из построения, а условие 2 вытекает из неравенства

$$|f(x) - f(x^*)| \le |x - x^*| L_1 \le d_0 L_1 < \varepsilon.$$

Таким образом, в одномерном случае алгоритм построен.

Теперь предположим, что алгоритм построен для размерности $n-1$. Опишем его для размерности n.

Зафиксируем целое i: $1 \le i \le n$. Начнем с точки x^0 и рассмотрим k-й шаг. Пусть известна точка x^κ. Проведем через нее гиперплоскость $x_i = x_i^k$. На этой гиперплоскости построим $(n-1)$-мерный прямоугольник, внутри которого достигается минимум (на этой плоскости) функции, остальные значения отличаются от минимума на величину не более $\varepsilon_0 = (n-1)\varepsilon/n$. Вершины прямоугольника обозначим через $y^j, j = 1, ..., 2^{n-1}$ и вычислим в них значения производной $\partial f/\partial x_i$:

$$v^j = \partial f(y^j)/\partial x_i, \quad j = 1, ..., 2^{n-1}.$$

Если все v^j имеют один знак, то выберем точку x^{k+1} следующим образом:

$$x^{k+1} = x^k - \mathrm{sign}(v^1) d_k e_i, \tag{2.24}$$

где $d_k = \varepsilon(1.5)^k/(L_1 n)$, а e_i — базисный вектор $(0, ..., 1, ..., 0)$ с единицей в i-й позиции. Если хотя бы два значения v^j имеют противоположные знаки, то величину ε_0 будем уменьшать вдвое и строить на гиперплоскости $x_i = x_i^k$ прямоугольники меньшего размера до тех пор, пока все v^j не будут одного знака. Это можно сделать ввиду непрерывности производной $\partial f/\partial x_i$.

Далее на гиперплоскости $x_i = x_i^{k+1}$ найдем $(n-1)$-мерный прямоугольник с вершинами $z_j, j = 1, ..., 2^{n-1}$, аналогичный прямоугольнику, построенному на гиперплоскости $x_i = x_i^k$. Если

$$\mathrm{sign}(\partial f(y^1)/\partial x_i) = \mathrm{sign}(\partial f(z^1)/\partial x_i), \tag{2.25}$$

то переходим к шагу $k+1$. В противном случае построена полоса V_i, содержащая точку x^*:

$$V_i = \{x | a_i \le x_i \le b_i\},$$

$$a_i = \min(x_i^k, x_i^{k+1}), \quad b_i = \max(x_i^k, x_i^{k+1}). \tag{2.26}$$

После этого необходимо ширину полосы V_i довести до требуемой величины. Если $|x_i^k - x_i^{k+1}| > \varepsilon/(L_i n)$, то проведем гиперплоскость $x_i = (x_i^k + x_i^{k+1})/2$, на которой построим $(n-1)$-мерный прямоугольник, обладающий свойствами, аналогичными построенным прямоугольникам. Сравнивая знаки в вершинах, выберем полосу, содержащую x^*. Ясно, что через несколько шагов получим полосу V_I, содержащую точку x^*, с

шириной не более $\varepsilon/(L_i n)$. Ввиду произвольности i такие полосы строим для всех $i = 1, 2, ..., n$. Тогда искомый прямоугольник определяется их пересечением:

$$x^* \in U_* = \bigcap_{i=1}^{n} V_i. \tag{2.27}$$

Докажем, что при построении выполнено условие 1. Рассмотрим k-й шаг, предшествующий формуле (2.24). Согласно работе [103], в точке минимума y, определяемой равенством $f(y) = \min\limits_{\substack{x \in U \\ x_i = x_i^k}} f(x)$, направление вектора

$$-\operatorname{sign}(\partial f(y)/\partial x_i) e_i \tag{2.28}$$

указывает на полупространство, содержащее точку x^*. По построению все v^j имеют один знак. Следовательно, в силу выпуклости функции f

$$\operatorname{sign}(v^j) = \operatorname{sign}(\partial f(y)/\partial x_i), \quad j = 1, 2, ..., 2^{n-1}.$$

Тогда [103] $x^* \in V_i, i = 1, ..., n$ и условие 1 выполнено. Условие 2 вытекает из неравенства

$$|f(x) - f(x^*)| \leq \sum_{i=1}^{n} L_i |x_i - x_i^*| \leq \varepsilon.$$

2.6. Интервальные интерполяционные полиномы

В этом разделе мы рассмотрим задачу интерполирования интервальных функций интервальными аналогами интерполяционных полиномов.

Обратимся к задаче интерполирования функций интервальными расширениями интерполяционных полиномов Лагранжа.

Пусть задана функция f и дан $(n+1)$ интервал $x_0, x_1, \ldots, x_n$, причем $x_i \neq x_j$. Предположим, что известны следующие оценки:

$$x_i \in \boldsymbol{x}_i, f(x_i) \in \boldsymbol{f}_i, \boldsymbol{x}_i \subseteq [a, b] i = 0, ..., n.$$

Определим интервальный интерполяционный полином Лагранжа как интервальное расширение

$$\boldsymbol{L}(\boldsymbol{x}) = \{L(x) | x_i \in \boldsymbol{x}_i, f(x_i) \in \boldsymbol{f}_i, i = 0, ..., n, x \in \boldsymbol{x}\},$$

где $L(x) = \sum_{i=0}^{n} f_i w_i(x)$ — интерполяционный полином Лагранжа, проходящий через точки (x_i, f_i). Заметим, что построенный интервальный полином существует только при условии

a) $\boldsymbol{x}_i \cap \boldsymbol{x}_j = \emptyset$

и кроме того

b) $\boldsymbol{L}(\boldsymbol{x}_i) \supset \boldsymbol{f}_i$.

Равенство в b) возможно только в случае вырожденных интервалов $\mathrm{wid}\boldsymbol{x}_i = 0,\ i = 0, 1, ..., n$. Оценка погрешности существенным образом зависит не только от $||f^{(n+1)}||_{C[a,b]}$, но и от $\mathrm{wid}(\boldsymbol{f}_i)$, $\mathrm{wid}(\boldsymbol{x}_i)$.

Описанная ниже модификация интервального интерполяционного полинома Лагранжа позволяет в некоторых случаях избежать этих неудобств.

Пусть задана интервальная функция $\boldsymbol{f}$ и дан $(n+1)$ интервал $\boldsymbol{x}_0$, $\boldsymbol{x}_1$, $\ldots$, $\boldsymbol{x}_n$, причем $\underline{x}_i \neq \underline{x}_j$, $\overline{x}_i \neq \overline{x}_j$.

Рассмотрим задачу приближения $\boldsymbol{f}$ интервальной функцией $\boldsymbol{L}_n$:

$$\boldsymbol{f}(\boldsymbol{x}_i) = \boldsymbol{L}_n(\boldsymbol{x}_i), i = 0, 1, \ldots, n. \tag{2.29}$$

В дальнейшем нам понадобится специальная арифметика над интервальными числами и двумерными векторами:

$$(x_1, x_2) \oplus (y_1, y_2) = (x_1 + y_1, x_2 + y_2),$$

$$(x_1, x_2) \ominus (y_1, y_2) = (x_1 - y_1, x_2 - y_2),$$

$$(x_1, x_2) \otimes (y_1, y_2) = (x_1 y_1, x_2 y_2),$$

$$(x_1, x_2) \oslash (y_1, y_2) = (x_1/y_1, x_2/y_2).$$

Оператор $\Psi : R^2 \to \boldsymbol{R}$ выполняет вложение двумерных векторов в множество интервальных чисел:

$$\Psi(v_1, v_2) = [\min\{v_1, v_2\}, \max\{v_1, v_2\}],$$

$\psi(v_1, v_2) = (v_2, v_1)$. Приведем в качестве $\boldsymbol{L}_n$ аналог интерполяционной функции Лагранжа:

$$\boldsymbol{L}_n(\boldsymbol{x}) = \Psi(\sum_{i=0}^{n} \boldsymbol{w}_i(\boldsymbol{x}) \otimes \boldsymbol{f}(\boldsymbol{x}_i), \tag{2.30}$$

где

$$\boldsymbol{w}_i(\boldsymbol{x}) = [\frac{\Pi_{j \neq i}(\underline{x} - \underline{x}_j)}{\Pi_{j \neq i}(\underline{x}_i - \underline{x}_j)}, \frac{\Pi_{j \neq i}(\overline{x} - \overline{x}_j)}{\Pi_{j \neq i}(\overline{x}_i - \overline{x}_j)}].$$

Несложно убедиться, что построенная функция $\boldsymbol{L}_n(\boldsymbol{x})$ удовлетворяет условиям (2.29), в случае вырожденных интервалов переходит в известный интерполяционный полином Лагранжа.

Для оценки погрешности интерполяционной формулы (2.30) введем следующее определение производной от интервальной функции.

Функция $\boldsymbol{f}$ называется πF-дифференцируемой в точке $\boldsymbol{x}$, если существует линейное отображение $dF : R^2 \to R^2$ и отображение $S : R^2 \to R^2$ с двумя свойствами [99]:

– $S(0,0) = (0,0), \lim_{\Delta x \to 0} \frac{||S(\Delta x)||}{||\Delta x||} = 0$; для любого вектора $\Delta x \in R^2$, лежащего в окрестности точки $(0,0)$;

– $\boldsymbol{f}(\boldsymbol{x} \oplus \Delta x) \ominus \boldsymbol{f}(\boldsymbol{x}) = dF(\boldsymbol{x} \otimes \Delta x \oplus S(\Delta x)$.

Определим два класса дифференцируемых функций. Будем говорить, что $\boldsymbol{f} \in \mathcal{D}_1$, если существует конечный предел

$$\boldsymbol{f}'(\boldsymbol{x}) = \lim_{\Delta x \to 0} (\boldsymbol{f}(\boldsymbol{x} \oplus \Delta x) \ominus \boldsymbol{f}(\boldsymbol{x})) \oslash \Delta x,$$

и $\boldsymbol{f} \in \mathcal{D}_2$, если существует конечный предел

$$\boldsymbol{f}'(\boldsymbol{x}) = \lim_{\Delta x \to 0} (\boldsymbol{f}(\boldsymbol{x} \oplus \Delta x) \ominus \boldsymbol{f}(\boldsymbol{x})) \oslash \psi(\Delta x).$$

Если $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{D}_i$ то справедливы следующие правила:

если $\boldsymbol{f} = \text{const}$, то $\boldsymbol{f}' = (0,0)$;

если $\boldsymbol{f} = \boldsymbol{x}$, то $\boldsymbol{f}' = (1,1)$;

если $\boldsymbol{f} + \boldsymbol{g} \in \mathcal{D}_i$, то $(\boldsymbol{f} + \boldsymbol{g})' = \boldsymbol{f}' + \boldsymbol{g}'$;

если $\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{f} \otimes \boldsymbol{g} \in \{calD_i$, то $(\boldsymbol{f} \otimes \boldsymbol{g})' = \boldsymbol{f}' \otimes \boldsymbol{g} + \boldsymbol{f} \otimes \boldsymbol{g}'$;

если $\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{f} \oslash \boldsymbol{g} \in \{calD_i$, то $(\boldsymbol{f} \oslash \boldsymbol{g})' = (\boldsymbol{f}' \otimes \boldsymbol{g} - \boldsymbol{f} \otimes \boldsymbol{g}') \oslash (\boldsymbol{g} \otimes \boldsymbol{g}; 0 \notin \boldsymbol{g}$.

В случае, когда функция $\boldsymbol{f}$ зависит от вещественного аргумента, $\mathcal{D}_i$ совпадают и производная определяется из предела

$$\boldsymbol{f}'(x) = \lim_{\Delta x \to 0} \frac{\boldsymbol{f}(x + \Delta x) - \boldsymbol{f}(x)}{\Delta x}$$

и совпадает с производной в [97].

Определение 5. *Пространство функций, имеющих непрерывные производные из $\mathcal{D}_i$, будем обозначать $\mathcal{D}(C)$.*

Теорема 19. *Пусть $\boldsymbol{f} \in \mathcal{D}(C)$, тогда функция $\boldsymbol{f}$ в некоторой окрестности точки $(x_1, x_2) \in R^2$ представима в одном из двух видов:*

$$\boldsymbol{f}(x_1, x_2) = [f_1(x_1), f_2(x_2)]$$

или

$$\boldsymbol{f}(x_1, x_2) = [f_1(x_2), f_2(x_1)].$$

Доказательство. Действительно, $\boldsymbol{f}(x_1, x_2) = [f_1(x_1, x_2), f_1(x_1, x_2)]$ и dF представимо в виде

$$dF(dx_1, dx_2) = F'(dx_1, dx_2)^T,$$

где F' — матрица с элементами $\{\partial f_i/\partial x_j\}$. В силу последнего равенства и определения производной получаем либо

$$\frac{\partial f_1}{\partial x_2} = \frac{\partial f_2}{\partial x_1} = 0,$$

либо

$$\frac{\partial f_2}{\partial x_1} = \frac{\partial f_1}{\partial x_2} = 0,$$

что и требовалось доказать. □

Далее мы можем ввести понятие производных высших порядков

$$\boldsymbol{f}^{n+1} = (\boldsymbol{f}^n)'.$$

Более того, для введенной производной естественным образом оказывается справедливым теорема о среднем:

$$\boldsymbol{f}(y_1, y_2) \ominus \boldsymbol{f}(x_1, x_2) = \boldsymbol{f}'(\xi_1, \xi_2) \otimes (y_1 - x_1, y_2 - x_2).$$

Используя понятие производной и теоремы о среднем, следуя аналогичным оценкам для вещественных полиномов Лагранжа, оценим погрешность интерполяционной формулы (2.30)

$$\boldsymbol{f}(\boldsymbol{x}) \ominus \boldsymbol{L}_n(\boldsymbol{x}) = \boldsymbol{w}(\boldsymbol{x})\boldsymbol{f}^{(n+1)}(\xi_1, \xi_2)/(n+1)!,$$

где

$$\boldsymbol{w}(\boldsymbol{x}) = \prod_{i=0}^{n} (\boldsymbol{x} \ominus \boldsymbol{x}_i),$$

$$\xi_1 \in (\underline{x}_0, \underline{x}_n), \xi_2 \in (\overline{x}_0, \overline{x}_n).$$

Мы можем построить аналогичным образом и интерполяционную формулу Ньютона

$$\boldsymbol{N}(\boldsymbol{x}) = \Psi(\mathbf{a}_0 \oplus \boldsymbol{a}_1 \otimes (\boldsymbol{x} \ominus \boldsymbol{x}_0) \oplus \boldsymbol{a}_2 \otimes (\boldsymbol{x} \ominus \boldsymbol{x}_0) \otimes (\boldsymbol{x} \ominus \boldsymbol{x}_1) \ldots \oplus \boldsymbol{a}_n \otimes (\boldsymbol{x} \ominus \boldsymbol{x}_0) \otimes \ldots \otimes (\boldsymbol{x} \ominus \boldsymbol{x}_n),$$

где

$\boldsymbol{a}_0 = \boldsymbol{f}(\boldsymbol{x}_0)$;

$\boldsymbol{a}_1 = (\boldsymbol{f}(\boldsymbol{x}_1) \ominus \boldsymbol{f}(\boldsymbol{x}_0)) \oslash (\boldsymbol{x}_1 \ominus \boldsymbol{x}_0)$;

$\boldsymbol{a}_2 = (\boldsymbol{f}(\boldsymbol{x}_2) \ominus \boldsymbol{f}(\boldsymbol{x}_0)) \oslash ((\boldsymbol{x}_2 \ominus \boldsymbol{x}_0) \otimes (\boldsymbol{x}_2 \ominus \boldsymbol{x}_1)) \ominus (\boldsymbol{f}(\boldsymbol{x}_1) \ominus \boldsymbol{f}(\boldsymbol{x}_0)) \oslash ((\boldsymbol{x}_2 \ominus \boldsymbol{x}_0) \otimes (\boldsymbol{x}_2 \ominus \boldsymbol{x}_1))$

и т. п.

Поскольку интерполяционная формула Ньютона имеет вид подобный $\boldsymbol{L}(\boldsymbol{x})$, следуя приемам классического математического анализа для $\boldsymbol{N}(\boldsymbol{x})$, можно получить оценки остаточных членов.

В качестве примера рассмотрим интерполяцию функции $\mathbf{sin}(\pi\mathbf{x})$ на отрезке $[-0.5, 0.5]$. Несложно убедиться, что $\mathbf{sin}(\pi\mathbf{x}) \in \mathcal{D}_1$.

Пусть заданы узлы интерполирования и значения функции на них:

$\boldsymbol{x}_0 = [-0.5, -1/6]$; $\boldsymbol{f}(\boldsymbol{x}_0) = [-1.0, -0.5]$;

$\boldsymbol{x}_1 = [-1/3, 0.0]$; $\boldsymbol{f}(\boldsymbol{x}_1) = [-\sqrt{3}/2, 0.0]$;

$\boldsymbol{x}_2 = [-1/6, 1/6]$; $\boldsymbol{f}(\boldsymbol{x}_2) = [-0.5, 0.5]$;

$\boldsymbol{x}_3 = [0.0, 1/3]$; $\boldsymbol{f}(\boldsymbol{x}_3) = [0.0, \sqrt{3}/2]$;

$\boldsymbol{x}_4 = [1/6, 0.5]$; $\boldsymbol{f}(\boldsymbol{x}_4) = [0.5, 1.0]$.

Непосредственное вычисление показывает, что

$$||\mathbf{sin} \ominus \boldsymbol{L}_4||_{C[-0.5,0.5]} = 2.39056e - 03.$$

Следующий пример демонстрирует интерполяцию функции $\boldsymbol{x}^2$ полиномом $\boldsymbol{L}_2$ на отрезке $[-1.0, 1.0]$ со специальным выбором узлов. Несложно убедиться, что интервальная функция $\boldsymbol{x}^2$ при различных значениях аргумента принадлежит различным классам $\mathcal{D}_i$.

Зададим узлы интерполирования и значения функции $\boldsymbol{x}^2$ на них:

$\boldsymbol{x}_0 = [0.0, 0.0]$; $\boldsymbol{f}(\boldsymbol{x}_0) = [0.0, 0.0]$;

$\boldsymbol{x}_1 = [-0.5, 0.5]$; $\boldsymbol{f}(\boldsymbol{x}_1) = [0.0, 0.25]$;

$\boldsymbol{x}_2 = [-1.0, 1.0]$; $\boldsymbol{f}(\boldsymbol{x}_2) = [0.0, 1.0]$.

Представим интервальный интерполяционный полином Лагранжа в следующем виде:

$$\boldsymbol{L}_2(\boldsymbol{x}) = \sum_{i=0}^{2} \Psi(\boldsymbol{w}_i(\boldsymbol{x}))\boldsymbol{f}(\boldsymbol{x}_i).$$

Заметим, что условие (2.29) для построенного полинома полностью выполнено и $\boldsymbol{L}_2 \equiv \boldsymbol{x}^2$ для всех $\boldsymbol{x} \ni 0$, но это свойство сохраняется только для этого выбора узлов.

Данный пример показывает направления дальнейших обобщений построения специальных интервальных интерполяционных полиномов.

2.7. Интервальные сплайны

В этом разделе мы затронем вопросы построения интервальных сплайнов. Они вводятся как интервальные расширения по определенным параметрам соответствующих вещественных сплайнов. При небольшой априорной информации об интерполируемой функции строятся ее двусторонние приближения.

Для частного вида интервальных функций введем следующие понятия. Пусть интервальная функция имеет вид

$$\boldsymbol{f}(x) = \sum_{i=1}^{n} \boldsymbol{a}_i g_i(x), \quad \text{где} \quad g_i \in C^m[a,b].$$

Тогда формальную производную от $\boldsymbol{f}(x)$ определим таким образом:

$$\partial^k \boldsymbol{f}(x) = \sum_{i=1}^{n} \boldsymbol{a}_i g_i^{(k)}(x), \quad k = 0, ..., m.$$

Соответственно, функцию

$$f(x) = \sum_{i=1}^{n} a_i g_i(x), \quad \text{где} \quad a_i \in \boldsymbol{a}_i,$$

будем называть *сужением функции* $\boldsymbol{f}$ по константам $\boldsymbol{a}_i$.

Приведем необходимые для дальнейшего изложения элементы теории сплайнов [36]. Пусть на отрезке $[a,b]$ задана сетка

$$\omega = \{x_i | a = x_0 < x_1 < ... < x_N = b\}$$

с целым $N \geq 2$ и шагами $h_i = x_{i+1} - x_i, \quad h = \max\limits_{0 \leq i \leq N-1} h_i$.

Функцию s называют сплайном степени n дефекта k (k — целое, $1 \leq k \leq n$) с узлами на ω, если:

$$1) \qquad s(x) = \sum_{i=0}^{n} a_{ij}(x - x_j)^i \quad \text{на} \quad [x_j, x_{j+1}], \quad j = 0, ..., N-1; \tag{2.31}$$

$$2) \qquad s \in C^{n-k}[a,b]. \tag{2.32}$$

Множество сплайнов, удовлетворяющих этим условиям, обозначим через S_n^k.

Обратимся к вопросу интерполирования заданных функций сплайнами нечетных степеней. Положим целое $m = (n-1)/2$ и будем считать дефект $k \leq m+1$.

Пусть $f \in C^{n-k}[a,b]$. Поставим задачу определения сплайна $s \in S_n^k$, интерполирующего функцию f в следующем смысле:

$$s^{(j)}(x_i) = f^{(j)}(x_i),\ i = 1, ..., N-1,\ j = 0, ..., k-1. \tag{2.33}$$

Дополнительно задается еще по $m+1$ краевому условию $[a,b]$. Часто они берутся в виде

$$s^{(j)}(a) = f^{(j)}(a), \quad s^{(j)}(b) = f^{(j)}(b), \quad j = 0, ..., m. \tag{2.34}$$

Условия (2.33), (2.34) приводят, соответственно, к $2(N-1)k$ и $(n+1)$ уравнениям для коэффициентов a_{ij}. Еще $(n-2k+1)(N-1)$ уравнений вытекают из условия непрерывности производных в соответствии с (2.32). В итоге получается система линейных алгебраических уравнений

$$Aa = b \tag{2.35}$$

с квадратной матрицей $A \in R^{N(n+1)\times N(n+1)}$, вектором неизвестных $a \in R^{N(n+1)}$, известной правой частью $b \in R^{N(n+1)}$. Не исследуя общего случая, в дальнейшем выпишем эту систему для конкретных примеров и исследуем ее разрешимость.

Теперь перейдем к определению интервальных сплайнов вещественного аргумента. Необходимость их использования возникает, когда вместо точных значений функции f и ее производных известны только интервальные оценки, которым они принадлежат. Пусть известны интервальные константы:

$$f^{(j)}(x_i) \in \boldsymbol{f}_i^j, \quad i = 1, ..., N-1, \quad j = 0, ..., k-1; \tag{2.36}$$

$$f^{(j)}(a) \in \boldsymbol{f}_0^j, \quad f^{(j)}(b) \in \boldsymbol{f}_N^j, \quad j = 0, ..., m \tag{2.37}$$

Объединенным интервальным сплайном назовем интервальную функцию $\boldsymbol{s}_u : [a,b] \to \boldsymbol{R}$ с значениями:

$$\boldsymbol{s}_u(x) = \{s(x) | s \in S_n^k, \quad s^{(j)}(x_i) \in \boldsymbol{f}_i^j, \quad i = 1, .., N-1,$$
$$j = 0, ..., k-1; \quad s^{(j)}(a) \in \boldsymbol{f}_0^j, \quad s^{(j)}(b) \in \boldsymbol{f}_N^j, \quad j = 0, ..., m\}. \tag{2.38}$$

Эта интервальная функция является объединенным расширением соответствующего множества вещественных сплайнов. При этом она не имеет представления вида (2.31) и вычисляется довольно сложно.

В предположении разрешимости системы (2.35) построим более простую функцию $\boldsymbol{s}$ такую, что

$$\boldsymbol{s}_u(x) \subset \boldsymbol{s}(x) \quad \forall x \in [a,b]. \tag{2.39}$$

Функцию $\boldsymbol{s}$ будем искать в следующем виде:

$$\boldsymbol{s}(x) = \sum_{i=0}^{n} \boldsymbol{c}_{ij}(x - x_j)^i \quad \text{на} \quad [x_j, x_{j+1}]. \tag{2.40}$$

В качестве вектора $\boldsymbol{c} = \{\boldsymbol{c}_{ij}\}$ возьмем произвольный интервальный вектор, содержащий объединенное множество решений

$$X = \{x | x = A^{-1}b, \quad b \in \boldsymbol{b}\},$$

где вектор $\boldsymbol{b}$ — интервальное расширение вектора b из системы (2.35), полученное заменой $f^{(j)}(x_i)$ для $\boldsymbol{f}_i^j$. Методы получения и уточнения векторов мы подробно обсудим в разделе, посвященном системам линейных алгебраических уравнений.

Заметим, что в общем случае сужение интервальной функции (2.38) по интервальным константам $\{\boldsymbol{c}_{ij}\}$ не всегда будет сплайном из S_n^k, поскольку константы $c_{ij} \in \boldsymbol{c}_{ij}$ после сужения могут не удовлетворять системе (2.35).

Если известна оценка нормы $\|f\|_{p,\infty}$, то можно построить интервальные функции $\boldsymbol{r}_j(x)$, учитывающие ошибки аппроксимации сплайнов, такие что

$$f^{(j)}(x) \in \boldsymbol{s}^{(j)}(x) + \boldsymbol{r}_j(x), \quad x \in [a, b]. \tag{2.41}$$

Для получения функций $\boldsymbol{r}_j(x)$ воспользуемся следующим результатом [11].

Теорема 20. *Пусть $f \in W_\infty^p[a, b]$, $1 \le p \le n + 1$ и сплайн $s \in S_n^k$ интерполирует f в смысле (2.33),(2.34). Тогда*

$$\|\partial^j(f - s)\|_\infty \le K_j h^{p-j} \|f\|_{p,\infty},$$

где K_j — константы, не зависящие от f и h. □

Положим $\varepsilon = f - s$ и разложим $\varepsilon(x)$ в окрестности точки x_i, $i = 0, ..., N - 1$, $x \in [x_i, x_{i+1}]$:

$$\varepsilon(x) = \varepsilon(x_i) + \varepsilon'(x_i)(x - x_i)/1! + ...$$

$$... + \varepsilon^{(k-1)}(x_i)(x - x_i)^{k-1}/(k-1)! + \varepsilon^{(k)}(\xi)(x - x_i)^k/k!,$$

где $\xi \in [x_i, x_{i+1}]$. Поскольку $\varepsilon^{(j)}(x_i) = 0, \quad 0 \le j \le k - 1$, то

$$|\varepsilon(x)| \le (x - x_i)^k/k! \, \|\partial^k(f - s)\|_{\infty,[x_i,x_{i+1}]},$$

$$|\varepsilon^{(j)}(x)| \le (x-x_i)^{k-j}/(k-j)!\ \|\partial^k(f-s)\|_{\infty,[x_i,x_{i+1}]}.$$

Поэтому на основании теоремы вложения из W_2^p в L_∞

$$\begin{aligned} \boldsymbol{r}_j(x) = [-1,1]\min\{&K_j h^{p-j}\|f\|_{p,\infty},\\ &K_k h^{p-k}(x-x_i)^{k-j}/(k-j)!\ \|f\|_{p,2}\} \\ \forall x \in [x_i, x_{i+1}], &\quad i = 0,1,...,N-1. \end{aligned} \tag{2.42}$$

Тем самым найдены интервальные функции $\boldsymbol{r}_j,\ j = 0,1,...,p-1$, позволяющие строить полосы, содержащие в себе значения j-й производной функции f.

Введенное определение интервальных сплайнов проиллюстрируем на примере сплайнов степени $n = 1, 3$.

Пусть на сетке ω заданы значения $f_i \in \boldsymbol{f}_i,\ i = 0,1,...,N$. На интервале $[x_i, x_{i+1}]$ сплайн первой степени имеет вид

$$s(x) = f_i(x_{i+1}-x)/h_i + f_{i+1}(x-x_i)/h_i.$$

Поэтому интервальный сплайн первой степени можно представить следующим образом:

$$\boldsymbol{s}(x) = \boldsymbol{f}_i(x_{i+1}-x)/h_i + \boldsymbol{f}_{i+1}(x-x_i)/h_i, x \in [x_i, x_{i+1}]. \tag{2.43}$$

Отметим, что здесь сужение сплайна по интервальным константам $\boldsymbol{f}_i$ будет вещественным сплайном.

Перейдем к интервальным кубическим сплайнам. Рассмотрим следующую интервальную функцию $\boldsymbol{s}_u : [a,b] \to \boldsymbol{r}$ со значениями

$$\begin{aligned} \boldsymbol{s}_u(x) = \{s(x) | s \in S_3^1, &\quad s(x_i) \in \boldsymbol{f}_i^0, \quad i = 0,...,N, \\ &s''(a) \in \boldsymbol{f}_0^2, \quad s''(b) \in \boldsymbol{f}_N^2\}. \end{aligned}$$

Заметим, что вместо условий

$$s'(a) = f'(a), \quad s'(b) = f'(b),$$

вытекающих из (2.34), мы взяли более употребительные краевые условия [36]

$$s''(a) = f''(a), \quad s''(b) = f''(b), \tag{2.44}$$

что не меняет сути изложения. В итоге кубический сплайн на отрезках $[x_{j-1}, x_j], j = 1,...,N$ имеет два представления [4]:

$$s(x) = M_{j-1}(x_j - x)^3/(6h_j) + M_j(x-x_{j-1})^3/(6h_j) +$$

$$+ (f_{j-1} - M_{j-1}h_j^2/6)(x_j - x)/h_j + (f_j - M_j h_j^2/6)(x - x_{j-1})/h_j, \quad (2.45)$$

или

$$\begin{gathered} s(x) = m_{j-1}(x_j - x)^2(x - x_{j-1})/h_j^2 - \\ m_j(x - x_{j-1})^2(x_j - x)/h_j^2 + \\ + f_{j-1}(x_j - x)^2(2(x - x_{j-1}) + h_j)/h_j^3 + \\ + f_j(x - x_{j-1})^2(2(x_j - x) + h_j)/h_j^3, \end{gathered} \quad (2.46)$$

где $M_j = s''(x_j)$, $m_j = s'(x_j)$, $f_j = f(x_j)$. Заменяя M_j, m_j, f_j на $\boldsymbol{M}_j$, $\boldsymbol{m}_j$, $\boldsymbol{f}_j \equiv \boldsymbol{f}_j^0$, мы получаем соответствующее представление интервальной функции, содержащей интервальный кубический сплайн. Выпишем интервальные системы линейных алгебраических уравнений:

для $\boldsymbol{M}_j$

$$\mu_j \boldsymbol{M}_{j-1} + 2\boldsymbol{M}_j + \lambda_j \boldsymbol{M}_{j+1} = \boldsymbol{D}_j, \quad (2.47)$$

$$\boldsymbol{M}_0 = \boldsymbol{f}_0^2, \quad \boldsymbol{M}_N = \boldsymbol{f}_N^2,$$

$$\boldsymbol{D}_j = 6((\boldsymbol{f}_{j+1} - \boldsymbol{f}_j)/h_{j+1} - (\boldsymbol{f}_j - \boldsymbol{f}_{j-1})/h_j)/(h_j + h_{j+1}),$$

$$\lambda_j = h_{j+1}/(h_j + h_{j+1}), \quad \mu_j = 1 - \lambda_j;$$

для m_j

$$\lambda_j \boldsymbol{M}_{j-1} + 2\boldsymbol{M}_j + \mu_j \boldsymbol{m}_{j+1} = \boldsymbol{D}_j, \quad (2.48)$$

$$2\boldsymbol{M}_0 + \boldsymbol{M}_1 = 3(\boldsymbol{f}_1 - \boldsymbol{f}_0)/h_1 - h_1 \boldsymbol{f}_0^2/2,$$

$$2\boldsymbol{M}_N + \boldsymbol{M}_{N-1} = 3(\boldsymbol{f}_N - \boldsymbol{f}_{N-1})/h_N + h_N \boldsymbol{f}_N^2/2,$$

$$\boldsymbol{D}_j = 3\lambda_j(\boldsymbol{f}_j - \boldsymbol{f}_{j-1})/h_j + 3\mu_j(\boldsymbol{f}_{j+1} - \boldsymbol{f}_j)/h_{j+1}. \quad j = 1, ..., N-1.$$

Матрицы этих систем имеют вещественные элементы, а правые части содержат интервальные числа. Кроме того, при упорядочении уравнений и неизвестных по возрастанию j, матрицы становятся трехдиагональными и строго диагонально преобладающими. Это позволяет использовать простой метод, сводящий вычисление интервального вектора к определению его границ из двух систем уравнений с исходной матрицей, но вещественными правыми частями.

С помощью найденных интервальных чисел построим интервальную функцию $\boldsymbol{s}(x) : [a, b] \to \boldsymbol{R}$ со значениями

$$\begin{gathered} \boldsymbol{s}(x) = \boldsymbol{M}_{j-1}\left((x_j - x)^3/(6h_j) - (x_j - x)h_j/6\right) + \\ + \boldsymbol{M}_j((x - x_{j-1})^3/6h_j) - (x - x_{j-1})h_j/6) + \\ + \boldsymbol{f}_{j-1}(x_j - x)/h_j + \boldsymbol{f}_j(x - x_{j-1})/h_j, \end{gathered} \quad (2.49)$$

или

$$\begin{aligned} \boldsymbol{s}(x) = \boldsymbol{M}_{j-1}(x_j - x)^2(x - x_{j-1})/h_j^2 - \\ -\boldsymbol{M}_j(x - x_{j-1})^2(x_j - x)/h_j^2 \\ +\boldsymbol{f}_{j-1}(x_j - x)^2(2(x - x_{j-1}) + h_j)/h_j^3 + \\ +\boldsymbol{f}_j(x - x_{j-1})^2(2(x_j - x) + h_j)/h_j^3, \end{aligned} \tag{2.50}$$

при $x \in [x_{j-1}, x_j], j = 1, ..., N$. Полученные интервальные функции содержат интервальный кубический сплайн $\boldsymbol{s}_u$.

Заметим, что если в (2.49) брать сужение $\boldsymbol{f}_j$ по интервальным константам $\boldsymbol{M}_j$, то в общем случае мы не получим кубического сплайна, так как $\boldsymbol{M}_j$ могут не удовлетворять системе (2.47) и, как следствие этого, $s' \notin C[a, b]$. Если брать сужение по $\boldsymbol{M}_j$ в (2.50), то $s' \in C[a, b]$, но $s'' \notin C[a, b]$.

Этот факт выражает то обстоятельство, что при аппроксимации функций кубическими сплайнами на конкретных компьютерах в связи с ошибками округления могут получаться функции, не являющиеся кубическими сплайнами.

Рассмотрим интервальные эрмитовы кубические сплайны $\boldsymbol{s}_1 : [a, b] \to \boldsymbol{R}$. На каждом интервале $[x_{j-1}, x_j], j = 1, ..., N$, согласно формуле (2.46), эти сплайны представимы в виде:

$$\begin{aligned} \boldsymbol{s}_1(x) = \boldsymbol{f}_{j-1}v((x - x_{j-1})/h_j) + \boldsymbol{f}_{j-1}^1 w((x - x_{j-1})/h_j) + \\ +\boldsymbol{f}_j v((x - x_j)/h_j) + \boldsymbol{f}_j^1 w((x - x_j)/h_j), \end{aligned}$$

где $v(x) = (|x|-1)^2(2|x|+1)$, $w(x) = x(|x|-1)^2$. Обратим внимание на то, что интервальные эрмитовы сплайны в отличие от кубических сплайнов точно выражаются этой формулой, т. е. $\boldsymbol{s}_u(x) = \boldsymbol{s}_1(x)$. Кроме того, сужение $\boldsymbol{s}_1$ по константам $\boldsymbol{f}_j, \boldsymbol{f}_j^1$ будет эрмитовым кубическим сплайном. Следовательно, при аппроксимации функций эрмитовыми сплайнами наличие ошибок в данных не снижает гладкости сплайна.

Перейдем далее к вопросу получения двусторонних аппроксимаций функций. Предположим, что для функции f известны ее оценки в некоторой норме, например в нормах пространств $W_\infty^p[a, b]$ или $C^p[a, b]$. Тогда можно построить интервальные полосы, включающие точные значения интерполируемой функции. В общем случае интервальная функция погрешности дается формулой (2.42), но она довольно груба. Построим более узкие полосы, например, для сплайнов первой степени. Сначала пусть $f \in C^1[a, b]$ и известны константы K_l в оценке норм производной

$$\|f'\|_{\infty,[x_l,x_{l+1}]} \le K_l, \quad l = 0, 1, ..., N - 1.$$

Тогда $\boldsymbol{r}(x) = [-1,1]h_lK_l$. Следовательно, на каждом отрезке $[x_l, x_{l+1}]$ интерполируемая функция содержится в полосе:

$$f(x) \in \boldsymbol{s}(x) + [-1,1]h_lK_l.$$

Кроме того,

$$\operatorname*{int}_{[x_l,x_{l+1}]} f(x) \subset \boldsymbol{s}([x_l, x_{l+1}]) + [-1,1]h_lK_l.$$

Теперь пусть $f \in C^2[a,b]$ и известны константы $K_H \geq \|f''_-\|_{\infty,[a,b]}, K_B \geq \|f''_+\|_{\infty,[a,b]}$ для компонент разложения $f'' = f''_+ - f''_-$, где $f''_+(x) = \max\{0, f''(x)\|$, и $f''_-(x) = -\min\{0, f''(x)\|$. Тогда, следуя доказательству [36], получаем

$$\varphi(x) = f(x) - s(x) = ((x_{i+1} - x)\int_{x_i}^{x}(v - x_i)f''(v)dv +$$

$$+(x - x_i)\int_{x}^{x_{i+1}}(x_{i+1} - v)f''(v)dv)/h_i.$$

Отсюда

$$\varphi(x) \leq \left((x_{i+1} - x)\int_{x_i}^{x}(v - x_i)dv + \right.$$

$$\left. +(x - x_i)\int_{x}^{x_{i+1}}(x_{i+1} - v)dv\right)\|f''_+\|_{\infty,[x_i,x_{i+1}]}/h_i.$$

Обозначим

$$\sigma(x) = ((x_{i+1} - x)(x_i - x)^2 + (x - x_i)(x_{i+1} - x)^2)/2h_i.$$

Тогда $\varphi(x) \leq \sigma(x)\max\limits_{x\in[x_i,x_{i+1}]} f''_+(x)$ и аналогично, $\varphi(x) \geq \sigma(x)\min\limits_{x\in[x_i,x_{i+1}]} f''_-(x)$. Следовательно,

$$\boldsymbol{r}(x) = [-K_H, K_B]\sigma(x).$$

Таким образом, построена функция $\boldsymbol{r}$, гарантирующая включение

$$f(x) \in \boldsymbol{s}(x) + \boldsymbol{r}(x).$$

Перейдем к изучению интервальных сплайнов многих переменных. Наиболее распространены два подхода для обобщения понятия сплайна на случай многих переменных. Первый способ состоит в тензорном

произведении одномерных сплайнов, он достаточно прост и будет использован в последующих главах. Здесь остановимся на втором подходе, развиваемом в теории конечных элементов.

Предположим, что Ω — ограниченная односвязная область в R^2. Триангулируем ее [55], т.е. построим замкнутый односвязный многоугольник $\Omega_h \subset \bar{\Omega}$, представимый в виде объединения прямосторонних замкнутых треугольников T_i:

$$\Omega_h = \bigcup_{i=1}^{N} T_i.$$

Причем треугольники таковы, что пересечение двух различных T_i, T_j есть либо вершина, либо целиком сторона, либо T_i, T_j вообще не пересекаются. Рассмотрим полный полином порядка m

$$P_m(x,y) = \sum_{h+l=0}^{m} a_{kl} x^k y^l.$$

В линейном случае $(m = 1)$ он имеет вид $P_1(x,y) = a_1 + a_2 x + a_3 y$ и для его однозначного определения достаточно знать значения функции f в вершинах $z_i, i = 1,2,3$ треугольника T.

Полином второй степени имеет шесть неопределенных коэффициентов:

$$P_2(x,y) = a_1 + a_2 x + a_3 y + a_4 x^2 + a_5 xy + a_6 y^2.$$

Пусть $z_i, i = 4,5,6$— середины сторон треугольника T. Легко видеть, что существует единственный полином $p_2(x,y)$, интерполирующий f в узлах $z_i : f(z_i) = R_2(z_i), i = 1,...,6$.

Можно построить аппроксимацию на треугольнике T для полных кубических полиномов. Кубический полином

$$P_3(x,y) = a_1 + a_2 x + a_3 y + a_4 x^2 + a_5 xy + a_6 y^2 +$$

$$+ a_7 x^3 + a_8 x^2 y + a_9 x y^2 + a_{10} y^3$$

определяется значениями в десяти узлах $z_i, i = 1,...,10$, причем $z_i, i = 4,...9$ расположены парами на сторонах треугольника так, что делят их на равные части, а z_{10} лежит на пересечении медиан.

Пусть в узлах $z_i, i = 1,...,N$ на треугольнике T заданы значения функции $f(z_i) = f_i \in \boldsymbol{f}_i$. Назовем *интервальным конечным элементом* степени m следующую функцию:

$$\boldsymbol{P}_m(x,y) = \{P_m(x,y) | P_m(z_i) \in \boldsymbol{f}_i\}.$$

Пусть $\psi_i(x,y)$ полные полиномы степени m такие, что $\psi_i(z_i)=\delta_{ij}, i,j=1,...,N$. Тогда интервальный конечный элемент можно представить как

$$\boldsymbol{P}_m(x,y)=\sum_{i=1}^{N}\boldsymbol{f}_i\psi_i(x,y).$$

Поскольку $\psi_i(x,y)$ — вещественные функции, сужение по интервальным константам $\{\boldsymbol{f}_i\}$ дает конечный элемент, аппроксимирующий $f(x,y)$.

Перейдем непосредственно к кусочно-линейным элементам. Пусть треугольник T имеет вершины $z_i=(x_i,y_i)$; тогда базисный элемент с вершиной в точке (x_i,y_i) записывается в виде

$$\psi_1(z)=\frac{A(z,z_2,z_3)}{A(z_1,z_2,z_3)},\quad \psi_2(z)=\frac{A(z_1,z,z_3)}{A(z_1,z_2,z_3)},$$

$$\psi_3(z)=\frac{A(z_1,z_2,z)}{A(z_1,z_2,z_3)},$$

$$z=(x,y),\quad A(z_1,z_2,z_3)=\begin{vmatrix}x_1 & y_1 & 1\\ x_2 & y_2 & 1\\ x_3 & y_3 & 1\end{vmatrix}.$$

Следовательно,

$$\boldsymbol{s}_1(z)=\sum_{i=1}^{3}\boldsymbol{f}_i\psi_i(z).$$

Построим полосу, содержащую функцию f. Пусть $f\in C^2(\Omega)$, тогда [36]

$$s_1(z)-f(z)=\sum_{i=1}^{3}\psi_i(z)R_i(z)\equiv\varphi(z),$$

где

$$R_i(z)=\int_x^{x_i}(x_i-x)\partial_{11}f(v,y_i)dv+$$

$$+\int_y^{y_i}(y_i-v)\partial_{22}f(x,v)dv+(x_i-x)\int_y^{y_i}\partial_{12}f(x,y)dv.$$

Отсюда

$$\varphi(z)\in\sum_{i=1}^{3}\psi_i(z)[\underline{R}_i(z),\bar{R}_i(z)],$$

где

$$\underline{R}_i(z)=\frac{1}{2}(x_i-x)^2\min_{\Omega}\partial_{11}f_-+\frac{1}{2}(y_i-y)^2\min_{\Omega}\partial_{22}f_-+$$

$$+(x_i - x)(y_i - y) \min_{\Omega} \partial_{12} f_-,$$

$$\overline{R}_i(z) = \frac{1}{2}(x_i - x)^2 \max_{\Omega} \partial_{11} f_+ + \frac{1}{2}(y_i - y)^2 \max_{\Omega} \partial_{22} f_+ +$$
$$+(x_i - x)(y_i - y) \max_{\Omega} \partial_{12} f_+.$$

2.8. Интервальные интегралы

Интегрирование непрерывных функций

Пусть $f(x)$, $x \in [a, b]$ — непрерывная вещественная функция, имеющая в качестве интервального расширения функцию $\boldsymbol{f}(x)$, заданную при $x \subset [a, b]$, монотонную по включению и удовлетворяющую условию типа Липшица, т. е. $\mathrm{wid}(\boldsymbol{f}(x)) \leq L\ \mathrm{wid}(\boldsymbol{x})$, где $L \geq 0$ — некоторая постоянная. По теореме о среднем

$$\int_{\underline{x}}^{\overline{x}} f(t)dt = \int_{[\underline{x}, \overline{x}]} f(t)dt = f(\xi)\mathrm{wid}(\boldsymbol{x}),$$

здесь $\xi \in \boldsymbol{x}$. Следовательно,

$$\int_{\boldsymbol{x}} f(t)dt = \boldsymbol{f}(x)\mathrm{wid}(\boldsymbol{x}). \tag{2.51}$$

Возьмем произвольное натуральное число n и рассмотрим разбиение $[a, b]$ на n подынтервалов $\boldsymbol{x}_i = [\underline{x}_i, \overline{x}_i]$, $i = 1, ..., n$, т.е. $[\underline{x}_i, \overline{x}_i] \cap [\underline{x}_j, \overline{x}_j] = \emptyset$, $i \neq j$, $[a, b] = \cup_{i=1}^{n} \boldsymbol{x}_i$. Из свойств аддитивности вытекает

$$\int_a^b f(x)dx = \sum_{i=1}^{n} \int_{\boldsymbol{x}_i} f(t)dt. \tag{2.52}$$

АДля функций удовлетворяющих условию Липшица справедливо утверждение [90].

Теорема 21. *Существует константа L, не зависящая от n, такая, что*

$$\int_a^b f(x)dx \in \sum_{i=1}^{n} f(\boldsymbol{x}_i)\ \mathrm{wid}\ \boldsymbol{x}_i$$

и

$$\mathrm{wid}(\sum_{i=1}^{n} f(\boldsymbol{x}_i)\ \mathrm{wid}\ \boldsymbol{x}_i) \leq L. \sum_{i=1}^{n} \mathrm{wid}(\boldsymbol{x}_i)^2$$

□

Определим

$$\boldsymbol{I}_n = \sum_{i=1}^{n} f(\boldsymbol{x}_i) \operatorname{wid} \boldsymbol{x}_i.$$

Пусть $h = \max \operatorname{wid} \boldsymbol{x}_i$. Из теоремы 21 следует

$$\operatorname{wid} \boldsymbol{I}_n \leq Lh^2.$$

Пусть непрерывная, монотонная по включению интервальная функция $\boldsymbol{f}(t)$ представима в виде

$$\boldsymbol{f}(t) = [\underline{f}(t), \overline{f}(t)].$$

В этом случае [89]

$$\int_a^b \boldsymbol{f}(t)dt = [\int_a^b \underline{f}(t)dt, \int_a^b \overline{f}(t)dt].$$

Если выполнено включение $\boldsymbol{f}(t) \subseteq \boldsymbol{g}(t)$, $\forall t \in [a, b]$, тогда

$$\int_a^b \boldsymbol{f}(t)dt \subseteq \int_a^b \boldsymbol{g}(t)dt.$$

В частности, если $\boldsymbol{f}(t)$ — интервальное расширение $f(t)$, тогда

$$\int_a^b f(t)dt \in \int_a^b \boldsymbol{f}(t)dt.$$

Пример 6. Предположим, что мы хотим вычислить следующий интеграл:

$$I = \int_0^b e^{-t^2} dt.$$

Для функции e^{-t^2} несложно получить интервальное расширение, основанное на рядах Тейлора:

$$e^{-t^2} = 1 - t^2 + \frac{1}{2}t^4 - \frac{1}{3!}t^6 + ... + (-1)^n \frac{1}{n!}t^{2n} + (-1)^{n+1} \frac{1}{(n+1)!} e^{-\theta t^2} t^{2(n+1)}.$$

где $\theta \in [0, 1]$. $e^{-\theta t^2} \in [e^{-t^2}, 1]$. Ограничимся $n = 3$,

$$e^{-t^2} \in 1 - t^2 + \frac{1}{2}t^4 - \frac{1}{3!}t^6 + \frac{1}{4!}[e^{-t^2}, 1]t^8$$

$$(1 - \frac{1}{4!}t^8)e^{-t^2} \geq 1 - t^2 + \frac{1}{2}t^4 - \frac{1}{3!}t^6.$$

$$e^{-t^2} \geq \frac{1 - t^2 + \frac{1}{2}t^4 - \frac{1}{3!}t^6}{(1 - \frac{1}{4!}t^8)} = \underline{f}(t).$$

$$e^{-t^2} \le 1 - t^2 + \frac{1}{2}t^4 - \frac{1}{3!}t^6 + \frac{1}{4!}t^8 = \overline{f}(t).$$

Итак, мы можем вычислить интегралы от функций $\underline{f}(t)$, $\overline{f}(t)$.

Перейдем к случаю вычисления интервальных интегралов, основанные на численных квадратурах. Пусть $f(x)$ — непрерывная функция с интервальной константой Липшица $\boldsymbol{L} = [\underline{L}, \overline{L}]$:

$$f(x) - f(y) \in \boldsymbol{L}(x - y).$$

Предположим, что известны значения $\boldsymbol{f}_i = f(\boldsymbol{x}_i)$, $i = 1, 2, ..., n$.

Предположим также, что функция не является линейной, т. е. $[\underline{L} \ne \overline{L}]$, $\boldsymbol{x}_i \cap \boldsymbol{x}_{i+1} = \emptyset$. Заметим, что на отрезке $[\overline{x}_i, \underline{x}_{i+1}]$ функция $f(x)$ не превосходит величины $\min\{\overline{f}(\overline{x}_i) + \overline{L}(x - \overline{x}_i), \overline{f}(\underline{x}_{i+1}) + \underline{L}(x - \underline{x}_{i+1})\}$. Следовательно, выполнено неравенство

$$\int_{\overline{x}_i}^{\underline{x}_{i+1}} f(t)dt \le (\overline{f}(\overline{x}_i) + \overline{L}(\xi - \overline{x}_i)/2)(\xi - \overline{x}_i) + (\overline{f}(\underline{x}_{i+1}) + \underline{L}(\xi - \underline{x}_{i+1})/2)(\underline{x}_{i+1} - \xi).$$

где $\xi = (\overline{f}(\underline{x}_{i+1}) + \overline{f}(\overline{x}_i) + \overline{L}\overline{x}_i - \underline{L}\underline{x}_{i+1})/(\overline{L} - \underline{L})$.

Метод трапеций

Пусть в узлах сетки $a = x_0 < x_1 <, ..., < x_n = b$ известны значения $f(x_i) \in \boldsymbol{f}_i$ и известна оценка второй производной функции $f''(x) \in [\underline{F}_i^{(2)}, \overline{F}_i^{(2)}]$, $\forall x \in [x_{i-1}, x_i]$. Тогда справедливо включение

$$\int_a^b f(t)dt \in \frac{1}{2}\sum_{i=1}^{n}(x_i - x_{i-1})(\boldsymbol{f}_i + \boldsymbol{f}_{i-1}) + \boldsymbol{R},$$

where $\boldsymbol{R} = -\frac{1}{12}\sum_{i=1}^{n}(x_i - x_{i-1})^3 \boldsymbol{F}(2)_i$.

В случае равномерной сетки $x_i - x_{i-1} = h$ эти формулы упрощаются:

$$\int_a^b f(t)dt \in h\sum_{i=1}^{n-1}\boldsymbol{f}_i + \frac{h}{2}(\boldsymbol{f}_0 + \boldsymbol{f}_n) + \boldsymbol{R},$$

where $\boldsymbol{R} = -\frac{b-a}{12}h^2[\min_i \underline{F}_i^{(2)}, \max_i \overline{F}_i^{(2)}]$.

Применение сплайнов

Обратимся к вопросу о применении изложенной теории аппроксимации для интегрирования функций, заданных с ошибкой. Пусть $f^{(j)}(x_i) \in \boldsymbol{f}_{i,j}$, $i = 0, 1, ..., N; j = 0, 1, ..., l - 1$ и известна априорная информация об

f, позволяющая построить $\boldsymbol{r}(x)$. Кроме того, пусть $\boldsymbol{s}(x)$ — сплайн, аппроксимирующий f и такой, что

$$f(x) \in \boldsymbol{s}(x) + \boldsymbol{r}(x), \quad x \in [a, b]. \tag{2.53}$$

Тогда

$$\int_a^b f(x)dx \in \int_a^b (\boldsymbol{s}(x) + \boldsymbol{r}(x))dx = [\underline{I}, \bar{I}].$$

Интеграл в правой части, обычно, можно вычислить точно, особенно просто — в случае полиномиальных сплайнов, когда на каждом отрезке $[x_l, x_{l+1}]$

$$\boldsymbol{s}(x) + \boldsymbol{r}(x) = \sum_{i=0}^{2m-1} \boldsymbol{a}_i x^i. \tag{2.54}$$

Интеграл от интервального полинома снова будет интервальным полиномом вида

$$\int_a^b \sum_{i=0}^{2m-1} \boldsymbol{a}_i x^i dx = \sum_{i=0}^{2m-1} (\boldsymbol{a}_i x^{i+1}/(i+1))|_a^b.$$

Используя развитый ранее прием интерполяции, можно находить интегралы многомерных функций, которые заданы в конечном наборе точек и для которых известны оценки старших производных. В отличие от теории интегрирования, развитой в [39, 54], данный подход не требует привлечения такого понятия, как интервальное расширение функций, заменяя его априорной информацией о производных. В ряде случаев это бывает полезным, например, когда интервальное расширение не известно или очень грубое.

Обратимся теперь к вопросу апостериорного оценивания ошибок интегрирования вещественных функций. Пусть

$$I(x) = \int_a^x f(\tau)d\tau, \quad x \in (a, b). \tag{2.55}$$

Далее обозначим через $I^h(x)$ приближенное значение интеграла (2.55), найденное каким-либо численным методом. Оценим погрешность $I(x) - I^h(x)$. Для этого построим полином $\mathcal{P}_n(a, b)$, интерполирующий функцию $I^h(x)$ следующим образом:

$$s(a) = I^h(a) = 0, \quad s(b) = I^h(b), \tag{2.56}$$

$$\frac{ds}{dx}(\xi_i) = f(\xi_i), \quad i = 1, 2, ..., n-2, \tag{2.57}$$

где $\xi_i \in [a,b]$, $\{a = \xi_1 < \xi_2 < ... < \xi_{n-2} = b\}$ — набор внутренних узлов.

Обозначим $\varepsilon(x) = I(x) - s(x)$. Поскольку

$$\frac{dI(x)}{dx} = f(x), \quad x \in (a,b),$$

то

$$\frac{d\varepsilon(x)}{dx} = \frac{dI(x)}{dx} - \frac{ds(x)}{dx} = \varphi(x), \tag{2.58}$$

где $\varphi(x) = f(x) - \frac{ds(x)}{dx}$. Следовательно,

$$(x-a)\min\varphi(x) \le \varepsilon(x) \le (x-a)\max\varphi(x). \tag{2.59}$$

Заметим, что в определение φ входят конкретные функции f и s.

Остановимся на вопросе точности двустороннего неравенства (2.59). Для этого нам понадобятся оценки φ и ε:

$$|\varepsilon^{(\nu)}(x)| = |I^{(\nu)}(x) - s^{(\nu)}(x)| \le |I^{(\nu)}(x) - s_T^{(\nu)}(x)| +$$

$$+ |s_T^{(\nu)}(x) - s^{(\nu)}(x)|, \quad \nu = 0,1. \tag{2.60}$$

Здесь $s_T \in \mathcal{P}_n(a,b)$— полином, интерполирующий $I(x)$ по формулам (2.56)-(2.58). Первое слагаемое в правой части неравенства (2.60) оценивается так:

$$|I^{(\nu)}(x) - s_T^{(\nu)}(x)| \le K(b-a)^{n+1-\nu}\|f\|_{n,\infty}.$$

Для оценки второго слагаемого представим полином $s_T - s$ в виде

$$s_T(x) - s(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n.$$

Заметим, что в силу построения s_T, s справедлива оценка $|a_i| \le K\varepsilon(b)/(b-a)^i$. Следовательно,

$$|\varphi| \le K_1(b-a)^n\|f\|_{n,\infty} + K_2\varepsilon(b)/(b-a)$$

и

$$|I(x) - s(x)| \le (K_1(b-a)^n\|f\|_{n,\infty} + K_2\varepsilon(b)/(b-a))(x-a).$$

Для того чтобы построенная апостериорная оценка погрешности интеграла соответствовала истинной погрешности, необходимо согласовать порядки величин $(b-a)^{n+1}\|f\|_{n,\infty}$ и $\varepsilon(b)$, т.е. степень полинома должна соответствовать порядку точности метода интегрирования.

2.9. Вопросы и упражнения

1. Постройте естественное и объединенное интервальное расширения для следующих функций:

а) $f(x) = 2x^2 + 3x - 5$;

б) $f(x_1, x_2) = x_1 x_2 + x_1 + x_2 + 1$.

2. Будет ли следующая функция интервальным расширением некоторой функции $f(x)$. Если да, то какой?

а) $\boldsymbol{f}(\boldsymbol{x}) = 2[\underline{x}, \overline{x}] + [-1, 1]$;

b)

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{cases} [0, \max\{\underline{x}^2, \overline{x}^2\}] + 2 & \text{если } \underline{x}\overline{x} < 0; \\ [\underline{x}^2, \overline{x}^2] + 2 & \text{если } \underline{x} \geq 0; \\ [\overline{x}^2, \underline{x}^2] + 2 & \text{если } \underline{x} \leq 0. \end{cases}$$

3. Найдите интервальный интеграл на $[a, b]$ для функции

$$\boldsymbol{f}(x) = [1, 2]x^2 + [-1, 1]x + [1, 0].$$

4. Используя формулу Тейлора, найдите интервальную оценку интеграла

$$\int_0^{\pi/2} sin(x)dx.$$

5. Постройте интервальный эрмитов кубический сплайн на отрезке $[0, 1]$, если $\boldsymbol{f}_0 = [0.5, 1]$, $\boldsymbol{f}_1 = [0, 0.5]$, $\boldsymbol{f}'_0 = [-0.5, -0.5]$, $\boldsymbol{f}'_1 = [-0.5, -0.5]$.

6. Найдите сумму двух гистограммных чисел $\boldsymbol{a}= \{x \in [1, 3], p_a = 0.5\}$, $\boldsymbol{b}= \{x \in [2, 4], p_b = 0.5\}$.

Глава 3

Алгебраические задачи

3.1. Нормированные пространства

Пусть $\mathcal{L}$ — линейное пространство, $x, y \in \mathcal{L}$ — вектора, $\lambda \in R$.

Определение 6. *Функция $\varphi : \mathcal{L} \to R$ называется нормой, если выполнены следующие условия:*
1) $\varphi(x) > 0$ для всех $x \neq \emptyset$;
2) $\varphi(\lambda x) = |\lambda|\varphi(x)$;
3) $\varphi(x + y) \leq \varphi(x) + \varphi(y)$.

Линейное пространство, в котором задана норма, называется *нормированным*, норму вектора обозначают $||x||$. Заметим, что из 2) вытекает $\varphi(\emptyset) = 0$.

В нормированном пространстве мы можем определить *расстояние* $\rho(x, y)$ между векторами x и y как $||x - y||$.

Рассмотрим примеры норм в R^n:
1) $||x||_1 = \sum_i |x_i|$ — l-норма;
2) $||x||_2 = (\sum_i |x_i|^2)^{1/2}$ — евклидова норма;
3) $||x||_p = \sum_i |x_i|^p)^{1/p}$ — норма Гельдера, или l_p-норма;
4) $||x||_\infty = \max_i |x_i|$ — c-норма.

В R^n можно ввести скалярное произведение $(x, y) = \sum_i x_i y_i$, тогда $(x, x) = ||x||_2^2$.

Эквивалентность норм

Пусть φ — норма в линейном пространстве $\mathcal{L}$, тогда $\psi = \alpha\varphi$ $(\alpha > 0)$ также является нормой.

Будем говорить, что норма φ *мажорирует* норму ψ: $\psi \leq \varphi$, если $\forall x \in \mathcal{L}$ выполнено неравенство $\psi(x) \leq \varphi(x)$.

Нормы ψ и φ называются *эквивалентными*, если существуют положительные α_1 и α_2 такие, что

$$\alpha_1\varphi \le \psi \le \alpha_2\varphi.$$

Теорема 22. *В R^n l-норма, c-норма и евклидова норма эквивалентны.*

Доказательство. Для любой нормы $||\cdot||$ справедливо

$$||x|| = ||\sum_i x_i \boldsymbol{e}_i|| \le \alpha \sum_i |x_i| = \alpha||x||_1, \tag{3.1}$$

где $\boldsymbol{e}_i$ — базисные вектора и $\alpha = \max_i ||\boldsymbol{e}_i||$. Но (3.1) можно оценить по-другому:

$$||x|| \le \max_i |x_i| \sum_i ||\boldsymbol{e}_i|| = \beta||x||_\infty, \tag{3.2}$$

где $\beta = \sum_i ||\boldsymbol{e}_i||$. Из (3.1) и (3.2) получаем

$$\alpha^{-1}||x||_\infty \le ||x||_1 \le \beta||x||_\infty.$$

Доказана эквивалентность l-нормы и c-нормы. Из оценок

$$\max_i |x_i| \le \left(\sum_i |x_i|^2\right)^{1/2} \le \sqrt{n}\max_i |x_i|$$

вытекает эквивалентность c-нормы и евклидовой нормы. Теорема доказана. □

Справедлива более общая теорема [9].

Теорема 23. *В R^n все нормы эквивалентны.*

Нормы матриц

Для матриц как элементов линейного пространства могут быть введены различные нормы.

Определение 7. *Матричная норма называется согласованной с векторной нормой, если для любой матрицы и вектора выполнено неравенство*

$$||Ax|| \le ||A||||x||. \tag{3.3}$$

Можно построить согласованные нормы следующим образом:

$$||A|| = \sup_{||x||=1} ||Ax||. \tag{3.4}$$

Определение 8. *Норма, определенная формулой (3.4), называется индуцированной.*

Теорема 24. *Каждая согласованная норма мажорирует индуцированную.*

Если $||E|| = 1$, то говорят, что "норма сохраняет единицу". Нормы, удовлетворяющие условию

$$||AB|| \leq ||A|| \cdot ||B||, \tag{3.5}$$

называются *кольцевыми нормами.* Для кольцевых норм

$$||A|| \leq ||E|| \cdot ||A||,$$

следовательно, $||E|| \geq 1$. Кроме того, справедливо

$$||A^k|| \leq ||A||^k, k = 1, 2, ...$$

и

$$||A^{-1}|| \leq ||A||^{-1}.$$

Теорема 25. *Любая индуцированная норма сохраняет единицу и обладает кольцевым свойством.*

Первая часть очевидна. Вторая вытекает из оценки:

$$||AB|| = \sup_{||x||=1} ||A(Bx)|| = ||A|| \sup_{||x||=1} ||Bx|| = ||A|| \cdot ||B||.$$

Определение 9. *Спектральным радиусом матрицы A называется число*

$$\rho(A) = |\lambda_{max}|,$$

где $|\lambda_{max}|$ — максимальное собственное число матрицы A.

Будем обозначать индуцированные нормы аналогично векторным нормам. Наиболее употребительны следующие нормы [16]:

$$||A||_\infty = \max_i \sum_j |a_{ij}|,$$

$$||A||_1 = \max_j \sum_i |a_{ij}|.$$

Матричная норма, индуцированная евклидовой, равна

$$||A||_2 = \sqrt{\rho(AA^*)} = \alpha_{max},$$

где α_{max} — максимальное сингулярное число. В случае самосопряженных матриц $A = A^*$

$$||A||_2 = \rho(A).$$

Для невырожденной матрицы

$$||A^{-1}||_2 = \alpha_{min}^{-1},$$

где α_{min} — минимальное сингулярное число. Часто используется *евклидова* норма матрицы

$$||A||_E = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}.$$

Евклидова норма обладает кольцевым свойством и является согласованной евклидовой нормой:

$$||Ax||_2 \leq ||A||_E ||x||_2.$$

Не существует векторной нормы, которая индуцировала бы евклидову матричную норму.

Заметим, что для любой индуцированной (согласованной) матричной нормы и v — собственного вектора справедливо

$$||\lambda v|| = ||Av|| \leq ||A||||v||.$$

Следовательно, наибольшее по модулю собственное число не превосходит матричную норму:

$$\max_i |\lambda_i| \leq ||A||.$$

Поскольку $\max_i |\lambda_i| = \rho(A)$, то последнее неравенство можно переписать в виде

$$\rho(A) \leq ||A||.$$

Определение 10. *Матрица $A = (a_{ij})$ называется M-матрицей, если $a_{ij} \leq 0$, $i \neq j$ и выполнено одно из эквивалентных условий:*

1) $detA \neq 0$ и $A^{-1} \geq 0$;

2) $a_{ii} > 0$, $i = 1, 2, ..., n$ и $\rho(E - D^{-1}A) < 1$, где D — диагональ матрицы A;

3) все собственные значения матрицы A имеют положительные вещественные части;

4) для каждого вектора x: $Ax \geq 0$ вытекает $x \geq 0$.

Если B — M-матрица и $A \geq B$, то A — M-матрица.

3.2. Прямые методы

Рассмотрим систему линейных алгебраических уравнений

$$Ax = b, \tag{3.6}$$

где $b = (b_i), i = 1, \ldots, n$ — известный вектор, $A = (a_{i,j}), i, j = 1, \ldots, n$ — невырожденная матрица. Тогда вектор $x = A^{-1}b$ — решение системы (3.6). Предположим, что A и b содержат ошибки и известно, что их элементы принадлежат соответствующим интервальным числам

$$a_{i,j} \in \boldsymbol{a}_{i,j},$$

$$b_i \in \boldsymbol{b}_i, i, j = 1, \ldots, n.$$

Пусть также, все матрицы из множества $\boldsymbol{A}$ не вырождены. Множество векторов

$$\mathcal{X} = \{x | Ax = b, A \in \boldsymbol{A},\ b \in \boldsymbol{b}\}$$

назовем *множеством решений системы*

$$\boldsymbol{Ax} = \boldsymbol{b}. \tag{3.7}$$

Пример 7. Пусть необходимо решить систему интервальных линейных алгебраических уравнений

$$\boldsymbol{Ax} = \boldsymbol{b}$$

с матрицей $\boldsymbol{A}$ и правой частью $\boldsymbol{b}$.

$$\boldsymbol{A} = \begin{pmatrix} [2,4] & [-2,1] \\ [-1,2] & [2,4] \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} [-2,2] \\ [-2,2] \end{pmatrix}. \tag{3.8}$$

Множество векторов $\mathcal{X}$ для этой задачи изображено на рис. 3.1.

Минимальным интервальным вектором, содержащим множество ее решений, является интервальный вектор $\boldsymbol{x} = ([-4,4],[-4,4])^T$.

Множество $\mathcal{X}$ может быть описано следующим образом [72]:

$$\mathcal{X} = \{x | x \in R^n, \boldsymbol{A}x \cap \boldsymbol{b} \neq \emptyset\} \tag{3.9}$$

или

$$\{x | x \in R^n, 0 \in \boldsymbol{A}x - \boldsymbol{b}\}.$$

Справедливо следующее утверждение [95].

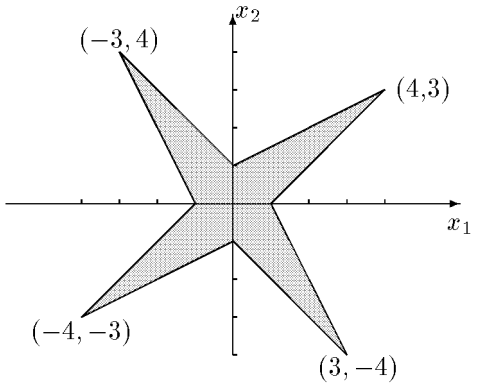

Рис. 3.1. Множество решений системы (3.8)

Замечание 1.

$$x \in \mathcal{X} \Leftrightarrow |\operatorname{mid} Ax - \operatorname{mid} b| \leq \operatorname{rad}(\boldsymbol{b}).$$

Поставим задачу найти интервальный вектор $\boldsymbol{x} \in \boldsymbol{R}^n$, содержащий множество $\mathcal{X}$.

Самым простым способом для нашего примера будет метод Крамера. Действительно, для СЛАУ $A = (a_{ij})$, $b = (b_i)$, $i, j = 1, 2$. Решение находится по формулам

$$\boldsymbol{x}_1 = \frac{\Delta_1}{\Delta}, \boldsymbol{x}_2 = \frac{\Delta_2}{\Delta},$$

где

$$\Delta = a_{11}a_{22} - a_{12}a_{21},$$

$$\Delta_1 = b_1a_{22} - b_2a_{21},$$

$$\Delta_2 = a_{11}b_2 - a_{12}b_1.$$

Для решения СЛАУ с интервальными коэффициентами из примера (3.8) сделаем интервальное расширение формул Крамера. Непосредственными вычислениями получаем $\boldsymbol{x}_1 = [-6, 6]$, $\boldsymbol{x}_2 = [-6, 6]$.

Как видим, ответ несколько шире оптимального. Этот факт легко объясняется, поскольку рациональные выражения для $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ не удовлетворяют условию теоремы 13, т.е. содержат переменные более одного раза.

Применять формулы Крамера для решения больших систем крайне не рационально. В вычислительной математике для этих целей используют прямые методы: метод Гаусса, LU- и QR-разложения и т. д.

Рассмотрим еще ряд примеров.

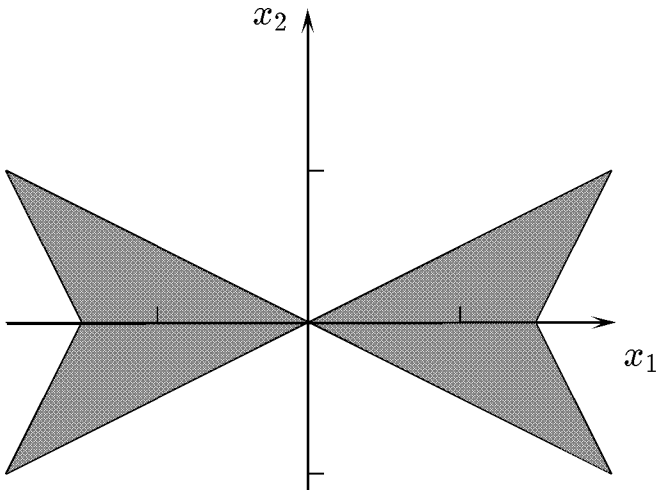

Рис. 3.2. Множество решений системы (3.10)

Пример 8. Пусть необходимо решить систему интервальных линейных алгебраических уравнений

$$\boldsymbol{A} = \begin{pmatrix} [2,4] & [-1,1] \\ [-1,1] & [2,4] \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} [-3,3] \\ [0,0] \end{pmatrix}. \tag{3.10}$$

Множество векторов $\mathcal{X}$ этой задачи согласно замечанию 1

$$\mathcal{X} = \{x | x \in R^2, 2|\boldsymbol{x}_2| \leq |\boldsymbol{x}_1|, 2|\boldsymbol{x}_1| \leq 3 + |\boldsymbol{x}_2|\}$$

и изображено на рис. 3.2.

Пример 9. Пусть

$$\boldsymbol{A} = \begin{pmatrix} 2 & \alpha \\ \beta & 2 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} 1.2 \\ -1.2 \end{pmatrix}, \quad \alpha, \beta \in [0,1]. \tag{3.11}$$

Заметим, что по правилу Крамера

$$\boldsymbol{x}_1 = 1.2(2-\alpha)/(4-\alpha\beta),$$

$$\boldsymbol{x}_2 = -1.2(2-\beta)/(4-\alpha\beta).$$

Следовательно,

$$\square\mathcal{X} = \begin{pmatrix} [0.3, 0.6] \\ [-0.6, -0.3] \end{pmatrix}.$$

Применяя формально правило Крамера к системе (3.11), мы получаем интервальный вектор $([0.3, 1.2], [-0.8, 1.2])$.

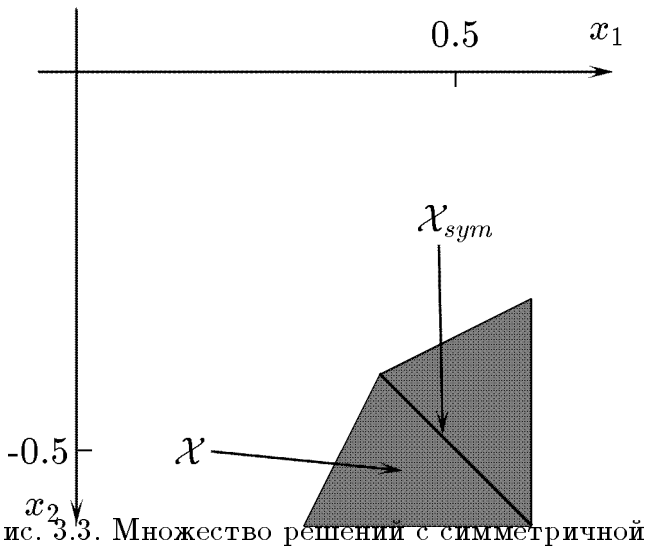

Рис. 3.3. Множество решений с симметричной матрицей

Найдем обратную матрицу

$$A^{-1} = \square \left\{ \frac{1}{4-\alpha\beta} \begin{pmatrix} 2 & \alpha \\ \beta & 2 \end{pmatrix} \;\middle|\; \alpha, \beta \in [0,1] \right\} =$$

$$= \begin{pmatrix} [1/2, 2/3] & [0, 1/3] \\ [0, 1/3] & [1/2, 2/3] \end{pmatrix}.$$

Несложно убедиться, что

$$A^{-1}b \neq \square\mathcal{X}.$$

Пусть A — симметричная матрица

$$A = \begin{pmatrix} 2 & \alpha \\ \alpha & 2 \end{pmatrix}, \alpha \in [0,1].$$

Множество решений системы линейных алгебраических уравнений с симметричной матрицей представляет собой отрезок с концами $(0.4, -0.4)$, $(0.6, -0.6)$. Интервальная оболочка множества решений с симметричной матрицей

$$\square\mathcal{X}_{sym} = \begin{pmatrix} [0.4, 0.6] \\ [-0.6, -0.4] \end{pmatrix} \neq \square\mathcal{X}.$$

Метод Гаусса и LU-разложение

Интервальный метод Гаусса представляет корректное интервальное расширение метода Гаусса для СЛАУ.

Рассмотрим систему интервальных линейных алгебраических уравнений с квадратной матрицей размерности n:

$$\begin{pmatrix} \boldsymbol{a}_{11}, & \boldsymbol{a}_{12}, & \dots & \boldsymbol{a}_{1n} \\ \boldsymbol{a}_{21}, & \boldsymbol{a}_{22}, & \dots & \boldsymbol{a}_{2n} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{a}_{n1}, & \boldsymbol{a}_{n2} & \dots & \boldsymbol{a}_{nn} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}. \tag{3.12}$$

Метод Гаусса состоит из двух ходов: прямого и обратного. На прямом ходе с помощью элементарных преобразований будем последовательно стремиться привести систему (3.12) к верхнетреугольному виду.

Первый шаг. Преобразуем систему к виду

$$\begin{pmatrix} \boldsymbol{a}_{11}, & \boldsymbol{a}_{12}, & \dots & \boldsymbol{a}_{1n} \\ 0, & \boldsymbol{a}_{22}^{(1)}, & \dots & \boldsymbol{a}_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0, & \boldsymbol{a}_{n2}^{(1)} & \dots & \boldsymbol{a}_{nn}^{(1)} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2^{(1)} \\ \vdots \\ \boldsymbol{b}_n^{(1)} \end{pmatrix}, \tag{3.13}$$

где

$$\boldsymbol{a}_{ik}^{(1)} = \boldsymbol{a}_{ik} - \boldsymbol{l}_{i1}\boldsymbol{a}_{1k},$$

$$\boldsymbol{b}_i^{(1)} = \boldsymbol{b}_i - \boldsymbol{l}_{i1}\boldsymbol{b}_1, \boldsymbol{l}_{i1} = \boldsymbol{a}_{i1}/\boldsymbol{a}_{11}.$$

Далее переходим к подсистеме с матрицей $\boldsymbol{A}^{(1)}$ и правой частью $\boldsymbol{b}^{(1)}$ размерности $n-1$. Применим к этой системе первый шаг и т. д. Окончательно приводим систему (3.12) к верхнетреугольному виду

$$\begin{pmatrix} \boldsymbol{a}_{11} & \boldsymbol{a}_{12} & \dots & \boldsymbol{a}_{1n} \\ 0 & \boldsymbol{a}_{22}^{(*)} & \dots & \boldsymbol{a}_{2n}^{(*)} \\ \vdots & \vdots & & \vdots \\ 0, & 0 & \dots & \boldsymbol{a}_{nn}^{(*)} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2^{(*)} \\ \vdots \\ \boldsymbol{b}_n^{(*)} \end{pmatrix}. \tag{3.14}$$

Обратный ход. Далее последовательно вычисляем неизвестные $\boldsymbol{x}_n, \boldsymbol{x}_{n-1}, \dots, \boldsymbol{x}_1$:

$$\boldsymbol{x}_n = \boldsymbol{b}_n^{(*)}/\boldsymbol{a}_{nn}^{(*)},$$

$$\boldsymbol{x}_i = \frac{\boldsymbol{b}_i - \sum_{j=n}^{i+1} \boldsymbol{a}_j^{(*)} \boldsymbol{x}_j}{\boldsymbol{a}_{ii}^{(*)}} \quad i = n-1, \dots, 1.$$

Интервальный метод Гаусса не всегда можно реализовать даже для регулярных матриц $\boldsymbol{A}$.

Пример Райхмана

Рассмотрим интервальную матрицу

$$S(a) = \begin{pmatrix} 1 & [0,a] & [0,a] \\ [0,a] & 1 & [0,a] \\ [0,a] & [0,a] & 1 \end{pmatrix}.$$

Проводя прямой ход метода Гаусса, получаем

$$S(a) \sim \begin{pmatrix} 1 & [0,a] & [0,a] \\ 0 & 1 & [-a^2/(1-a^2), a/(1-a^2)] \\ 0 & 0 & [1-a^2-a^2/(1-a^2), 1+a^3(1-a^2)] \end{pmatrix}.$$

Несложно убедиться, для любого $a \in [(\sqrt{5}-1)/2, 1]$ следует, что $0 \in [1-a^2-a^2/(1-a^2), 1+a^3(1-a^2)]$. Таким образом, метод Гаусса не может быть закончен.

Теорема 26. *Если матрица* $\boldsymbol{A}$ *— M-матрица или имеет диагональное преобладание, то решение* $\boldsymbol{x}_G$ *по методу Гаусса существует и*

$$\square\mathcal{X} \subseteq \boldsymbol{x}_G.$$

Согласно методу Холецкого матрицу $\boldsymbol{A}$ можно представить в виде произведения двух треугольных матриц $\boldsymbol{L}, \boldsymbol{U}$.

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{l}_{11}, & 0 & \dots & 0 \\ \boldsymbol{l}_{21}, & \boldsymbol{l}_{22} & \dots & 0 \\ \vdots & & & \vdots \\ \boldsymbol{l}_{n1}, & \boldsymbol{l}_{n2} & \dots & \boldsymbol{l}_{nN} \end{pmatrix}, U = \begin{pmatrix} 1, & \boldsymbol{u}_{12} & \dots & \boldsymbol{u}_{1n} \\ 0, & 1 & \dots & \boldsymbol{u}_{2n} \\ \vdots & & & \vdots \\ 0, & 0 & \dots & 1 \end{pmatrix}.$$

Коэффициенты матриц $\boldsymbol{L}, \boldsymbol{U}$ находят по формулам:

$$\boldsymbol{l}_{i1} = \boldsymbol{a}_{i1} \quad \text{для} \quad i = 1, ..., n,$$

$$\boldsymbol{u}_{1i} = \boldsymbol{a}_{1i}/\boldsymbol{l}_{11} \quad \text{для} \quad i = 2, ..., n.$$

Далее справедливы рекуррентные соотношения:

$$\boldsymbol{l}_{is} = (\boldsymbol{a}_{is} - \sum_{k=1}^{s-1} \boldsymbol{l}_{ik}\boldsymbol{u}_{ki}), i = s, ..., n; s = 2, 3, ..., n,$$

$$\boldsymbol{u}_{si} = (\boldsymbol{a}_{is} - \sum_{k=1}^{s-1} \boldsymbol{l}_{ik}\boldsymbol{u}_{ki})/\boldsymbol{l}_{ss}, i = s+1, ..., n; s = 2, ..., n-1.$$

Если матрицы $\boldsymbol{L}, \boldsymbol{U}$ построены, предлагается следующий алгоритм. Сначала решают вспомогательную систему

$$\boldsymbol{L}y = \boldsymbol{b}.$$

Так как матрица $\boldsymbol{L}$ треугольная, то нетрудно выписать рекуррентные соотношения

$$\boldsymbol{y}_1 = \boldsymbol{b}_1/\boldsymbol{l}_{11}, \boldsymbol{y}_2 = (\boldsymbol{b}_2 - \boldsymbol{l}_{12}\boldsymbol{y}_1)/\boldsymbol{l}_{22}, ...,$$

$$\boldsymbol{y}_n = (\boldsymbol{b}_n - \sum_{i=1}^{n-1} \boldsymbol{l}_{ni}\boldsymbol{y}_i)/\boldsymbol{l}_{nn}.$$

Зная вектор $\boldsymbol{y}$, находят решение системы (3.12) с помощью треугольной матрицы $\boldsymbol{U}$.

$$\boldsymbol{U}x = \boldsymbol{y}.$$

Аналогично выписывают рекуррентные соотношения

$$\boldsymbol{x}_n = \boldsymbol{y}_n, \boldsymbol{x}_{n-1} = \boldsymbol{y}_{n-1} - \boldsymbol{u}_{n-1,n}\boldsymbol{x}_n, ...,$$

$$\boldsymbol{x}_1 = \boldsymbol{y}_1 - \sum_{i=2}^{n} \boldsymbol{u}_{1i}\boldsymbol{x}_i.$$

Метод LU-разложения наиболее предпочтителен, если систему (3.12) необходимо решать для разных правых частей b. Один раз нужно найти матрицы L, U, а затем пользоваться обратным ходом нахождения векторов Y, X.

Пример 10. Рассмотрим

$$\boldsymbol{A} = \begin{pmatrix} 2 & [-1,0] \\ [-1,0] & 2 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} 1.2 \\ -1.2 \end{pmatrix}. \tag{3.15}$$

$\boldsymbol{A}$ — M-матрица с LU-разложением

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 \\ [-0.5,0] & 1 \end{pmatrix}, \boldsymbol{U} = \begin{pmatrix} 2 & [-1,0] \\ 0 & [1.5,2] \end{pmatrix}.$$

Следовательно,

$$\boldsymbol{x} = \begin{pmatrix} [0.2,0.6] \\ [-0.8,-0.3] \end{pmatrix}.$$

3.3. Итерационные методы

Сначала предварительно рассмотрим итерационные методы решения систем линейных алгебраических уравнений

$$Ax = b, \tag{3.16}$$

где
$x = (x_i), i = 1, \ldots, n$ — вектор решения,
$b = (b_i), i = 1, \ldots, n$ — известный вектор,
$A = (a_{i,j}), i, j = 1, \ldots, n$ — невырожденная матрица.
Для применения итерационных методов решения система (3.16) должна быть представлена в виде

$$x = Bx + c. \tag{3.17}$$

Приведение системы (3.16) к виду (3.17) можно выполнить следующим образом. Пусть H — невырожденная матрица, умножим обе части равенства (3.16) на $-H$ и перенесем b в левую часть

$$-H(Ax - b) = 0$$

Добавляя x в обе части последнего равенства, получаем

$$x - H(Ax - b) = x$$

систему эквивалентную (3.17).

Для нахождения вектора x применим метод простой итерации

$$x_{i+1} = Bx_i + c, i = 0, 1, 2... \tag{3.18}$$

Теорема 27. *Для сходимости метода простой итерации (3.18) при любом начальном приближении x_0 необходимо и достаточно, чтобы*

$$\rho(B) < 1. \tag{3.19}$$

Следствие 1. *Достаточным условием сходимости метода простой итерации (3.18) при любом начальном приближении x_0 является условие $||B|| < 1$.*

Пример 11. Пусть

$$B = \begin{pmatrix} 0.5 & 0.5 \\ -0.5 & 0.5 \end{pmatrix}, c = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

Тогда спектральный радиус $\rho(B) = 0.5$, но $||B||_\infty = 1$.

Перейдем к интервальным системам линейных алгебраических уравнений

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{3.20}$$

Для преобразования системы (3.20) к виду (3.17) можно использовать матрицу $H = (\mathrm{mid}\,\boldsymbol{A})^{-1}$, получаем

$$\boldsymbol{x}_{i+1} = \boldsymbol{B}\boldsymbol{x}_i + \boldsymbol{c}, i = 0, 1, 2... \tag{3.21}$$

Необходимое и достаточное условие сходимости интервальных методов простой итерации дает теорема [3].

Теорема 28. *Метод простой итерации (3.21) сходится к единственной неподвижной точке уравнения*

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{c}$$

при любом начальном $\boldsymbol{x}_0$ *тогда и только тогда, когда*

$$\rho(|\boldsymbol{B}|) < 1. \tag{3.22}$$

Следует обратить внимание на отличие условия (3.22) от (3.19).

Заметим, что, поскольку $\rho(|B|) = 1$, для $\boldsymbol{x}_0 = ([0, 2], [0, 2])$ и матрицы B из последнего примера интервальный метод простой итерации 3.21 не сходится и $\boldsymbol{x}_i = \boldsymbol{x}_0$, $i = 1, 2, ...$

Из неравенства $\rho(|\boldsymbol{B}|) \leq ||\boldsymbol{B}||_\infty$ вытекает достаточное условие сходимости метода простой итерации $||\boldsymbol{B}||_\infty < 1$.

Заметим, что если множество решений интервальной СЛАУ $\mathcal{X} \subseteq \boldsymbol{x}_0$, то $\mathcal{X} \subseteq \boldsymbol{x}_i, i = 1, 2, 3....$

3.4. Уточнение решений

Предлагаемый метод основан на минимизации специальных функционалов. С помощью функционалов последовательно находятся границы интервального решения системы линейных интервальных алгебраических уравнений. Метод нахождения границ интервальных решений рассматривался также в работе Хансена [83]. В ней для определения границ интервального решения приходилось дополнительно решать системы линейных алгебраических уравнений, но более простой структуры. В

отличие от этой работы предлагаемый метод позволяет уточнять известные интервальные решения до необходимой точности.

Рассмотрим систему линейных алгебраических уравнений

$$Ax = b, \tag{3.23}$$

где $b = (b_i), i = 1, \ldots, n$— известный вектор, $A = (a_{i,j}), i, j = 1, \ldots, n$— невырожденная матрица. Тогда вектор $x = A^{-1}b$ является решением системы (3.23). Предположим, что A и b содержат ошибки и известно, что их элементы принадлежат соответствующим интервальным числам

$$a_{i,j} \in \mathbf{a}_{i,j},$$
$$b_i \in \mathbf{b}_i, i, j = 1, \ldots, n.$$

Пусть все матрицы из множества $\boldsymbol{A}$ не вырождены. Множество векторов

$$\mathcal{X} = \{x | Ax = b, A \in \boldsymbol{A},\ b \in \boldsymbol{b}\}$$

назовем множеством решений системы

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \tag{3.24}$$

Поставим задачу найти интервальный вектор $\boldsymbol{x} \in \boldsymbol{R}^n$, содержащий множество $\mathcal{X}$.

Решение задачи (3.23) доставляет минимум функционалу

$$\Phi(x, A, b) = (Ax - b, Ax - b),$$

где $(,)$ — евклидово скалярное произведение. Определим функционал

$$T(\mathbf{v}, \mathbf{w}) = \min_{v \in \mathbf{v}, v \in \mathbf{v}} \sum_{i=1}^{n} (v_i - w_i)^2.$$

Обратим внимание, что $T(Ax, b) = \Phi(x, A, b)$. Поэтому очевиден следующий результат.

Лемма 5. *Равенство $T(\boldsymbol{A}x, \boldsymbol{b}) = 0$ — необходимое и достаточное условие принадлежности произвольного вектора x множеству $\mathcal{X}$.*

Обозначим

$$\rho = \sup_{A \in \boldsymbol{A}, ||x||=1} (Ax, Ax) > 0,$$
$$\gamma = \inf_{A \in \boldsymbol{A}, ||x||=1} (Ax, Ax) > 0.$$

Зафиксируем некоторое целое i: $1 \le i \le n$ и введем функционалы

$$D^R(x,\alpha,i) = -x_i + \alpha T(\boldsymbol{A}x, \boldsymbol{b}),$$

$$D^L(x,\alpha,i) = x_i + \alpha T(\boldsymbol{A}x, \boldsymbol{b}), \tag{3.25}$$

где $\alpha > 0$ — параметр, который мы выберем позднее. Пусть $x^R(\alpha,i)$ — один из векторов, минимизирующий $D^R(x,\alpha,i)$, т.е.

$$D^R(x^R(\alpha,i),\alpha,i) = \min_{y\in R^n} D^R(y,\alpha,i).$$

Аналогично определяется вектор $x^L(\alpha,i)$, минимизирующий $D^L(x,\alpha,i)$. Пусть $x_j^R(\alpha,i)$ — компонента с номером j вектора $x^R(\alpha,i)$.

Теорема 29. *Для векторов x^R, x^L выполняются соотношения*
1) $x_i^R(\alpha,i) \ge \overline{x}_i, \forall\alpha > 0$,
2) $\lim_{\alpha\to\infty} x_i^R(\alpha,i) = \overline{x}_i$,
1) $x_i^L(\alpha,i) \le \underline{x}_i, \forall\alpha > 0$,
2) $\lim_{\alpha\to\infty} x_i^L(\alpha,i) = \underline{x}_i$.

Доказательство. Остановимся на свойствах 1), 2). Свойства 3), 4) доказываются аналогично. Покажем, что $D^R(x,\alpha,i)$ не может принимать своего минимального значения на $\mathcal{X}$. Заметим, что $\forall x \in \mathcal{X}, D^R(x,\alpha,i) = -x_i$. Рассмотрим $x^* \in \mathcal{X}$ такой, что $x_i^* = \overline{x}_i$. Возьмем вектор x^ϵ с компонентами

$$x_j^\epsilon = \epsilon\delta_{i,j} + x_j^*, j = 1,\dots,n,$$

где ϵ будет выбрано позднее. Вектор x^* будет в силу выбора ближайшим вектором из $\mathcal{X}$ к вектору x^ϵ, т.е.

$$(x^\epsilon - x^*, x^\epsilon - x^*) = \min_{x\in\mathcal{X}}(x - x^*, x - x^*). \tag{3.26}$$

Поскольку $x^* \in \mathcal{X}$, то найдутся $A^* \in \boldsymbol{A}, b^* \in \boldsymbol{b}$, для которых выполнено равенство

$$A^*x^* = b^*.$$

Следовательно,

$$D^R(x^\epsilon,\alpha,i) \le -x_i^* - \epsilon + \alpha(A^*(x^\epsilon - x^*), A^*(x^\epsilon - x^*)) \le$$

$$-x_i^* - \epsilon + \alpha\rho\epsilon^2.$$

Положим $\epsilon = 1/(4\rho\alpha)$, тогда

$$D^R(x^\epsilon,\alpha,i) \le D^R(x^*,\alpha,i) = -\overline{x}_i.$$

Вследствие, этого минимум функционала D^R меньше $-x_i$, что может произойти только при условии

$$x_i^R \geq \overline{x}_i.$$

Свойство 1) доказано.

Далее, пусть $A^t \in \boldsymbol{A}, b^t \in \boldsymbol{b}$ такие, что

$$D^R(x^\epsilon, \alpha, i) = -x_i^* - \epsilon + \alpha T(A^t x^\epsilon, b^t).$$

Тогда найдется $x^t \in \mathcal{X}$, для которого выполнено равенство

$$A^t x^t = b^t.$$

Тогда

$$D^R(x^\epsilon, \alpha, i) = -x_i^* - \epsilon + \alpha(A^t(x^\epsilon - x^t), A^t(x^\epsilon - x^t)) \geq$$

$$-x_i^* - \epsilon + \alpha\gamma(x^\epsilon - x^t, x^\epsilon - x^t).$$

Отсюда на основании (3.26) следует, что

$$D^R(x^\epsilon, \alpha, i) = -x_i^* - \epsilon + \alpha\gamma(x^\epsilon - x^t, x^\epsilon - x^t) =$$

$$= -x_i^* - \epsilon + \alpha\rho\epsilon^2. \tag{3.27}$$

Из (3.27) непосредственно вытекает свойство 2). Теорема доказана. Для приближенного нахождения компонент вектора $\boldsymbol{x} \supseteq \mathcal{X}$ достаточно минимизировать функционалы (3.25) при некотором $\alpha > 0$. За начальное приближение можно взять вектор

$$x^0 = (A^0)^{-1} b^0, \tag{3.28}$$

где $A^0 = \mathrm{mid}\boldsymbol{A}$, $b^0 = \mathrm{mid}\boldsymbol{b}$.

Поскольку в общем случае функционалы (3.25) не дифференцируемы, то для их минимизации следует использовать методы, основанные на использовании субградиентов [12]. Заметим, что в достаточно малой окрестности своих точек минимума функционалы (3.25) выпуклые, дифференцируемые и представляются в виде

$$D^R(x, \alpha, i) = -x_i - \epsilon + \alpha(A^t x - b^t), A^t x - b^t)),$$

где A^t, b^t постоянны для всех векторов x из достаточно малой окрестности точки минимума, и выбраны, следуя работе [83], из соотношений

$$(\sum_{j=1}^{n} a_{ij}^t x_j - b_i^t)^2 = \min_{a_{ij} \in \boldsymbol{A}_{ij}, b_i \in \boldsymbol{b}_i} (\sum_{j=1}^{n} a_{ij} x_j - b_i)^2, i = 1, \ldots, n.$$

Тогда [83]

$$A^t x^* = b^t \tag{3.29}$$

и для нахождения x^* достаточно решить полученную вещественную систему. Подобную методику можно успешно применять для уточнения уже известных интервальных решений систем линейных уравнений. В этом случае в качестве начального приближения для минимизации функционалов (3.25) необходимо взять известные границы решения.

Проиллюстрируем это на решении системы интервальных линейных алгебраических уравнений (3.8) [82]. Множество векторов $\mathcal{X}$ для этой задачи изображено на рис. 3.1.

Минимальным интервальным вектором, содержащим множество ее решений, является интервальный вектор $\boldsymbol{x} = ([-4,4],[-4,4])^T$.

В соответствии с построенным методом оценим границу первой компоненты вектора $\boldsymbol{x}$, т.е. $\overline{x}_1$. Для этого положим $i = 1$ и минимизируем соответствующий функционал D^R, при этом начальные значения возьмем согласно (3.28) $x_0 = (0,0)^T$, $\alpha = 10$. Тогда получаем следующее приближенное значение точки минимума функционала D^R: $x^R = (4.618, 3.207)^T$.

Для определения $\overline{x}_1$ умножим интервальную матрицу $\boldsymbol{A}$ на вектор x^R. Получаем

$$\boldsymbol{A}\begin{pmatrix} x_1^R \\ x_2^R \end{pmatrix} = \begin{pmatrix} [2x_1^R - 2x_2^R, 4x_1^R + 1x_2^R] \\ [-1x_1^R + 2x_2^R, 2x_1^R + 4x_2^R] \end{pmatrix} = \begin{pmatrix} [2.822, 21.679] \\ [1.796, 22.064] \end{pmatrix}. \tag{3.30}$$

Поскольку правая граница первой компоненты вектора $\mathbf{x}$ определяется из решения некоторой конкретной системы линейных алгебраических уравнений, то из сравнения интервального вектора (3.30) с $\mathbf{b}$ в силу близости x_1^R и $\overline{x}_1$ вытекает, что в системе (3.28)

$$A^t = \begin{pmatrix} 2 & -2 \\ -1 & 2 \end{pmatrix}, b^t = \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$$

Решая систему (3.29), получаем $x^* = (4,3)^T$, т.е. достигается правая точная граница: $\overline{x}_1 = x^R = 4$. Аналогично находятся другие компоненты вектора $\boldsymbol{x}$.

Следовательно, рассмотренный метод позволяет находить точные границы решений для систем линейных алгебраических уравнений с интервальными коэффициентами.

3.5. Вопросы и упражнения

1. Вычислите $||\boldsymbol{A}||_\infty$

$$\boldsymbol{A} = \begin{pmatrix} [2,3], & [-1,0] & [0,0] \\ [-2,0], & [2,2] & [0,1] \\ [0,1], & [-1,0] & [-3,-2] \end{pmatrix}.$$

2. Найти множество решений системы линейных алгебраических уравнений

$$\boldsymbol{Ax} = \boldsymbol{b}$$

со следующей матрицей $\boldsymbol{A}$ и правой частью $\boldsymbol{b}$:

$$\boldsymbol{A} = \begin{pmatrix} [2,3] & [-1,1] \\ [-1,1] & [2,3] \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} [0,2] \\ [0,2] \end{pmatrix}.$$

3. Решить систему линейных алгебраических уравнений методом Гаусса

$$\boldsymbol{Ax} = \boldsymbol{b},$$

где

$$\boldsymbol{A} = \begin{pmatrix} [2,3] & [-1,1] \\ [-1,1] & [2,3] \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} [0,2] \\ [0,2] \end{pmatrix}.$$

4. Будет ли матрица

$$\boldsymbol{A} = \begin{pmatrix} [1,2] & [-1,0] \\ [-1,0] & [1,2] \end{pmatrix}$$

интервальной M-матрицей?

5. Будут ли сходиться следующие интервальные итерационные процессы

$$\boldsymbol{X}_{i+1} = A\boldsymbol{X}_i + d,$$

с матрицами

a)

$$A = \begin{pmatrix} 0.7 & -0.4 \\ 0.4 & 0.7 \end{pmatrix},$$

b)

$$A = \begin{pmatrix} 0.5 & 0.4 \\ 0.4 & 0.5 \end{pmatrix},$$

c)

$$A = \begin{pmatrix} 0.5 & -0.4 \\ -0.4 & 0.5 \end{pmatrix}?$$

Какой интервальный процесс дает оптимальные границы решения?

Глава 4

Нелинейные уравнения

4.1. Метод простой итерации

Пусть в некоторой области $\Omega \subset R^n$ задано отображение $\varphi : R^n \to R^n$, причем для любых $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ выполнено неравенство

$$\rho(\varphi(x) - \varphi(y)) \leq \alpha\rho(\boldsymbol{x} - \boldsymbol{y}), \tag{4.1}$$

где $0 \leq \alpha < 1$. Отображение, удовлетворяющее свойству (4.1), называется *сжимающим отображением*. Точка $\boldsymbol{x}^*$ называется *неподвижной точкой* отображения φ, $\varphi(\boldsymbol{x}^*) = \boldsymbol{x}^*$. Другими словами, неподвижная точка — это решение уравнения

$$\varphi(\boldsymbol{x}) = \boldsymbol{x}.$$

Теорема 30 (о принципе сжимающих отображений). *Всякое сжимающее отображение имеет одну и только одну неподвижную точку* $\boldsymbol{x}^*$, *причем*

$$\boldsymbol{x}^* = \lim_{k\to\infty} \boldsymbol{x}^k, \tag{4.2}$$

где $\boldsymbol{x}_0 \in \Omega$ — *произвольное начальное приближение,*

$$\boldsymbol{x}^k = \varphi(\boldsymbol{x}^{k-1}). \tag{4.3}$$

Доказательство. Покажем, что последовательность $\{\boldsymbol{x}^k\}$ фундаментальная. Действительно, рассмотрим процесс, который называется *методом простой итерации*:

$$\boldsymbol{x}^1 = \varphi(\boldsymbol{x}^0),\ x^2 = \varphi(\boldsymbol{x}^1) = \varphi^2(\boldsymbol{x}^0); \boldsymbol{x}^n = \varphi(\boldsymbol{x}^{n-1}) = \varphi^n(\boldsymbol{x}^0).$$

Будем считать для определенности $n \leq m$, имеем

$$\rho(\boldsymbol{x}^n - \boldsymbol{x}^m) = \rho(\varphi^n(\boldsymbol{x}) - \varphi^m(\boldsymbol{x})) \leq \alpha^n\rho(\boldsymbol{x}^0 - \boldsymbol{x}^{m-n}) \leq$$

$$\leq \alpha^n(\rho(\boldsymbol{x}^0 - \boldsymbol{x}^1) + \rho(\boldsymbol{x}^1 - \boldsymbol{x}^2) + \ldots + \rho(\boldsymbol{x}^{m-n-1} - \boldsymbol{x}^{m-n})) \leq$$

$$\leq \alpha^n \rho(\boldsymbol{x}^0 - \boldsymbol{x}^1)(1 + \alpha + \alpha^2 + \ldots + \alpha^{m-n-1}) \leq \alpha^n \rho(\boldsymbol{x}^0 - x^1)/(1 - \alpha).$$

Поскольку $\alpha < 1$, то при достаточно большом n величина $\rho(\boldsymbol{x}^n - \boldsymbol{x}^m)$ сколь угодно мала, то последовательность $\{\boldsymbol{x}_k\}$ фундаментальная и предел (4.2) существует. Тогда в силу непрерывности φ

$$\varphi(\boldsymbol{x}^*) = \varphi(\lim_{k\to\infty} \boldsymbol{x}^k) = \lim_{k\to\infty} \varphi(\boldsymbol{x}^k) = \lim_{k\to\infty} \boldsymbol{x}^{k+1} = \boldsymbol{x}^*.$$

Итак, существование неподвижной точки доказано, покажем ее единственность. Если

$$\varphi(\boldsymbol{x}) = \boldsymbol{x},\ \varphi(\boldsymbol{y}) = \boldsymbol{y},$$

то неравенство (4.1) принимает вид

$$\rho(\boldsymbol{x} - \boldsymbol{y}) \leq \alpha\rho(\boldsymbol{x} - \boldsymbol{y}).$$

Следовательно, $\rho(\boldsymbol{x} - \boldsymbol{y}) = 0$, или $\boldsymbol{x} = \boldsymbol{y}$. Теорема доказана.

Рассмотрим простейшие применения принципа сжимающих отображений к решению нелинейных уравнений.

Пусть задано уравнение с одной неизвестной x:

$$x = \varphi(x), \tag{4.4}$$

где $\varphi(x)$ — заданная функция. Уравнение (4.4) может иметь различное число корней или совсем не иметь решений.

Определение 11. *Функция φ удовлетворяет на отрезке $[a, b]$ условию Липшица с постоянной α, если для любых $x_1, x_2 \in [a, b]$ выполняется неравенство*

$$|\varphi(x_1) - \varphi(x_2)| \leq \alpha|x_1 - x_2|. \tag{4.5}$$

Замечание 2. *Если функция $\varphi(x)$ дифференцируема на отрезке $[a, b]$, то она удовлетворяет на $[a, b]$ условию Липшица с постоянной*

$$\alpha \leq \max_{[a,b]} |\varphi'(x)|. \tag{4.6}$$

Теорема 31. *Пусть функция φ удовлетворяет на отрезке $[a, b]$ условию Липшица с постоянной α, причем*

$$0 \leq \alpha < 1. \tag{4.7}$$

Тогда уравнение (4.4) имеет на отрезке $[a, b]$ *единственное решение*

$$x_* = \lim_{k\to\infty} x_k, \tag{4.8}$$

где x_0 *— начальное приближение из отрезка* $[a, b]$,

$$x_k = \varphi(x_{k-1}). \tag{4.9}$$

Доказательство непосредственно вытекает из принципа сжимающих отображений.

Оценим скорость сходимости (4.9). Имеем

$$x_* = \varphi(x_*). \tag{4.10}$$

Вычтем (4.9) из (4.10), тогда

$$|x_* - x_k| \le \alpha^k |x_* - x_0| \le \alpha^k (b - a). \tag{4.11}$$

Таким образом, мы видим, что расстояние между точным решением нашего уравнения и приближенным убывает в геометрической прогрессии.

На практике уравнения вида (4.4) встречаются довольно редко. Обычно нелинейные уравнения задаются в виде

$$f(x) = 0. \tag{4.12}$$

Приведем способ преобразования (4.12) к виду (4.4). Заметим, что для любого $k \neq 0$ следующее уравнение эквивалентно предыдущему:

$$\frac{f(x)}{k} = 0 \tag{4.13}$$

и

$$x = x - \frac{f(x)}{k}. \tag{4.14}$$

Следовательно, мы можем положить $\varphi(x) = x - \frac{f(x)}{k}$. В силу замечания 2 для построения сходящегося процесса нам необходимо выполнение условия (4.7):

$$|\varphi'(x)| = |1 - \frac{f'(x)}{k}| \le 1.$$

Предположим, что $f'(x)$ на интервале $[a, b]$ не меняет знак и для определенности $f'(x) > 0$. Тогда k можно положить

$$k = \max_{[a,b]} f'(x).$$

Рассмотрим систему линейных алгебраических уравнений вида

$$\boldsymbol{x} = \mathcal{B}\boldsymbol{x} + \boldsymbol{b}, \tag{4.15}$$

где B — заданная матрица n-го порядка, $\boldsymbol{b} \in R^n$ — заданный вектор.

Приведем условия, при которых отображение $\mathcal{B}\boldsymbol{x} + \boldsymbol{b}$ является сжимающим.

$$\rho(\mathcal{B}\boldsymbol{x} - \mathcal{B}\boldsymbol{y}) = \rho(\mathcal{B}\boldsymbol{x} - \boldsymbol{y}) \leq \alpha\rho(\boldsymbol{x} - \boldsymbol{y}),$$

другими словами, необходимо выполнение условия:

$$\sup_{x,y \in R^n} \frac{\rho(\mathcal{B}\boldsymbol{x} - \boldsymbol{y})}{\rho(\boldsymbol{x} - \boldsymbol{y})} = \sup_{v \neq 0} \frac{\rho(\mathcal{B}\boldsymbol{v} - 0)}{\rho(\boldsymbol{v} - 0)} \leq \alpha,$$

или

$$\sup_{v \neq 0} \frac{||(\mathcal{B}\boldsymbol{v}||}{||\boldsymbol{v}||} \leq \alpha. \tag{4.16}$$

В пространстве квадратных матриц мы можем задать норму следующим образом:

$$||\mathcal{B}|| = \sup_{v \neq 0} \frac{||(\mathcal{B}\boldsymbol{v}||}{||\boldsymbol{v}||}. \tag{4.17}$$

Можно показать, что

$$||A|| \leq (\sum_{i,j=1}^{n} a_{ij}^2)^{1/2}. \tag{4.18}$$

Следовательно, для сходимости метода простой итерации достаточно показать, что норма матрицы $||\mathcal{B}||$ удовлетворяет следующему неравенству $||\mathcal{B}|| \leq \alpha$ или выполнено более сильное условие

$$(\sum_{i,j=1}^{n} b_{ij}^2)^{1/2} \leq \alpha.$$

Системы нелинейных уравнений

Рассмотрим систему нелинейных уравнений с n неизвестными:

$$\begin{aligned} x_1 &= \varphi_1(x_1, x_2, \ldots, x_n), \\ x_2 &= \varphi_2(x_1, x_2, \ldots, x_n), \\ &\ldots\ \ldots \\ x_n &= \varphi_n(x_1, x_2, \ldots, x_n), \end{aligned} \tag{4.19}$$

или в векторном виде

$$\boldsymbol{x} = \varphi(\boldsymbol{x}).$$

Если отображение φ сжимающее, то для нахождения решения системы (4.19) мы можем применить метод простых итераций

$$\boldsymbol{x}^1 = \varphi(\boldsymbol{x}^0),\ \boldsymbol{x}^2 = \varphi(\boldsymbol{x}^1), \ldots, \boldsymbol{x}^n = \varphi(\boldsymbol{x}^{n-1}). \tag{4.20}$$

Предположим, что вектор-функция φ имеет непрерывные частные производные по $x_1, x_2, \ldots, x_n$, и обозначим $A = (a_{ij})$, где

$$a_{ij} = \max_{\Omega} |\partial\varphi_i/\partial x_j|. \tag{4.21}$$

Пусть $\boldsymbol{x}, \boldsymbol{y} \in \Omega$. Согласно формуле конечных приращений Лагранжа, имеем

$$\varphi_i(\boldsymbol{x}) - \varphi_i(\boldsymbol{y}) = \sum_{j=1}^{n} \frac{\partial\varphi_i(\xi^j)}{\partial x_j}(x_j - y_j).$$

Следовательно,

$$\varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y}) = A(\boldsymbol{x} - \boldsymbol{y})$$

и

$$\rho(\varphi_i(\boldsymbol{x}) - \varphi_i(\boldsymbol{y})) = ||\varphi_i(\boldsymbol{x} - \varphi_i(\boldsymbol{y})|| = ||A(\boldsymbol{x} - \boldsymbol{y})|| \leq$$
$$\leq ||A||||\boldsymbol{x} - \boldsymbol{y}|| = ||A||\rho(\boldsymbol{x} - \boldsymbol{y}),$$

т. е.

$$\rho(\varphi_i(\boldsymbol{x}) - \varphi_i(\boldsymbol{y})) \leq ||A||\rho(\boldsymbol{x} - \boldsymbol{y}).$$

Если $||A|| \leq 1$, то оператор φ является сжимающим и итерационный процесс (4.20) сходится.

В интервальном виде метод простой итерации записывается как

$$\boldsymbol{x}^{i=1} = \boldsymbol{\varphi}(\boldsymbol{x}^i) \cap \boldsymbol{x}^i,\ i = 0, \ldots$$

Причем для начального приближения $\boldsymbol{x}^0$ должно выполняться включение $\boldsymbol{x}^* \subseteq \boldsymbol{x}^0$.

4.2. Метод Ньютона

Пусть $f \in C^2[a, b]$ — некоторая дважды непрерывно дифференцируемая функция. Рассмотрим задачу нахождения корня уравнения:

$$f(x) = 0. \tag{4.22}$$

Теорема 32. *Пусть на отрезке $[a,b]$ производная $f'(x)$ не меняет знак и, кроме того, $f(a)f(b) < 0$, тогда на отрезке существует корень x_* уравнения (4.22), причем единственный.*

Доказательство этой теоремы очевидно из геометрических соображений.

Приведем следующую цепочку тождественных преобразований:

$$-f(x_0) = f(x_*) - f(x_0) = f'(\xi)(x_* - x_0),\ x_0 \in [a,b].$$

Разделив обе части равенства на $f'(\xi)$, получаем

$$x_* - x_0 = -f(x_0)/f'(\xi),$$

или

$$x_* = x_0 - f(x_0)/f'(\xi),\ \xi \in [x_*, x_0].$$

Таким образом, если точка x_0 близка к x_*, то в последнем равенстве можно заменить неизвестное значение ξ на x_0. Получаем приближенное равенство

$$x_* \approx x_0 - f(x_0)/f'(x_0).$$

На основе этого приближения мы можем построить следующий итерационный процесс:

$$x_i = x_{i-1} - f(x_{i-1})/f'(x_{i-1}), i = 1, 2, \ldots, \tag{4.23}$$

который называют *методом Ньютона* или — поскольку (4.23) геометрически означает уравнение касательной к графику функции f в точке x_{i-1} — *методом касательных.*

Оценим скорость сходимости. Согласно формуле Тейлора имеем

$$0 = f(x_*) = f(x_{i-1}) - f'(x_{i-1})(x_* - x_{i-1}) + f''(\xi)\frac{(x_* - x_{i-1})^2}{2}$$

или

$$x_* = x_{i-1} - f(x_{i-1})/f'(x_{i-1}) + (x_* - x_{i-1})^2\frac{f''(\xi)}{2f'(x_{i-1})},\ \xi \in [x_*, x_{i-1}].$$

Вычитая из последнего равенства (4.23) и предполагая

$$\frac{f''(\xi)}{2f'(x_{i-1})} \le \beta,$$

получаем

$$|x_* - x_i| \le \beta(x_* - x_{i-1})^2.$$

Отсюда следует, что если выполнено неравенство

$$\beta|x_* - x_i| < 1, \tag{4.24}$$

то погрешность убывает очень быстро по квадратичному закону. Через n итераций будем иметь

$$|x_* - x_{i+n}| \le \frac{1}{\beta}(\beta(x_* - x_i))^{2^n}.$$

В качестве примера рассмотрим следующее уравнение:

$$f(x) = x^2 - a = 0, f'(x) = 2x.$$

Метод Ньютона запишется в виде

$$x_{i+1} = x_i - \frac{x_i^2 - a}{2x_i} = \frac{1}{2}(x_i - a/x_i), \tag{4.25}$$

за начальное приближение можно взять точку $x_0 = a$. Тем самым мы получили известную формулу для вычисления квадратных корней. Заметим, что условие (4.24) может не выполняться, однако мы можем интерпретировать формулу (4.25) как метод простой итерации. Несложно убедиться, что условие сходимости выполнено и итерационный процесс (4.25) будет сходиться. Вначале скорость сходимости будет как у метода простой итерации. В дальнейшем при выполнении условия (4.24) скорость сходимости будет квадратичная и процесс очень быстро сойдется.

Перейдем к интервальным методам Ньютона. Пусть $f \in C^1$ и известно, что $x^* \in \boldsymbol{x}^0$ — корень уравнения $f(x^*) = 0$. Пусть $f'(\xi) \neq 0$, $\forall \xi \in \boldsymbol{x}^0$.

Интервальный метод Ньютона можно записать в виде

$$\operatorname{mid} \boldsymbol{x}^{k+1} = (\operatorname{mid} \boldsymbol{x}^k - f(\operatorname{mid} \boldsymbol{x}^k)/f'(\boldsymbol{x}^k)) \cap \boldsymbol{x}^k. \tag{4.26}$$

Последовательность $\{\boldsymbol{x}^k\}$, вычисленная по формулам (4.26), обладает свойствами:

$$\boldsymbol{x}^0 \supset \boldsymbol{x}^1 \supset \boldsymbol{x}^2, \ldots,$$

$$\lim_{k\to\infty} \boldsymbol{x}^k = x^*, \quad x^* \in \boldsymbol{x}^k, \forall k.$$

4.3. Метод Кравчика

Предположим, что функция f удовлетворяет условиям метода Ньютона. Рассмотрим оператор K [85]:

$$K(\boldsymbol{x}, x, y) = x - yf(x) + (1 - yf'(\boldsymbol{x}))(\boldsymbol{x} - x).$$

Теорема 33. *Пусть* $x^* \in \boldsymbol{x}^0$. *Определим последовательность* $\{\boldsymbol{x}^k\}$

$$\boldsymbol{x}^{k+1} = K(\boldsymbol{x}^k, x^k, y^k) \cap \boldsymbol{x}^k,$$

где $x^k \in \boldsymbol{x}^k$, y^k *произвольны. Тогда:*

1) $x^* \in \boldsymbol{x}^k, \quad \forall k;$

2) $\operatorname{wid} \boldsymbol{x}^k \to 0$ *при условии*

$$1 - yf'(\boldsymbol{x}^k) \supset [0, q], \quad q < 1, \quad \forall k.$$

На практике в качестве x^k часто берут $\operatorname{mid} \boldsymbol{x}^k$, а $y^k \approx (\operatorname{mid} f'(\boldsymbol{x}^0))^{-1}$ или $(\operatorname{mid} f'(\boldsymbol{x}^k))^{-1}$. Заметим, что метод Кравчика не требует обращения интервальных матриц.

4.4. Двусторонние методы

Раздел посвящен двусторонним итерационным методам решения систем нелинейных уравнений и их связи с интервальной математикой. Показано, что многие двусторонние и интервальные итерационные процессы эквивалентны. Более того, аппарат интервальной математики можно с успехом применять для построения эффективных двусторонних итерационных процессов, нахождения границ двусторонних решений, близких к оптимальным.

Проблеме построения двусторонних методов решения систем нелинейных уравнений посвящено большое число работ. В основном работы базируются на следующих подходах: теореме Брауэра о неподвижной точке, теореме Миранда, принципе сжатых отображений, методе Ньютона. В монографии [92] рассматриваются интервальные методы решения систем нелинейных уравнений. В монографии [44] изложены итерационные двусторонние методы, основанные на теореме Брауэра. В работе [82] приводятся итерационные методы, использующие теорему Миранда, в [91] — методы, основанные на принципе сжимающих отображений. Подробная библиография по итерационным методам есть в монографиях [3, 92].

Рассмотрим систему нелинейных уравнений

$$x = f(x, k), \tag{4.27}$$

где $f = \{f_i\}_{i=1}^n$, $f_i = f_i(x, k)$; $k \in R^m$ — вектор параметров, $k \in \boldsymbol{k}$; $x \in R^n$ — вектор неизвестных.

Множество всех решений системы (4.27) определим таким образом:

$$\mathcal{X} = \{x | x = f(x, k), k \in \boldsymbol{k}\}. \tag{4.28}$$

Приведем несколько вспомогательных результатов из монографии [44]. Пусть существуют вещественные функции $F^l(y_1, \ldots, y_n, z_1, \ldots, z_n)$, $l = 1, 2$ такие, что при $\underline{y} \leq \overline{y}$, $\underline{z} \leq \overline{z}$ выполнены неравенства

$$F_i^l(\underline{y}, \overline{z}) \leq F_i^l(\overline{y}, \underline{z}), l = 1, 2; i = 1, \ldots, n, \tag{4.29}$$

кроме того, выполнены неравенства

$$F_i^1(x, x) < f_i(x, k) < F_i^2(x, x). \tag{4.30}$$

Пусть вектора $\underline{x}, \overline{x} \in R^n$ являются решением системы

$$\begin{aligned} \underline{x} &= F^1(\underline{x}, \overline{x}), \\ \overline{x} &= F^2(\overline{x}, \underline{x}), \end{aligned} \tag{4.31}$$

тогда любое решение x системы (4.27) удовлетворяет оценкам

$$\underline{x} \leq x \leq \overline{x}. \tag{4.32}$$

Заметим, что в силу (4.29)

$$F^1(\underline{x}, \overline{x}) \leq \min_{k \in \boldsymbol{k}} f(x, k), F^2(\overline{x}, \underline{x}) \geq \max_{k \in \boldsymbol{k}} f(x, k).$$

Покажем, что в качестве F можно взять границы монотонного по включению интервального расширения функций f. Действительно, пусть $\boldsymbol{F}(x) = [\underline{F}(x), \overline{F}(x)]$ — интервальное расширение функции f, $x = [\underline{x}, \overline{x}]$, тогда функции $\underline{F}, \overline{F}$ полностью удовлетворяют условиям (4.29), (4.30); условию (4.29) — поскольку интервальное расширение монотонно по включению, условию (4.30) — в силу свойств интервальных расширений.

Таким образом, систему (4.31) можно переписать в виде

$$\underline{x} = \underline{F}(\boldsymbol{x}), \overline{x} = \overline{F}(\boldsymbol{x}) \ \textit{или} \ \boldsymbol{x} = \boldsymbol{F}(\boldsymbol{x}). \tag{4.33}$$

4.5. Построение вектора начальных приближений

Рассмотрим вопрос о построении вектора начальных приближений для итерационного процесса решения системы (4.33).

$$\boldsymbol{x}^{i+1} = \boldsymbol{F}(\boldsymbol{x}^i), i = 0, 1, 2, \ldots \tag{4.34}$$

Пусть $\boldsymbol{x}^0$ — вектор начальных приближений, $\boldsymbol{x}^*$ — решение системы уравнений (4.33). Тогда для сходимости итерационного процесса (4.34) необходимо выполнение условия [44]:

$$\boldsymbol{x}^1 = \boldsymbol{F}(x^0) \subseteq \boldsymbol{x}^0. \tag{4.35}$$

Это включение влечет за собой следующие свойства [3]:

$$\mathcal{X} \subseteq \boldsymbol{x}^i, i = 0, 1, 2, \ldots, . \tag{4.36}$$

Пусть $\boldsymbol{F}$ удовлетворяет в некоторой норме $||\cdot||$ условию Липшица с константой $L < 1$ в области Ω

$$||\boldsymbol{F}(\boldsymbol{x}_1) - \boldsymbol{F}(\boldsymbol{x}_2)|| < L||\boldsymbol{x}_1 - \boldsymbol{x}_2||, \boldsymbol{x}_1, \boldsymbol{x}_2 \subseteq \Omega, \tag{4.37}$$

тогда, в силу принципа сжимающих отображений, итерационный процесс (4.34) сходится для любого начального приближения $x_0 \in \Omega$.

В интервальном анализе итерационный процесс (4.34) часто записывается в виде

$$\boldsymbol{x}^{i+1} = \boldsymbol{F}(\boldsymbol{x}^i) \cap \boldsymbol{x}^i, i = 0, 1, 2, \ldots, \tag{4.38}$$

при этом на вектор начальных приближений накладывается условие 4.35.

Пусть x^* — решение системы нелинейных уравнений при некотором конкретном значении параметров k. Тогда $x^* \in \boldsymbol{x}^0$ и x^* можно использовать в качестве начального приближения для решения системы (4.31). Если выполнено условие (4.37), то итерационный процесс (4.34) сходится к $\boldsymbol{x}^*$. При этом выполнение условий (4.36) не гарантируется. Заметим, что решения систем (4.31) и (4.33) совпадают, следовательно, в силу близости некоторого $\boldsymbol{x}^i$ к $\boldsymbol{x}^*$, можно построить $\boldsymbol{x}^0$ следующим образом:

$$\boldsymbol{x}^0 = \mathrm{mid}(\boldsymbol{x}) + [-1, 1]\mathrm{wid}(\boldsymbol{x})d, \quad d > 1,$$

где d — некоторый параметр, который можно оценить следуя работе [82] или непосредственно проверить выполнение условия (4.35). В случае неудачи можно или продолжить итерации (4.34), или увеличить d.

Таким образом, проблема нахождения вектора начальных приближений для интервальных итерационных процессов сводится к приближенному решению системы (4.31) с вещественным начальным вектором $x^0 \in \Omega$.

4.6. Уточнение решений

Пусть $\boldsymbol{x} \supset \mathcal{X}$, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ — некоторое двустороннее решение. Рассмотрим следующий интервальный вектор:

$$\boldsymbol{g} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, g, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_n).$$

Он представляет интервальный вектор $\boldsymbol{x}$, у которого i компонента заменена на число g. Если мы умеем определять, пересекается ли вектор $\boldsymbol{g}$ с множеством $\mathcal{X}$, то оптимальное значение для $\underline{x_i}$

$$\underline{x_i} = \inf\{g | \mathcal{X} \cap \boldsymbol{g} \neq 0\}.$$

Однако на практике проще определить, когда

$$\mathcal{X} \cap \boldsymbol{g} = \emptyset.$$

В этом случае

$$\underline{x_i} = \sup\{g | \mathcal{X} \cap \boldsymbol{g} = \emptyset, x_i > g, \forall x \in \mathcal{X}\}. \tag{4.39}$$

Перейдем к вопросу об уточнении $\boldsymbol{x}$. Подставим $\boldsymbol{g}$ в уравнение (4.34):

$$\boldsymbol{g}^1 = \boldsymbol{F}(\boldsymbol{g}),$$

при этом i-я компонента вектора $\boldsymbol{g}$ перейдет в некоторое интервальное число $\boldsymbol{g}_i$.

Лемма 6. *Пусть* $\boldsymbol{g}^1 \cap \boldsymbol{g} = \emptyset$ *или, что то же самое* $g \notin \boldsymbol{g}_i$. *Тогда*

$$\mathcal{X} \cap \boldsymbol{g} = \emptyset.$$

Доказательство. Предположим противное, пусть z — некоторый вектор

$$z \in \mathcal{X} \cap \boldsymbol{g},$$

но тогда z является неподвижной точкой некоторого оператора $F \in \boldsymbol{F}$

$$Fz = z \in \boldsymbol{g}$$

и, следовательно, в силу свойств интервальных итерационных процессов

$$Fz \in \boldsymbol{g}^1.$$

Значит, $\boldsymbol{g}^1 \cap \boldsymbol{g} \neq \emptyset$. Полученное противоречие доказывает лемму.

Таким образом, мы можем построить итерационное уточнение по следующей процедуре:

ПРОЦЕДУРА 1 {уточнение i-й нижней грани}.
$\boldsymbol{g}^0 := (x_1, \ldots, x_{i-1}, \underline{x}_i, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_n)$;
Пока $\boldsymbol{g}^1 \cap \boldsymbol{g} = \emptyset$ цикл
$\boldsymbol{g}^1 := \boldsymbol{F}(g^l)$,
$\boldsymbol{g}^0 := (x_1, \ldots, x_{i-1}, \underline{z}_i, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_n)$;
конец цикла;
$\boldsymbol{x}_i := [\underline{z}_i, \overline{x}_i]$;
конец процедуры.

Аналогично строится процедура уточнения верхней грани.

Рассмотрим еще одну процедуру уточнения интервального решения. Обратимся непосредственно к уравнению (4.27). Пусть как и прежде нам необходимо уточнить нижнюю границу i-й координаты. Для этого удалим из него i-ю строку и зафиксируем i-ю компоненту.

$$\begin{aligned}
\boldsymbol{x}_1 &= f_1(\boldsymbol{x}_1, \ldots, \underline{x}_i, \ldots, \boldsymbol{x}_n) \qquad (4.40)\\
&\ldots\\
\boldsymbol{x}_{i-1} &= f_{i-1}(x_1, \ldots, \underline{x}_i, \ldots, \boldsymbol{x}_n)\\
\boldsymbol{x}_{i+1} &= f_{i+1}(x_1, \ldots, \underline{x}_i, \ldots, \boldsymbol{x}_n)\\
&\ldots\\
\boldsymbol{x}_n &= f_n(x_1, \ldots, \underline{x}_i, \ldots, \boldsymbol{x}_n)
\end{aligned}$$

Решим полученную систему итерационным способом. Сходимость (4.40) непосредственно вытекает из сходимости исходной системы (4.33). Пусть $\boldsymbol{z} = (z_1, \ldots, z_{i-1}, \underline{x}_i, \boldsymbol{z}_{i+1}, \ldots, z_n)$ — решение системы (4.40), тогда справедлива следующая лемма.

Лемма 7. *Выполнение условия*

$$0 \notin \underline{x}_i - \boldsymbol{f}_i(z) \qquad (4.41)$$

влечет за собой

$$\mathcal{X} \cap \boldsymbol{z} = \emptyset.$$

Доказательство проведем от противного. Пусть $\exists z \in \mathcal{X} \cap \boldsymbol{z}$, но тогда z является решением системы (4.33):

$$\underline{x}_i = \boldsymbol{z}_i \in \boldsymbol{f}_i(z_1, \ldots, \underline{x}_i, \ldots, \boldsymbol{z}_n).$$

Таким образом, $0 \in \underline{x}_i - \boldsymbol{f}_i(z)$, что противоречит условию (4.41). Лемма доказана.□

Замечание 3. *Из выполнения условия* $0 \in \boldsymbol{z}_i - f_i(z)$ *в общем случае не вытекает, что* $\mathcal{X} \cap \boldsymbol{z} = \emptyset$.

Следуя (4.39), для уточнения $\underline{x}_i$, несложно получить различные процедуры, основанные на алгоритмах нахождения экстремумов функций, например, на методе деления отрезка пополам, методе золотого сечения и т.п.

Численные примеры

Рассмотрим численные примеры. Начнем с линейных алгебраических систем с интервальными коэффициентами [82]:

$$\boldsymbol{A}x = \boldsymbol{b},$$

где $\boldsymbol{A}$ — интервальная матрица, $\boldsymbol{b}$ — интервальный вектор:

$$\boldsymbol{A} = \begin{pmatrix} [2,4] & [-1,1] \\ [-1,1] & [2,4] \end{pmatrix}.$$

Приведем систему к виду (4.27). Для этого перепишем систему следующим образом:

$$x = (I - Y\boldsymbol{A})x - Y\boldsymbol{b},$$

где Y — матрица вида

$$\begin{pmatrix} 1/3 & 0 \\ 0 & 1/38 \end{pmatrix}.$$

Далее, на рис. 4.1, 4.2 показаны множества решений этой системы при различных векторах $\boldsymbol{b}$, приведены границы решений полученных в работе [82]. Применение указанных выше методик позволило построить интервальные решения, близкие к оптимальным.

Рассмотрим теперь нелинейную систему:

$$\begin{aligned} x_1 &= -k_0 x_2 + k_1, \\ x_2 &= 0.1 x_1 x_2 + k_2, \end{aligned} \tag{4.42}$$

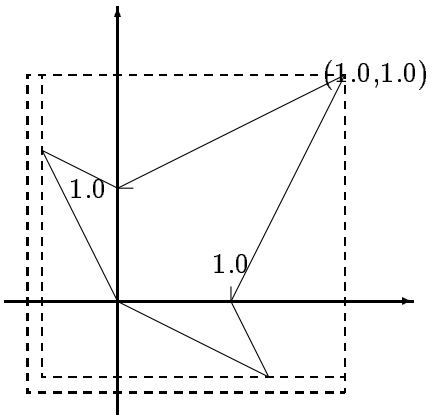

Рис. 4.1. b =([0.0,2.0],[0.0,2.0])

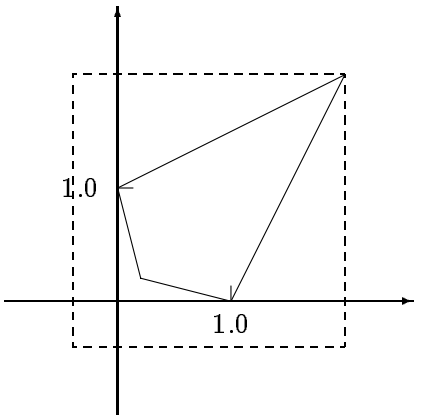

Рис. 4.2. b =([1.0, 2.0],[1.0,2.0])

где $\boldsymbol{k}_0 = [0.1, 0.2], \boldsymbol{k}_1 = [0.6, 1.0], \boldsymbol{k}_2 = [0.0, 0.45]$. Интервальный итерационный процесс сходится к вектору

$$\boldsymbol{x} = ([0.5, 1.0], [0.0, 0.5]),$$

оптимальные границы множества решений

$$x = ([0.505, 1.0], [0.0, 0.49725]).$$

В результате использования процедуры, основанной на лемме 7, получены оптимальные границы множества решений системы.

На рис. 4.3 представлено точное множество решений системы (4.42) и решения с помощью интервального метода простой итерации, а также уточненные границы решений.

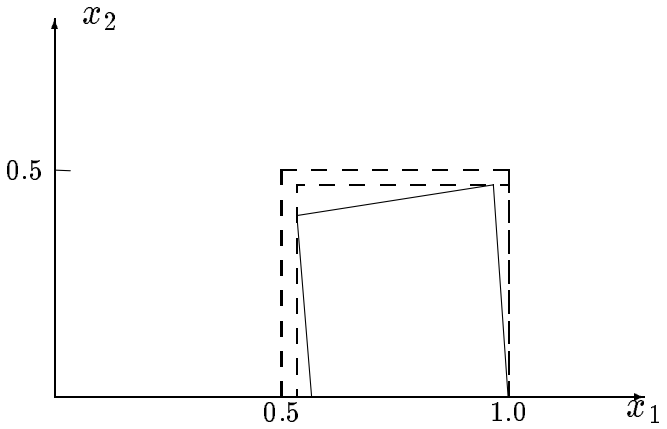

Рис. 4.3. Уточнение границ нелинейных уравнений

4.7. Вопросы и упражнения

1. Привести уравнение к виду пригодному для метода простой итерации

$$x^2 - 3 = 0.$$

Сделать две итерации. Чему равна константа Липшица?

2. Записать системы нелинейных уравнений в виде интервального метода простой итерации

$$\begin{cases} \sin(x+1) - y &= 1, \\ 2x + \cos y &= 2. \end{cases}$$

$$\begin{cases} x^2 + x + 3y &= 1, \\ 2x + y^2 &= 2. \end{cases}$$

Найти векторы начальных приближений.

3. Написать интервальный метод Ньютона для уравнения

$$x^3 - 5 = 0.$$

Найти начальное приближение и сделать две итерации.

Глава 5

Задачи Коши для систем ОДУ

В этой главе мы обратимся к задачам Коши для систем обыкновенных дифференциальных уравнений. Основное изложение проведем для скалярного уравнения (5.1), но в каждом случае укажем способы обобщения на системы уравнений.

$$y' = f(x, y), \quad x \geq 0, \tag{5.1}$$

$$y(0) = y_0. \tag{5.2}$$

В разд. 5.1 обсуждаются двусторонние методы с полярными остаточными членами, появившиеся одними из первых [26, 30]. Следует отметить, что они не двусторонние в строгом смысле, так как не дают гарантии принадлежности точного решения построенному коридору, но обладают этим свойством в асимптотическом смысле: при стремлении шага разностной сетки в нуль построенный коридор все менее отличается от гарантированного. Все последующие методы дают двустороннее решение в строгом смысле. Наиболее детально изложен метод апостериорного оценивания, в котором для двусторонней оценки использовались мажорантные приемы Е. М. Лозинского [49, 50].

При решении систем ОДУ важно иметь надежные численные методы. Рассмотрим задачу Коши

$$y' = y^2, \quad t \in (0, 1),$$

с начальным условием

$$y(0) = 1.$$

Точное решение этого уравнения $y(t) = 1/(1 - x)$. Как видно решение существует только на интервале $(0, 1)$, в точке $t = 1.0$ решение уходит "на бесконечность". Решим численно эту задачу методом Рунге-Кутта с шагом $h = 0.1$.

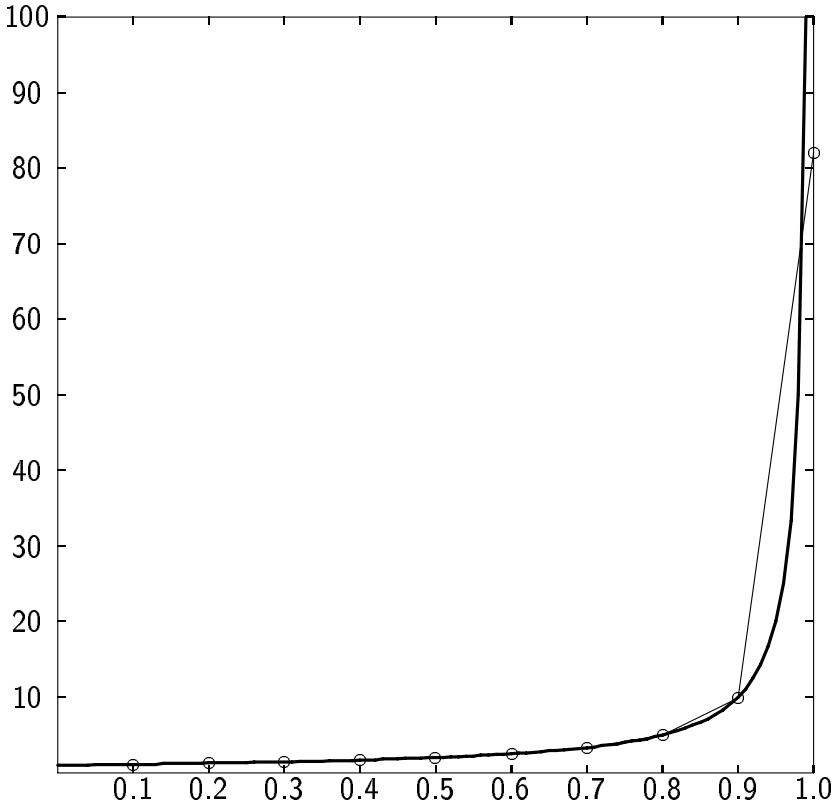

Рис. 5.1. ∘ — численное решение

Как видно из рис. 5.1, численное решение в точке $t = 1.0$ существует. Таким образом, при численном решении ОДУ необходимо использовать надежные методы и следить за погрешностью численного решения. Для этой цели можно применять широкий спектр двусторонних и интервальльных методов. Однако эти методы обладают рядом особенностей.

Перед описанием двусторонних и интервальных методов приведем пример Р. Е. Мура [89], иллюстрирующий так называемый эффект упаковывания (wrapping effect). В литературе он также часто упоминается как эффект раскрутки или эффект Мура. Этот эффект проявляется в чрезмерном увеличении ширины интервального решения системы дифференциальных уравнений по сравнению с истинным. Этот эффект связан только с внутренними свойствами интервальных методов безотносительно к ошибкам численных решений.

Приведем систему дифференциальных уравнений [89]

$$\begin{aligned} y_1' &= y_2, \\ y_2' &= -y_1, \quad t \geq 0 \end{aligned} \tag{5.3}$$

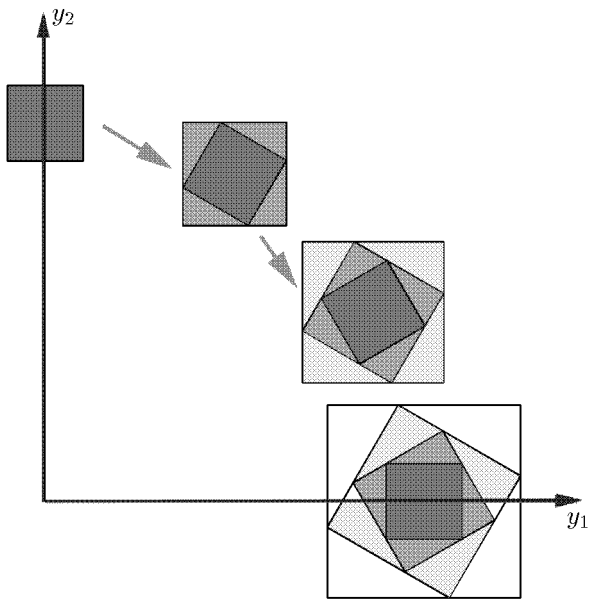

Рис. 5.2. Эффект упаковывания

с начальными условиями

$$y_1(0) \in 1 + [-\varepsilon, \varepsilon], \quad y_2(0) \in [-\varepsilon, \varepsilon].$$

Множество решений этой задачи на фазовой плоскости (y_1, y_2) вращается относительно начала координат. Смоделируем численное решение в моменты времени $t_j = j\tau,\ j = 0, 1, ...$, и построим интервальные решения, содержащие в себе множество решений (рис. 5.2). В результате мы видим возрастание ширины интервального решения.

Для преодоления этого эффекта были предложены различные методы, в частности:

- автоматическое преобразование координат [88];
- преобразование системы ОДУ к удобному виду [105];
- ограничение множества решений параллелепипедами, эллипсоидами и другими видами областей [93],[65];
- аналитическое решение системы ОДУ и последующее построение интервального расширения [44], [98].

Эффект упаковывания обсуждался также в работах [13],[94], [90], [87] и др.

Подробный обзор по интервальным методам решения систем ОДУ имеется в [94], в монографии [44] рассматриваются двусторонние мето-

ды, использующие разложение оператора на изотонную и антитонную составляющие. В монографии [34] для построения двусторонних решений исследованы апостериорные оценки погрешности.

5.1. Разностные методы с полярными остаточными членами

Приведем асимптотически двусторонние методы решения задачи Коши. Они основаны на построении пары формул интегрирования с противоположными по знаку главными членами погрешности.

Вернемся к задаче Коши:

$$y' = f(x, y), \quad x \geq 0, \tag{5.4}$$

$$y(0) = y_0. \tag{5.5}$$

Основную идею изложим на примере метода Рунга — Кутта. Предположим, что нам известно точное значение $y(x_0)$ в некоторой точке $x_0 \geq 0$, и попытаемся найти значение $y(x_0 + h)$. Для этого введем следующие величины [7]:

$$k_j = hf(\xi_j, \eta_j), \quad \xi_j = x_0 + \alpha_j h, \tag{5.6}$$

$$\eta_j = y(x_0) + \beta_{j,1}k_1 + \ldots + \beta_{j,j-1}k_{j-1}, \quad j = 1, ..., m, \tag{5.7}$$

$$y^h(x_0 + h) = y(x_0) + p_1k_1 + \ldots + p_mk_m. \tag{5.8}$$

Здесь $\alpha = \{\alpha_j\}$, $\beta = \{\beta_{j,l}\}$, — наборы числовых параметров, выбираемых из условий аппроксимации и устойчивости [7]. Помимо этих обычных требований, попытаемся выбрать параметры так, чтобы локальная ошибка имела специальный вид:

$$y^h(x_0 + h) - y(x_0 + h) = h^{q+1}\gamma\psi(f) + O(h^{q+2}). \tag{5.9}$$

Здесь ψ — определенный оператор, выражающийся через функцию f и ее частные производные и не зависящий от α, β, p, а γ, наоборот, выражается только через α, β, p. Целое $q > 0$ будем называть степенью метода. Первое слагаемое в правой части (5.9), очевидно, является главным членом локальной ошибки. Представим себе, что удастся найти два набора параметров $\alpha^{(1)}, \beta^{(1)}, p^{(1)}$ и $\alpha^{(2)}, \beta^{(2)}, p^{(2)}$ таких, что

$$\gamma(\alpha^{(1)}, \beta^{(1)}, p^{(1)}) = \gamma = -\gamma(\alpha^{(2)}, \beta^{(2)}, p^{(2)}). \tag{5.10}$$

В результате будем иметь две независимые формулы численного интегрирования на отрезке $[x_0, x_0 + h]$, которые дают два приближенных значения: $y^h_{(1)} = (x_0 + h)$ и $y^h_{(2)}(x_0 + h)$, причем

$$\begin{aligned} y^h_{(1)}(x_0 + h) &= y(x_0 + h) + h^{q+1}\bar{\gamma}\psi(f) + O(h^{q+2}), \\ y^h_{(2)}(x_0 + h) &= y(x_0 + h) - h^{q+1}\bar{\gamma}\psi(f) + O(h^{q+2}). \end{aligned} \tag{5.11}$$

Пусть, для определенности, $\bar{\gamma}\psi(f) > 0$. Тогда при достаточно малом h погрешность в первой формуле будет положительной, а во второй — отрицательной. В результате мы получаем двустороннее приближение такое, что

$$y(x_0 + h) \in \mathbf{y}(x_0 + h) = [y_{(2)}(x_0 + h), y_{(1)}(x_0 + h)]. \tag{5.12}$$

Аналогичная ситуация имеет место при $\bar{\gamma}\psi(f) < 0$, но в (5.12) пределы переставляются местами. Ситуация остается неопределенной при обращении в нуль $\psi(f)$. Тогда ни при каком малом h нет гарантии, что первое слагаемое будет определять знак локальной погрешности. В итоге приходится надеяться, что неопределенность выливается в погрешность более высокого порядка малости $O(h^{q+2})$ и происходит на сравнительно небольшом участке близости к нулю функции $\psi(f)$.

Этот анализ иллюстрирует нестрогость оценок, предоставляемых асимптотически двусторонними методами в случае крупного шага h или смены знака у функции $\psi(f)$.

Мы наложили на параметры α, β, p столько требований, что возникает естественный вопрос о существовании таких наборов. Для примера выпишем формулы степени 2 и поясним реализацию алгоритма в целом. Для этого введем разностную сетку $\omega_h = \{x_i = ih, i = 0, 1, ...\}$ с шагом $h > 0$ и рассмотрим две трехстадийные формулы Рунге — Кутта с начальным условием y_0 в точке x_0 [30]:

$$\begin{aligned} &k_1 = hf(x_0, y_0), \quad k_2 = hf(x_0 + h/3, y_0 + k_1/3), \\ &k_3^{(1)} = hf(x_0 + h/2, y_0 + k_2/2), \\ &k_3^{(2)} = hf(x_0 + 5h/6, y_0 + 5k_2/6), \\ &y^h_{(1)}(x_0 + h) = y_0 + k_3^{(1)}, \\ &y^h_{(2)}(x_0 + h) = y_0 + 2/5\ k_1 + 3/5\ k_3^{(2)}. \end{aligned} \tag{5.13}$$

Процесс интегрирования происходит следующим образом. Пусть в некоторой точке $x \in \omega_h$ известны границы асимптотически двустороннего решения $\boldsymbol{y}(x)$. Взяв в качестве начального условия y_0 сначала $\bar{y}(x + h)$,

а затем $\underline{y}(x+h)$, по формулам (5.13) получаем два интервала: $\boldsymbol{y}_1(x+h)$ и $\boldsymbol{y}_2(x+h)$. Их интервальная оболочка дает асимптотически двустороннее решение $\boldsymbol{y}(x+h)$. Полагая последовательно $x = 0, h, 2h, ...$, мы можем вычислить асимптотически двусторонние оценки $\boldsymbol{y}(x)$ в любом узле сетки ω_h.

Формулы с полярными погрешностями аппроксимации можно строить и на основе метода Адамса. Однако начальные значения в этом многошаговом методе должны быть насчитаны предварительно каким-либо двусторонним методом.

5.2. Ряды Тейлора

Излагаемый в начале раздела метод — естественное интервальное обобщение метода Эйлера. Он появился в работах Р. Е. Мура [89] одним из первых среди методов, которые в отличие от алгоритмов разд. 5.1 дают гарантированные двусторонние оценки. Мы изложим его на примере одного уравнения, но обобщение на случай системы очевидно. Ширина получаемого интервального решения для точных данных — величина $O(h)$, что вытекает из первого порядка точности метода Эйлера на сетке с шагом h. В конце раздела мы приведем модификации метода более высокого порядка точности и остановимся на алгоритме предварительного прогноза границ решения. Рассмотрим задачу Коши

$$y' = f(y), \quad x \in (0, l), \tag{5.14}$$

$$y(0) = y_0 \in \boldsymbol{y}_0. \tag{5.15}$$

Пусть $y(x) \in \Delta_y = [\underline{\Delta}, \bar{\Delta}]$ и f определена на Δ_y, кроме того, f имеет интервальное расширение $\boldsymbol{f}$ со следующими свойствами:

1) $\boldsymbol{f}$ — непрерывная функция;
2) $\boldsymbol{f}$ монотонна по включению;
3) существует $L > 0$ такое, что

$$\text{wid}\,(\boldsymbol{f}(\mathbf{y})) \le L\text{wid}\,(\boldsymbol{y}) \subset \Delta_y.$$

Введем разностную сетку $\omega_h = \{0 = x_0 < x_1 < \ldots < x_n = l\}$ с шагами $h_i = x_{i+1} - x_i$. Предположим, что уже вычислено интервальное значение $\boldsymbol{y}(x_i) \supset y(x_i)$. Тогда $\forall x \in [x_i, x_{i+1}]$ определим

$$\begin{aligned} &\boldsymbol{y}(x) = \boldsymbol{y}(x_i) + (x - x_i) f(\boldsymbol{z}), \\ &\boldsymbol{z} = \boldsymbol{y}(x_i) + [0, h_i]\boldsymbol{f}(\Delta_y), \quad i = 0, 1, \ldots, n-1. \end{aligned} \tag{5.16}$$

В качестве начального значения берем

$$\boldsymbol{y}(0) = \boldsymbol{y}_0. \tag{5.17}$$

Полагая в (5.16) $x = x_{i+1}$, получаем значение $\boldsymbol{y}(x_{i+1})$ и продолжаем процесс для следующего i. В итоге построена интервальная функция $\boldsymbol{y}(x)$ с кусочно-линейными границами $\underline{y}(x)$ и $\bar{y}(x)$. Для этого метода справедливо следующее утверждение.

Теорема 34. *Задача Коши (5.14), (5.15) при условиях 1–3 имеет единственное решение $y \in \boldsymbol{y}$. Кроме того, существуют константы c_1, c_2 такие, что*

$$\text{wid } (\mathbf{y}(x)) \leq c_1 h + c_2 \text{wid } (\mathbf{y}_0).$$

Формулы (5.16) носят название метода первого порядка. Можно построить методы более высоких порядков. В самом деле, пусть $y \in C^k[0, l]$. Рассмотрим разложение в ряд Тейлора с остаточным членом в форме Лагранжа:

$$y(x) = y(0) + \sum_{j=1}^{k-1} f^{(i-1)} x^j / j! + f(y(\theta)) x^k / k!, \quad \theta \in (0, l).$$

Если существуют интервальные расширения $\boldsymbol{f}^{(j)}$, то формулы метода k-го порядка записываются в следующем виде:

$$\boldsymbol{y}(x) = \boldsymbol{y}(x_i) + \sum_{j=1}^{k-1} \boldsymbol{f}^{(j-1)}(\boldsymbol{y}(x_i)) \frac{(x - x_i)^j}{j!} +$$

$$+ \boldsymbol{f}^{(k-1)}(\Delta_y) \frac{(x - x_i)^k}{k!}$$

При этом считаем, что выполнено условие $\boldsymbol{y}([x_i, x_i + x]) \subset \Delta_y$. Из свойства 3 непосредственно вытекает, что

$$\text{wid } (\boldsymbol{y}(x)) \leq (c_{k-1}/k!) \text{wid } (\Delta_y) h_i^k +$$

$$+ \left(1 + \sum_{j=1}^{n-1} c_{j-1} h^j\right) \text{wid } (\mathbf{y}(x)).$$

Таким образом, ширина интервального решения является величиной $O(h^k)$ для точных данных.

Пример 12. Рассмотрим задачу Коши

$$y' = y^2, \quad x \geq 0,$$

$$y(0) = 1.$$

Выпишем разложение в ряд Тейлора

$$\boldsymbol{y}(x+h) = y(x) + y^2(x)h + \boldsymbol{y}^3 h^2.$$

Положим $\Delta = [1, 2]$. Тогда

$$\underline{y}(h) = 1 + h + \underline{\Delta} h^2 = 1 + h + h^2,$$

$$\overline{y}(h) = 1 + h + \overline{\Delta} h^2 = 1 + h + 8h^2.$$

Заметим, что на h накладывается ограничение $\overline{y}(h) = 1 + h + 8h^2 \leq 2$. Таким образом, $h < 0.296\ldots$ Положим $h = 0.25$ и вычислим $\boldsymbol{y}(0.25)$:

$$\underline{y}(0.25) = 1 + 0.25 + (0.25)^2 = 1.3125,$$

$$\overline{y}(0.25) = 1 + 0.25 + 8(0.25)^2 = 1.75.$$

Как нетрудно убедиться, точное решение $y(0.25) = 1.3333\ldots \in [1.3125, 1.75]$.

5.3. Метод последовательных приближений Пикара

Запишем задачу (5.4), (5.5) в виде интегрального уравнения Вольтерра и решим его итеративно. Пусть $y^{(0)}(t)$ — начальное приближение, тогда

$$y^{(i+1)}(t) = y_0 + \int_0^t f(\tau, y^{(i)}(\tau))d\tau, \quad i = 0, 1, ... \tag{5.18}$$

Интервальный метод Пикара запишем в следующем виде. Пусть $y \in \boldsymbol{y}^{(0)}(t)$ — начальное приближение, $y^{(0)}(0) = y_0$. Положим

$$\boldsymbol{y}^{(i+1)}(t) = \boldsymbol{y}_0 + \int_0^t \boldsymbol{f}(\tau, \boldsymbol{y}^{(i)}(\tau))d\tau, \quad i = 0, 1, ..., \tag{5.19}$$

где $\boldsymbol{y}^{(i)}$ — интервальные функции, содержащие точное решение $y(t)$, а $\boldsymbol{f}$ — интервальное расширение функции f.

Для решения задачи (5.4), (5.5) интервальным методом (5.19) необходимо прежде всего задать начальное приближение $\boldsymbol{y}^{(0)}(t) \ni y(t)$. Кроме того, функция $\boldsymbol{f}$ должна обладать свойствами монотонности по включению и Липшиц-непрерывности:

wid $(\boldsymbol{f}(x, \boldsymbol{y})) \leq L$wid $(\boldsymbol{y})$. Тогда итерационный процесс сходится.

Для вычисления интеграла в (5.19) следует воспользоваться одним из методов интегрирования интервальных функций. Сходимость алгоритма вытекает из сходимости соответствующего процесса для интервального уравнения Вольтерра.

Пример 13. Рассмотрим задачу Коши

$$y' = y, \quad x \geq 0,$$

$$y(0) = 1.$$

Перепишем задачу в виде метода последовательных приближений Пикара

$$\boldsymbol{y}^{(i+1)}(t) = \boldsymbol{y}_0 + \int_0^t \boldsymbol{f}(\tau, \boldsymbol{y}^{(i)}(\tau))d\tau, \quad i = 0, 1, ...$$

Начальное приближение возьмем в виде $\boldsymbol{y}_0 = [1, 2]$. Тогда получаем

$$\boldsymbol{y}_1 = 1 + \int_0^x [1, 2]dx = [1 + x, 1 + 2x],$$

далее,

$$\boldsymbol{y}_2 = 1 + \int_0^x [1 + x, 1 + 2x]dx = [1 + x + x^2/2, 1 + x + x^2].$$

5.4. Двусторонние методы

Рассмотрим следующую систему

$$\begin{aligned} x_i' &= f_i(t, x, k), i = 1, ..., n, t \in (0, l), \\ x(0) &= x_0, \end{aligned} \tag{5.20}$$

где $x_0 \in R^n$ — вектор начальных значений, $x_0 \in \boldsymbol{x}_0$,
$k \in R^n$ — вектор параметров, $k \in \boldsymbol{k}$.
$x \in R^n$ — вектор неизвестных.

Далее будем считать, что x есть функция t, k, x :

$$x = x(t, k, x_0). \tag{5.21}$$

Обозначим через $\mathcal{X}(t)$ множество решений системы ОДУ:

$$\mathcal{X}(t) = \{x(t, k, x_0) | x_0 \in \boldsymbol{x}_0, \quad k \in \boldsymbol{k}\}.$$

Определение 12. *Минимальное по ширине двустороннее решение x задачи (5.20) будем называть оптимальным .*

Пусть существуют вещественные функции [44] $F^l(t, y_1, \ldots, y_n, z_1, \ldots, z_n)$, $l = 1, 2$; такие, что при $\underline{y} \leq \overline{y}, \underline{z} \leq z$ выполнены следующие неравенства:

$$F_i^l(t, \underline{y}, \overline{z}) \leq F_i^l(t, \overline{y}^{[\underline{y}_i]}, \underline{z}), l = 1, 2; i = 1, \ldots, n, \tag{5.22}$$

где $y^{[z_i]} = (y_1, \ldots, y_{i-1}, z_i, y_{i+1}, \ldots y_n)$. Кроме того, выполнены неравенства

$$F_i^1(x, x) < f_i(x, k) < F_i^2(x, x). \tag{5.23}$$

Теорема 35. *Пусть вектор-функции $\underline{x}, \overline{x} \in R^n$ удовлетворяют соотношениям*

$$\begin{aligned} \underline{x'} &\leq F^1(t, \underline{x}, \overline{x}), \\ \overline{x'} &\geq F^2(t, \overline{x}, \underline{x}), \\ \underline{x}(0) &\leq \underline{x}_0, \\ \overline{x}(0) &\geq \overline{x}_0. \end{aligned} \tag{5.24}$$

Тогда любое решение x системы (5.20) с начальным условием $\underline{x}_0 \leq x(0) \leq \overline{x}_0$ удовлетворяет оценкам

$$\underline{x}_t \leq x(t) \leq \overline{x}_t.$$

Заметим, что в силу (5.22)

$$F^1(t, \underline{x}, \overline{x}) \leq \inf f(t, x, k), F^2(t, \overline{x}, \underline{x}) \geq \sup f(t, x, k)$$

и в качестве F можно взять границы монотонного по включению интервального расширения функций f.

Определение 13. *Функция f называется изотонной, если выполняется условие*

$$x \leq y \Rightarrow f(x) \leq f(y),$$

и антитонной, если

$$x \leq y \Rightarrow f(x) \geq f(y).$$

Функция называется монотонной, если она изотонная или антитонная.

Замечание 4. *Пусть функция $f(x, y)$ изотонная по x и антитонная по y, тогда интервальное расширение $f(x, y)$ имеет вид:*

$$\underline{f}(\boldsymbol{x}, \boldsymbol{y}) = f(\underline{x}, \overline{y}), \overline{f}(\boldsymbol{x}, \boldsymbol{y}) = f(\overline{x}, \underline{y}).$$

Систему (5.24) можно переписать таким образом:

$$\begin{aligned} \underline{x'}_i &\leq \underline{f}_i(t, \boldsymbol{x}^{[\underline{x}_i]}, \boldsymbol{k}), \\ \overline{x'} &\geq \overline{f}_i(t, \boldsymbol{x}^{[\overline{x}_i]}, \boldsymbol{k}), \\ \underline{x}(0) &\leq \underline{x}_0, \\ \overline{x}(0) &\geq \overline{x}_0. \end{aligned} \tag{5.25}$$

Решение построенной системы дает в общем случае более широкое, чем оптимальное, двустороннее решение. Однако в некоторых частных случаях решение системы (5.25) дает оптимальные границы. Рассмотрим эти случаи.

Пусть система (5.25) имеет следующий вид:

$$\begin{aligned} \underline{x'} &= f(t, \underline{x}, k^1), \\ \overline{x'} &= f(t, \overline{x}, k^2), \\ \underline{x}(0) &= \underline{x}_0, \\ \overline{x}(0) &= \overline{x}_0, \end{aligned} \tag{5.26}$$

где $k^i \in \boldsymbol{k}$. В этом случае $\underline{x}, \overline{x}$ являются некоторыми частными решениями исходной системы и, следовательно, они оптимальны.

Сформулируем достаточные условия для представления системы (5.25) как (5.26).

Лемма 8. *Пусть выполнены следующие условия:*

$$\frac{\partial f_i}{\partial x_j} \geq 0, i \neq j,\ i, j = 1, 2, \ldots, n, \tag{5.27}$$

$$sign \frac{\partial f_i}{\partial k_j} = const, j : \mathrm{wid}(\boldsymbol{k}_j) \neq 0, \forall k \in \boldsymbol{k}, \forall x \in \boldsymbol{x}. \tag{5.28}$$

Тогда система (5.25) имеет вид (5.26).

Доказательство. Действительно, условия (5.27), (5.28) гарантируют представление интервального расширения функции f как в (5.26) в силу монотонности f по x и k. □

Пример 14.

$$\begin{aligned} x_1' &= -k_1 x_1, \qquad x_1(0) = 1, \\ x_2' &= k_1 x_1 - k_2 x_2, \quad x_2(0) = 0, \end{aligned} \tag{5.29}$$

где

$$k_i = k_i^0 \exp(-E_i/RT) \in \boldsymbol{k}_i, \quad i = 1, 2,$$

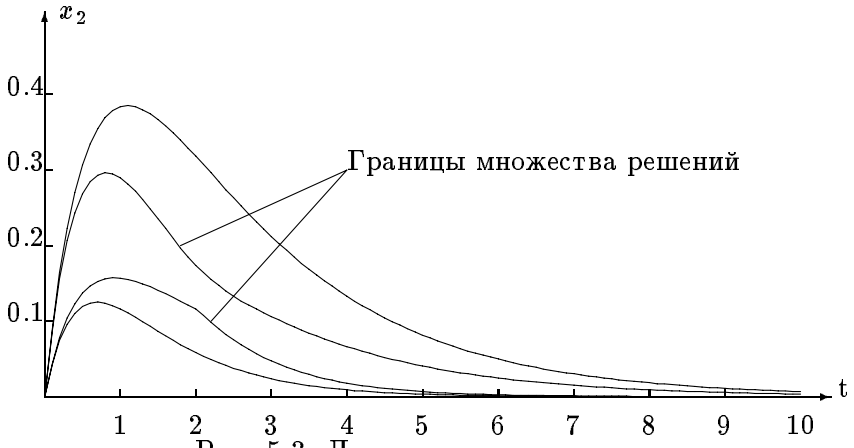

Рис. 5.3. Двустороннее решение

E_i, R — константы и $T = T(t) \in \boldsymbol{T}$.

Эта система ОДУ моделирует простейшую химическую реакцию. Для уравнений химкинетики характерно выполнение условий (5.27) [10].

Однако условия (5.28), как правило, не выполняются. Система ОДУ для двустороннего решения выглядит следующим образом:

$$\begin{aligned} \underline{x}_1' &= -\overline{k}_1\underline{x}_1, \\ \underline{x}_2' &= \underline{k}_1\underline{x}_1 - \overline{k}_2\underline{x}_2, \\ \overline{x}_1' &= -\underline{k}_1\overline{x}_1 \\ \overline{x}_2' &= \overline{k}_1\overline{x}_1 - \underline{k}_2\overline{x}_2. \end{aligned} \tag{5.30}$$

Построенная система распадается на две независимые подсистемы, но поскольку не выполнено условие (5.28), то двустороннее решение получается шире истинного. В данном примере ширина двустороннего решения зависит от $\mathrm{wid}(\boldsymbol{k}_1)$, и если $\mathrm{wid}(\boldsymbol{k}_1) = 0$, то получаем оптимальные границы для множества решений.

На рис. 5.3 представлено двустороннее решение системы (5.29).

Предположим, что известно дополнительное дифференциальное уравнение

$$T' = g(t, x, T). \tag{5.31}$$

Тогда, решая совместно системы (5.30) и (5.31), мы можем ожидать, что ширина двустороннего решения значительно уменьшится. Этот пример показывает, что знание дополнительной информации может быть направлено на уменьшение ширины двусторонних решений.

Решение системы (5.25) в аналитическом виде в общем случае невозможно, поэтому для нахождения решения можно воспользоваться чис-

ленными методами. В некоторых случаях погрешностями метода интегрирования можно пренебречь, тогда систему (5.25) можно решить методом, наиболее удобным для данной задачи. Если необходимы гарантированные оценки, то можно воспользоваться апостериорными оценками погрешности имеющегося численного решения.

Перейдем к вопросу о построении двустороннего решения с учетом погрешностей численного метода.

5.5. Апостериорная оценка погрешности

Рассмотрим методы, основанные на апостериорных оценках погрешностей сглаженного разностного решения исходной задачи. Причем начальные данные и правые части системы могут быть заданы как интервальные. Такая ситуация возникает при неточном задании данных, ошибки которых лежат в некоторых известных пределах, например в пределах ошибки измерения, в пределах небольшого динамического изменения и т.п.

Методы, основанные на мажорантах Лозинского

Рассмотрим систему уравнений

$$\frac{dy}{dx} = f(x,y), \quad x \in (0,l), \tag{5.32}$$

$$y(0) = g. \tag{5.33}$$

Здесь $y(x) = (y_1(x), y_2(x), ..., y_n(x))$, $g = (g_1, ..., g_n)$ — векторы с n компонентами, $f = (f_1, f_2, ..., f_n)$ — вектор-функция $n+1$ аргументов:

$$f_i(x,y) = f_i(x, y_1, y_2, ..., y_n), \quad i = 1, ..., n,$$

причем $f_i \in C^k([0,l] \times R^n), k \geq 3$.

Пусть компоненты правой части уравнения (5.32) являются представителями некоторых интервальных функций $\mathbf{f}_i$: $f_i(x,y) \in \mathbf{f}_i(x,y)$, начальные значения $g_i \in \mathbf{g}_i$, где $\mathbf{g}_i = [\underline{g}_i, \overline{g}_i]$ — интервальные числа. Относительно решения задачи (5.32), (5.33) предположим, что оно существует, единственно и $y_i \in C^{k+1}[0,l], i = 1, ..., m$.

Для построения двустороннего решения задачи с интервальными данными сначала приближенно решим задачу (5.32),(5.33) для конкретных представителей $f_i, g_i, i = 1, ..., m$, используя, например, метод Рунге

— Кутта четвертого порядка [7] на равномерной разностной сетке

$$\omega_h = \{x_i = ih, i = 0, 1, ..., n\}, \quad h = 1/n,$$

где n — целое. В результате на сетке ω_h получается приближенное решение $y_i^h(x), i = 1, ..., n$. С помощью уравнения (5.32) в узлах ω_h можно вычислить приближенные значения. Проведем через точки $y_i^h(x), x \in \omega_h$ эрмитовы сплайны третьей степени $s \in S_3^2$ и определим следующие функции:

$$\varphi_i(x, s) = f_i(x, s) - ds_i/dx, s = (s_1, ..., s_m).$$

Отметим, что выбранный метод решения (Рунге — Кутта четвертого порядка) не играет особой роли. Вместо него можно было взять любой другой метод, обеспечивающий точность приближенного решения порядка h^3.

Решим численно две задачи

$$\frac{du}{dx} = Wu + w, \quad x \in (0, l), \tag{5.34}$$

$$u(0) = 0$$

и

$$\frac{dv}{dx} = Wv, \quad x \in (0, l), \tag{5.35}$$

$$v(0) = z,$$

где векторы w, z имеют компоненты $w_i = 1$ и $z_i = (\bar{g}_i - \underline{g}_i)/2, i = 1, ..., m$. Матрица W состоит из элементов

$$W_{ii} = \frac{\partial f_i}{\partial y_i}(x, s), \quad i = 1, ..., m,$$

$$W_{ij} = \left|\frac{\partial f_i}{\partial y_j}(x, s)\right|, \quad i \neq j, \quad i, j = 1, ..., m.$$

Построим эрмитовы сплайны $s_i^{(1)}, s_i^{(2)}$, аппроксимирующие полученные численные решения $u_i^h, v_i^h, i = 1, ...m$.

Будем искать двустороннее решение в виде

$$\boldsymbol{y}_i = s_i + [-1, 1]s_i^{(2)} + \boldsymbol{a}s_i^{(1)} \tag{5.36}$$

с некоторой интервальной константой $\boldsymbol{a} = [-a, a]$.

Введем следующие интервальные функции:

$$\frac{\partial f_i}{\partial y_j}(x,\theta) \in \boldsymbol{f}_{y,i,j}(x,[\underline{\theta},\overline{\theta}]) = [\underline{f}_{y,i,j}, \overline{f}_{y,i,j}], \quad \theta \in [\underline{\theta},\overline{\theta}],$$

$$i,j = 1,2,...,m,$$

$$[\underline{\varphi}_i, \bar{\varphi}_i] = \boldsymbol{f}_i(x,s) - ds_i/dx.$$

Далее, пусть $\delta_i > 0$ — априорно заданные константы такие, что

$$\boldsymbol{y}_i(x) \in s_i(x) + [-1,1]s_i^{(2)}(x) + [-\delta_i,\delta_i]s_i^{(1)}(x).$$

Эти константы можно выбрать, используя работы [39],[89]. Тогда положим $\boldsymbol{r}_i = s_i + [-1,1]s_i^{(2)} + [-\delta_i,\delta_i]s_i^{(1)}$ и определим

$$\tilde{f}_{y,i,i}(x) = \bar{f}_{y,i,i}(x,\boldsymbol{r}), \quad i = 1,...,m,$$

$$\tilde{f}_{y,i,j}(x) = \max(|\bar{f}_{y,i,j}(x,\boldsymbol{r})|,$$

$$|\underline{f}_{y,i,j}(x,\boldsymbol{r})|), \quad i \neq j, \quad i,j = 1,...,m,$$

$$\Phi_i(x) = \max(|\underline{\varphi}_i(x,s)|, |\overline{\varphi}_i(x,s)|) - ds_i^{(2)}(x)/dx +$$

$$+ \sum_{j=1}^{m} \tilde{f}_{y,i,j}(x)s_j^{(2)}(x),$$

$$\Psi_i(x) = ds_i^{(1)}(x)/dx - \sum_{j=1}^{m} \tilde{f}_{y,i,j}s_j^{(1)}(x).$$

Для обеспечения двусторонних оценок выберем a следующим образом:

$$a = \max_{\substack{x\in[0,l]\\ i=1,...,m}} (\Phi_i(x)/\Psi_i(x), 0). \tag{5.37}$$

Оценим ширину полученного двустороннего решения. Предположим, что численный метод, с помощью которого была решена исходная задача (5.32), (5.33), имеет точность

$$|\varepsilon_i(x)| = |y_i(x) - y_i^h(x)| \leq c_1 h^k. \tag{5.38}$$

В случае метода Рунге — Кутта четвертого порядка $k = 4$. Пусть s_i^T — эрмитов сплайн третьей степени, интерполирующий y_i на сетке ω_h. Тогда существует константа c, не зависящая от h, y, такая, что [36]

$$\|d^\nu(y_i - s_i^T)/dx^\nu\|\infty \leq ch^{4-\nu}\|y_i\|_{4,\infty}, \quad \nu = 0,1. \tag{5.39}$$

Теорема 36. *Пусть s_i интерполирует приближенное решение y_i^h. Тогда существует константа c, не зависящая от h, y_i, такая, что*

$$\|d^\nu(y_i - s_i)/dx^\nu\|_\infty \le c(h^{4-\nu}\|y_i\|_{4,\infty} + h^k), \quad \nu = 0, 1. \tag{5.40}$$

Доказательство. Оценим норму

$$\|d^\nu(y_i - s_i)/dx^\nu\|_\infty \le \|d^\nu(y - s_i^T)\|_\infty +$$

$$+\|d^\nu(s_i^T - s_i)/dx^\nu\|_\infty, \quad \nu = 0, 1.$$

В силу предположения (5.38) приближенные значения производных $y_i'(x),\ i = 1, ..., m$ будут определены с точностью $O(h^k)$:

$$\|ds_i(x)/dx - y_i'(x)\|_\infty \le c_3 h^k \quad \forall x \in \omega_h. \tag{5.41}$$

Представим $s_i^T - s_i$ на отрезке $[x_j, x_{j+1}]$ в виде $s_i^T - s_i = a_3 t^3 + a_2 t^2 + a_1 t + a_0,\ t = x - x_j$. Тогда из (5.38), (5.41) вытекает, что

$$|a_0|, |a_1| \le c_4 h^k, \quad |a_2| \le c_5 h^{k-1}, \quad |a_3| \le c_6 h^{k-2}.$$

Следовательно,

$$\|d^\nu(s_i^T - s_i)/dx^\nu\|_\infty \le c_7 h^k, \quad \nu = 0, 1. \tag{5.42}$$

Объединяя оценки (5.39), (5.42), получаем (5.40). □

Покажем, что построенное двустороннее решение вида (5.36) содержит точное решение задачи (5.32), (5.33). Для доказательства этого утверждения исследуем уравнение для ошибки $y_i - s_i$. Действительно, s_i удовлетворяют следующим равенствам:

$$ds_i/dx = f_i(x, s) - \varphi_i(x, s). \tag{5.43}$$

Вычтем (5.43) из (5.32):

$$d(y_i - s_i)/dx = f_i(x, y) - f_i(x, s) + \varphi_i(x, s).$$

Используя теорему о среднем, приходим к равенству

$$d(y_i - s_i)/dx = \sum_{j=1}^{n} \frac{\partial f_i}{\partial y_j}(x, r_i)(y_j - s_j) + \varphi_i,$$

где r_i — некоторые функции, принимающие в точке x значения из интервалов с концами $y_i(x), s_i(x)$.

Приведем вспомогательную задачу для вектор-функции: $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$:

$$\frac{d\varepsilon}{dx} = V\varepsilon + \zeta, \quad x \in (0,1), \tag{5.44}$$
$$\varepsilon(0) = \varepsilon_0,$$

где матрица $V = \{V_{ij}\}_{ij=1}^n$, причем

$$V_{ii}(x) \geq \frac{\partial f_i}{\partial y_i}(x, r_i), \quad i = 1, ..., n,$$

$$V_{ij}(x) \geq \left|\frac{\partial f_i}{\partial y_j}(x, r_i)\right|, \quad i, j = 1, ..., n \quad i \neq j, \tag{5.45}$$
$$\varepsilon_0 = (\varepsilon_1^0, \varepsilon_2^0, ..., \varepsilon_n^0), \quad \varepsilon_i^0 \geq |g_i - s_i(0)|,$$
$$\zeta = (\zeta_1, ..., \zeta_n), \quad \zeta_i(x) \geq |\varphi_i(x, s)|, \quad x \in (0,1).$$

Для решения задачи (5.44) при условиях (5.45) справедлив следующий результат [48]:

$$\varepsilon_i(x) \geq |y_i(x) - s_i(x)|, \quad x \in [0, l].$$

Выберем матрицу V, полагая $V_{ij} = \tilde{f}_{y,i,j}$. Заметим, что матрица V удовлетворяет условиям (5.45). Следовательно, из (5.42) вытекает неравенство

$$(d(s^{(2)} + as^{(1)})/dx - V(s^{(2)} + as^{(1)}))_i =$$
$$= (ds^{(2)}/dx - Vs^{(2)} + a(ds^{(1)}/dx - Vs^{(1)}))_i \geq |\varphi_i|.$$

Поэтому $s_i^{(2)} + as_i^{(1)} \geq |y_i - s_i|$, следовательно, $y_i \in s_i + [-1, 1]s_i^{(2)} + as_i^{(1)}$ $\forall x \in [0, l]$.

Покажем, что всегда существует интервал $(0, l)$, на котором a определено. Для этого рассмотрим вспомогательные функции

$$\tilde{\Psi}_i = ds_i^{(1)}/dx - \sum_{j=1}^m W_{ij}s_j^{(1)}.$$

Заметим, что в силу теоремы 36

$$\tilde{\Psi}_i(x) \geq 1 - ch^3 \quad \forall x \in (0, l), \quad i = 1, ..., m.$$

Кроме того, предположим, что существует константа c_8 такая, что [36]:

$$|\tilde{f}_{y,i,j} - W_{ij}| \leq c_8 \sum_{j^1=1}^m \left(s_{j^1}^{(2)} + \delta_i s_{j^1}^{(1)}\right).$$

Тогда

$$|\Psi_i - \tilde{\Psi}_i| \le c_8 \sum_{j=1}^{m} \left(\sum_{j^1=1}^{m} \left(s_{j^1}^{(2)} + \delta_i s_{j^1}^{(1)} \right) s_j^{(1)} \right).$$

Поскольку $s_i^{(1)}(0) = 0,\ i = 1, ..., n$, и $s_i^{(1)}$ интерполирует решение задачи (5.34), предположим, что существует константа c_9 такая, что

$$|s^{(1)}| \le c_9 l \quad \forall i = 1, ..., n.$$

Поэтому можно выбрать такие h, l , что $\Psi_i(x) > 0$, и, следовательно, a определено.

Оценим ширину двустороннего решения в случае $\boldsymbol{f}_i = f_i, \boldsymbol{g}_i \equiv g_i$. Заметим, что ширина полученного двустороннего решения зависит от величин $a, |s_i^{(1)}|, |s_i^{(2)}|, i = 1, ..., m$.

Из формулы (5.37) вытекает оценка $a \le c \max\limits_{x \in (0,l)} |\varphi_i(x, s)|$.

По теореме 36 φ_i можно оценить следующим образом:

$$\|\varphi_i\|_\infty \le c \left(h^3 \|y_i\|_{4,\infty} + h^4 \max_{i=1,...,m} \|y_i\|_{4,\infty} \right),$$

следовательно,

$$a = c \max_{i=1,...,m} (h^3 \|y_i\|_{4,\infty} + h^k).$$

Ширина $\rho(x)$ двустороннего решения ограничена таким образом:

$$\rho(x) \le c \max_{i=1,...,m} (h^3 \|y_i\|_{4,\infty} + h^k) s_i^{(1)}(x).$$

Полученное двустороннее решение можно уточнить, если воспользоваться вычисленным a и положить $\delta_i = a$. Тогда можно вычислить новые значения a.

Пусть $\boldsymbol{f}_{y,i,j}, \boldsymbol{f}$ монотонны по включению, т. е. из условия $\tilde{\boldsymbol{r}}_i \subset \boldsymbol{r}_i, i = 1, ..., n$ вытекает, что $f(x, \tilde{\boldsymbol{r}}) \subset f(x, \boldsymbol{r})$. Тогда $\tilde{a} \le a$ и новое двустороннее решение

$$\boldsymbol{y}_i = s_i + [-1, 1] s^{(2)} + \tilde{a} s_i^{(1)}$$

имеет меньшую ширину.

Рассмотрим модельную задачу

$$\frac{dy}{dx} = by + c, \quad x \in (0, 1), \tag{5.46}$$

$$y(0) = y_0, \quad y_0 \in \boldsymbol{y}_0, \quad b \in \boldsymbol{b}, \quad c \in \boldsymbol{c}.$$

Точное решение имеет вид $y = (y_0 + c/b)\exp(bx)$. Вспомогательные задачи имеют вид:

$$\frac{dy_1}{dx} = by_1 + 1, \quad x \in (0,1), \tag{5.47}$$
$$y_1(0) = 0$$

и

$$\frac{dy_2}{dx} = by_2, \quad x \in (0,1), \tag{5.48}$$
$$y_2(0) = (\bar{y}_0 - \underline{y}_0)/2 = \varepsilon_0.$$

Их точные решения $y_1 = \exp(bx)b - 1, y_2 = \varepsilon_0 \exp(bx)$. Поскольку задача линейна, для вычисления двустороннего решения нет необходимости строить функцию $\boldsymbol{r}$ и, следовательно, априорно задавать δ.

Пусть $s, s^{(1)}, s^{(2)}$— сплайны, интерполирующие численные решения задач (5.46)-(5.48). Тогда

$$\Phi = |bs + c - ds/dx| - ds^{(2)}/dx + bs^{(2)},$$

$$\Psi = -bs^{(1)} + ds^{(1)}/dx.$$

Заметим, что в силу теоремы 36 $\Psi(x) \geq 1 - ch^3 > 0$. Таким образом, a определено. Двустороннее решение имеет вид

$$\boldsymbol{y} = s + [-1,1]s^{(2)} + [-a,a]s^{(1)}.$$

Его ширина удовлетворяет неравенству

$$\rho(x) \leq 2s^{(2)}(x) + 2as^{(1)}(x),$$

причем $\rho(0) = -2\varepsilon_0$. Поскольку

$$a \leq ch^3, s^{(1)} \leq ch, s^{(2)} \leq \varepsilon_0 \exp(bh),$$

то

$$\rho(h) \leq 2\varepsilon_0 \exp(bh) + ch^4.$$

Для $b \leq b_0 \leq 0$ $\rho(h) \leq \rho(0)$, т.е. ширина двустороннего решения убывает.

В качестве численного примера была решена задача:

$$\frac{dy_1}{dx} = -(1 + [-\varepsilon,\varepsilon])y_1 + (3 + [-\varepsilon,\varepsilon])y_3 + (1 + [-\varepsilon,\varepsilon])y_2y_3 -$$
$$-3e^{-3x} - e^{-5x} + [-\varepsilon,\varepsilon](e^{-3x} + e^{-5x} - e^{-x}),$$

$$\frac{dy_2}{dx} = (1 + [-\varepsilon, \varepsilon])y_1 + (2 + [-\varepsilon, \varepsilon])y_2 + (1 + [-\varepsilon, \varepsilon])y_1 y_2 -$$

$$-e^{-x} - e^{-4x} + [-\varepsilon, \varepsilon](e^{-x} - e^{-2x} + e^{4x}),$$

$$\frac{dy_3}{dx} = (2 + [-\varepsilon, \varepsilon])y + (3 + [-\varepsilon, \varepsilon])y_3 + (1 + [-\varepsilon, \varepsilon])y_2 y_1 - 2e^{-2x} -$$

$$-e^{-3x} + [-\varepsilon, \varepsilon](e^{-2x} + 2e^{-3x}),$$

$$y_1(0) - y_2(0) = y_3(0) = 1.0, \quad \varepsilon = 0.001.$$

В ней коэффициенты правой части являются интервальными числами. Точное решение при $\varepsilon = 0$ имеет вид

$$y_1(x) = e^{-x}, \quad y_2(x) = e^{-2x}, \quad y_3(x) = e^{-3x}.$$

Исходная задача решалась методом Рунге — Кутта четвертого порядка с шагом $h = 0.1$. После этого приближенное решение интерполировалось эрмитовыми сплайнами третьей степени. Полученное сглаженное решение использовалось для определения невязки. Затем по невязке строилась мажоранта для ошибки сглаженного решения. Задача получения двусторонней оценки решалась на каждом отрезке $[jh, (j+1)h], j = 0, 1, ..., N-1$. В качестве начальных значений принимались либо двусторонние решения, вычисленные на отрезке $[(j-1)h, jh]$, если $j \geq 1$, или $y_i, i = 1, 2, 3$ если $j = 0$. Для нахождения a на каждом отрезке $[jh, (j+1)h]$ использовались интервальные расширения функций Φ_i, ξ_i по x. Тогда, обозначая

$$[\underline{\xi}_i(x), \bar{\xi}_i(x)] = \boldsymbol{\Phi}_i(x)/\boldsymbol{\psi}_i(x),$$

определим a:

$$a = \max_{\substack{i=1,\ldots,m \\ j=0,1,\ldots,n-1}} \left(|[\underline{\xi}_i([jh, (j+1)h])|, \ |\bar{\xi}_i([jh, (j+1)h])| \right).$$

После этого двустороннее решение вычислялось по формуле (5.34). Приведем результаты численного эксперимента:

компоненты	Нижняя граница	Точное решение	Верхняя граница	Ширина $\times 100$
$x = 0.1$				
$i = 1$	0.904400	0.904837	0.905315	0.0916
$i = 2$	0.818347	0.818730	0.819149	0.0802
$i = 3$	0.740437	0.744081	0.744234	0.0897
$x = 0.5$				
$i = 1$	0.604271	0.606530	0.608983	0.4712
$i = 2$	0.366810	0.367879	0.369035	0.2225
$i = 3$	0.222048	0.223130	0.224246	0.1198
$x = 0.9$				
$i = 1$	0.401540	0.406559	0.411946	1.0406
$i = 2$	0.164171	0.165248	0.166503	0.2932
$i = 3$	0.066152	0.067205	0.068330	0.1178

Сначала ширина компонент интервального решения нарастает. При дальнейшем увеличении x происходит стабилизация ширины каждой компоненты, а затем идет ее уменьшение.

Использование дифференциальных неравенств

Представим задачу (5.25) в следующем виде:

$$\begin{aligned} \underline{x'}_i &= \underline{f}_i(t, \boldsymbol{x}^{[\underline{x}_i]}, \mathbf{k}), \\ \overline{x'} &= \overline{f}_i(t, \boldsymbol{x}^{[\overline{x}_i]}, \boldsymbol{k}), t \in (0, l), \\ \underline{x}(0) &= \underline{x}_0, \\ \overline{x}(0) &= \overline{x}_0. \end{aligned} \tag{5.49}$$

Предположим, что задача (5.49) решена численно с использованием некоторого метода интегрирования точности p на сетке

$$w_h = \{x_j, j = 1, 2, \ldots, N\}, \text{N} \text{ — целое.}$$

В результате, на сетке w_h имеем приближенные решения. С помощью уравнения (5.49) в узлах сетки w_h можно вычислить приближенные значения производных $x'(t)$. Используя полученные значения, проведем через точки $\underline{x}_i^h(t), \overline{x}_i^h(t), t \in w_h$ эрмитовы сплайны $\underline{s}_i, \overline{s}_i$ степени $r - 1$. Справедлива теорема [77].

Теорема 37. *Пусть s интерполирует приближенное решение x^h . Тогда существует константа C , не зависящая от h, x^h , такая, что*

$$||d^\nu(x-s)/dx^\nu||_{L_\infty[0,l]} \le C(h^{r-\nu}||x||_{W^r_\infty[0,l]} + h^p), \nu = 0, 1, \ldots, r-1.$$

Решение системы (5.25) будем искать в виде

$$\boldsymbol{s} = s + \boldsymbol{\alpha} s^1, \tag{5.50}$$

где $\boldsymbol{\alpha}$ — константа, $\boldsymbol{s}$ — сплайны, интерполирующие численное решение следующей системы ОДУ:

$$\underline{z}'_i = \sum_{j=1}^{n} (\partial \underline{f}_i(s, \mathbf{k})/\partial \underline{x}_j)\underline{z}_j + (\underline{f}_i(s, \mathbf{k}) - \underline{s}'_i)_-,$$

$$\overline{z}'_i = \sum_{j=1}^{n} (\partial \overline{f}_i(s, \mathbf{k})/\partial \overline{x}_j)\overline{z}_j + (\overline{f}_i(s, \mathbf{k}) - \overline{s}'_i)_+,$$

$$\underline{z}(0), \overline{z}(0) = 0,$$

где

$$(f)_+ = \begin{cases} f \ если \ f \ge 0 \\ 0 \ иначе \end{cases},$$

$$(f)_- = \begin{cases} f \ если \ f \le 0 \\ 0 \ иначе \end{cases}.$$

Подставив (5.50) в систему (5.25), получаем, что $\boldsymbol{\alpha}$ должна удовлетворять следующей системе неравенств:

$$(\underline{s} + \underline{\alpha s}_1)'_i \le \underline{f}_i(t, (\underline{s} + \underline{\alpha s}_1, \overline{s} + \overline{\alpha s}_1)^{[(\underline{s}+\underline{\alpha s}_1)_i]}, \boldsymbol{k}), t \in (0, l),$$

$$(\overline{s} + \overline{\alpha s}_1)'_i \ge \overline{f}_i(t, (\underline{s} + \underline{\alpha s}_1, \overline{s} + \overline{\alpha s}_1)^{[(\overline{s}+\overline{\alpha s}_1)_i]}, \mathbf{k}),$$

$$(\underline{s} + \underline{\alpha s}_1)_i \le \underline{x}_0,$$

$$(\overline{s} + \overline{\alpha s}_1)_i \ge \overline{x}_0.$$

Эту систему неравенств можно переписать в виде системы нелинейных уравнений относительно $\underline{\alpha}, \overline{\alpha}$:

$$\underline{\alpha} = \Phi_1(\underline{\alpha}, \overline{\alpha}),$$

$$\overline{\alpha} = \Phi_2(\underline{\alpha}, \overline{\alpha}).$$

Построенную систему нелинейных уравнений можно решить методом простой итерации

$$\begin{aligned} \underline{\alpha}_{i+1} &= \Phi_1(\underline{\alpha}_i, \overline{\alpha}_i), \\ \overline{\alpha}_{i+1} &= \Phi_2(\underline{\alpha}_i, \overline{\alpha}_i). \end{aligned} \tag{5.51}$$

Начальное приближение можно выбрать следующим:

$$\underline{\alpha}_0, \overline{\alpha}_0 = 1.0. \tag{5.52}$$

Несложно видеть, что итерационный процесс (5.51) с начальным приближением (5.52) сходится при достаточно малом l.

Следующее неравенство оценивает ширину построенного двустороннего решения при $\mathrm{wid}(\boldsymbol{k}) = 0$ и $\mathrm{wid}(\boldsymbol{x}_0) = 0$:

$$\rho(t) \le C \max_{i=1,\dots,n} \{h^r ||\underline{x}||_{W_\infty^{r+1}} + h^p)\underline{s}^1(t), h^r ||\overline{x}||_{W_\infty^{r+1}} + h^p)\overline{s}^1(t), \}, \tag{5.53}$$

где r — степень сплайна, p — порядок точности численного метода.

Теорема 38. *Пусть система ОДУ (5.25) может быть представлена в виде (5.26), x^* - оптимальное решение. Тогда*

$$|\underline{x}(t) - \underline{x}^*(t)| \le Ch^\sigma,$$

$$|\overline{x}(t) - \overline{x}^*(t)| \le Ch^\sigma,$$

где $\sigma = \min(r-1, p)$.

Доказательство непосредственно вытекает из формулы (5.53).

Тем самым мы можем избежать эффекта упаковывания, если можем представить ситему (5.25) в виде (5.26).

5.6. Анализ чувствительности

Основная идея подхода заключается в анализе частных производных решения по параметрам. Этот подход во многом пересекается со стандартным анализом чувствительности, и для его реализации используют аппарат интервального анализа. Поэтому далее будем называть его методом интервального анализа чувствительности (МИАЧ).

Предположим, что мы хотим оценить $\overline{x}^{(i)}$ — верхнюю границу $\mathbf{x}(t)$ по i-й координате $\overline{x}^{(i)} \ge x_i$.

Рассмотрим систему:

$$\begin{aligned} x_i' &= f_i(t, x, k),\ t \in (0, l), \\ x_i(0) &= x_{0i},\ i = 1, \dots, n, \end{aligned} \tag{5.54}$$

где

$x \in R^n$ — вектор неизвестных переменных,
$x_0 \in R^n$ — вектор начальных данных, $x_0 \in \boldsymbol{x}_0$,
$k \in R^m$ — вектор параметров, $k \in \boldsymbol{k}$.

Пусть решение x — функция от t, k и x_0:

$$x = x(t, k, x_0). \tag{5.55}$$

Обозначим через $\mathcal{X}(t)$ — решение системы ОДУ

$$\mathcal{X}(t) = \{x(t, k, x_0) | x_0 \in \boldsymbol{x}_0, k \in \boldsymbol{k}\}.$$

Для оценки $\overline{x}_i$ рассмотрим систему ОДУ

$$\begin{aligned} \tilde{x}' &= f(t, \tilde{x}, \tilde{k}),\ \tilde{k} \in \tilde{\boldsymbol{k}}, \\ \tilde{x}(0) &= \tilde{x}_0 \in \boldsymbol{x}_0. \end{aligned} \tag{5.56}$$

Здесь

$$\tilde{\boldsymbol{k}}_j = \begin{cases} \overline{k}_j, & \text{если } \boldsymbol{x}^k_{ij}(t) \le 0, \\ \underline{k}_j, & \text{если } \boldsymbol{x}^k_{ij}(t) \ge 0, \\ \boldsymbol{k}_j, & \text{если } \boldsymbol{x}^k_{ij}(t) \ni 0, \end{cases}$$

и

$$\tilde{\boldsymbol{x}}_0 = \begin{cases} \overline{x}_{0j}, & \text{если } \boldsymbol{x}^0_{ij}(t) \le 0, \\ \underline{x}_{0j}, & \text{если } \boldsymbol{x}^0_{ij}(t) \ge 0, \\ \boldsymbol{x}_0, & \text{если } \boldsymbol{x}^0_{ij}(t) \ni 0, \end{cases}$$

где $\boldsymbol{x}^k_{ij}(t)$ — интервальное расширение $\partial x_i / \partial k_j$ и $\boldsymbol{x}^0_{ij}(t)$ — интервальное расширение $\partial x_i / \partial x_{0j}$. Если интервалы $\boldsymbol{x}^k_{ij}(t)$ и $\boldsymbol{x}^0_{ij}(t)$ не содержат в себе 0, то система (5.56) не содержит интервальных параметров и решается интервальными или двусторонними методами с произвольной точностью [78, 31].

Интервальную функцию $\boldsymbol{x}^k_{ij}(t)$ и $\boldsymbol{x}^0_{ij}(t)$ можно определить, одновременно решая систему (5.54) и системы ОДУ:

$$x^{k\prime}_{ij} = \sum_{l=1}^{n} \frac{\partial f_i}{\partial x_l}(t, x, k) x^k_{lj} + \frac{\partial f_i}{\partial k_j}(t, x, k), \tag{5.57}$$

$$x^k_{ij}(0) = 0, i, j = 1, 2, ..., n.$$

$$x^{0\prime}_{ij} = \sum_{l=1}^{n} \frac{\partial f_i}{\partial x_l}(t, x, k) x^0_{lj}, i, j = 1, 2, ..., n. \tag{5.58}$$

$$x^0_{ij}(0) = \delta_{ij},$$

где δ_{ij} — символ Кронекера.

Теорема 39. *Пусть*

$$0 \notin \frac{\partial f_i}{\partial k_j}(0, \boldsymbol{x}^0, \boldsymbol{k}),$$

$$0 \notin \frac{\partial f_i}{\partial x_k}(0, \boldsymbol{x}^0, \boldsymbol{k}).$$

Тогда существует $t_0 > 0$ *такое, что*

$$0 \notin \boldsymbol{x}_{ij}^k(t), \quad 0 \notin \boldsymbol{x}_{ij}^0(t).$$

Доказательство. Проверяем, что при выполнении условий теоремы при $t = 0$ правые части систем обыкновенных дифференциальных уравнений (5.57), (5.58) не содержат нулей и выполнены соотношения $0 \notin \boldsymbol{x}_{ij}^{k\prime}(0)$, $0 \notin \boldsymbol{x}_{ij}^{0\prime}(0)$. Следовательно, существует окрестность начальной точки $t = 0$, в которой $0 \notin \boldsymbol{x}_{ij}^k(t)$, $0 \notin \boldsymbol{x}_{ij}^0(t)$.□

Таким образом, до момента $t_0 > 0$ мы можем построить оптимальные границы множества решений системы ОДУ.

Приведем пример:

$$\begin{aligned} x_1' &= kx_2, x_1(0) = x_{01} \in [-0.1, 0.1], \\ x_2' &= -kx_1, x_2(0) = x_{02} \in [0.9, 1.1], k \in \boldsymbol{k} = [1.0, 2.0]. \end{aligned} \tag{5.59}$$

На рис. 5.4, сравнивается двустороннее решение, полученное методом интервального анализа чувствительности, и точное решение. Это сравнение показывает, что предложенный метод построил оптимальные границы интервального решения до времени $t \approx 0.71$.

5.7. Теория огибающих

В этом разделе мы распространим теорию огибающих [37] на решения систем ОДУ с параметрами. *Семейством* решений системы ОДУ назовем решение задачи (5.54)

$$x(t, k), k \in \boldsymbol{k}. \tag{5.60}$$

Участком огибающей семейства (5.60) назовем кривую $r(t)$, если при каждом значении t она касается хотя бы одного из решений семейства (5.60). Соответствие между t и k задается функцией

$$k = k(t), k(t) \not\equiv const. \tag{5.61}$$

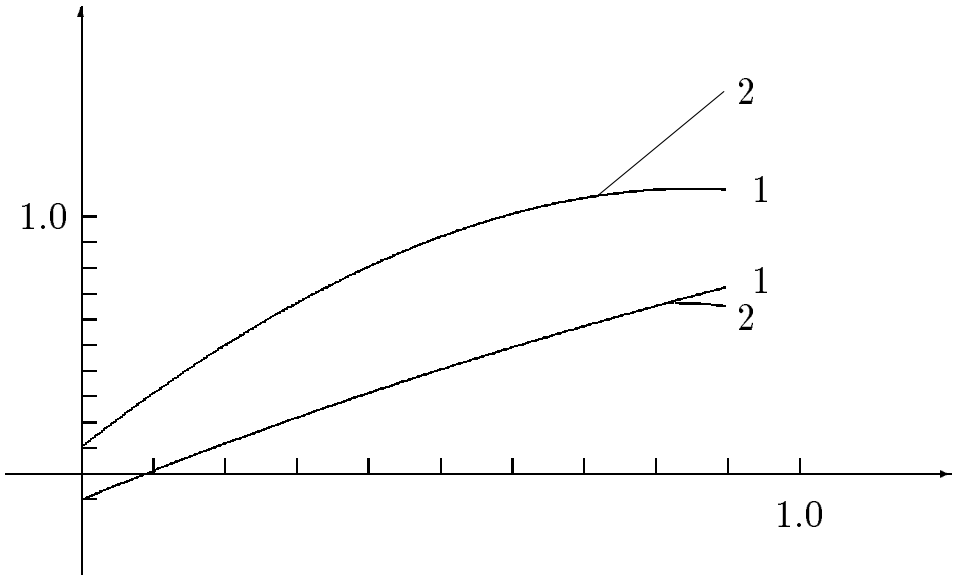

Рис. 5.4. 1 — точное решение, 2 — метод интервального анализа чувствительности

Функции (5.61) называются *законом прикрепления.*

Для огибающих семейств решений систем ОДУ можно сформулировать необходимый признак огибающей.

Теорема 40. *Если у семейства (5.60) есть огибающая* $r(t)$, *то*

$$\frac{\partial x(t,k)}{\partial k} = 0, \;\; u \;\; r(t) = x(t, k(t)).$$

Доказательсво полностью повторяет соответствующее доказательство из [37].

Построение огибающих

Для определенности рассмотрим построение огибающей семейства решений $x_1 = x_1(t,k)$, зависящей от одного параметра k. Таким образом, для приближенного построения огибающей систему (5.54) решаем совместно с (5.57), (5.58). Далее находим множества $T = \{(t,k)\}$, при которых $x^k = 0$. Пусть T — одно из таких множеств, $t \in [t_1, t_2]$ и k_1 соответствует t_1, k_2 — t_2. Тогда, исходя из свойств огибающих, получаем, что для огибающей $r(t)$ семейства решений на отрезке $t \in [t_1, t_2]$

выполнены следующие соотношения:

$$\begin{aligned} x_1(t_1,k_1) &= r(t_1); \\ x_1'(t_1,k_1) &= f(t_1,x,k_1) = r'(t_1); \\ x_1(t_2,k_2) &= r(t_2); \\ x_1'(t_2,k_2) &= f(t_2,x,k_2) = r'(t_2). \end{aligned} \tag{5.62}$$

Из (5.62) несложно построить аппроксимацию огибающей на отрезке $t \in [t_1, t_2]$, например используя эрмитовы сплайны. Подобным образом можно построить для огибающих в точках t_1, t_2 производные высоких порядков и, соответственно, получить более точную аппроксимацию границ множества решений.

Численные примеры

В качестве примера рассмотрим кинетическую модель, которая является осциллятором:

$$\begin{aligned} \frac{dx}{dt} &= k_1(1-x-y) - k_{-1}x - k_2x(1-x-y)^2, \\ \frac{dy}{dt} &= k_3(1-x-y) - k_{-3}y. \end{aligned} \tag{5.63}$$

При значениях параметров $k_1 \in [0.12, 0.125]$, $k_{-1} = 0.01$, $k_2 = 1$,$k_3 = 0.0032$, $k_{-3} = 0.002$ система (5.63) имеет автоколебания. Далее система (5.63) решалась при значениях параметра $k_1 = 0.12$ и $k_1 = 0.125$ совместно с системой (5.57).

Для иллюстрации на рис. 5.5 даны также несколько решений x_1 системы (5.63). В тех случаях, когда $0 \notin \boldsymbol{x}_1^k(t)$, в качестве границ двустороннего решения выбирали соответственно частные решения при значениях параметра $k_1 = 0.12$ или $k_1 = 0.125$. В противном случае находились отрезки $[t_1, t_2]$, на которых $0 \in \boldsymbol{x}_1^k(t)$, и на них строились огибающие. На рисунке они показаны более жирной линией, другую границу двустороннего решения выбирали из частных решений при $k_1 = 0.12$ или $k_1 = 0.125$.

Мы видим, что при сравнительно небольших вычислительных затратах можно построить сходящееся двустороннее решение с приемлемой точностью.

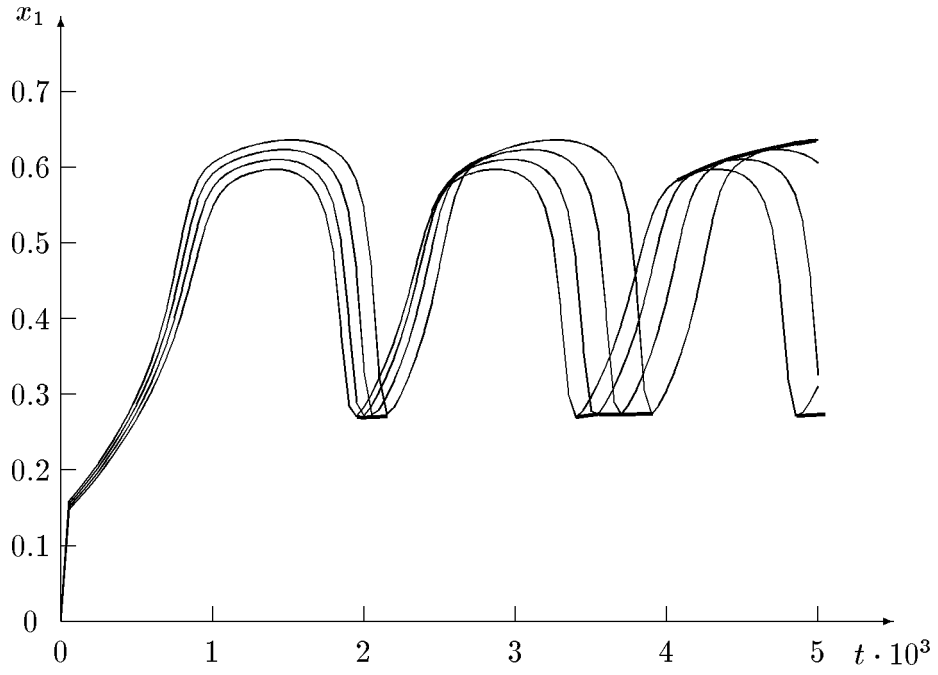

Рис. 5.5. Двустороннее решение системы (5.61)

5.8. Построение областей, содержащих множества решений

Обозначим через $\mathcal{R}$ множество всех n-мерных областей. Элементы этого множества будем обозначать так же, как и интервальные числа. Будем рассматривать только такие $\boldsymbol{x} \in \mathcal{R}$, которые можно однозначно описать конечным набором параметров, например, сферы, эллипсоиды, многогранники и т. п.

Перейдем к задаче оценки в фазовом пространстве множества решений $\{x(t,k,x_0)|k \in \boldsymbol{k}\}$ системы обыкновенных дифференциальных уравнений n-мерным параллелепипедом $\boldsymbol{x}(t) \in \mathcal{R}$

$$\{x(t,k,x_0)|k \in \boldsymbol{k}, x_0 \in \boldsymbol{x}(0)\} \subseteq \boldsymbol{x}(t),$$

где $x(t,k,x_0)$ — некоторое конкретное решение системы (5.49) с параметрами k и начальным условием $\boldsymbol{x}(0)$ в момент времени $t = 0$.

Любой параллелепипед можно задать некоторой вершиной O и ребрами, выходящими из нее: $e_i, i = 1, 2, \ldots, n$. Следовательно, каждому такому элементу $\boldsymbol{x} \in \mathcal{R}$ сопоставим вектор параметров v.

Заметим, что начальное состояние системы (5.54) представляет собой некоторый параллелепипед. Предположим, что в некоторый момент

времени t нам известен параллелепипед $\boldsymbol{x}(t)$, содержащий множество решений исходной системы ОДУ. Построим множество $\mathcal{X}$

$$\mathcal{X}(t,\tau,\boldsymbol{k},\boldsymbol{x}_0) \supseteq \{x(t+\tau,k,x_0)|k \in \mathbf{k}, x_0 \in \boldsymbol{x}(t)\}.$$

Это множество определим приближенно, используя методы численного интегрирования, например метод Эйлера

$$\mathcal{X}(t,\tau,\boldsymbol{k},x_0) \supseteq \{x^h(t+\tau,k,x_0)|x^h(t+\tau,k,x_0) = x_0 + tf(t,x_0,k), k \in \boldsymbol{k}\}.$$

Ясно, что в данном случае граница $\mathcal{X}(t,\tau,\boldsymbol{k},x_0)$ отличается от истинной на величину, не превышающую $O(\tau^2)$. Таким образом, зная в некоторый момент времени t $\boldsymbol{x}(t,\boldsymbol{k},\boldsymbol{x}_0)$, можем построить область $\boldsymbol{x}(t+\tau,\boldsymbol{k},\boldsymbol{x}_0)$

$$\boldsymbol{x}(t+\tau,\mathbf{k},\mathbf{x}_0) \supseteq \{\mathcal{X}(t,\tau,\boldsymbol{k},z_0|z_0 \in \boldsymbol{x}(t,\boldsymbol{k},\boldsymbol{x}_0)\}. \tag{5.64}$$

Существует известный произвол при построении параллелепипеда, обладающего свойствами (5.64). Будем стремиться строить его таким образом, чтобы он имел наименьший объем.

Если $\forall t, \tau > 0$ известны векторы $v(t)$ и $v(t+\tau)$, описывающие поведение параллелепипеда $\boldsymbol{x}$, то можно предельным переходом построить систему ОДУ, описывающую поведение v:

$$v_i'(t) = lim_{\tau \to 0}(v_i(t+\tau) - v_i(t))/\tau.$$

Следовательно, можно построить систему ОДУ, описывающую поведение параллелепипеда, содержащего множество решений исходной системы

$$\begin{aligned} v' &= g(t,v,k), \\ v(0) &= v_0. \end{aligned} \tag{5.65}$$

Если мы построили систему ОДУ (5.65) так, что по $v(t)$ можно восстановить некоторые траектории $x(t,k,x)$, то можно попытаться построить оптимальные границы множества решений.

Приведем несколько численных примеров. Для системы (5.29) на рис. 5.6 приведены параллелепипедные оценки множества решений в различные моменты времени.

На рис. 5.7 представлены параллелепипедные оценки множества решений для задачи (5.59).

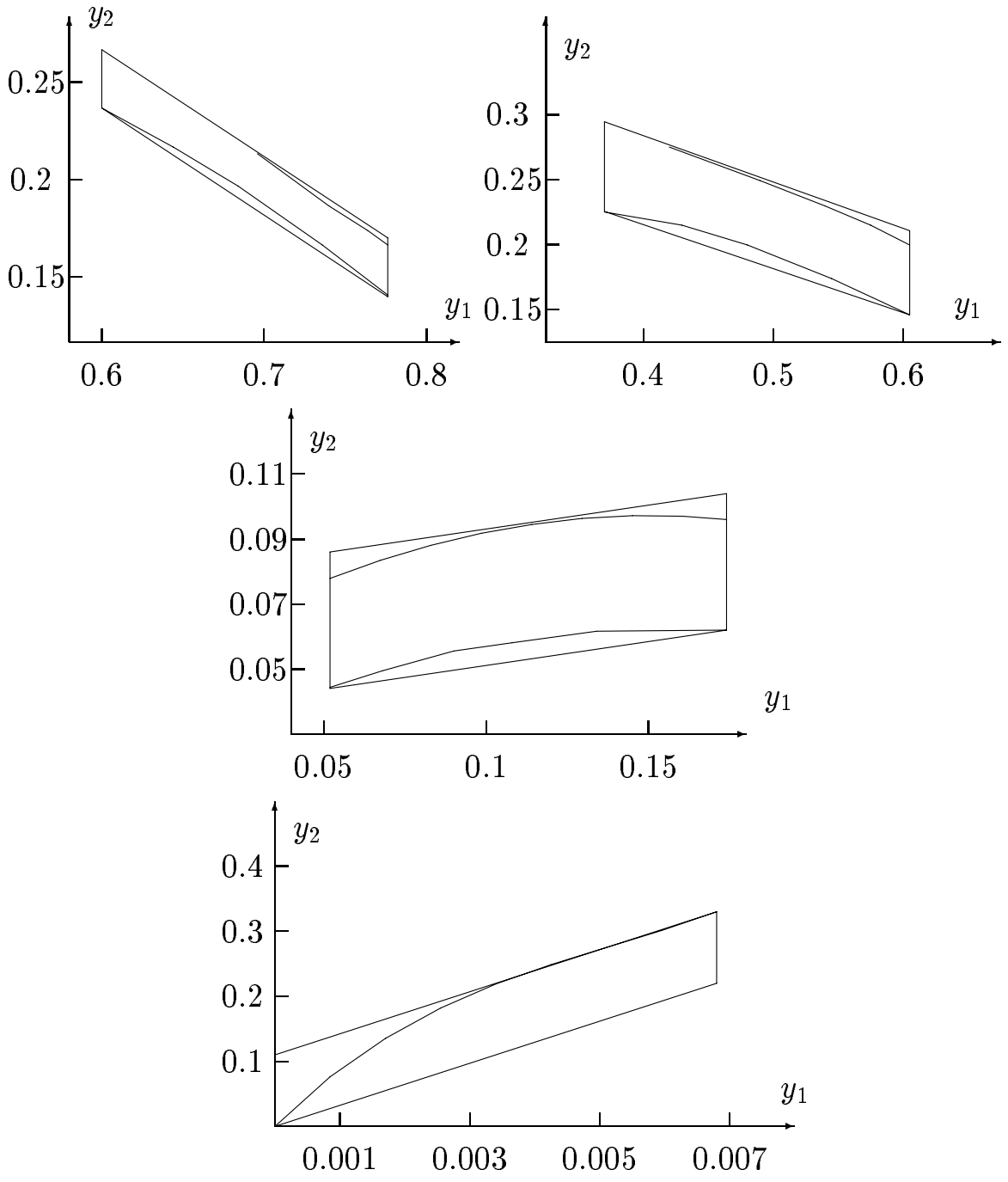

Рис. 5.6. Параллелепипедные оценки множества решений системы (5.29)

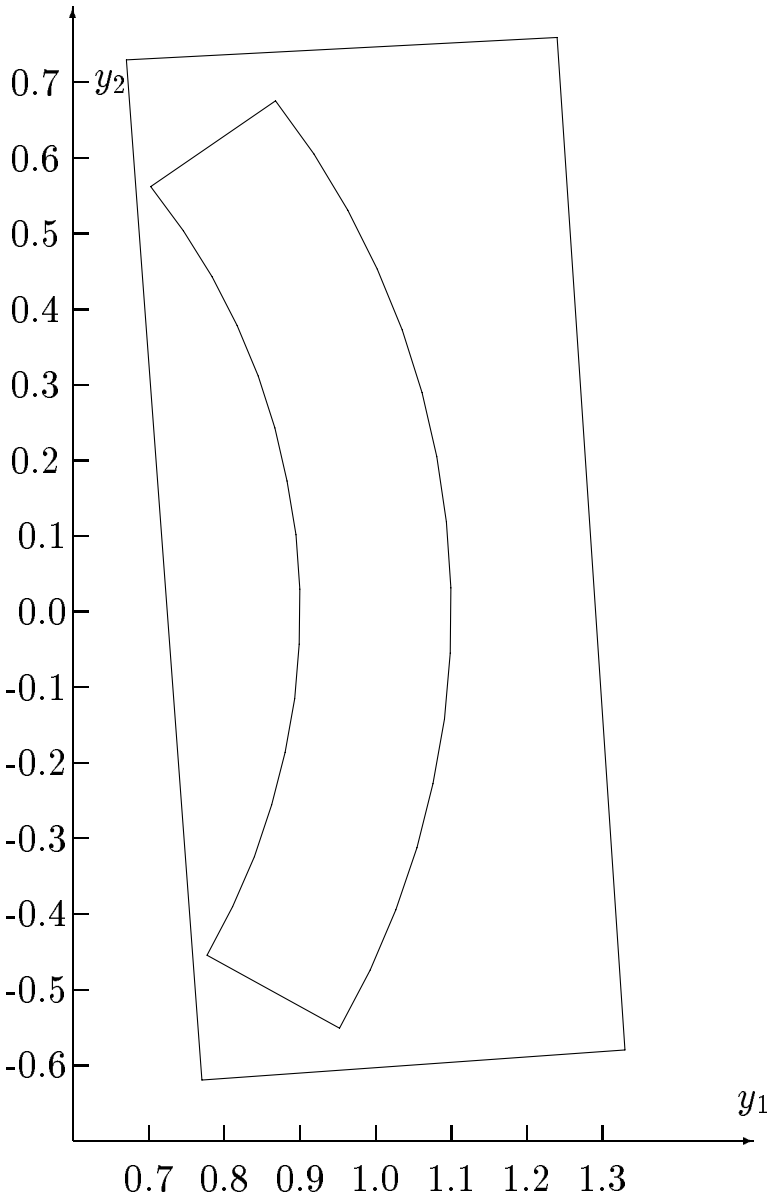

Рис. 5.7. Параллелепипедные оценки множества решений системы (5.59)

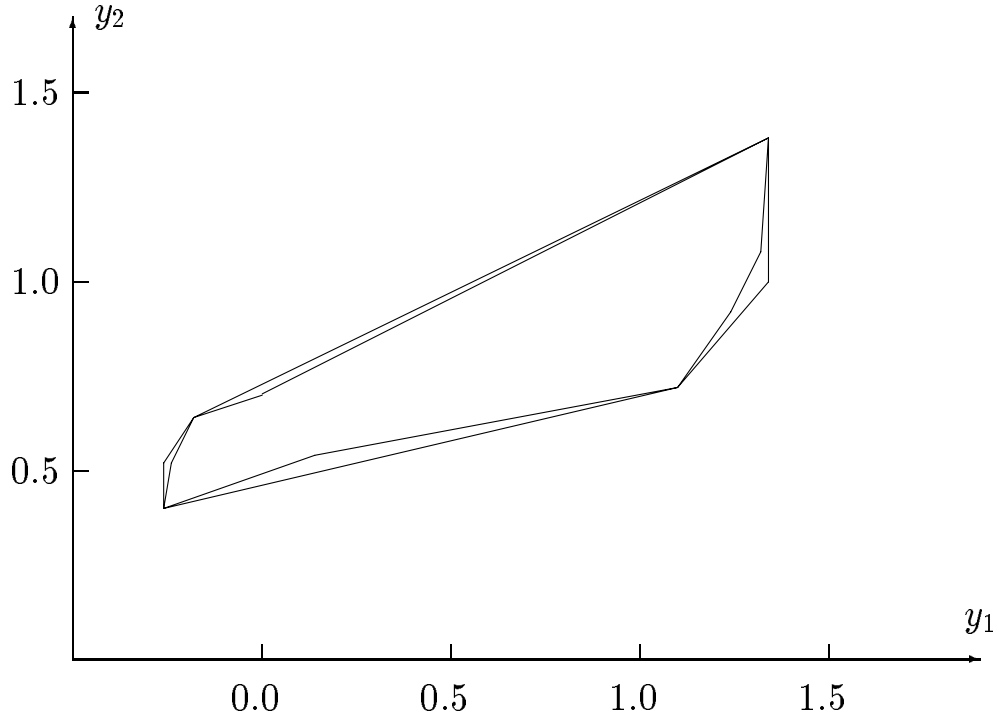

Рис. 5.8. Пересечение многогранников, $t = 0.5$

Исследуем применение областей в виде многогранников на примере следующей системы ОДУ:

$$\begin{aligned} x_1' &= x_1, \\ x_2' &= -k_1 x_1 - k_2 x_2 \end{aligned} \tag{5.66}$$

$$k_1 \in \boldsymbol{k} = [0.25, 1.0], k_2 \in \boldsymbol{k} = [0.0, 0.6].$$

Начальные условия для этой задачи — отрезок с концами (-0.5,0.5), (0.5,1.5). Введем еще один параметр $k \in \boldsymbol{k} = [0.0, 1.0]$. Тогда начальные условия можно представить в параметрическом виде:

$$x(0) = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix} + k \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}.$$

Поскольку выбор области в виде многогранника не однозначен, то для представления множества решений использованы два различных многогранника. Они выбраны таким образом, что некоторые их вершины всегда лежат на границе множества решений. Таким образом, пересечение этих многогранников не отрывается от множества решений. На рис. 5.8, 5.9 показаны пересечения многогранников в моменты $t = 0.5, 1.0$.

Как видно из рис. 5.8, 5.9, несмотря на то, что каждый многогранник имеет большие размеры, их пересечение не обладает эффектом упаковывания.

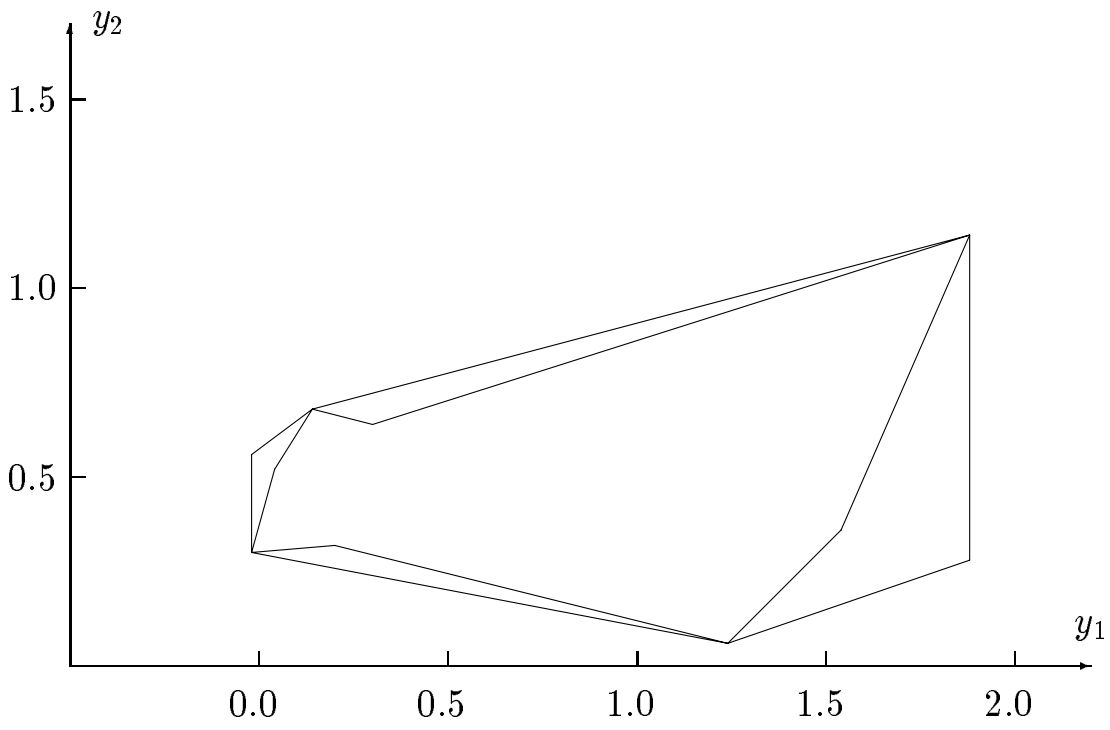

Рис. 5.9. Пересечение многогранников, 1.0

5.9. Преобразование системы ОДУ

Как было показано, более благоприятен вариант для построения оптимальных двусторонних решений когда система (5.25) представляется в виде (5.26). Добиться этого можно, в частности, заменой переменных.

Пусть

$$x = x(y), k = k(p). \tag{5.67}$$

Тогда, подставив эти соотношения в систему (5.25), получаем преобразованную систему ОДУ

$$\begin{aligned} y' &= g(y,p), \\ y(0) &= y. \end{aligned} \tag{5.68}$$

Зависимости (5.67) должны быть подобраны такими, что функция g удовлетворяет условиям (5.27), (5.28).

Следующая теорема говорит о возможном выборе преобразования (5.67) [5].

Теорема 41. *В достаточно малой окрестности неособой точки x можно выбрать систему координат $(y_1, y_2, \ldots, y_n)$ такую, что в этих*

координатах уравнение (5.67) запишется в виде

$$y_1' = 1,$$

$$y_2', y_3', \ldots, y_n' = 0.$$

Рассмотрим пример (5.59). Сделаем следующее преобразование:

$$x_1 = rcos(\phi),$$

$$x_2 = rsin(\phi).$$

Тогда в новых переменных система (5.59) запишется в виде

$$\phi' = \boldsymbol{k}, \phi(0) = \boldsymbol{c}_0,$$

$$r' = 0, r(0) = \boldsymbol{r}_0.$$

Несложно убедиться, что преобразованная система не обладает эффектом упаковывания, хотя и ее решение несколько шире точного.

5.10. Оценки областей достижимости

В качестве примера применения параллелепипедных оценок для решения систем ОДУ обратимся к одной из важных задач управления: построение оценок областей достижимости. Этой задаче посвящена монография [65].

Рассмотрим управляемую систему, описываемую векторным дифференциальным уравнением

$$x' = f(t, x, u), t \in (0, l), \tag{5.69}$$

где $u \in R^m$ — вектор управляющих функций, на который наложены в общем случае ограничения

$$u(t) \in \mathcal{U}(t, x). \tag{5.70}$$

Кроме того, мы будем предполагать, что в начальный момент выполнены ограничения

$$x(0) \in \mathcal{X}_0. \tag{5.71}$$

Если принять, что $\mathcal{U}$, $\mathcal{X}_l$ принадлежат пространству параллелепипедов, то мы приходим к стандартной постановке построения оценок множества решений для задач с неопределенностями.

Определение 14. *Множеством достижимости* $\mathcal{D}(t, \boldsymbol{x}_0)$ *управляемой системы при* $t \geq 0$ *называется совокупность концов* $x(t)$ *всех траекторий этой системы, начинающихся в точках начального множества* $\mathcal{M}$*:*

$$\mathcal{D} = \{x(t) | u(t) \in \mathbf{u}(t), x(0) \in \boldsymbol{x}_0\}.$$

Множества достижимости играют важную роль в теории управляемых систем, они используются при решении большого количества прикладных задач управления [65]:

- Оценка возможностей управления. Зная множество достижимости $\mathcal{D}$ управляемой системы, можно оценить возможности управления.
- Управляемость и интегральные воронки. Рассмотрим множество $\Delta(\boldsymbol{x}_0)$, являющееся объединением всех множеств достижимости $\mathcal{D}(t, \boldsymbol{x}_0)$ при всех $t \geq 0$. Это множество иногда называют интегральной воронкой. Если $\Delta(\boldsymbol{x}_0)$ совпадает со всем фазовым пространством, то систему называют вполне управляемой.
- Оценка возмущений и задача о накоплении возмущений.
- Задача оптимального управления.

В качестве примера рассмотрим двумерную управляемую систему

$$\begin{aligned} x_1' &= x_2, \\ x_2' &= u, u \in \mathbf{u} = [-1, 1], \\ x_1(0) &= x_2(0) = 0. \end{aligned} \tag{5.72}$$

Согласно работе [65] для задачи (5.72) точное множество достижимости в автомодельных переменных

$$\xi_1 = x_1/t^2, \quad \xi_2 = x_2/t$$

имеет вид

$$\mathcal{D} = (1 + \xi_2)^2/4 - 0.5 \leq \xi_1 \leq 0.5 - (\xi_2 - 1)^2/4,$$
$$|\xi_2| \leq 1.$$

На рис. 5.10 показано сравнение параллелепипедных и эллипсоидальных оценок множества достижимости. Отношение площадей эллипсоидов — 0.1976, параллелепипедов — 0.4. В этом случае параллелепипедные оценки несколько лучше, чем эллипсоидальные. Это обусловлено тем, что параллелепипеды были выбраны таким образом, что были жестко привязаны к множеству достижимости.

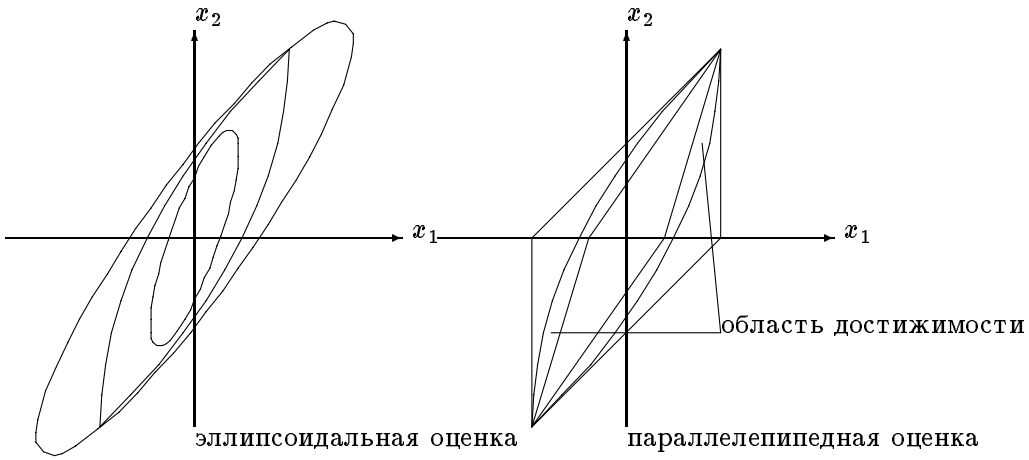

Рис. 5.10. Сравнение оценок областей достижимости

5.11. Решение жестких систем ОДУ

Обратимся к одной из задач, возникающих в химической кинетике [77]:

$$\begin{aligned} x_1' &= -20000x_1, \\ x_2' &= 20000x_1 - x_2 + x_3, \\ x_3' &= x_2 - x_3, \end{aligned} \tag{5.73}$$

$$x_1(0) = 1.0, y_2(0) = y_3(0) = 0.$$

В точке $t = 0$ решение задачи (5.73) имеет особенность. Собственные числа матрицы, составленной из коэффициентов этой задачи,— $\lambda_1 = 0$, $\lambda_2 = -2$, $\lambda_3 = -20000$. Мы будем строить двустороннее решение с учетом поведения решения в этой точке. На первом этапе мы решим численно задачу (5.73) специальным методом для жестких задач. В частности, для этой задачи, применялся неявный метод Рунге-Кутта. Задача (5.73) интегрировалась на сетке ω_h с переменным шагом $h_k = t_{k+1} - t_k$ $h_k \in [0.001, 0.1]$.

Используя полученное численное решение, мы построили в зоне погранслоя специальный нелинейный сплайн

$$r_i(t, c^k, \lambda^k) = s_i(t) + c_i^k \exp(\lambda_i^k t), \quad t \in [t_k, t_{k+1}], k = 0, 1, \ldots, N-1.$$

Здесь s_i — эрмитовы кубические сплайны. Константы c^k, λ^k находились

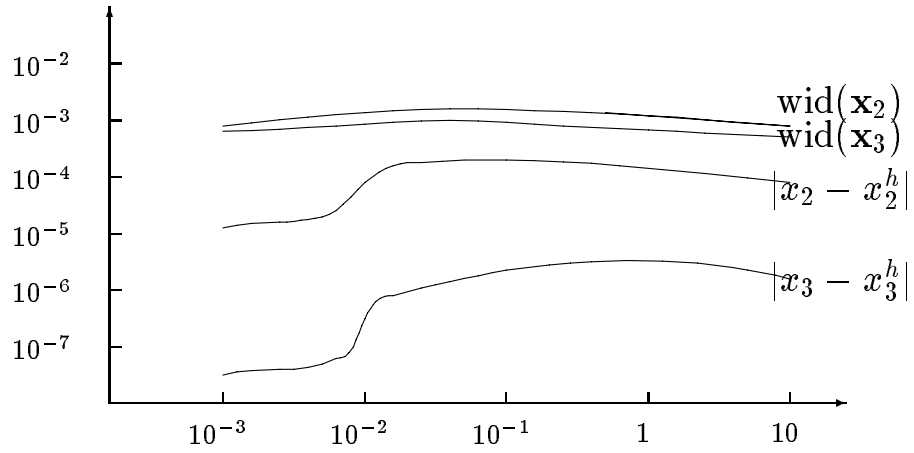

Рис. 5.11. Сравнение ширины двустороннего решения и ошибок численного решения

численно, для их нахождения минимизировался следующий функционал:

$$\Phi(c,\lambda) = \sum_{\xi}(r'(\xi,c,\lambda) - f(\xi,r))^2. \tag{5.74}$$

Задача минимизации (5.74) сводится к решению системы нелинейных уравнений

$$\frac{\partial\Phi}{\partial c} = 0,$$
$$\frac{\partial\Phi}{\partial\lambda} = 0.$$

Вне погранслоя использовался обычный эрмитовый кубический сплайн.

Результаты расчетов показаны на рис. 5.11. Как можно видеть, ширина двустороннего решения вначале возрастает, затем стабилизируется и убывает.

Глава 6

Краевые задачи для обыкновенных дифференциальных уравнений

В этом разделе мы исследуем задачу Дирихле для обыкновенного дифференциального уравнения второго порядка. Излагаемые методы слабо привязаны к типу краевого условия, и поэтому очевиден перенос результатов на корректные краевые условия Неймана и Ньютона. Получены оценки ширины двустороннего решения краевой задачи для обыкновенного дифференциального уравнения. Этот анализ выполнен также для уравнения с малым параметром при старшей производной. Несмотря на решение задачи типа погранслоя, обоснована не зависящая от ε ширина коридора двустороннего решения.

Изложены интервальный метод решения для квазилинейного уравнения и двусторонний метод для уравнений с малым параметром при старшей производной. Основной вывод, который можно сделать из результатов сравнения интервальных и апостериорных методов в этой и трех предыдущих главах, состоит в том, что существует два подхода, обладающих разными свойствами и дополняющих друг друга при решении различных задач.

6.1. Апостериорное оценивание

Рассмотрим апостериорное оценивание погрешности предварительно найденного приближенного разностного решения, восполненного эрмитовым сплайном из S_3^2 на весь отрезок $[0,1]$. Недостающие значения производной строим с помощью разностных отношений.

Обоснование качества двустороннего приближения существенно отличается от предыдущих разделов. Стандартная разностная схема дает

второй порядок точности приближенного решения u^h:

$$|u - u^h| = O(h^2) \quad \text{на} \quad \bar{\omega}_h. \tag{6.1}$$

При определении через u^h второй производной, участвующей в вычислении невязки, за счет разностного дифференцирования получаем точность порядка $O(1)$. Поэтому ожидаемая величина невязки также характеризуется величиной $O(1)$. А поскольку на основании теоремы сравнения ширина двустороннего решения является величиной одного порядка с невязкой, качество двустороннего решения может оказаться неутешительным. Вместе с тем результаты численных экспериментов вполне удовлетворительные.

Поэтому для обоснования точности следует использовать более тонкие свойства сплайнов и разностных схем. Главный член дифференциальной невязки является ограниченным функционалом от разностной невязки приближенного решения. Отсюда ширина двустороннего решения имеет такой же порядок малости, как порядок аппроксимации и сходимости.

Рассмотрим задачу Дирихле для уравнения

$$Lu \equiv -(pu')' + qu = f, \quad x \in (0,1), \tag{6.2}$$

$$u(0) = g_0, \quad u(1) = g_1. \tag{6.3}$$

Относительно функций p, q, f предположим, что они непрерывны на $[0,1]$ и

$$p(x) \geq c_1 > 0, \quad q(x) \geq 0 \quad \forall x \in (0,1). \tag{6.4}$$

Для численного решения поставленной задачи построим равномерную сетку $\overline{\omega}_h = \{x_i = ih; i = 0,1,...,N\}$ с шагом $h = 1/N$ и целым $N \geq 2$. Обозначим через $\omega_h = \{x_i = ih; \quad i = 1,...,N-1\}$ множество внутренних точек этой сетки.

Каждой внутренней точке поставим в соответствие разностное уравнение

$$-(pu^h_{\overset{\circ}{x}})_{\overset{\circ}{x}} + qu^h = f \quad \text{на} \quad \omega_h \tag{6.5}$$

с сеточной функцией $u^h(x)$, определенной на $\overline{\omega}_h$. Здесь по определению для произвольной функции v

$$v_{\overset{\circ}{x}}(x) \equiv [v(x + h/2) - v(x - h/2)]/h,$$

поэтому

$$(pv_{\overset{\circ}{x}})_{\overset{\circ}{x}}|_{x=x_i} = [p_{i+1/2}(v_{i+1} - v_i) - p_{i-1/2}(v_i - v_{i-1})]/h^2.$$

Здесь и далее, где это не вызывает двойного толкования, мы будем использовать обозначение $p_v = p(x_v)$ для любой функции p в узлах $\overline{\omega}_h$ или средних точках. Добавление к уравнениям (6.5) двух краевых условий

$$u^h(0) = g_0, \quad u^h(1) = g_1 \tag{6.6}$$

дает $N+1$ уравнений для отыскания $N+1$ значений сеточной функции в узлах $\bar{\omega}_h$.

Хорошо известно (см., например, [51, 53, 59]), что разностная задача (6.4)-(6.5) устойчива и имеет второй порядок аппроксимации. Это влечет второй порядок сходимости, что и записано в (6.1). Решение указанной задачи методом прогонки также не вызывает затруднений. В итоге мы приходим к приближенному решению $\tilde{u}^h$, близкому к u^h (с точностью до влияния ошибок округления).

Перейдем к построению двустороннего решения. Восполним полученное разностное решение $\tilde{u}^h$ на весь отрезок $[0,1]$ с помощью эрмитова сплайна $s \in S_3^2$. Для этого в каждом узле $x_i \in \bar{\omega}_h$ необходимо задать значения $s(x_i)$ и $s'(x_i)$. Первое естественно положить равным $\tilde{u}^h(x)$, а в качестве $s'(x)$ следует взять разностную производную, удовлетворяющую двум условиям. Во-первых, при любом $x \in \bar{\omega}_h$ ее узлы должны состоять только из точек $\bar{\omega}_h$. Во-вторых, для обеспечения нужного порядка точности двустороннего решения она должна, как мы увидим далее, обеспечивать третий порядок аппроксимации на решении $u \in C^4[0,1]$. Этим условиям удовлетворяет, например, разностная производная

$$\partial_x^h v(x) = \frac{1}{6h} \begin{cases} -11v(x) + 18v(x+h) - 9v(x+2h) + 2v(x+3h), & \text{при } x = 0; \\ v(x-2h) - 6v(x-h) + 3v(x) + 2v(x+h), & \text{при } x = 1-h; \\ -2v(x-3h) + 9v(x-2h) - 18v(x-h) + 11v(x), & \text{при } x = 1; \\ -2v(x-h) - 3v(x) + 6v(x+h) - v(x+2h), & \text{в остальных узлах.} \end{cases} \tag{6.7}$$

На основании раздела 2.7 равенства

$$s = \tilde{u}^h \quad \text{and} \quad s' = \partial^h \tilde{u}^h \quad \text{на} \quad \bar{\omega}_h \tag{6.8}$$

однозначно определяют сплайн $s \in S_3^2$. Он аппроксимирует решение u на всем отрезке $[0,1]$. Найдем его невязку

$$\varphi = Ls - f \quad \text{на} \quad [0,1]. \tag{6.9}$$

Теперь рассмотрим два приема построения двусторонних решений разной алгоритмической сложности и разной точности.

Алгоритм I. Найдем число α такое, что

$$|\varphi| \le \alpha \quad \text{на} \quad [0,1]. \tag{6.10}$$

Это можно сделать с помощью алгоритмов, изложенных в разд. 2.5 и [22]. Тогда двустороннее решение $\boldsymbol{u}$, содержащее u, записывается в виде

$$\boldsymbol{u} = s + \frac{\alpha}{c_1}[-1, 1], \tag{6.11}$$

где константа c_1 взята из (4). Представление (12) можно записать иначе:

$$s - \alpha/c_1 \le u \le s + \alpha/c_1. \tag{6.12}$$

Для доказательства этих неравенств используем следующий результат.

Лемма 9. *При выполнении условий (6.4) для задачи (6.2),(6.3) справедлива оценка*

$$||u||_\infty \le \frac{1}{c_1}||f||_\infty + \max\{|g_0|, |g_1|\}. \tag{6.13}$$

Доказательство. Для разностной задачи (6.5),(6.6) аналогичная оценка доказана в [53]:

$$||u^h||_\infty^h \le \frac{1}{c_1}\max_{\omega_h}|f| + \max\{|g_0|, |g_1|\}. \tag{6.14}$$

Переходя к пределу при $n \to \infty$ и используя непрерывность f и u, получаем оценку (6.13), поскольку $u^h \to u$.

Перепишем равенство (6.9) в виде

$$Lv = \varphi \quad \text{на} \quad (0,1), \quad \text{где} \quad v = s - u, \tag{6.15}$$

и добавим краевые условия $v = s - u$ с $x = 0, 1$. Применяя лемму (9) к этой краевой задаче с решением $v = s - u$, получаем оценку

$$||v||_\infty \le \frac{1}{c_1}||\varphi||_\infty + \max_{x=0,1}|s - u|,$$

откуда с помощью (6.10) вытекает неравенство (6.12).

Алгоритм II. Рассмотрим задачу

$$Lv = 1, \quad x \in (0,1), \tag{6.16}$$

$$v(0) = v(1) = 0.$$

Повторим для нее всю последовательность операций, приводящих к сплайну, аппроксимирующему решение на $[0, 1]$. А именно, методом прогонки решим разностную задачу

$$-(pv_{\overset{\circ}{x}}^h)_{\overset{\circ}{x}} + qv^h = 1 \quad \text{на} \quad \omega_h, \tag{6.17}$$

$$v^h(0) = v^h(1) = 0.$$

В итоге на сетке $\bar{\omega}_h$ мы получим приближенное решение $\tilde{v}^h$. Используя разностное отношение (6.7), построим сплайн $s_1 \in S_3^2$, удовлетворяющий равенствам

$$s_1 = \tilde{v}^h, \quad s_1' = \partial^h \tilde{v}^h \quad \text{на} \quad \bar{\omega}_h. \tag{6.18}$$

Далее предположим, что для него справедливо неравенство

$$Ls_1 \geq c_2 > 0. \tag{6.19}$$

Это требование вполне естественно при достаточно малых h, поскольку $Ls_1 = 1 + O(h^2)$, как мы покажем далее.

На основании (6.19) функция φ/Ls_1 определена на каждом интервале (x_i, x_{i+1}) и в узлах x_i имеет разрывы первого рода. Это позволяет найти уже упоминавшимися методами работы [22] и разд. (2.5) константы

$$\bar{\alpha} \geq \max_{[0,1]}(-\varphi/Ls_1), \quad \underline{\alpha} \leq \min_{[0,1]}(-\varphi/Ls_1). \tag{6.20}$$

Тогда двустороннее решение, содержащее u, записывается в виде

$$\boldsymbol{u} = s + \boldsymbol{\alpha} s_1 \tag{6.21}$$

или, что то же самое:

$$s + \underline{\alpha} s_1 \leq u \leq s + \bar{\alpha} s_1. \tag{6.22}$$

Для обоснования (6.22) используем следующую теорему сравнения, устанавливаемую на основании принципа максимума [45].

Теорема 42. *Пусть в задаче (6.2), (6.3) правая часть* $f \in L_2(0,1)$ *и* $f \geq 0$ *почти всюду на* $(0,1)$, $g_0 \leq 0, \quad g_1 \leq 0$. *Тогда* $u \geq 0$ *на [0,1].*

Рассмотрим функцию $v = s + \overline{\alpha} s_1 - u$. На основании (6.10), (6.20)

$$Lv = Ls + \bar{\alpha} Ls_1 - Lu \geq \varphi + \bar{\alpha} Ls_1 \geq 0 \quad \text{на} \quad (0,1).$$

Кроме того, из определения s и s_1 следует, что $v(0) \geq 0, v(1) \geq 0$. Поэтому на основании теоремы (42) $v \geq 0$ на $[0,1]$. Отсюда вытекает правое неравенство в (6.22). Левое неравенство получается применением этой же теоремы к функции $v = u - s - \underline{\alpha} s_1$.

Теперь исследуем довольно важный вопрос о ширине построенных коридоров в (6.13), (6.22). Условия

$$p \in C^3[0,1], \quad q, f \in C^2[0,1] \tag{6.23}$$

обеспечивают достаточную гладкость решения:

$$u \in C^4[0,1]. \tag{6.24}$$

Это влечет для разностной задачи (6.5), (6.6) второй порядок аппроксимации и сходимости. Поэтому естественно от ширины коридора требовать такого же порядка малости. Сначала мы покажем, что оба алгоритма дают именно такие коридоры без учета вычислительной погрешности при решении разностной задачи, т. е. когда $\tilde{u}^h \equiv u^h$ константы $\alpha, \underline{\alpha}_1, \bar{\alpha}_2$ в (6.10) и (6.20) определяются точно:

$$\alpha = ||\varphi||_\infty, \quad \bar{\alpha} = \max_{[0,1]} \varphi / Ls_1, \quad \underline{\alpha} = \min_{[0,1]} \varphi / Ls_1. \tag{6.25}$$

Затем мы проследим, как влияет вычислительная погрешность на увеличение ширины коридора. На основании устойчивости разностной схемы она имеет незначительную величину. Но в выражении $\varphi = Ls - f$ сплайн s дважды дифференцируется, что в принципе может увеличить в h^{-2} раз ошибку и, следовательно, ширину $\mathbf{u}$. Но фактически вклад этой погрешности в ширину коридора остается на том же уровне, на каком она входит в невязку разностной задачи.

Мы докажем оба факта в случае $p \equiv 1$. Это существенно упрощает выкладки и позволяет наглядно проследить вклад различных погрешностей. Вместе с тем общность не теряется, поскольку сохраняются все основные этапы обоснования.

Лемма 10. *Для сплайна $s \in S_3^2$ со значениями (6.9) справедливы оценки*

$$||s||_\infty \le c||\tilde{u}^h||_\infty \quad \text{и} \quad ||s''||_\infty \le c \max_{\omega_h} \left|\tilde{u}^h_{\circ\circ}\right|$$

с константами c, не зависящими от $h, \tilde{u}^h, s$.

Доказательство. Рассмотрим сплайн s на отрезке $[x_i, x_{i+1}]$, $1 \le i \le n-2$. На основании (6.9) и (6.7) сплайн является линейной комбинацией пяти значений $\tilde{u}^h$:

$$s = \sum_{j=-1}^{3} \sum_{k=0}^{3} c_{j,k} \left(\frac{x - x_i}{h}\right)^k \tilde{u}^h_{i+j} \tag{6.26}$$

с весами $c_{j,k}$, не зависящими от x, h, i. После дифференцирования получаем

$$s'' = \sum_{j=-1}^{3} \frac{d_j(x)}{h^2} \tilde{u}^h_{i+j}, \tag{6.27}$$

где линейные функции d_j равномерно ограничены:

$$|d_j| \le c_1 \quad \forall j = -1, ..., 3 \quad \text{на} \quad [x_i, x_{i+1}]. \tag{6.28}$$

Покажем, что

$$s'' = \beta_0 \tilde{u}^h_{\circ\circ}(x_i) + \beta_1 \tilde{u}^h_{\circ\circ}(x_{i+1}) + \beta_2 \tilde{u}^h_{\circ\circ}(x_{i+2}) \quad \text{на} \quad [x_i, x_{i+1}]. \tag{6.29}$$

Из сравнения (6.27) и (6.29) вытекает система уравнений

$$\beta_0 = d_{-1}, \ -2\beta_0 + \beta_1 = d_0,$$
$$\beta_0 - 2\beta_1 + \beta_2 = d_1,$$
$$\beta_1 - 2\beta_2 = d_2, \beta_2 = d_3.$$

Из первого и двух последних уравнений следует, что

$$\beta_0 = d_{-1}, \beta_1 = d_2 + 2d_3, \beta_2 = d_3. \tag{6.30}$$

Выясним, будут ли выполняться два других уравнения, которые в результате замены (6.30) приобретают вид

$$-2d_{-1} + d_2 + 2d_3 = d_0 \quad \text{и} \quad d_{-1} - 2d_2 - 3d_3 = d_1. \tag{6.31}$$

Рассмотрим сплайн s в конкретной ситуации, когда $\tilde{u}^h = \alpha x + \beta$ на всем отрезке [0, 1]. Поскольку на многочленах степени меньше $4 \partial \tilde{u}^h \equiv \partial \tilde{u}^h$, сплайн s точно воспроизводит $\tilde{u}^h$, т.е. $s = \alpha x + \beta$ на $[x_i, x_{i+1}]$. Поэтому

$$s'' \equiv 0 \quad \text{на} \quad [x_i, x_{i+1}]. \tag{6.32}$$

Выберем две пары α, β:

$$\alpha = 1/h, \beta = -x_i/h \quad \text{и} \quad \alpha = -1/h, \beta = -x_{i+1}/h.$$

Тогда на основании (6.27) мы приходим к двум функциональным равенствам

$$\frac{1}{h^2}(-d_{-1} + d_1 + 2d_2 + 3d_3) = 0 \quad \text{и} \quad \frac{1}{h^2}(-2d_{-1} - d_0 +$$
$$+d_2 + 2d_3) = 0 \quad \text{на} \quad [x_i, x_{i+1}],$$

эквивалентным (6.31). Поэтому тождество (6.29) действительно справедливо. Из него вытекает оценка

$$\max_{[x_i, x_{i+1}]} |s''| \le c_2 \max_{x = x_i, x_{i+1}, x_{i+2}} \left| \tilde{u}^h_{\circ\circ}(x) \right| \tag{6.33}$$

с константой c_2, не зависящей от $h, i, x, s, \tilde{u}^h$. Доказательство аналогичных утверждений для $i = 0, n-1, n$ несколько проще, поскольку линейная комбинация (6.27) в этих случаях содержит только четыре значения. Объединяя все четыре ситуации, приходим ко второму неравенству леммы.

Первое неравенство вытекает непосредственно из оценки правой части в (6.26) и в аналогичных ему представлениях на других отрезках. □

Теперь обоснуем второй порядок малости ширины коридора в алгоритме 1 сначала без учета вычислительных погрешностей, когда выполнено (6.25).

Теорема 43. *При выполнении условий (6.23)–(6.25) константа* α *и ширина* $\mathrm{wid}\, \boldsymbol{u}$ *равномерно на [0, 1] являются величинами* $O(h^2)$.

Доказательство. Представим разность $u - s$ в виде суммы $u - s = (u - s_2) + (s_2 - s_3) + (s_3 - s)$, где s_2, s_3— сплайны из S_3^2,, интерполирующие u в следующем смысле:

$$s_2 = u, \quad s_2' = u' \quad \text{на} \quad \bar{\omega}_h, \tag{6.34}$$

$$s_3 = u, \quad s_3' = \partial^h u \quad \text{на} \quad \bar{\omega}_h. \tag{6.35}$$

Для разности $u - s_2$ на основании [36] справедлива оценка

$$||\partial^j (u - s_2)||_\infty = O(h^{4-j}), \quad j = 0, ..., 3. \tag{6.36}$$

На основании (6.24) разностное отношение $\partial^h u$ аппроксимирует u' с третьим порядком точности, поэтому

$$|s_2' - s_3'| = O(h^3) \quad \text{на} \quad \bar{\omega}_h.$$

Из (6.34), (6.35), кроме того, следует, что $s_2 = s_3 \text{на} \bar{\omega}_h$. С учетом этих свойств, используя вид базисных функций эрмитового сплайна $s_2 - s_3$, приходим к оценке

$$||\partial^j (s_2 - s_3)||_\infty = O(h^{4-j}), \quad j = 0, ..., 3. \tag{6.37}$$

Разность $z = s_3 - s$ также является сплайном из S_3^2, построенным по значениям

$$z = u - u^h \quad \text{и} \quad z' = \partial^h (u - u^h) \quad \text{на} \quad \bar{\omega}_h. \tag{6.38}$$

Поэтому на основании леммы (10) справедливы оценки

$$||z||_\infty \le c||u - u^h||_\infty = O(h^2), \tag{6.39}$$

$$||z''||_\infty \le c \max_{\omega_h} \left|(u-u^h)_{\overset{\circ}{x}\overset{\circ}{x}}\right|. \tag{6.40}$$

Из разностного уравнения (6.5) при $p \equiv 1$ вытекает равенство

$$-(u-u^h)_{\overset{\circ}{x}\overset{\circ}{x}} + q(u-u^h) = O(h^2),\ x \in \omega_h.$$

С его помощью (6.40) преобразуется к виду

$$||z''||_\infty = O(h^2). \tag{6.41}$$

Поэтому на основании (6.9) и (6.36), (6.37)

$$\alpha = ||\varphi||_\infty = ||L(s-u)||_\infty \le ||(s-u)''||_\infty + \\ + ||q||_\infty ||s-u||_\infty = O(h^2).$$

Вследствие этого ширина двустороннего решения (6.12) или (6.11) действительно является величиной $O(h^2)$ на [0,1].

Теперь рассмотрим вопросы точности в алгоритме II. Вновь предположим отсутствие вычислительной погрешности, т. е. сплайн s строится непосредственно по значениям u^h и справедливо (6.23).

Сначала обратим внимание на то, что из доказательства теоремы 43 для сплайна s, построенного с учетом (6.18), справедлива оценка $||L(s_1-v)||_\infty = O(h^2)$ и, следовательно, $||Ls_1 - 1||_\infty = O(h^2)$. Поэтому при достаточно малых h условие (6.19) действительно выполняется. Из этой же теоремы вытекает, что $||\varphi||_\infty = O(h^2)$. Поэтому, $\overline{\alpha} = O(h^2), \underline{\alpha} = O(h^2)$, и ширина двустороннего решения в случае (6.25) оказывается равной $O(h^2)$.

В заключение раздела обратимся к вопросу о применении вместо конечно-разностного подхода метода конечных элементов. Сначала рассмотрим случай гладких коэффициентов p, q, f.

Использование простейших кусочно-линейных базисных функций на равномерной сетке не вносит никаких изменений ни в алгоритмы, ни в обоснование их точности. Единственное изменение коснется разностного уравнения (6.5). Новая система также может быть записана в трехдиагональной форме и решена методом прогонки [52, 53].

Аналогичные изменения на этапе формирования системы сеточных уравнений происходят при использовании лагранжевых конечных элементов степени 2 и выше. В таком случае порядок точности приближенного решения u^h повышается, и поэтому для выравнивания с ним порядка малости ширины коридора двустороннего решения следует использовать эрмитовы сплайны степени 5 и выше.

Наиболее интересный эффект получается при использовании эрмитовых конечных элементов для формирования сеточной задачи (6.5), например, эрмитовых сплайнов степени 3. Мы не будем останавливаться на формировании системы линейных алгебраических уравнений для вычисления u^h. В какой-то мере это изложено в разд. 6.3 для квазилинейного уравнения (см. также [51]). Матрица системы становится пятидиагональной. Главный эффект состоит в том, что получаемое решение u^h(или$\tilde{u}^h$) уже принадлежит области определения дифференциальной задачи, что достигается принадлежностью решения классу $W_2^2(0,1)$. Поэтому отпадает необходимость в гладких интерполянтах и приближенное решение само играет роль сплайна s. В итоге алгоритмы существенно упрощаются.

В случае разрывных (кусочно-гладких) коэффициентов p, q, f необходимо скорректировать алгоритмы. Во-первых, необходимо сделать сетку ω_h кусочно-равномерной, согласовав ее узлы с точками разрыва функций p, q, f и их соответствующих производных, влияющих на точность приближенного решения. Во-вторых, для $i = 0, 1..., n$ разностная производная $\partial^h v_i$ выбирается по формулам (6.7) так, чтобы все узлы лежали в одной зоне гладкости. В-третьих, дифференциальная задача в точках разрыва ξ функции p имеет дополнительные условия согласования потока [53]: $(pu')|_{\xi+0} = (pu')|_{\xi-0}$. В результате необходимо строить кусочно-гладкий интерполянт s, согласуя значения производной ps' в точках разрыва ξ. Один из способов такого согласования детально описан в разд. 7.4.

6.2. Интервальное решение

Рассмотрим другой подход к решению задачи (6.2),(6.3), отличный от обсуждаемого в разд. 6.1. Он по-прежнему использует стандартную разностную схему. Но на этот раз вместо приближенной разностной схемы мы сформулируем точную разностную схему, в которой учтем ошибку аппроксимации. Поскольку для нее известны только пределы изменения, приходим к решению системы линейных алгебраических уравнений с интервальной правой частью. Отметим, что в силу свойств разностной схемы поиск интервального решения сводится к решению двух вещественных систем линейных алгебраических уравнений с трехдиагональной матрицей.

Таким образом, алгоритм определения интервального решения, со-

держащего точное решение дифференциальной задачи, состоит из четырех этапов. Сначала находим априорные оценки решения и его нескольких производных. Затем конструируем точную разностную схему, содержащую в правой части неизвестную погрешность аппроксимации. От нее переходим к алгебраической задаче с известной интервальной правой частью. И наконец определяем обе границы интервального решения. Итак, вновь рассмотрим задачу

$$Lu \equiv -(pu')' + qu = f \quad \text{на} \quad (0,1), \tag{6.42}$$

$$u(o) = g_0, \quad u(1) = g_1 \tag{6.43}$$

с условиями (6.4). Сначала оценим ее решение и четыре производных через известные данные. При этом предположим выполненными следующие условия гладкости:

$$p \in C^3[0,1], \quad q, f \in C^2[0,1]. \tag{6.44}$$

Тогда из [53] следует, что

$$u \in C^4[0,1]. \tag{6.45}$$

Лемма 11. *Для решения задачи (6.42),(6.43) с условиями (6.1) справедливы оценки*

$$||u^{(j)}||_\infty \le V_j(p, q, f, g_0, g_1), \quad j = 0, 1, ..., 4,$$

где V_j — конкретные формулы, содержащие значения

$$||p^{(k)}||_\infty, ||q^{(k)}||_\infty, ||f^{(k)}||_\infty, g_0, g_1.$$

Доказательство. На основании леммы (36)

$$||u||_\infty \le V_0 \equiv \frac{1}{c_1}||f||_\infty + \max\{|g_0|, |g_1|\}.$$

По теореме Лагранжа на отрезке [0,1] ввиду краевых условий (6.43) найдется точка t, в которой $u'(t) = g_1 - g_0$. Используем тождество, полученное интегрированием уравнения (6.42):

$$-p(t)u'(t) + p(x)u'(x) = \int_t^x (q(\xi)u(\xi) - f(\xi))d\xi.$$

Проводя очевидные преобразования, приходим к неравенству

$$p(x)u'(x)| \le p(t)|u'(t)| + \int_0^1 |q(\xi)u(\xi) - f(\xi)|d\xi.$$

Далее получаем

$$||u'||_\infty \le V_1 \equiv \frac{1}{c_1}(||p||_\infty |g_1 - g_0| + ||q||_\infty ||u||_\infty + ||f||_\infty).$$

Выразим u'' из (6.42):

$$u'' = (-p'u' + qu - f)/p. \tag{6.46}$$

Следовательно,

$$||u''||_\infty \le V_2 \equiv \frac{1}{c_1}(||p||_\infty ||u'||_\infty + ||q||_\infty ||u||_\infty + ||f||_\infty).$$

Аналогично рекуррентным образом оцениваются третья и четвертая производные (а также $(pu')'''$ и $(pu''')'$) на основании тождеств, вытекающих из дифференцирования (6.46).

Отметим, что полученные неравенства представляют собой один из приемов оценки решения и его производных. Конкретные задачи предоставляют больший выбор, особенно если необходимо учесть локальные неоднородности.

Теперь построим точную разностную схему. Для этого подставим решение дифференциальной задачи в левую часть разностного уравнения (6.5). В результате

$$L^h u \equiv -(pu_{\overset{\circ}{x}})_{\overset{\circ}{x}} + qu = f + \xi \quad \text{на} \quad \omega_h, \tag{6.47}$$

$$u(0) = g_0, \quad u(1) = g_1,$$

где $\xi(x)$ — погрешность аппроксимации, определенная на ω_h. Оценим максимальный разброс ее значений. Для этого с учетом гладкости (6.44),(6.45) разложим в ряд Тейлора функцию u в точках $x \pm h/2$, а затем используем результаты гл. 7 книги [53]. В итоге

$$|\xi(x)| \le \frac{h^2}{192} \max_{[x-h,x+h]} |u^{IV}| \max_{[x-h,x+h]} |p| + \tag{6.48}$$

$$+\frac{h^2}{24} \max_{[x-h,x+h]} |(pu')'''| + \frac{h^2}{24} \max_{[x-h,x+h]} |(pu''')'|.$$

Из леммы (11) нам известны интервалы, в которых лежат функции, входящие в правую часть этого неравенства. Таким образом, мы указали интервальную сеточную функцию $\boldsymbol{a}(x)$, содержащую значения ошибки аппроксимации:

$$\xi(x) \in \boldsymbol{a}(x) \quad \forall x \in \omega_h, \tag{6.49}$$

причем ширина каждого из этих интервалов является величиной $O(h^2)$.

Перейдем к разностной задаче с интервальной правой частью:

$$L^h \boldsymbol{u} = f + \boldsymbol{a} \quad \text{на} \quad \omega_h, \tag{6.50}$$

$$\boldsymbol{u}(0) = g_0, \quad \boldsymbol{u}(1) = g_1. \tag{6.51}$$

Поскольку правая часть точной задачи (6.47) принадлежит покомпонентно правой части задачи (6.51), на основании определения внешнего интервального решения $\boldsymbol{u}$ справедливо включение

$$u \in \boldsymbol{u} \quad \forall x \in \bar{\omega}_h. \tag{6.52}$$

Вместо поиска внешнего интервального решения $\boldsymbol{u}$ задачи (6.51) рассмотрим решение двух задач с вещественными правыми частями

$$L^h v^h = f + \mathrm{med}\boldsymbol{a} \quad \text{на} \quad \omega_h, \tag{6.53}$$

$$v^h(0) = g_0, \quad u^h(1) = g_1,$$

$$L^h w^h = \frac{1}{2}\mathrm{wid}\boldsymbol{a} \quad \text{на} \quad \omega_h, \tag{6.54}$$

$$w^h(0) = w^h(1) = 0.$$

Тогда для точного решения (в узлах разностной сетки) справедливы оценки

$$v^h - w^h \le u \le v^h + w^h \quad \text{на} \quad \bar{\omega}_h, \tag{6.55}$$

т. е.

$$u \in \boldsymbol{u} = v^h + [-1, 1]w^h \quad \forall x \in \overline{\omega}_h.$$

Докажем два последних соотношения. Напомним, что для разностного оператора задачи (6.47) справедлив разностный принцип максимума, из которого вытекает следующая теорема сравнения [53].

Теорема 44. *Пусть в задаче (6.5),(6.6)* $f \ge 0$ *на* $\omega_h, g_0 \ge 0, g_1 \ge 0$. *Тогда*

$$u^h \ge 0 \quad \text{on} \quad \bar{\omega}_h. \tag{6.56}$$

Рассмотрим функцию $z = v^h + w^h - u$ на $\bar{\omega}_h$. При ее подстановке в оператор разностных задач (6.53),(6.54) получаем

$$L^h z = \mathrm{med}\boldsymbol{a} + \frac{1}{2}\mathrm{wid}\boldsymbol{a} - \xi \ge 0 \quad \text{on} \quad \omega_h,$$

$$z(0) = z(1) = 0.$$

Поэтому из теоремы 44 следует, что $z \ge 0$. Отсюда и из определения функции z вытекает правое неравенство в (6.55). Аналогично, полагая $z = u - v^h + w^h$, из теоремы 44 получим левое неравенство.

Теорема 45. *При выполнении условий гладкости (6.44) ширина построенного интервального решения* $\boldsymbol{u}$ *есть величина* $O(h^2)$ *равномерно на* $\bar{\omega}_h$.

Доказательство. Из построений леммы 11 и условий гладкости (6.44), (6.45) вытекает, что $||u^{(j)}||_\infty$ и $||p^{(k)}||_\infty, j = 0, ..., 4, k = 0, ..., 3$ — ограниченные величины, не зависящие от h. Поэтому на основании неравенства (6.48) имеем $\mathrm{wid}\boldsymbol{a}(x) \le ch^2$ на ω_h.

Используя оценку (6.15) применительно к задаче (6.54), приходим к выводу, что $||w^h||_\infty \le ch^2/c_1$. А поскольку $\mathrm{wid}\boldsymbol{u}(x) = w^h(x)$, то

$$\mathrm{wid}\boldsymbol{u} \le \frac{c}{c_1}h^2.$$

Несмотря на такую (асимптотическую) оценку при решении конкретных задач с умеренным h может наблюдаться чрезмерно большая ширина интервального решения. Она, как правило, вызвана неудачными априорными оценками функции u и ее производных. Один из приемов улучшения этих оценок состоит в использовании полученной информации.

6.3. Метод Ньютона для квазилинейного уравнения

Проиллюстрируем применение интервальных сплайнов. Двустороннее решение строится с использованием небольшой априорной информации о правой части квазилинейного обыкновенного дифференциального уравнения. Метод позволяет определять полосы, включающие точное решение. Рассмотрим следующую задачу:

$$-u''(x) + f(x, u, u') = 0 \quad \text{на} \quad (0,1), \tag{6.57}$$

$$u(0) = u(1) = 0. \tag{6.58}$$

Здесь $f(x, \theta, \varphi)$ — непрерывная функция по $x \in [0,1]$ и θ, φ на $(-\infty, \infty)$. Введем в пространстве $\overset{\circ}{W}{}_2^1[0,1]$ квазилинейную форму

$$B(u,v) \equiv \int_0^1 u'v'dx + \int_0^1 f(x,u,u')vdx.$$

Будем говорить, что краевая задача (6.57),(6.58) имеет обобщенное решение $u \in \overset{\circ}{W}{}_2^1[0,1]$, если

$$B(u,v) = 0 \quad \forall v \in \overset{\circ}{W}{}_2^1[0,1]. \tag{6.59}$$

Справедливы следующие теоремы [76].

Теорема 46. *Пусть существует непрерывная неотрицательная функция g на R такая, что*

$$|f(x,\theta,\varphi)| \le g(|\theta|)(1+|\varphi|^2) \tag{6.60}$$

для почти всех $x \in [0,1]$, $-\infty < \theta, \varphi < \infty$, и пусть

$$\begin{aligned}(f(x,\theta_1,\varphi_1) - f(x,\theta_2,\varphi_2))(\theta_1-\theta_2) \ge \\ \ge a(\theta_1-\theta_2)^2 - b(\varphi_1-\varphi_2)(\theta_1-\theta_2)\end{aligned} \tag{6.61}$$

для $x \in [0,1]$, $-\infty < \theta_1, \theta_2, \varphi_1, \varphi_2 < \infty$, где константы a, b удовлетворяют соотношениям

$$\max(-a,0)/\pi^2 + |b|/4 < 1.$$

Тогда задача (6.57),(6.58) имеет единственное обобщенное решение $u \in \overset{\circ}{W}{}_2^1[0,1]$.

Теорема 47. *[76] Пусть $f(x,\theta,\varphi) \in C^1([o,1]\times R\times R)$ и существуют константы A, B такие, что*

$$\frac{\partial f}{\partial \theta}(x,\theta,\varphi) \ge A, \quad 0 \le x \le 1, \quad -\infty < \theta, \varphi < \infty;$$

$$\left|\frac{\partial f}{\partial \varphi}(x,\theta,\varphi)\right| < B, \quad 0 \le x \le 1, \quad -\infty < \theta, \varphi < \infty;$$

и

$$K \equiv \max(-A,0)/\pi^2 + B/4 < 1.$$

Тогда для любого решения и задачи (6.57),(6.58) справедливы неравенства

$$||u||_\infty \le 0.5||u'||_2 \le M/2\pi(1-K) \equiv B_0, \tag{6.62}$$

$$||u'||_\infty \le (4M_1^2 + (2+4\beta^2)M^2/(\pi^2(1-K^2)))^{1/2} \equiv B_1, \tag{6.63}$$

где

$$M = \sup_{x\in[0,1]} |f(x,0,0)|, M_1 = \sup_{\substack{x\in[0,1]\\ |\theta|\le\beta_0}} |f(x,\theta,0)|.$$

Замечание 5. *Если $f \in C^m([0,1]\times R\times R)$, то, как видно из (6.57), любое решение из $\overset{\circ}{W}{}_2^1[0,1]$ будет принадлежать также $C^{m+2}[0,1]$, причем*

$$\partial^\mu u(x) = \alpha_\mu(x)u(x) + \beta_\mu(x)u'(x) + \gamma_\mu(x), \tag{6.64}$$

где $\alpha_\mu, \beta_\mu, \gamma_\mu$ — определенные многочлены от $f, \partial^v \quad f/\partial x^v, \partial^v f/\partial x^\alpha \partial\theta^\beta$, и т.д.

Рассмотрим вопрос о применении интервальных эрмитовых сплайнов к двустороннему решению задачи (6.57),(6.58). Возьмем в качестве конечномерного подпространства $\overset{\circ}{W}{}_2^1[0,1]$ пространство эрмитовых сплайнов S_3^2. Для этого введем равномерную сетку с шагом $h = 1/N, N \geq 2, \omega_h = \{x_i = ih, i = 0, ..., N\}$.

В пространстве S_3^2 введем базис для аппроксимации функции

$$w_i = \begin{cases} (|t|-1)^2(2|t|+1), & |t| \leq 1, \\ 0, & |t| > 1, \end{cases}$$

$$i = 1, ..., N-1; \quad t = (x - x_i)/h,$$

и ее производной

$$w_i = \begin{cases} t(|t|-1)^2, & |t| \leq 1, \\ 0, & |t| > 1, \end{cases}$$

$$i = N, ..., 2N; \quad t = (x - x_{i-N})/h.$$

Приближенное решение будем искать в виде

$$u_h(x) = \sum_{i=1}^{2N} c_i w_i(t).$$

Коэффициенты c_i в соответствии с методом Бубнова — Галеркина будем определять из соотношений

$$\int_0^1 (u_h' w_i' + f(x, u_h, u_h'),\ w_i) dx = 0, \quad i = 1, ..., 2N. \tag{6.65}$$

Следующая теорема дает ответ на вопрос о разрешимости сформулированной задачи и, кроме того, оценку для $u - u_h$.

Теорема 48. *В предположениях (6.60),(6.61) существует единственный элемент $u_h \in S_3^2$, удовлетворяющий (6.65). Кроме того, существует константа K_1 такая, что*

$$||u - u_h||_\infty \leq 0.5||u' - u_h'||_2 \leq K_1 \inf_{\tilde{u} \in S_3^2} ||u' - \tilde{u}'||_2.$$

Равенства (6.65) можно переписать в форме системы нелинейных алгебраических уравнений относительно вектора коэффициентов $C = \{c_1, ...c_{2N}\}$:

$$AC + F(C) = 0,$$

$$A = \{a_{ij}\}_{i,j=1}^{2N}, \quad a_{ij} = \int_0^1 w_i w_j dx,$$

$$F(C) = \{f_i(C)\}_{i=1}^{2N},$$

$$f_i(C) = \int_0^1 f\left(x, \sum_{i=1}^{2N} c_i w_i, \sum_{i=1}^{2N} c_i w_i'\right) w_j dx.$$

Данную систему будем решать интервальным методом Ньютона. Запишем его в виде

$$\mathbf{C}_{(k+1)} = \mathbf{C}_{(k)} - (\mathbf{B})^{-1}(AC_{(k)}^m - F(C_{(k)}^m)), \quad k = 0, 1, 2, ...,$$

где $\mathbf{C}_{(k)} = \left\{\left[\underline{C}_{(k)i}, \bar{C}_{(k)i}\right]\right\}_{i=1}^{2N}$ — интервальный вектор,

$$C_{(k)}^m = \left\{\left(\underline{C}_{(k)i} + \bar{C}_{(k)i}\right)/2\right\}_{i=1}^{2N}$$

— центр интервального вектора $\mathbf{C}_{(k)}$, $\mathbf{B} = \left\{\left[\underline{b}_{i,j}, \bar{b}_{ij}\right]\right\}_{i,j=1}^{2N}$ —- интервальная матрица с элементами

$$\left[\underline{b}_{ij}, \bar{b}_{ij}\right] = a_{ij} - \int_0^1 \left(\frac{\partial f}{\partial u}(x, \mathbf{U}_{(k)}, \mathbf{U}'_{(k)}) w_i, w_j + \right.$$

$$\left. + \frac{\partial f}{\partial u'}(x, \mathbf{U}_{(k)}, \mathbf{U}'_{(k)}) w_j' w_i\right) dx,$$

$$\mathbf{U}_{(k)} = \sum_{i=1}^{2N} \mathbf{C}_{(k)i} w_i(x), \quad \mathbf{U}'_{(k)} = \sum_{i=1}^{2N} \mathbf{C}_{(k)i} w_i'(x).$$

Вектор $\mathbf{D} = (\mathbf{B})^{-1}(AC_{(k)}^m - F(C_{(k)}^m))$ будем искать, решая систему линейных алгебраических уравнений с интервальными коэффициентами

$$\mathbf{BD} = AC_{(k)}^m - F(C_{(k)}^m).$$

Это можно сделать, например, методом, описанным в [3]. Начальное приближение построим, используя априорные оценки (6.62),(6.63). Положим

$$\left[\underline{C}_{(0)i}, \overline{C}_{(0)i}\right] = \begin{cases} [-B_0, B_0], & i = 1, ..., N-1, \\ [-B_1, B_1], & i = N, ..., 2N. \end{cases}$$

Итерации продолжим до тех пор, пока не выполнится условие

$$\max_{i=1,...,2N} |\bar{C}_{(k)i} - \underline{C}_{(k)i}| \le \varepsilon,$$

где ε будем выбирать согласованным с точностью получаемого приближенного решения. Используя интервальный вектор $\mathbf{C}_{(k)}$, построим интервальный эрмитовый сплайн

$$\mathbf{s}(x) = \sum_{i=1}^{2N} \mathbf{C}_{(k)i} w_i(x).$$

Заметим, что этот сплайн содержит точное решение системы (6.65), т.е. $u_h \in \mathbf{s}$. Тогда

$$u(x) \in \mathbf{s}(x) + \vec{\varphi}(x), \tag{6.66}$$

где $\vec{\varphi}(x) = K_1[-K_2, K_2]h^3||u^{(4)}||_\infty, K_2 = \sqrt{3}/216$. Действительно, согласно теореме 48

$$||u - u_h||_\infty \le 0.5||u' - u'_h||_2 \le K_1 \inf_{\tilde{u}\in S_3^2} ||u' - \tilde{u}'||_2.$$

Оценим правую часть неравенства, используя аппроксимационные свойства эрмитовых сплайнов [36]

$$\inf_{\tilde{u}\in S_3^2} ||u' - \tilde{u}'||_2 \le ||u' - s'_T||_2 \le ||u' - s'_T||_\infty \le K_2 h^3 ||u^{(4)}||_\infty, \tag{6.67}$$

где s_T — эрмитовый сплайн, интерполирующий u. Константу K_1 можно найти, следуя работе [76]:

$$K_1 = 2(1 + |B_2|/\pi^2 + B/4)/(1 - K),$$

где

$$B_2 = \max_{\substack{x\in(0,1)\\ \theta\in[-B_0,B_0]\\ \varphi\in[-B_1,B_1]}} (f'_i(x,\theta,\varphi)).$$

Для определения $||u^{(4)}||_\infty$ воспользуемся априорными оценками (6.62),(6.63) и замечанием 5:

$$||u^{(4)}||_\infty \le ||\alpha_4||_\infty ||u||_\infty + ||\beta_4||_\infty ||u'||_\infty + ||\gamma_4||_\infty. \tag{6.68}$$

Ввиду того, что априорные оценки могут быть слишком грубыми, рассмотрим вопрос их апостериорного уточнения. Из соотношений (6.66), (6.68) вытекают неравенства

$$||u||_\infty \le \max_{s\in\mathbf{s}} ||s||_\infty + K_1 K_2 h^3 ||u^{(4)}||_\infty,$$

$$||u'||_\infty = ||s'_T - s'_T + u'||_\infty \le ||s'_T||_\infty + ||u' - s'_T||_\infty \le$$
$$\le \max_{s\in\mathbf{s}} ||s'||_\infty + K_2 h^3 ||u^{(4)}||_\infty.$$

Привлекая эти оценки в (6.68), можно существенно уменьшить ширину коридора в (6.66).

Проиллюстрируем этот прием на следующей задаче:

$$-u'' + u^3 + 1/(1 + (u')^2) +$$

$$+\sin(\pi x)(\pi^2+\sin^2(\pi x))(1+\pi^2\cos(\pi x))+1)/$$
$$/(1+\pi^2\cos(\pi x))=0, \tag{6.69}$$
$$u(0)=u(1)=0. \tag{6.70}$$

Точное решение этой задачи равно $\sin(\pi x)$. Для численного решения использовалась равномерная сетка с шагом $h=0.25$. Начальное приближение для метода Ньютона выбрано следующим: $[\underline{C}_{(0)i},\bar{C}_{(0)i}]=[-2,2], \quad i=1,...,7$. Количество итераций для метода Ньютона равно четырем, при этом достигнута точность

$$\max_{i=1,...,7}(\bar{C}_{(4)i}-\underline{C}_{(4)i})<0.0001.$$

При использовании полученных значений $\mathbf{C}_{(4)}$ был построен эрмитовый сплайн $\mathbf{s}$, аппроксимирующий решение задачи (6.69), (6.70). Для него была найдена интервальная функция $\vec{\varphi}(x)$, содержащая ошибки аппроксимации эрмитового сплайна

$$u(x)\in\mathbf{s}(x)+\vec{\varphi}(x), \quad x\in[0,1]. \tag{6.71}$$

Для построения φ использовались аппроксимационные оценки (6.62), (6.63) и формула (6.66). Результаты счета приведены в табл. 1.

Таблица 1

x	Нижняя граница	Точное решение	Верхняя граница	Уточненная граница	
				нижняя	верхняя
0.1	0.2603	0.3090	0.3575	0.3031	0.3147
0.2	0.5391	0.5877	0.6363	0.5819	0.5935
0.3	0.7605	0.8090	0.8577	0.7968	0.8214
0.4	0.9021	0.9510	0.9993	0.9384	0.9630
0.5	0.9518	1.0000	1.0490	0.9881	1.0127
0.6	0.9021	0.9510	0.9993	0.9384	0.9630
0.7	0.7605	0.8090	0.8577	0.7968	0.8214
0.8	0.5391	0.5877	0.6363	0.5819	0.5935
0.9	0.2603	0.3090	0.3575	0.3031	0.3147

После построения интервального решения (6.71), содержащего точное решение задачи (6.69), (6.70), была уточнена $||u||_2, ||u'||_2$, а затем с использованием (6.71) построена уточненная оценка $||u^{(4)}||_2$ и, соответственно, φ. Уточненное интервальное решение задачи (6.69), (6.70) также представлено в табл. 1. Как видно, несмотря на грубую сетку, интервальное решение достаточно хорошо аппроксимирует решение задачи (6.69), (6.70).

6.4. Краевая задача для уравнения с малым параметром при старшей производной

Вновь обратимся к краевой задаче для уравнения второго порядка. Наличие малого параметра при старшей производной приводит к решению типа погранслоя, и стандартная разностная схема дает плохие результаты [35]. Поэтому необходимо использовать одну из специальных разностных схем [38, 47], в качестве которой мы возьмем схему экспоненциальной подгонки [35]. Отличие такого метода от изложенного в разд. 6.1 состоит в том, что разностное решение исходной задачи интерполируется кубическими эрмитовыми сплайнами и специальными функциями, учитывающими погранслой. Затем по невязке построенного интерполянта находится уклонение приближенного решения от точного на основе модификации алгоритма из разд. 6.1.

Рассмотрим задачу

$$Lu \equiv -\varepsilon^2 u'' + qu = f \quad \text{на} \quad (0,1), \tag{6.72}$$

$$u(0) = u(1) = 0. \tag{6.73}$$

Здесь $\varepsilon \ll 1$ — положительный параметр. Предположим, что

$$q(x) \geq \alpha > 0, \quad f \in C^2[0,1]. \tag{6.74}$$

Для нахождения двусторонних оценок сначала решим численно задачу (6.72),(6.73) разностным методом. Для этого построим разностную схему экспоненциальной подгонки [35]

$$L^h u^h \equiv \varepsilon^2 \sigma_v u^h_{\overset{\circ}{x}\overset{\circ}{x}} + qu^h = f \text{ на } \omega_h, \tag{6.75}$$

$$u^h(0) = u^h(1) = 0, \tag{6.76}$$

где

$$\sigma_\nu = \frac{\nu^2 q(x)}{4sh^2(\nu\sqrt{q(x)}/2)}, \quad \nu = h/\varepsilon. \tag{6.77}$$

В работе [35] доказана следующая

Теорема 49. *Пусть* $a'(0) = a'(1) = 0$. *Тогда решение и задачи (6.72),(6.73) и решение* u^h *задачи (6.75), (6.76) удовлетворяют неравенству*

$$||u - u^h||_\infty \leq ch^2,$$

где c *не зависит от* i, h *и* ε.

Несмотря на такую точность разностного решения u^h в узлах сетки, интерполяция кубическими сплайнами дает плохие результаты и тем хуже, чем ближе к концам отрезка. Поэтому необходимо использовать дополнительную информацию о решении. С такой целью введем функции типа погранслоя

$$\Pi_0(x) = \alpha_0 \exp(-x\sqrt{q(0)}/\varepsilon),$$

$$\Pi_1(x) = \alpha_1 \exp(-(1-x)\sqrt{q(1)}/\varepsilon),$$

где $\alpha_0 = f(0)/q(0), \quad \alpha_1 = f(1)/q(1)$. Пусть функция z является решением задачи

$$Lz = f - L(\Pi_0 + \Pi_1) \quad \text{на} \quad (0,1),$$

$$z(0) = -\alpha_0 - \alpha_1\Pi_1(0),$$

$$z(1) = -\alpha_0\Pi_0(1) - \alpha_1,$$

тогда $u = z + \Pi_0 + \Pi_1$. Относительно z справедлива следующая [35]

Лемма 12. *Предположим, что* $a'(0) = a'(1) = 0$. *Тогда*

$$|z^{(j)}| \le c(1 + \varepsilon^{2-j}), \quad j = 0, ..., 4.$$

Положим $z^h = u^h - \Pi_0 - \Pi_1$. Тогда на основании [35] существует константа c, не зависящая от h, ε и такая, что

$$||z^h - z||_\infty \le ch^2. \tag{6.78}$$

Решим задачу (6.75),(6.76) методом прогонки. В результате приближенное решение $\tilde{u}^h$, отличается от u^h вкладом ошибок округления. Вместо z^h получаем $\tilde{z}^h = \tilde{u}^h - \Pi_0 - \Pi_1$. Восполним $\tilde{z}^h$ на весь отрезок $[0,1]$ эрмитовым сплайном $s \in S_3^2$. На основании разд. 2.7 он целиком определяется значениями

$$s = \tilde{z}^h \quad \text{и} \quad s' = \partial^h \tilde{z}^h \quad \text{на} \quad \bar{\omega}_h. \tag{6.79}$$

Разностная производная $\partial^h v$ введена в разд. 6.1. Возьмем функцию

$$r = s + \Pi_0 + \Pi_1 \tag{6.80}$$

в качестве "осевого" приближенного решения. С его помощью построим границы двустороннего решения

$$\underline{r} = r + \underline{\beta}\varphi, \quad \bar{r} = r + \bar{\beta}\varphi, \tag{6.81}$$

где φ имеет характер функции погранслоя:

$$\varphi(x) = 1 - (\exp(-\sqrt{\alpha}x/\varepsilon) + \exp(\sqrt{\alpha}(x-1)/\varepsilon))/$$
$$/(1 + \exp(-\sqrt{\alpha}/\varepsilon)),$$
$$\alpha = \min_{[0,1]} q.$$

$\underline{\beta}, \bar{\beta}$ выберем следующим образом:

$$\bar{\beta} = \max_{[0,1]}(f - Lr)/L\varphi, \quad \underline{\beta} = \min_{[0,1]}(f - Lr)/L\varphi. \tag{6.82}$$

Обратим внимание на то, что

$$L\varphi \geq \alpha \quad \text{на} \quad [0,1], \tag{6.83}$$

поэтому числа $\bar{\beta}, \underline{\beta}$ определены и конечны.

Докажем, что точное решение содержится в интервале, определяемом функциями $\underline{r}, \bar{r}$:

$$r + \underline{\beta}\varphi \leq u \leq r + \bar{\beta}\varphi. \tag{6.84}$$

Действительно

$$L\bar{r} = L(s + \bar{\beta}\varphi) \geq f \quad \text{на} \quad [0,1],$$
$$L\underline{r} = L(s + \underline{\beta}\varphi) \leq f \quad \text{на} \quad [0,1].$$

Кроме того,

$$s(0) + \bar{\beta}\varphi(0) = u(0) = 0, \quad s(1) + \underline{\beta}\varphi(1) = u(1) = 0.$$

Поэтому на основании теоремы 42

$$\underline{r} \leq u \leq \bar{r} \quad \text{на} \quad [0,1]. \tag{6.85}$$

Таким образом, для решения u построены верхняя и нижняя границы.

Оценим ширину $\rho = \bar{r} - \underline{r}$ этого двустороннего решения. Сначала рассмотрим ситуацию без учета вычислительных погрешностей, когда

$$\tilde{u}^h \equiv u^h \tag{6.86}$$

и константы вычислены по формулам (6.82) точно.

Теорема 50. *При выполнении условий (6.82), (6.86) константы* $\underline{\beta}, \bar{\beta}$ *и ширина* ρ *равномерно на* $[0,1]$ *являются величинами* $O(h^2)$.

Доказательство. Представим разность $z - s$ в виде суммы $z - s = (u - s_2) + (s_2 - s_3) + (s_3 - s)$, где s_2, s_3, — сплайны из S_3^2, интерполирующие z в следующем смысле:

$$s_2 = z, s_2' = z' \quad \text{на} \quad \bar{\omega}_h, \tag{6.87}$$

$$s_3 = z, s_3' = \partial^h z \quad \text{на} \quad \bar{\omega}_h. \tag{6.88}$$

Для разности $r - s_2$ на основании [35, 36] и теоремы 49 справедливы оценки

$$||z - s_2||_\infty \le ch^2 ||z''||_\infty = O(h^2),$$
$$||(z - s_2)''||_\infty \le ch^2 ||z^{(4)}||_\infty = O(h^2 \varepsilon^{-2}).$$

Отсюда

$$||L(z - s_2)||_\infty = O(h^2). \tag{6.89}$$

На основании (6.24) разностное отношение $\partial^h z$ аппроксимирует z' с третьим порядком точности, поэтому равномерно на $\bar{\omega}_h$

$$|(s_2 - s_3)'| \le ch^3 ||z^{(4)}||_\infty = O(h^3 \varepsilon^{-2}) \tag{6.90}$$

и

$$|(s_2 - s_3)'| \le ch ||z''||_\infty = O(h). \tag{6.91}$$

Кроме того, $s_2 = s_3$ на $\bar{\omega}_h$. Следовательно, эрмитов сплайн $s_2 - s_3$ удовлетворяет следующим оценкам: на основании (6.91) $||s_2 - s_3||_\infty = O(h^2)$, а на основании (6.90) $||(s_2 - s_3)''||_\infty = O(h^2 \varepsilon^{-2})$. В итоге

$$||L(s_2 - s_3)||_\infty = O(h^2). \tag{6.92}$$

Разность $y = s_3 - s$ также является сплайном из S_3^2, построенным по значениям

$$y = z - z^h \quad \text{и} \quad y' = \partial^h (z - z^h) \quad \text{на} \quad \bar{\omega}_h. \tag{6.93}$$

Поэтому на основании леммы 10 справедливы оценки

$$||y||_\infty \le c||z - z^h||_\infty = O(h^2) \tag{6.94}$$

и

$$||y''||_\infty \le c \max_{\omega_h} |(z - z^h)_{\overset{\circ}{x}\overset{\circ}{x}}|. \tag{6.95}$$

Из разностного уравнения (6.75) и леммы 12, вытекает оценка

$$-\varepsilon^2 \sigma_\nu y_{\overset{\circ}{x}\overset{\circ}{x}} + qy = O(h^2) \ \text{ на } \omega_h. \tag{6.96}$$

Здесь использована также равномерная ограниченность функции σ_ν при $\nu \to 0$. С помощью (6.94), (6.96) преобразуем (6.95) к виду $||y''||_\infty = O(h^2\varepsilon^{-2})$. Поэтому с учетом (6.94) $||Ly||_\infty = O(h^2)$. Объединяя это равенство с (6.89), (6.92), с учетом (6.83) получаем, что

$$|\underline{\beta}| \le ||f - Lr||_\infty/\alpha \le (||L(z - s_2)||_\infty +$$
$$+||L(s_2 - s_3)||_\infty + ||Ly||_\infty)/\alpha = O(h^2).$$

Аналогично $|\bar{\beta}| = O(h^2)$ и в силу ограниченности функции φ

$$||\rho||_\infty = (\bar{\beta} - \underline{\beta})||\varphi||_\infty = O(h^2).$$

Теперь рассмотрим влияние вычислительных погрешностей в следующей форме. Пусть вместо u^h вычислено приближенное решение $\tilde{u}^h$, удовлетворяющее разностной задаче

$$L^h\tilde{u}^h = f + \delta \quad \text{на} \quad \omega^h, \tag{6.97}$$
$$\tilde{u}^h(0) = \tilde{u}^h(1) = 0.$$

Максимальный уровень невязки обозначим через Δ:

$$\Delta = \max_{\omega_h} |\delta|.$$

Тогда для разности $w^h = \tilde{u}^h - u^h$ справедливы равенства

$$L^h w^h = f + \delta \quad \text{на} \quad \omega^h, \tag{6.98}$$
$$w^h(0) = w^h(1) = 0.$$

Наряду с ними рассмотрим равенства для постоянной функции $w = \Delta/\alpha$ на $[0,1]$

$$L^h w = \Delta q/\alpha \quad \text{на} \quad \omega_h, \quad w(0) = w(1) = \Delta/\alpha.$$

На основании разностной теоремы сравнения для этих двух задач вытекает оценка

$$|w^h| \le w \quad \text{на} \quad \bar{\omega}_h.$$

Поэтому

$$||z^h - \tilde{z}^h||_\infty = ||u^h - \tilde{u}^h||_\infty \le \Delta/\alpha.$$

По значениям разности $z^h - \tilde{z}^h$ построим сплайн $s_4 \in S_3^2$ такой, что $s_4 = w^h$ и $\partial s_4 = \partial^h w^h$ на $\bar{\omega}_h$. Тогда на основании леммы (10)

$$||s_4||_\infty \le c\Delta \quad \text{и} \quad ||s_4''||_\infty \le c \max_{\omega_h} |w^h_{\circ\circ xx}|.$$

Поэтому с учетом (6.98) и теоремы 50 получаем, что

$$||Lr - f||_\infty = ||L(s - z)||_\infty \leq ch^2 + c\Delta.$$

Таким образом, вычислительная погрешность приводит к такому же вкладу в ширину двустороннего решения, как и в невязку приближенного решения. Практические алгоритмы нахождения величин $\underline{\beta}, \bar{\beta}$ дают, соответственно, заниженную и завышенную оценки. Влияние такого рода погрешности на ширину двустороннего решения прослеживается очевидно.

В качестве численного примера была решена следующая задача:

$$Lu = -\varepsilon^2 u'' + u = \cos^2(\pi x) + 2\varepsilon\pi^2 \cos(2\pi x), \tag{6.99}$$

$$u(0) = u(1) = 0.$$

Непосредственной подстановкой легко убедиться, что точное решение имеет вид

$$u(x) = -[\exp(-(1-x)/\varepsilon) + \exp(-x/\varepsilon)]/(1+$$

$$+ \exp(-1/\varepsilon)) + \cos^2(\pi x).$$

Таблица 2

x	Нижняя граница	Точное решение	Верхняя граница	Ширина	Ошибка чис. реш.
0.1	0.90299	0.90446	0.90910	6.11	-1.58
0.2	0.65206	0.65450	0.65817	6.11	-0.61
0.3	0.34182	0.34549	0.34794	6.11	0.61
0.4	0.09084	0.09549	0.09695	6.11	1.59
0.5	-0.00109	0.0	0.00502	6.11	1.96

Задача (6.99) была решена с помощью разностной схемы (6.75), (6.76) с параметрами $\varepsilon = 0.01, h = 0.1$. Результаты счета приведены в табл. 2, в которой аргументы $x > 0.5$ опущены в силу симметрии. Вне пограничного слоя ширина коридора стабилизируется. Это характерная черта построенного алгоритма. Для более точного двустороннего приближения можно аналогичным образом модифицировать алгоритм II из разд. (6.1), но, как уже обсуждалось, повышение точности в нем сопровождается увеличением сложности.

6.5. Вопросы и упражнения

1. Сделать две итерации интервального метода Пикара для задачи Коши

$$y' = y^2,$$
$$y(0) = 1.$$

Начальное приближение взять $\boldsymbol{y}_0 = [1, 2]$. На каком интервале $(0, l)$ можно построить решение?

2. Написать интервальный метод второго порядка для задачи Коши

$$y' = y,$$
$$y(0) \in [1, 2].$$

Сделать два шага с $h = 0.1$.

3. Какая система дифференциальных неравенств мажорирует решение задачи Коши для системы ОДУ:

$$\begin{aligned} y_1' &= y_1 y_2 + [0, 1], \\ y_2' &= [1, 2] y_1 + [1, 1]? \end{aligned}$$

4. Оцените погрешность разностного решения краевой задачи на сетке $\omega_h = \{0, \pi/4, \pi/2\}$:

$$-u'' + u = 2\sin(x), \quad x \in (0, \pi/2),$$
$$u(0) = 0, \quad u(\pi/2) = 1.$$

Глава 7

Краевые задачи для уравнений в частных производных

7.1. Двусторонние методы

В данном разделе изучаются методы построения двусторонних решений для дифференциальных уравнений с нелинейной правой частью. Для этого предварительно находится приближенное решение исходной задачи каким-либо численным методом. Например, методом конечных элементов (МКЭ) с использованием кусочно-линейных базисных функций или с помощью разностных схем. Полученное приближенное решение сглаживается с помощью сплайнов или конечных элементов типа Клафа-Точера. Далее по величине невязки с помощью сравнения с решениями некоторых вспомогательных задач строится коридор, в котором лежит разница между построенным сглаженным решением и точным.

Рассмотрим задачу Дирихле для уравнения эллиптического типа

$$\begin{aligned} Lu &= f(x,u),\ x \in \Omega, \\ u(x) &= 0,\ x \in \partial\overline{\Omega}, \end{aligned} \tag{7.1}$$

где Ω — открытый многоугольник из R^2 с границей $\partial\overline{\Omega}$,

$$Lu = -\sum_{i=1}^{2} \frac{\partial}{\partial x_i}(a_i \frac{\partial}{\partial x_i} u) + qu.$$

Правая часть f — непрерывная функция на $\overline{\Omega} \times (-\infty, \infty)$, для которой выполняется условие

$$\partial f(x,\eta)/\partial u \ \le \ q(x) \le 0 \tag{7.2}$$

и, кроме того, существует константа K такая, что

$$|f(x,\eta)| \le K(1 + |\eta|), \quad \forall x \in \Omega. \tag{7.3}$$

Для численного решения задачи (7.1) перейдем к обобщенной формулировке. Введем билинейную форму

$$\mathcal{L}(u,v) = \int_\Omega \sum_{i=1}^{2} \partial_i u \partial_i v d\Omega \quad \forall u, v \in \overset{o}{W}{}_2^1(\Omega).$$

Тогда обобщенным решением задачи (7.1) из $\overset{o}{W}{}_2^1(\Omega)$ называется функция u, удовлетворяющая тождеству

$$\mathcal{L}(u,v) = (f,v) \quad \forall v \in \overset{o}{W}{}_2^1(\Omega),$$

где $(\cdot,\cdot)$ — скалярное произведение в L_2

$$(u,v) = \int_\Omega uvd\Omega.$$

На основании работы [76] задача (7.1) имеет единственное обобщенное решение в $\overset{o}{W}{}_2^1(\Omega)$.

Теорема 51. *Пусть выполнены условия (7.2), (7.3), тогда задача (7.1) имеет единственное обобщенное решение* $u \in \overset{o}{W}{}_2^1(\Omega)$.

□

Следуя методу конечных элементов, приближенное решение u^h задачи (7.1) мы определим как функцию из S_1^1, удовлетворяющую уравнению

$$L(u^h, v^h) = (f, v^h), \forall v^h \in S_1^1, \tag{7.4}$$
$$u_h = 0, \quad на\ \partial\overline{\Omega}.$$

Используя полученное численное решение, мы можем построить специальную функцию $s \in S_2^3(\overline{\Omega})$ и вычислить следующую невязку:

$$\phi(x,s) = Ls - f(x,s),\ x \in \Omega. \tag{7.5}$$

Используя МКЭ, найдем приближенное решение следующей задачи:

$$L_1 u_1 = 1,\ x \in\ \Omega, \tag{7.6}$$
$$u_1(x) = 0,\ x \in\ \partial\overline{\Omega}, \tag{7.7}$$

где $L_1 = Lu - qu$.

Построим аппроксимацию более гладкой функцией $s_1 \in S_2^3(\overline{\Omega})$ аналогично тому, как это было сделано выше.

Тогда двустороннее решение будем искать в виде

$$\boldsymbol{u} = S + [\underline{\alpha}, \overline{\alpha}]S_1 + [\underline{\beta}, \overline{\beta}],$$

где

$$\overline{\alpha} = \max_{\overline{\Omega}}(\phi/L_1 s_1, 0), \underline{\alpha} = \min_{\overline{\Omega}}(\phi/L_1 s_1, 0),$$

$$\overline{\beta} = \max_{\partial\Omega}(-\overline{\alpha}s_1 - s, 0), \underline{\beta} = \min_{\partial\Omega}(-\underline{\alpha}s_1 - s, 0).$$

Для доказательства принадлежности точного решения полученному двустороннему исследуем вспомогательную задачу:

$$L_1 u = f, \ x \in \ \Omega, \tag{7.8}$$

$$u(x) = \gamma, \ x \in \ \partial\overline{\Omega}, \tag{7.9}$$

где $f \in L_2(\Omega), \gamma \in W_2^2(\Omega)$. Далее нам понадобится теорема [21].

Теорема 52. *Пусть $u \in W_2^2(\Omega)$ — решение задачи (7.8) и почти всюду на Ω выполняются неравенства*

$$\gamma, \ f \geq 0,$$

тогда

$$u \geq 0 \quad на \Omega. \square$$

Рассмотрим две краевые задачи:

$$Lu_1 - qu_1 = f_1, \ x \in \ \Omega,$$

$$u_1 = 0, \ x \in \ \partial\overline{\Omega}.$$

$$Lu_2 = f_2, \ x \in \ \Omega,$$

$$u_2 = \gamma, \ x \in \ \partial\overline{\Omega}.$$

Из теоремы 52 непосредственно вытекает следующее

Замечание 6. *Пусть $f_1, f_2 \in L_2(\Omega)$, $\gamma \in W_2^2(\Omega)$ и почти всюду на Ω выполнены неравенства*

$$f_2 \geq f_1 \geq 0, \gamma \geq 0.$$

Тогда

$$u_2 \geq u_1, x \in \ \overline{\Omega}. \square$$

Пусть

$$Ls \equiv f(x,s) - \phi. \tag{7.10}$$

Вычитая (7.10) из (7.1), получаем

$$L(u-s) = \phi + f(x,u) - f(x,s).$$

Применяя теорему о среднем, приходим к уравнению

$$L(u-s) = f'_u(x,\xi)(u-s) + \phi,$$

где ξ — функция, принимающая значения, заключенные в интервале с концами u, s. Заметим, что в силу выбора $\overline{\alpha}, \overline{\beta}, \underline{\alpha}, \underline{\beta}$ справедливы неравенства

$$L_1(\overline{\alpha}s_1 + \overline{\beta}) \geq \phi \quad \forall x \in \Omega,$$
$$\overline{\alpha}s_1 + \overline{\beta} \geq -s \quad \forall x \in \partial\Omega,$$
$$L_1(\underline{\alpha}s_1 + \underline{\beta}) \leq \phi \quad \forall x \in \Omega,$$
$$\underline{\alpha}s_1 + \underline{\beta} \leq -s \quad \forall x \in \partial\Omega.$$

Таким образом в силу (52) и (6) доказана следующая теорема [34].

Теорема 53. *Для построенных сплайнов s, s_1 справедливы неравенства*

$$s + \underline{\alpha}s_1 + \underline{\beta} \leq u \leq s + \overline{\alpha}s_1 + \overline{\beta}.\square$$

Оценим ширину полученного двустороннего решения. В целях упрощения выкладок рассмотрим задачу для уравнения Пуассона, заданного в области $\Omega = (0,1) \times (0,1)$

$$-\Delta u = f \ \textit{в} \ \Omega, \tag{7.11}$$
$$u = 0 \ \textit{на} \ \partial\Omega,$$

решение которой удовлетворяет условию:

$$u \in W_\infty^4(\Omega). \tag{7.12}$$

Для численного решения этой задачи построим в $\overline{\Omega}$ квадратную разностную сетку. Для этого проведем параллельные прямые

$$x_{1,i} = ih, x_{2,j} = jh, i, j = 0, 1, \ldots, n,$$

где n — целое, $h = 1/n$ — шаг сетки. Для аппроксимации уравнения (7.11) используем стандартную пятиточечную разностную схему,

которая в данном случае практически совпадает с разностной схемой, построенной по МКЭ.

Численное решение u^h интерполируем эрмитовыми сплайнами s [36], для этого в каждом узле сетки необходимо задавать значения $s, \partial_1 s, \partial_2 s, \partial_{12} s$. Для их определения можно воспользоваться разностными производными.

Теорема 54. *Пусть* $u \in W_\infty^4(\Omega)$, *тогда*

$$\mathrm{wid}(\boldsymbol{u}(x)) \le Kh^2,$$

где h — *характерный размер сетки,* K — *константа, независящая от* h. □

Напомним, что ошибка в приближенном решении мажорируется приближенным решением задачи (7.8). В случае, если она распределена неравномерно, например имеет в некоторой подобласти значительно большие значения, это приведет к тому, что двустороннее решение будет иметь на всей области ширину, характеризующую ошибку лишь на небольшой подобласти. Такая ситуация может возникнуть в случае уравнения с малым параметром при старшей производной, точечных особенностей в коэффициентах, в угловых точках областей. Поэтому для уменьшения ширины двустороннего решения можно использовать решение еще двух дополнительных задач для лучшего мажорирования ошибки.

Решим две дополнительные задачи:

$$Lu_i = f_i,\ x \in \Omega, \tag{7.13}$$

$$u_i = \gamma_i,\ x \in \partial\overline{\Omega}, i = 2, 3, \tag{7.14}$$

где

$$f_2(x) = \max_{y \in \Omega_x} \phi(y), f_3(x) = \min_{y \in \Omega_x} \phi(y),$$

$$\gamma_2(x) = \max_{y \in \partial\Omega_x} (-s(y)), \gamma_3(x) = \min_{y \in \partial\Omega_x} (-s(y)),$$

$$\Omega_x = [x_1 - h, x_1 + h] \times [x_2 - h, x_2 + h] \cap \Omega,$$

$$\partial\Omega_x = [x_1 - h, x_1 + h] \times [x_2 - h, x_2 + h] \cap \partial\Omega.$$

Тогда справедливы неравенства:

$$s_3 + \underline{\alpha} s_1 + \underline{\beta} \le u - s \le s_2 + \overline{\alpha} s_1 + \overline{\beta},$$

здесь $\overline{\alpha}, \overline{\beta}, \underline{\alpha}, \underline{\beta}$ определяются так:

$$\overline{\alpha} = \max_{\overline{\Omega}}(f - L(s+s_2)/L_1 s_1, 0), \underline{\alpha} = \min_{\overline{\Omega}}(f - L(s+s_3)/L_1 s_1, 0),$$

$$\overline{\beta} = \max_{\partial\Omega}(-\overline{\alpha} s_1 - s - s_2, 0), \underline{\beta} = \min_{\partial\Omega}(-\underline{\alpha} s_1 - s - s_3, 0).$$

Полученное двустороннее решение будет более точно отражать реальное поведение ошибки. В качестве численного примера решена следующая задача:

$$\Delta u = \exp(u) - \exp(\sin(\pi x_1)\sin(\pi x_2)) - \pi^2 \sin(\pi x_1)\sin(\pi x_2) \text{ в } \Omega,$$

$$u = 0 \text{ на } \partial\Omega,$$

где $\Omega = (0,1) \times (0,1)$, точное решение:

$$u = \sin(\pi x_1)\sin(\pi x_2).$$

Решение искали с использованием пакета программ метода конечных элементов МОК-1 [2]. После этого строили восполнение элементами Клафа-Точера. Триангуляции области строили равномерные, шаги квадратных сеток выбирались равными 0.25, 0.125. Результаты расчетов приведены в табл. 3. В силу симметрии решения часть значений опущена.

Таблица 3

(x_1, x_2)	Нижняя граница	Точное решение	Верхняя граница	Ширина $\times$ 100	Ошибка $\times$ 100
$h = 0.25$					
(0.25, 0.25)	0.4895	0.5	0.5265	3.70	2.65
(0.5, 0.25)	0.6975	0.7071	0.7446	5.71	3.74
(0.5, 0.5)	0.9924	1.0	1.0530	6.06	5.03
$h = 0.125$					
(0.125, 0.125)	0.1448	0.1464	0.1483	0.35	0.19
(0.25, 0.125)	0.2685	0.2705	0.2741	0.55	0.35
(0.375, 0.125)	0.3515	0.3535	0.3581	0.65	0.46
(0.5, 0.125)	0.3807	0.3826	0.3876	0.68	0.49
(0.25, 0.25)	0.4975	0.5	0.5064	0.89	0.65
(0.375, 0.25)	0.6510	0.6532	0.6617	1.06	0.84
(0.5, 0.25)	0.7050	0.7071	0.7162	1.12	0.91
(0.375, 0.375)	0.8516	0.8535	0.8646	1.29	1.10
(0.5, 0.375)	0.9221	0.9238	0.9358	1.36	1.19
(0.5, 0.5)	0.9984	1.0	1.0129	1.44	1.29

Сопоставляя результаты численных расчетов, видно, что при уменьшении шага сетки в 2 раза ширина двустороннего решения уменьшилась примерно в 4 раза. Заметим, что значение ширины двустороннего решения сопоставимо с точностью разностной схемы.

7.2. Квазилинейные эллиптические уравнения

Рассмотрим краевую задачу для квазилинейного уравнения второго порядка

$$L(u)u = 0, \qquad x \in \Omega, \tag{7.15}$$

$$u = 0, \qquad x \in \partial\Omega, \tag{7.16}$$

где Ω — ограниченная, связанная область пространства R^n, $L(u)$ — квазилинейный оператор второго порядка вида

$$L(u) = \sum_{i,j=1}^{2} a_{ij}(x, u, Du)D_{ij} + b(x, u, Du), \tag{7.17}$$

$$a_{ij} = a_{ji}.$$

Мы будем предполагать, что $u \in H^2(\Omega) \cap H_0^1(\Omega)$ и a_{ij}, b — непрерывные функции, заданные на $\Omega \times R \times R^n$.

Пусть $\mathcal{U}$ — подмножество $\Omega \times R \times R^n$. Оператор называется *эллиптическим на множестве* $\mathcal{U}$, если матрица коэффициентов $[a_{ij}(x, z, p)]$ положительно определена для всех $(x, z, p) \in \mathcal{U}$ и $\xi = (\xi_1, \ldots, \xi_n) \in R^n - \{0\}$:

$$0 < \lambda(x, z, p)|\xi|^2 \leq a_{ij}(x, z, p)\xi_i\xi_j \leq \Lambda(x, z, p)|\xi|^2.$$

Оператор $L(u)$ называется оператором *монотонного типа,* если из

$$-L(u)u \leq -L(v)v, \qquad x \in \Omega,$$

и

$$u \leq v, \qquad x \in \partial\Omega$$

вытекает

$$u \leq v, \qquad x \in \Omega.$$

Определим интервальное расширение оператора $L(v)$ следующим образом:

$$L(\mathbf{v})s = \{L(v)s | v \in \mathbf{v}\}.$$

Предположим, что задача (7.15), (7.16) решена численным методом

$$L^h(u^h)u^h = 0, \qquad x \in \Omega_h, \tag{7.18}$$

$$u^h = 0, \qquad x \in \partial\Omega_h. \tag{7.19}$$

Здесь $L^h(u^h)$ — разностная аппроксимация оператора $L(u)$ на сетке $\Omega_h = \{x_i\}, i = 1, \ldots, n$. Необходимо, используя известное численное решение u^h задачи (7.18), (7.19), оценить погрешность $u - u^h$.

Операторы монотонного типа

В этом разделе мы рассмотрим апостериорные оценки погрешности решений краевых задач с операторами монотонного типа. Условия, при которых оператор $L(u)$ монотонного типа, сформулированы в следующей теореме [24]:

Теорема 55. *Пусть выполнены условия:*

1) оператор $L(u)$ локально равномерно эллиптичен на функции u;

2) коэффициенты $a_{ij}(x, z, p)$ не зависят от z;

3) коэффициент b является неубывающей функцией аргумента z в каждой точке $(x, p) \in \Omega \times R^n$;

4) коэффициенты a_{ij}, b являются непрерывно-дифференцируемыми функциями переменных p.

Тогда оператор $L(u)$ — монотонного типа.

Для нахождения апостериорной оценки погрешности нам необходимо численно решить вспомогательную задачу:

$$-L^h(u_1^h)u_1^h = \sigma, \qquad x \in \Omega_h$$

$$u_1^h = 0, \qquad x \in \partial\Omega_h.$$

Далее $s, s_1 \in H^2(\Omega) \cap H_0^1(\Omega)$ — аппроксимации численных решений u^h, $(u_1^h - u^h)/\sigma$ специальными конечными элементами высоких степеней [34], [79].

Мы будем искать интервал, содержащий ошибку $\varepsilon(x) = u(x) - s(x)$, в виде

$$\varepsilon(x) \in [\underline{\alpha}, \overline{\alpha}]s_1 + [\underline{\beta}, \overline{\beta}], \tag{7.20}$$

где $\underline{\alpha}, \overline{\alpha}, \underline{\beta}, \overline{\beta}$ — некоторые константы. Далее, пусть $\boldsymbol{s} = s + [\underline{\alpha}, \overline{\alpha}]s_1 + [\underline{\beta}, \overline{\beta}]$ и $\underline{\alpha}, \overline{\alpha}, \underline{\beta}, \overline{\beta}$ удовлетворяют следующим неравенствам:

$$-L(\overline{s})\overline{s} \geq 0, \qquad x \in \Omega, \tag{7.21}$$

$$\overline{s} \geq 0, \qquad x \in \partial\Omega, \tag{7.22}$$

$$-L(\underline{s})\underline{s} \leq 0, \qquad x \in \Omega. \tag{7.23}$$

$$\underline{s} \le 0, \qquad x \in \partial\Omega. \tag{7.24}$$

Тогда в силу монотонности оператора $L(u)$

$$\underline{s} \le u \le \overline{s}, \qquad x \in \Omega.$$

Константы $\underline{\alpha}, \overline{\alpha}, \underline{\beta}, \overline{\beta}$ будем искать поэтапно, сначала найдем $\underline{\alpha}, \overline{\alpha}$.

Для этого следующую систему неравенств будем решать методом последовательных приближений:

$$-L(\overline{s}_i)\overline{s}_{i+1} \ge 0, \quad x \in \Omega,$$

$$-L(\underline{s}_i)\underline{s}_{i+1} \le 0, \quad x \in \Omega, \quad i = 0, 1, \ldots,$$

где $\boldsymbol{s}_i = s + [\underline{\alpha}_i, \overline{\alpha}_i]s_1$ и $\boldsymbol{s}_0 = s$. Заметим, что процесс нахождения $[\underline{\alpha}, \overline{\alpha}]$ быстро сходится. Далее константы $\underline{\beta}, \overline{\beta}$ определяются на границе

$$\underline{\beta} = \min_{\partial\Omega} -s - \underline{\alpha}s_1, \quad \overline{\beta} = \max_{\partial\Omega} -s - \overline{\alpha}s_1.$$

Несложно убедиться в том, что построенные функции $\overline{s}, \underline{s}$ удовлетворяют системе неравенств (7.21),(7.22), (7.23),(7.24) и, следовательно, включение (7.20) выполнено.

Операторы немонотонного типа

Рассмотрим задачу, когда для оператора $L(u)$ не выполнены условия 2), 3) теоремы 55. Тогда оператор $L(u)$ может не быть оператором монотонного типа. В этом случае для оценки погрешностей численных решений эллиптических уравнений можно применить теорему Шаудера о неподвижной точке [24].

Мы вновь обратимся к (7.15), (7.16) с дифференциальным оператором Lu вида (7.17).

Пусть u — решение задачи (7.15),(7.16), предположим, что

1) оператор $L(u)$ локально равномерно эллиптичен на функции u;

2) коэффициенты a_{ij}, b являются непрерывно-дифференцируемыми функциями переменных p.

Определим оператор $\tilde{L}_(v, u)$, полученный из оператора $L(u)$ следующим образом:

$$\tilde{L}(v, u) = a_{ij}(x, v, Du)D_{ij} + b(x, v, Du).$$

Заметим, что построенный квазилинейный оператор $\tilde{L}$ обладает свойством монотонности, поскольку все свойства теоремы 1 выполнены.

Положим $v \equiv u$, следовательно, если найдена функция $\overline{s}$ такая, что

$$-\tilde{L}(v,\overline{s})\overline{s} \geq -L(u)u = 0, \qquad x \in \Omega,$$

$$\overline{s} \geq 0, \qquad x \in \partial\Omega,$$

то

$$\overline{s} \geq u, \qquad x \in \Omega.$$

Определим $\boldsymbol{s}$:

$$\boldsymbol{s} = s + [\underline{\alpha}, \overline{\alpha}]s_1 + [\underline{\beta}, \overline{\beta}],$$

где s, s_1 были определены выше.

Теорема 56. *Пусть выполнена следующая система неравенств:*

$$-\tilde{L}(\boldsymbol{s},\overline{s})\overline{s} \geq 0, \qquad x \in \Omega,$$

$$\overline{s} \geq 0, \qquad x \in \partial\Omega,$$

$$-\tilde{L}(\boldsymbol{s},\underline{s})\underline{s} \leq 0, \qquad x \in \Omega,$$

$$\underline{s} \leq 0, \qquad x \in \partial\Omega,$$

тогда $\mathbf{s}$ *содержит решение задачи (7.15),(7.16).*

Доказательство. Действительно, рассмотрим две функции $v, z \in H^2(\Omega) \cap H_0^1(\Omega), v \in \boldsymbol{s}$ такие, что

$$L(v)z = 0, \qquad x \in \Omega,$$

$$z = 0, \qquad x \in \partial\Omega.$$

Пусть T обратный к оператору L, заданный следующим образом: $Tv = z$, тогда

$$\boldsymbol{z} = T\boldsymbol{s}.$$

Заметим, что

$$-\tilde{L}(v,\overline{s})\overline{s} \geq 0, \qquad x \in \Omega,$$

$$\overline{s} \geq 0, \qquad x \in \partial\Omega.$$

Следовательно,

$$\overline{s} \geq \overline{z}, \qquad x \in \Omega.$$

Аналогично

$$\underline{s} \leq \underline{z} \qquad x \in \Omega.$$

Это равносильно утверждению

$$T\boldsymbol{s} \subseteq \boldsymbol{s},$$

из которого, в силу теоремы Шаудера о неподвижной точке [24], вытекает утверждение теоремы. □

Таким образом, ошибка $u - s$ содержится в интервале

$$u - s \in [\underline{\alpha}, \overline{\alpha}]s_1 + [\underline{\beta}, \overline{\beta}].$$

Для нахождения констант $\underline{\alpha}, \overline{\alpha}, \underline{\beta}, \overline{\beta}$ можно воспользоваться способом, описанным выше.

Численный пример

Рассмотрим квазилинейное эллиптическое уравнение второго порядка

$$L(u)u \equiv (1 - (u - r)^2)D_{11}u + (1 + (u - r)^2)D_{22}u = f, \qquad x \in \Omega,$$

$$u = 0, \qquad x \in \partial\Omega,$$

где $\Omega = [0,1] \times [0,1]$, $r(x_1, x_2) = \sin(\pi x_1)\sin(\pi x_2)$, $f = -2\pi^2 \sin(\pi x_1)\sin(\pi x_2)$. Точное решение этой задачи — $u = \sin(\pi x_1)\sin(\pi x_2)$. Очевидно, что оператор $L(u)$ не монотонного типа.

Предположим для определенности, что исходная задача решена разностным методом

$$L^h(u^h)u^h = f^h, \qquad x \in \Omega_h,$$

$$u^h = 0, \qquad x \in \partial\Omega_h,$$

где Ω_h — узлы прямоугольной равномерной сетки с шагом h.

В качестве вспомогательной задачи мы решим численно на сетке Ω_h следующее уравнение:

$$-\Delta u_1 = 1, \qquad x \in \Omega,$$

$$u_1 = 0, \qquad x \in \partial\Omega.$$

Пусть s, s_1 — аппроксимации эрмитовыми сплайнами третьей степени численных решений u^h, u_1^h [34]. Заметим, в нашем случае для s, s_1 граничные условия (7.16) выполнены точно.

Положим

$$\boldsymbol{s} = s + [\underline{\alpha}, \overline{\alpha}]s_1.$$

Мы будем искать константы $\underline{\alpha}, \overline{\alpha}$, удовлетворяющие следующей системе неравенств:

$$L(\boldsymbol{s})\overline{s} \geq f, \qquad x \in \Omega, \tag{7.25}$$

$$L(\boldsymbol{s})\underline{s} \leq f, \qquad x \in \Omega. \tag{7.26}$$

Для решения этой системы неравенств можно воспользоваться методом последовательных приближений, описанным выше.

Ниже приведено сравнение апостериорной оценки погрешности с ошибкой численного решения.

h	$\max_\Omega \lvert wid\boldsymbol{s}\rvert$	$\max_{\Omega_h} \lvert u - u^h\rvert$
0.25	9.43e-2	6.48e-2
0.125	1.92e-2	1.61e-2

7.3. Многосеточный метод

Формулировка задачи

Рассмотрим краевую задачу

$$\begin{aligned} Lu &= f,\ x \in \Omega, \\ u(x) &= 0,\ x \in \partial\overline{\Omega}, \end{aligned} \tag{7.27}$$

где Ω — ограниченная, связанная область пространства R^n, L — дифференциальный оператор второго порядка вида

$$Lu = -\sum_{i=1}^{2} \frac{\partial}{\partial x_i}(a_i \frac{\partial}{\partial x_i}u) + qu. \tag{7.28}$$

Предположим, что коэффициенты $a_i \in C^1(\Omega)$, $q, f \in C(\Omega)$, $g \in C(\partial\Omega)$ и что

$$a_i \geq c > 0, q \geq 0, x \in \Omega.$$

Предположим, что решение (7.27) существует, единственно и $u \in C^4(\Omega)$.

Пусть $\mathcal{T}_i, i : 0, 1, \ldots, m$ — последовательность разбиений Ω состоящей из элементов T_l и таких

$$\overline{\Omega} \subseteq \cup_i T_i, \quad T_i \cap T_j = \emptyset, \quad i \neq j.$$

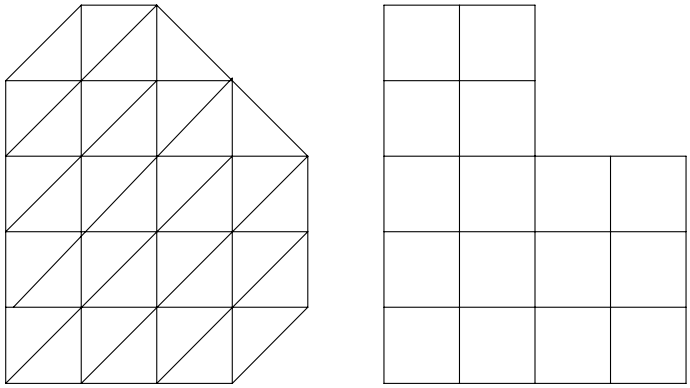

Рис. 7.1. Пример областей и разбиений

Элементы T_i, T_j могут иметь или общий угол или общую сторону рис. 7.1.

Если $\mathcal{T}_i$ — триангуляция, тогда пространство конечных элементов S_l^n определим введением кусочно-полиномиального базиса на $\mathcal{T}_i$:

$$S_l^n(\mathcal{T}_h) = \{s(x) | s \in W_2^l(\Omega), s|_T \in \mathcal{P}^n, T \in \mathcal{T}_h\}, \tag{7.29}$$

где $\mathcal{P}^n$ — пространство полиномов степени n.

Построение оператора аппроксимации

Опишем вспомогательные результаты по построению конечных элементов на различных разбиениях.

Рассмотрим задачу аппроксимации функции $s_i \in S_{l_1}^{n_1}(\mathcal{T}_i)$ конечными элементами $s_j \in S_{l_2}^{n_2}(\mathcal{T}_j)$ или, другими словами, определим оператор аппроксимации $A:\ S_l^n(\mathcal{T}_j) \to S_l^n(\mathcal{T}_i)$

$$s_i = A(\mathcal{T}_i, s_j). \tag{7.30}$$

Для построения оператора $A(\mathcal{T}_i, s_j)$ нам необходимо знать некоторое множество значений $s_i(x)$, $D_{i,j}s_i(x)$, $i, j = 0, 1, 2, \ldots$ для некоторых точек $x \in \overline{\Omega}$.

Например, пространство $S_2^5(\mathcal{T})$ определяется значениями $\{s,\ D_1 s,\ D_2 s,\ D_{11}s,\ D_{12}s,\ D_{22}s\}$ в узлах триангуляции и $D_n s$ в серединах сторон рис. 7.2.

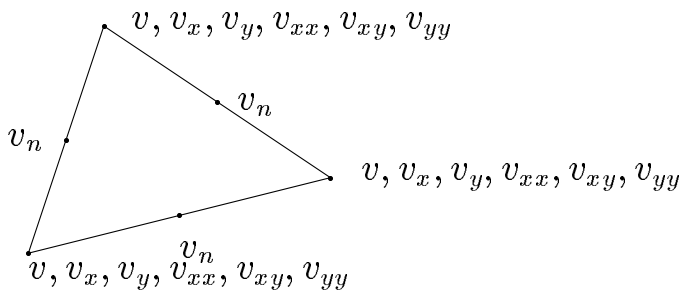

Рис. 7.2. Пространство $S_2^5(\mathcal{T})$

Пусть $x_0 = (x_{0,1}, x_{0,2})$ — одна из таких точек. В некоторых случаях необходимые значения производных можно вычислить непосредственно используя функцию s_j на разбиении $\mathcal{T}_j$. В большинстве случаев это затруднительно, тогда мы обратимся к результатам раздела 1.3.

Уменьшение ширины двустороннего решения

Как известно, ширина двустороннего решения $\mathrm{wid}(\mathbf{u})$ значительно зависит от значений производных в узлах сетки. Займемся уточнением производных.

Пусть $s \in S_2^n(\mathcal{T}_m)$ — некоторое приближенное решение задачи (7.27). Используя это решение как начальное приближение, мы построим новое решение s_{dec}, дающее двустороннее решение меньшей ширины.

Мы будем искать s_{dec} как приближенное решение следующей задачи:

$$\Phi_p(s_{dec}) = \min_{v \in S_2^n(\mathcal{T}_m)} \Phi_p(v), \tag{7.31}$$

где $\Phi_p(v)$

$$\Phi_p(v) = ||Lv - f||_{L_p(\Omega)} + K||v||_{L_p(\partial\Omega)}.$$

Заметим, что ширина двустороннего решения $\boldsymbol{s}$, построенного с использованием некоторой функции s, ограничена следующим неравенством:

$$\mathrm{wid}\ (\mathbf{s}) \le C \max_{\overline{\Omega}} |Ls - f| + \max_{\partial\Omega} |s|. \tag{7.32}$$

Следовательно, для уменьшения ширины двустороннего решения мы должны выбрать функцию s такую, что она минимизирует правую часть неравенства (7.32). Заметим, что функция $s = s(V)$ полностью определена своими параметрами V в узлах сетки $\mathcal{T}$. Предположим, что

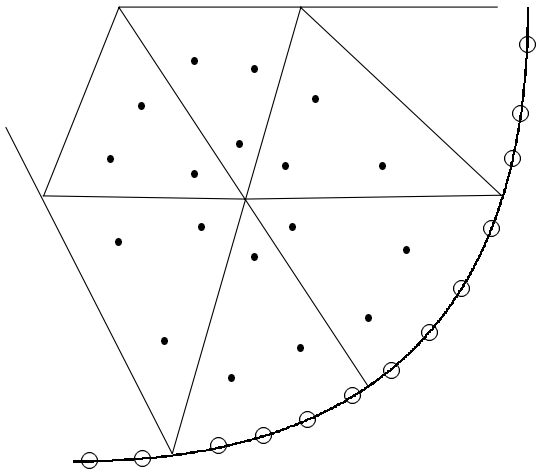

Рис. 7.3. Примеры вспомогательных сеток

V^* определяет лучшую функцию s в смысле ширины $\mathrm{wid}\boldsymbol{s}$, тогда

$$\Phi_\infty(V^*) = \min_V \Phi_\infty(V), \tag{7.33}$$

$$\Phi_\infty(V) = C \max_{\overline{\Omega}} |Ls - f| + \max_{\partial\Omega} |s|.$$

Поскольку задача (7.33) весьма сложная, мы будем искать ее приближенное решение. Как аппроксимацию функционала Φ_∞ рассмотрим семейство функционалов $\Phi_p = ||Ls - f||_{p,\omega_h} + ||s||_{p,\partial\omega_h}$, где $p > 0$ — целый параметр,

$$||v||_{p,\Omega_h} = \Big(\sum_{x\in\Omega_h} |v|^p\Big)^{1/p}, \tag{7.34}$$

ω_h, $\partial\omega_h$ — вспомогательные сетки рис. 7.3. Если вектор V отличается от оптимального V^*, мы можем исправить его, найдя V_1, выбирая соответствующий параметр p и вспомогательные сетки ω_h такие, что

$$\Phi_p(V_1) < \Phi_p(V) \Longrightarrow \Phi_\infty(V_1) < \Phi_\infty(V).$$

Таким образом, используя $s_{dec} = s(V_1)$, мы можем построить двустороннее решение меньшей ширины.

Рассмотрим для решения задачи специальный метод покоординатного спуска. Зафиксируем значения параметров V_j во всех узлах сетки $\mathcal{T}_m$, исключая узел сетки с номером i. Тогда функция s будет зависеть только от набора параметров V_i и мы обозначим s как $s(x, V, i)$.

Уточним V_i

$$\Phi_p^h(V_i) = \min_V \Phi_p^h(V).$$

Продолжим циклически уточнять V_i по всем узлам сетки, мы получаем последовательность функций $s^i, i = 1, 2, \ldots$ Процесс уточнения начинаем с узлов вблизи границы области $\partial\overline{\Omega}$ и движемся по направлению к центру. Такой выбор последовательности узлов обусловлен тем фактом, что вблизи границы $\partial\overline{\Omega}$ значение параметров V_i определяются точнее, поскольку нам известны значения функции на границе.

Многосеточный алгоритм

Пусть задана точность или допустимая ширина двустороннего решения $\varepsilon > 0$. Потребуем, чтобы ширина двустороннего решения не превосходила ε:

$$\mathrm{wid}(\mathbf{s}) < \varepsilon, \quad x \in \overline{\Omega}. \tag{7.35}$$

Построим функцию $s_M \in S_2^n(\mathcal{T}_M)$ такую, что $\mathrm{wid}(\mathbf{s})$ удовлетворяет соотношению (7.35).

При решении практических задач мы стремимся строить решение за минимальное число операций. Многосеточные методы — одни из таких алгоритмов.

Пусть $p_0 = 2 < p_1 < \ldots < p_m \leq \infty$ — последовательность целых чисел.

Предположим, что численное решение задачи (7.27) $s_0 \in S_2^n(\mathcal{T}_0)$ построено с использованием метода конечных элементов на сетке $\mathcal{T}_0$ и известно, что размерность пространства $S_2^n(\mathcal{T}_0)$ не велика.

Замечание 7. *Если нам известно численное решение u^h, полученное разностным методом или методом конечных элементов, с использованием кусочно-линейных элементов, начальное приближение $s_0 \in S_2^n$ на сетке $\mathcal{T}_0$ мы можем построить, используя аппроксимации конечными элементами высоких порядков, описанных в разделе 1.3.*

Пусть $C_0 = \Phi_{p_0}(V_0)$ и $C_i = \Phi_{p_i}(V_i)$. Рассмотрим переход с сетки $\mathcal{T}_{m-1}$ на сетку $\mathcal{T}_m$. Предположим, что мы имеем решение $s_{m-1} \in S_{p_{m-1}}^n(\mathcal{T}_{m-1})$. Построим решение s_m на $\mathcal{T}_m$ с

$$C_m \approx \left(\frac{h_m}{h_{m-1}}\right)^{n-1} C_{m-1}.$$

Шаг 1(аппроксимация). Используя оператор аппроксимации построим $s_{m,0}$

$$s_{m,0} = A(\mathcal{T}_m, s_{m-1}).$$

Шаг 2 (уточнение). Используя метод уточнения решений уточним $s_{m,0}$ и построим $s_{m,1}$.

Шаг 3(уменьшение). В заключение мы построим решение s_m на сетке $\mathcal{T}_m$, используя оператор уменьшения ширины двустороннего решения с $p = p_m$:

$$s_m = \mathrm{D}(s_{m,1}).$$

Вычислим ширину двустороннего решения. Если ширина построенного двустороннего решения удовлетворяет неравенству (7.35), следовательно, требуемое решение построено, в противном случае переходим на следующую сетку $\mathcal{T}_{m+1}$.

Замечание 8. *Один из возможных путей уменьшения ширины двустороннего решения — сгущение сетки в окрестности точек с большими значениями невязок.*

Пусть $u \in C^{n+1}(\Omega)$, тогда ширина окончательного двустороннего решения $\mathbf{s}$ удовлетворяет следующему соотношению:

$$||\mathrm{wid}(\mathbf{s})||_{L_\infty(\Omega)} \leq Kh^{n-1}, \tag{7.36}$$

где h — шаг сетки $\mathcal{T}_M$, K — константа, не зависящая от h.

Численный пример

Рассмотрим модельную задачу

$$\Delta u = f,\ x \in \Omega, \tag{7.37}$$

$$u_i = 0,\ x \in \partial\overline{\Omega}, \tag{7.38}$$

где

$$f = -2\pi^2 \sin(\pi x_1)\sin(\pi x_2), \Omega = [0,1] \times [0,1].$$

Точное решение этой задачи:

$$u = \sin(\pi x_1)\sin(\pi x_2).$$

Пусть нам необходимо построить двустороннее решение с шириной, не превосходящей величину $\varepsilon = 0.01$. Функцию s искали в виде эрмитовых

сплайнов третьей степени, сетки выбирали равномерными с шагами $h_i = 2^{-i}$, параметры $p_i = 2^i$.

Первая сетка имела всего одну внутреннюю точку. Сплайн $s_0(V_0)$ был построен как приближенное решение задачи

$$\Phi_2(V_0) = \min_V \Phi_2(V).$$

Вспомогательные сетки ω_h в (7.34) выбирали равномерными с $h = h_i/4$. Значения ширины $\mathrm{wid}(\mathbf{s}_i)$ построенных двусторонних решений приведены ниже.

Номер сетки	$\mathrm{wid}(\mathbf{s}_i)$
1	0.0629
2	0.0475
3	0.0136
4	0.0035

Для достижения требуемой точности, мы построили четыре сетки. Как видно из таблицы, асимптотическая сходимость ширины двустороннего решения соответствует теоретической оценке.

Заключительное замечание

Рассмотренный метод для построения двустороннего решения может быть с успехом применен для широкого класса уравнений в частных производных.

Как показано в монографии [34] оценка (7.36) может быть справедлива для s, построенных непосредственно как аппроксимации численных решений u^h на равномерных сетках, т.е. число сеток в многосеточном методе может быть равно единицам. Следовательно, при решении реальных задач, при условии, что начальные данные достаточно хорошие, достижение необходимой точности не потребует большого числа сеток.

7.4. Одномерное параболическое уравнение

Обратимся к одномерному параболическому уравнению теплопроводности с коэффициентом, имеющим разрыв первого рода. Его можно решить методом, основанным на вычислении невязки от специального сплайна, аппроксимирующего разностное решение. Сплайн построен

так, чтобы на линии разрыва коэффициента уравнения при фиксированном x выполнялись условия сопряжения (непрерывность решения и потока).

Рассмотрим уравнение

$$\partial_t u = \partial_x(p\partial_x u) - qu + f, \tag{7.39}$$

$$p(x,t) \geq c_0 > 0, q(x,t) \geq 0,$$

где $(x,t) \in Q = (0,1) \times (0,T)$. Определим для функции u начальные и краевые условия:

$$u(x,0) = 0, x \in [0,1], \tag{7.40}$$

$$u(0,t) = u_0(t), t \in [0,T],$$

$$u(1,t) = u_1(t), t \in [0,T],$$

Предположим, что коэффициенты уравнения и граничные функции удовлетворяют условиям

$$q, f \in C(\overline{Q}), u_0, u_1 \in C^1([0,T]).$$

Пусть коэффициент $p(x,t)$ имеет разрыв первого рода на прямой $x = \xi$, поэтому на линии разрыва вместо уравнения (7.39) выполняются условия

$$[u]_\xi = 0,\ [p\partial_x u]_\xi = 0,\ \forall t \in [0,T], \tag{7.41}$$

где $[f]_\xi = \lim_{x\to\xi-0} f(x) - \lim_{x\to\xi+0} f(x)$. Линия $x = \xi$ делит Q на две подобласти Q_1, Q_2, и для коэффициента p предполагаются выполненными следующие условия: $p, \partial_x p \in C^1(\overline{Q}_i), i = 1,2$, причем для каждой из подобластей $\overline{Q}_i$ на линии $x = \xi$ значение функции берется равным пределу по соответствующей подобласти. В этом смысле понимается условие

$$\partial_t u, \partial_x^2 u \in \overline{Q}_i, i = 1,2,$$

которое считается выполненным при дальнейшем изложении.

Введем равномерную сетку ω_h на отрезке $[0,1]$ с шагом $h = 1/n$, целым $n \geq 2$ и предположим, что $\xi \in \omega_h$. Это предположение не ограничивает общности, поскольку всегда можно сделать преобразование координат, линейное в каждой зоне гладкости и переводящее ξ в любую заданную точку. Кроме того, введем равномерные сетки по времени $\omega_\tau, \overline{\omega}_\tau$ и на прямоугольнике $\overline{Q} : \overline{\omega}_{h\tau} = \overline{\omega}_h \times \overline{\omega}_\tau$.

Приближенное решение будем искать в соответствии с явной схемой

$$u_t^\tau = Lu^\tau + f \text{ на } \omega_{h\tau}, \tag{7.42}$$

где разностные операторы вводятся по формулам

$$Lu^\tau = (pu^\tau_x)_x - qu^\tau,$$

$$u_x = \frac{u(x+h/2) - u(x-h/2)}{h},$$

$$u_t = \frac{u(t) - u(t-\tau)}{\tau}.$$

Дополним разностные уравнения начальными и краевыми условиями

$$u^\tau(x,0) = 0, x \in \omega_h,$$

$$u^\tau(0,t) = u_0(t), u^\tau(1,t) = u_1(t), t \in \omega_\tau.$$

В [59] приводится следующая оценка погрешности:

$$||u - u^\tau||_{\infty,\overline{\omega}_{h\tau}} \le c(h^2 + \tau), \tag{7.43}$$

где c не зависит от h, τ.

Рассмотрим вопрос о восполнении решения u^τ задачи 7.42 сплайном $s \in C^{1,0}(\overline{Q}_i)$. В качестве такого сплайна на сетке $\overline{\omega}_{h\tau}$ возьмем тензорное произведение одномерных сплайнов: на сетке $\overline{\omega}_\tau$ — кусочно-линейный сплайн, на сетке $\overline{\omega}_h$ — сплайн Эрмита третьей степени [36]. Поэтому для представления сплайна s на $\overline{Q}$ в каждом узле $(x_i, t_j) \in \overline{\omega}_h$ будем задавать значения функции s_{ij} и ее производных $\partial_x s_{ij}$, причем при $x = \xi$ задаются два значения производной $\partial_x s(\xi + 0, t_j, \partial_x s(\xi - 0, t_j$. Тогда на каждом прямоугольнике $[x_i, x_{i+1}] \times [t_j, t_{j+1}]$, следуя [36], сплайн s можно представить в виде

$$s(x,t) = \psi(y) F \theta(z),$$

где векторы $\psi, \theta \in R^2$ и матрица F:

$$\psi(y) = (1-y, y),$$

$$F = \begin{pmatrix} s_{ij} & s_{i+1,j} & \partial_x s_{ij} & \partial_x s_{i+1,j} \\ s_{ij+1} & s_{i+1,j+1} & \partial_x s_{ij+1} & \partial_x s_{i+1,j+1} \end{pmatrix}, x_i, x_{i+1} \ne \xi$$

$$\theta = \begin{pmatrix} (1-z)^2(1+2z) \\ (3-2z)z^2 \\ z(1-z)^2 \\ -z^2(1-z) \end{pmatrix},$$

$$y = (t - t_j)/\tau,\ z = (x - x_i)/h.$$

Для построения s определим в каждом узле значения:

$$s(x,t) = u^\tau(x,t), x = 0, h, \ldots, \xi, \ldots, 1;$$

$$\partial_x s(x,t) = \partial_x^h u^\tau(x,t), x = 0, h, \ldots, \xi - 0, \xi + 0, \ldots, 1; t \in \overline{\omega}_\tau.$$

Разностный оператор ∂_x определим таким образом:

$$\partial_x^h v(x) = \frac{1}{6h} \begin{cases} -11v(x) + 18v(x+h) - 9v(x+2h) + 2v(x+3h), & \text{при } x = 0, \xi + 0; \\ v(x-2h) - 6v(x-h) + 3v(x) + 2v(x+h), & \text{при } x = \xi - h, 1 - h; \\ -2v(x-3h) + 9v(x-2h) - 18v(x-h) + 11v(x), & \text{при } x = \xi - 0, 1; \\ 2v(x-h) - 3v(x) + 6v(x+h) - v(x+2h), & \text{в остальных узлах.} \end{cases}$$

На линии $x = \xi$ сплайн s построим так, чтобы для него выполнялись условия 7.41. Поэтому в прямоугольниках $[\xi, \xi + h] \times [t_j, t_{j+1}]$ матрицу F задаем так:

$$F = \begin{pmatrix} s(\xi, t_j) & s(\xi + h, t_j) & v_j(\xi + 0, t) & \partial_x s(\xi + h, t_j) \\ s(\xi, t_{j+1}) & s(\xi + h, t_j) & v_{j+1}(\xi + 0, t) & \partial_x s(\xi + h, t_{j+1}) \end{pmatrix},$$

а на прямоугольниках $[\xi - x, \xi] \times [t_j, t_{j+1}]$ полагаем

$$F = \begin{pmatrix} s(\xi - h, t_j) & s(\xi, t_j) & \partial_x s(\xi - h, t_j) & v_j(\xi - 0, t) \\ s(\xi - h, t_{j+1}) & s(\xi, t_{j+1}) & \partial_x s(\xi - h, t_{j+1}) & v_{j+1}(\xi - 0, t) \end{pmatrix}.$$

Здесь

$$v_l(\xi \pm 0, t) = \frac{\partial_x s(\xi - 0, t_l) p(\xi - 0, t) + \partial_x s(\xi + 0, t_l) p(\xi + 0, t)}{2p(\xi \pm 0, t)}, l = j, j + 1.$$

Отметим, что при таком определении функция s в полосе $x \in (\xi - h, \xi + h)$ будет линейной по t только для функций p, не зависящих от t. Поэтому в общем случае s является некоторым обобщением сплайна по t вблизи ξ, согласно построению выполнены соотношения

$$[s]_\xi = 0, \ [p\partial_x s]_\xi = 0, \ \forall t \in [0, T].$$

Следовательно, для s выполнены условия 7.41.

Всюду на Q, за исключением линии $x = \xi$, выполняется тождество

$$\partial_t s = \partial_x(p\partial_x s) - qs + f - \phi,$$

где

$$\phi \equiv -\partial_t s + \partial_x(p\partial_x s) - qs + f.$$

С помощью разностной схемы 7.42 на сетке $\omega_{h\tau}$ решим задачу

$$\partial_t \varepsilon = \partial_x(p\partial_x \varepsilon) - q\varepsilon + 1 \ на \ Q \setminus x = \xi \qquad (7.44)$$

с начальными и краевыми условиями

$$\varepsilon(x,0) = 0, x \in [0,1], \tag{7.45}$$

$$\varepsilon(0,t) = \varepsilon(1,t) = \psi(t), t \in [0,T]$$

и условиями сопряжения на линии $x = \xi$

$$[\varepsilon]_\xi = 0,\ [p\partial_x\varepsilon]_\xi = 0,\ \forall t \in [0,T].$$

По разностному решению ε^τ с помощью приведенного выше алгоритма построим сплайн s_1. Тогда двустороннее решение записывается в виде

$$\mathbf{u} = s + \alpha s_1 + \beta, \tag{7.46}$$

где α, β находим таким образом:

$$\overline{\alpha} = \max_{\overline{Q}}(f - L_1(s + \overline{\beta}))/L_1 s_1, \tag{7.47}$$

$$\underline{\alpha} = \min_{\overline{Q}}(f - L_1(s + \underline{\beta}))/L_1 s_1,$$

$$\overline{\beta} = \max_{[0,T]}\{u_0 - s(0,t), u_1 - s(1,t)\}, \tag{7.48}$$

$$\underline{\beta} = \min_{[0,T]}\{u_0 - s(0,t), u_1 - s(1,t)\}.$$

При достаточно малых h, τ сплайн s_1 аппроксимирует решение задачи (7.44), таким образом, $L_1 s_1 \approx 1$ на $\overline{Q}$. Поэтому величины $\overline{\alpha}, \underline{\alpha}$ определены и конечны.

Для доказательства включения (7.46) воспользуемся теоремой.

Теорема 57. *Пусть $f \in L_2(Q), u_0, u_1 \in L_2[0,T]$ и выполняются неравенства $f \geq 0$ почти всюду на Q, $0 \leq u_0, u_1 \leq c$ почти для всех $t \in [0,T]$. Тогда решение задачи (7.39) существует, единственно и удовлетворяет на Q неравенству $u \geq 0$.*

Доказательство. Существование и единственность решения задачи (7.39) показаны в [46]. Построим последовательность функций $P^m \in C^1(\overline{Q})$, $f^m \in C(\overline{Q})$, $u_0^m, u_1^m \in C^1[0,T]$ так, чтобы p^m, оставаясь равномерно ограниченными, сходились почти всюду к p, f^m сходились к f в норме пространства $L_2(Q)$, u_0^m, u_1^m сходились к u_0, u_1 в норме пространства $L_2[0,T]$. Кроме того, потребуем, чтобы

$$f^m \geq 0, 0 \leq u_0^m, u_1^m \leq c,$$

$$\partial_t u_0^m(0,0) = f^m(0,0), \partial_t u_1^m(1,0) = f^m(1,0).$$

Тогда в силу теоремы 4.5 из [46] решения u^m задач

$$\partial_t u^m = \partial_x(p^m \partial_x u^m) - q u^m + f^m,$$

$$u^m(x,0) = 0, x \in [0,1],$$

$$u^m(0,t) = u_0(t), t \in [0,T],$$

$$u^m(1,t) = u_1(t), t \in [0,T]$$

сходятся сильно в $L_2(Q)$ к решению задачи (7.39),(7.40). Заметим, что для каждого u^m в силу принципа максимума $u^m \geq 0$ на Q. Следовательно, для решения задачи (7.39),(7.40) выполнено неравенство $u \geq 0$ на Q. Теорема доказана. □

Покажем, что $w = s + \overline{\alpha} s_1 + \overline{\beta} \geq u$. Действительно, в области Q выполняется неравенство

$$L_1 w = L_1 s + \overline{\alpha} L_1 s_1 + L_1 \overline{\beta} \geq L_1 s + (f - L_1(s + \overline{\beta}) + L_1 \overline{\beta} \geq f.$$

На концах отрезка $[0,1]$ справедливо соотношение $s_1(0,t), s_1(1,t) \geq 0$, поэтому

$$w(0,t) = s_1(0,t) + \overline{\beta}) \geq s_1(0,t) + (u_0(t) - s_1(0,t)) = u_0(t),$$

$$w(1,t) = s_1(1,t) + \overline{\beta}) \geq s_1(1,t) + (u_1(t) - s_1(1,t)) = u_1(t).$$

Согласно теореме 57, $w \geq u$ почти всюду на $\overline{Q}$. Учитывая, что w, u непрерывны, получаем, что $w \geq u$ на $\overline{Q}$. Таким же образом можно показать, что $v = s + \underline{\alpha} s_1 + \underline{\beta} \leq u$.□

Далее оценим ширину полученного двустороннего решения. В целях упрощения выкладок дальнейшее обоснование проведем в случае $p = 1$ на $\overline{Q}$. Таким образом решаем уравнение

$$\partial_t u = \partial_x^2 u + f \tag{7.49}$$

с граничными условиями (7.40) и, естественно, без условий сопряжения (7.41). Разностная задача (7.42) упрощается соответствующим образом. Покажем, что производные находят с такой же точностью, как и само решение u. Для этого нам понадобится следующая

Теорема 58. *При достаточно гладком решении и задачи (7.49),(7.40) существует константа c, не зависящая от h, τ, такая, что*

$$\begin{aligned} |(u^\tau - u)_x(x,t)| &\le c(h^2+\tau), \\ |(u^\tau - u)_{xx}(x,t)| &\le c(h^2+\tau), \\ |(u^\tau - u)_t(x,t)| &\le c(h^2+\tau),\ (x,t) \in \omega_{h\tau}. \end{aligned} \tag{7.50}$$

Доказательство. Пусть $v(x,t) = u(x,t) - u^\tau(x,t)$ для $(x,t) \in \omega_{h\tau}$. Тогда

$$v_t = Lv + \phi \text{ на } \omega_{h\tau}, \tag{7.51}$$

$$v(x,0) = 0,\ x \in \omega_h,\ v(0,t) = v(1,t) = 0,\ x \in \omega_\tau,$$

где

$$\phi(x,t) = -\frac{\tau}{2}\partial_t^2 u(x, \eta_1 + p(x,t)\frac{h^2}{12}\partial_x^4 u(\eta_2,t) + \partial_x p(x,t)\frac{h^2}{3}\partial_x^3 u(\eta_3,t). \tag{7.52}$$

Далее положим $w(x,t) = v(x,t) - v(x,t-\tau)$. Поскольку $v(x,0) = 0$, то $w(x,t) = v(x,t)$. Оценим $w(x,t)$. Из (7.51) следует, что

$$w_{x,\tau}/\tau = Lw(x,\tau) + \phi(x,\tau).$$

Тогда, согласно [59], получаем

$$||w(\cdot,\tau)||_{\infty,\omega_h} \le \tau ||\phi(\cdot,\tau)||_{\infty,\omega_h}.$$

Далее оценим $w(x,t)$. Поскольку

$$w_{x,\tau}/\tau = Lw(x,\tau) + Lv(x,t-\tau) + \phi(x,\tau) =$$

$$= Lw(x,t) + w(x,t-\tau) + \phi(x,t) - \phi(x,t-\tau),$$

$$w(0,t) = w(1,t) = 0,$$

справедлива оценка [59]

$$||w(\cdot,t)||_{\infty,\omega_h} \le ||w(\cdot,t-\tau)||_{\infty,\omega_h} + \tau||\phi(\cdot,t) - \phi(\cdot,t-\tau)||_{\infty,\omega_h} \le ct\tau(h^2+\tau).$$

Окончательно, так как $t \le T$,

$$\frac{|v(x,t) - v(x,t-\tau)|}{\tau} \le c(h^2+\tau). \tag{7.53}$$

Перепишем разностное уравнение (7.51) при $t = const$ в следующем виде:

$$Lv = \phi_1 = -\phi + v_t \text{ на } \omega_h,$$

$$v(0,t) = v(1,t) = 0.$$

Поскольку в силу (7.52), (7.53)

$$||\phi_1||_{\infty,\omega_h} \le c(h^2+\tau),$$

то, согласно работе [59],

$$\begin{aligned} |v_x(x,t)| &\le c(h^2+\tau), x \in \omega_h, \\ |v_{xx}(x,t)| &\le c(h^2+\tau). \end{aligned} \tag{7.54}$$

Из оценок (7.53), (7.54) непосредственно вытекает утверждение теоремы.

□

Теперь обоснуем такой же порядок малости по τ, h ширины коридора двустороннего решения, какой имеет само приближенное решение.

Теорема 59. *Пусть* $\partial_t^2 u, \partial_t \partial_x^2 u, \partial_x^4 u \in C^1(\overline{Q})$ *и* τ, h *таковы, что* $L_1 s_1 \ge c > 0$ *на* $\overline{Q}$. *Тогда ширина двустороннего решения задачи (7.49), (7.40) является величиной* $O(h^2+\tau)$ *на* $\overline{Q}$.

Доказательство. Представим разность $u - s$ в виде суммы $u - s = (u - s_2) + (s_2 - s_3) + (s_3 - s)$, где s_2, s_3 - сплайны той же структуры, что и s (линейные по t и эрмитовы степени 3 по x), с данными

$$s_2 = u, \partial_x s_2 = \partial_x u, \text{ на } \overline{\omega}_{h\tau},$$

$$s_3 = u, \partial_x s_3 = \partial_x^h u, \text{ на } \overline{\omega}_{h\tau}$$

Для разности $u - s_2$ на основании [36] справедливы оценки

$$||\partial_x^2(u - s_2)||_{\infty,Q} = O(h^2), ||\partial_t(u - s_2)||_{\infty,Q} = O(\tau), \tag{7.55}$$

которые доказываются сначала на линиях t_j, x_i соответственно, а затем переносятся на все $\overline{Q}$. На основании гладкости u разностные отношения $\partial_x^h u$ аппроксимируют $\partial_x u$ с третьим порядком точности, поэтому

$$|s_2 - s_3| = 0, |\partial_x(s_2 - s_3)| = O(h^3) \text{ при } x \in \overline{\omega}_h, t \in (0,T).$$

С учетом величины базисных функций эрмитового сплайна приходим к оценке

$$||\partial_x^2(s_2 - s_3)||_{\infty,Q} = O(h^2). \tag{7.56}$$

Ввиду линейности сплайна $(s_2 - s_3)$ по t на каждом интервале $[t_j, t_{j+1}]$ справедливо равенство

$$\partial_t(s_2 - s_3) = (s_2 - s_3)_t, t \in \omega_\tau, \ x \in [0,1].$$

При фиксированном $t \in \omega_\tau$ функция $\partial_t(s_2 - s_3)$ является эрмитовым кубическим сплайном со значениями

$$\partial_t(s_2 - s_3) = 0, \partial_{xt}(s_2 - s_3) = \partial_x u_t - \partial_x^h u_t \text{ на } \overline{\omega}_h \times \omega_\tau.$$

На основании гладкости u разложение в ряд Тейлора дает равенство

$$|\partial_x u_t - \partial_x^h u_t| = O(h + \tau/h) \text{ на } \overline{\omega}_h \times \omega_\tau.$$

Из оценки величины базисных функций эрмитового сплайна следует, что

$$||\partial_t(s_2 - s_3)||_{\infty,Q} = O(h^2 + \tau). \tag{7.57}$$

Осталось рассмотреть разность $s_3 - s$. При фиксированном $t \in \overline{\omega}_\tau$ она является кубическим эрмитовым сплайном, для нее справедлива лемма 5.1.3 [34], из которой вытекает оценка

$$|\partial_x^2(s_3 - s)| \le c|(s_3 - s)_{xx}| = O(h^2 + \tau), \quad x \in [0, 1]. \tag{7.58}$$

Ее справедливость для произвольного $t \in [0, T]$ устанавливается линейной интерполяцией. При фиксированном $t \in \omega_\tau$ функция $z = \partial_t(s_3 - s) = (s_3 - s)_t$ — эрмитовый сплайн степени 3, удовлетворяющий условию

$$z = (u - u^\tau)_t, \partial_x z = \partial_x^h (u - u^\tau)_t \text{ на } \overline{\omega}_h \times \omega_\tau.$$

На основании теоремы 58 $|z(x,t)| = O(h^2 + \tau), (x,t) \in \omega_{h\tau}$. Заметим, что $|\partial_x z(x,t)| = O(h + \tau/h)$, $(x,t) \in \omega_{h\tau}$. С учетом величины базисных функций эрмитовых сплайнов имеем $|z(x,t)| = O(h^2 + \tau)$ при $x \in [0,1], t \in \omega_\tau$. Для вывода этой оценки с произвольным t используем линейную интерполяцию. В итоге

$$||\partial_t(s_3 - s)||_{\infty,Q} = O(h^2 + \tau). \tag{7.59}$$

Объединяя оценки (7.55)-(7.59), приходим к заключению, что

$$||L_1(u - s)||_{\infty,Q} = O(h^2 + \tau).$$

Следовательно, $\overline{\alpha}, \underline{\alpha}$, определяемые формулами (7.47), являются величинами порядка $O(h^2 + \tau)$. С учетом интерполирующих свойств кусочно-линейных функций вытекает, что $\overline{\beta}, \underline{\beta}$ из (7.48) равны $O(\tau^2)$. Доказательство ограниченности сплайна s_1 вытекает непосредственно из ограниченности сеточной функции ε^τ. Все это, вместе взятое, приводит к требуемой оценке

$$wid\mathbf{u} = (\overline{\alpha} - \underline{\alpha})s_1 + \overline{\beta} - \underline{\beta} = O(h^2 + \tau).$$

□

В качестве иллюстративного примера решена следующая задача:

$$\partial_t u = \partial_x(p\partial_x u) + f, (x,t) \in Q = (0,1)\times(0,1),$$

$$u(x,0) = 0, x \in [0,1],$$

$$u(0,t) = u(1,t) = 0, t \in [0,1],$$

$$p(x,t) = \begin{cases} 0.5 & x \le 0.5, \\ 1.0 & x > 0.5, \end{cases}$$

$$f(x,t) = \begin{cases} (1-2t)(-\pi^2x^2/2 + \pi^2x/2 + (8-\pi^2)/8)\sin(\pi x) - \\ (\pi^2\sin(\pi x) + 2(-\pi^2x + \pi^2/2)\pi\cos(\pi x)/2 - \\ (-\pi^2x^2/2 + \pi^2x/2 + (8-\pi^2)/8)\pi^2\sin(\pi x))(1-t)t, & x \le 0.5, \\ (1-2t)\sin(\pi x) - \pi^2\sin(\pi x)(1-t)t, & x > 0.5, \end{cases}$$

для которой выполнены условия сопряжения (7.41) в точке $\xi = 0.5$. Точное решение этой задачи выписывается аналитически:

$$u(x,t) = \begin{cases} (1-t)t(-\pi^2x^2/2 + \pi^2x/2 + (8-\pi^2)/8)\sin(\pi x) & x \le 0.5 \\ (1-t)t\sin(\pi x), & x > 0.5 \end{cases}.$$

Результаты расчетов приведены в табл. 4. При выбранном уменьшении шагов, на основании оценки (7.43), погрешность разностного решения уменьшается в 4 раза. Легко проверить, что этому условию удовлетворяет ширина полосы между верхней и нижней границей.

Таблица 4

x	Нижняя граница	Точное решение	Верхняя граница	Ширина полосы
$h_1, h_2 = 0.1,\ \tau = 0.1,\ t = 0.5$				
0.1	0.7477	0.7725	0.8083	0.0606
0.2	1.4222	1.4694	1.5350	0.1128
0.3	1.9575	2.0225	2.1105	0.1530
0.4	2.3012	2.3776	2.4792	0.1780
0.5	2.4196	2.5000	2.6060	0.1864
0.6	2.1876	2.2603	2.3569	0.1693
0.7	1.5711	1.6233	1.6939	0.1228
0.8	0.7957	0.8168	0.8533	0.0576
0.9	0.1573	0.1625	0.1700	0.0127

$h_1, h_2 = 0.05,\ \tau = 0.025,\ t = 0.5$

0.1	0.7662	0.7725	0.7797	0.0132
0.2	1.4574	1.4694	1.4828	0.0254
0.3	2.0060	2.0225	2.0407	0.0347
0.4	2.3582	2.3776	2.3988	0.0406
0.5	2.4796	2.5000	2.5221	0.0425
0.6	2.2418	2.2603	2.2804	0.0384
0.7	1.6100	1.6233	1.6379	0.0279
0.8	0.8101	0.8168	0.8242	0.0038
0.9	0.1612	0.1625	0.1640	0.0028

7.5. Двумерное параболическое уравнение

Продолжим изучение двусторонних методов решения параболических уравнений. Исходная дифференциальная задача, на этот раз с двумя пространственными переменными, решается методом расщепления. Затем полученное сеточное решение сглаживается с помощью сплайнов, являющихся тензорным произведением двумерных кубических сплайнов по пространственной переменной и линейных — по временной. Далее, по величине невязки путем сравнения с решением некоторой вспомогательной задачи определяется двустороннее решение.

Пусть $\Omega \subset R^2$ — ограниченная область с границей $\partial\Omega$, а Q — открытый цилиндр $\Omega \times (0, T)$ с боковой поверхностью S. Рассмотрим уравнение

$$\partial_t u = Lu + f \quad \text{в} \quad Q, \tag{7.60}$$

где $Lu = \sum\limits_{i=1}^{2} \partial_i(\alpha_i \partial u_i) - qu$. Для функции u определим начальные и граничные условия

$$u(x, 0) = u_0(x), \quad x \in \Omega,$$
$$u = g \quad \text{на} \quad S. \tag{7.61}$$

Коэффициенты уравнения удовлетворяют условиям

$$\alpha_i \geq \alpha > 0, \quad q \geq 0 \quad \text{на} \quad \bar{\Omega}.$$

Для применения метода расщепления построим разностную сетку в Ω так, как это сделано в [34] для области с криволинейной границей. Предполагается, что область Ω заключена в квадрат $[-b, b] \times [-b, b]$. Покроем его квадратной сеткой с шагом $h = 1/n$, образованной линиями $x_{1,i} = ih, x_{2,j} = jh, j = -n, ..., n$, и обозначим через ω_h множество узлов,

попадающих в Ω. Введем также множества $\Omega_i^r, \Omega_i^{ir}, \Gamma_i$. На временном отрезке $[0,T]$ введем сетку ω_τ и положим $\omega_{h\tau} = \omega_h \times \omega_\tau$.

Для разностной аппроксимации операторов $L_i u = \partial_i(\alpha_i \partial_i u)$ в узлах сетки $\omega_{h\tau}$ рассмотрим трехточечные формулы в регулярных узлах

$$L_1^h u(x) \equiv -\left(a_1 u_{\circ_{x_1}}\right)_{\circ_{x_1}}, \quad x \in \Omega_1^r$$

и четырехточечные формулы

$$L_1^h u(x) \equiv \frac{6a_1((x_1+\xi)/2, x_2)}{\delta(\delta+1)(\delta+2)h^2}(u(x) - u(\xi, x_2)) -$$

$$-\frac{(4-2\delta)a_1(x_1 \pm h/2, x_2)}{(1+\delta)h^2}(u(x) - u(x_1 \pm h, x_2)) +$$

$$+\frac{(1-\delta)a_1(x_1 \pm h, x_2)}{(2+\delta)h^2}(u(x) - u(x_1 \pm 2h, x_2)), \quad x \in \Omega_1^{ir}.$$

$$L_1^h u(x) \equiv \frac{6a_2(x_1, (x_2+\eta)/2)}{\rho(\rho+1)(\rho+2)h^2}(u(x) - u(x_1, \eta)) -$$

$$-\frac{(4-2\rho)a_2(x_1, x_2 \pm h/2)}{(1+\rho)h^2}(u(x) - u(x_1, x_2 \pm h)) +$$

$$+\frac{(1-\rho)a_2(x_1, x_2 \pm h)}{(2+\rho)h^2}(u(x) - u(x_1, x_2 \pm 2h)), \quad x \in \Omega_2^{ir}$$

в нерегулярных узлах. Мы будем изучать разностную схему расщепления, состоящую из циклически повторяющихся пространственно-одномерных задач [53]. Запишем ее, опуская у функций аргумент (x,t):

$$\frac{u^* - u^\tau(x, t-\tau)}{\tau} - L_1^h u^* - q u^* = f, \quad (x,t) \in \omega_{h\tau}, \tag{7.62}$$

$$u^* = g, \quad (x,t) \in \Gamma_1 \times \omega_\tau,$$

$$\frac{u^\tau - u^*}{\tau} - L_2^h u^\tau = 0, \quad (x,t) \in \omega_{h\tau}, \tag{7.63}$$

$$u^\tau = g, \quad (x,t) \in \Gamma_2 \times \omega_\tau.$$

Для начала процесса ставится условие

$$u^\tau(x,0) = u_0(x), \quad x \in \omega_h. \tag{7.64}$$

Последующий процесс решения сеточных задач (7.62), (7.63) состоит в применении некоторой модификации метода прогонки для систем с сильным диагональным преобладанием [53]. Влияние вычислительной погрешности здесь весьма мало, и мы на нем специально не

останавливаемся. В итоге после решения задач (7.62), (7.63) по слоям $t = \tau, 2\tau, ..., T$ получается сеточное решение u^τ, заданное в узлах $\omega_h \times \bar{\omega}_\tau$. В [60] показано, что при достаточной гладкости решения существует константа c, не зависящая от τ, h, и такая, что

$$||u - u^\tau||_{\infty,\omega_{h\tau}} \le c(h^2 + \tau).$$

Построим восполнение полученного численного решения u^τ с помощью тензорного произведения следующих сплайнов: на сетке $\bar{\omega}_\tau$ — кусочно-линейного сплайна, на сетке ω_h— двумерного эрмитова сплайна. Для построения эрмитова сплайна на сетке ω_h используем метод, изложенный в разд. 1.3. Напомним некоторые технические детали. Рассмотрим фиксированный слой $t \in \bar{\omega}_\tau$. Поскольку нам потребуются значения s на боковой поверхности S, необходимо определить s на всех квадратных ячейках, имеющих непустое пересечение Ω. Обозначим через Ω_h множество вершин всех таких ячеек. В этом случае для доопределения значений u^τ вне области следует использовать интерполяцию по четырем соседним узлам, лежащим на одной прямой и, возможно, включающим значения u^τ на S. Доопределенную на Ω_h сеточную функцию вновь обозначим через u^τ. Дальнейшее построение сплайна основано на формулах [36]

$$s = u^\tau, \quad \partial_1 s = \partial_1^h u^\tau, \quad \partial_2 s = \partial_2^h u^\tau, \quad \partial_{12} s = d_1^h d_2^h u^\tau,$$

$$t \in \bar{\omega}_\tau, \quad x \in \Omega_h.$$

Здесь ∂_i^h — сеточный оператор на подходящем четырехточечном шаблоне, аппроксимирующий производную ∂_i, с третьим порядком точности, а d_i^h — сеточный оператор на подходящем трехточечном шаблоне, аппроксимирующий производную ∂_i со вторым порядком точности. Образцы таких операторов имеются в разд. 1.3. После определения эрмитова сплайна $s(x,t)$ на каждом временном слое t используется линейная интерполяция: для всех $t \in [t_j, t_{j+1}]$

$$s(x,t) = \frac{t_{j+1} - t}{\tau} s(x,t_j) + \frac{t - t_j}{\tau} s(x,t_{j+1}).$$

Отметим, что полученный сплайн s входит в область определения дифференциального оператора уравнения (7.60). Следовательно, почти всюду на Q выполняется тождество

$$\partial_t s = Ls + f - \varphi,$$

где $\varphi = Ls + f - \partial_t s$.

С помощью сформулированной ранее локально-одномерной разностной схемы решим следующую задачу:

$$\partial_t \varepsilon = L\varepsilon + 1 \quad x \in Q,$$

с начальными и краевыми условиями

$$\varepsilon(x,0) = 0, \quad x \in \Omega,$$

$$\varepsilon(x,t) = \psi(t) \quad x \in S,$$

где функция $\psi(t) \geq 0$ выбирается для выполнения условия согласования порядка 1 и может быть взята, например, в виде $\psi(t) = t\chi(t)$ со срезающей функцией χ из [34]. По разностному решению ε^τ этой задачи с помощью приведенного выше алгоритма построим сплайн s_1. Тогда двустороннее решение записывается в виде

$$\mathbf{s} = s + \alpha s_1 + \beta \ni u, \tag{7.65}$$

где α, β находят следующим образом:

$$\bar{\beta} = \max\left\{\max_S(g-s),\ \max_{\bar{\Omega}}(u_0(x) - s(x,0))\right\}, \tag{7.66}$$

$$\underline{\beta} = \min\left\{\min_S(g-s),\ \min_{\bar{\Omega}}(u_0(x) - s(x,0))\right\},$$

$$\bar{\alpha} = \max_{\bar{\Omega}}(f - L_1(s+\bar{\beta}))/Ls_1,$$

$$\underline{\alpha} = \min_{\bar{\Omega}}(f - L_1(s+\underline{\beta}))/Ls_1,$$

$$L_1 v = \partial_t v - Lv.$$

Доказательство существования $\underline{\alpha}, \bar{\alpha}$ и включения (7.65) повторяет аналогичное доказательство из раздела 7.4.

Оценку ширины построенного двустороннего решения проведем для модельной задачи

$$\partial_t u = \Delta u + f, \quad x \in \quad \Omega, \tag{7.67}$$

$$u(x,0) = 0, \quad x \in \Omega, \tag{7.68}$$

$$u = 0 \quad x \in S.$$

Далее будем считать, что решение задачи (7.67), (7.68) обладает следующей гладкостью:

$$\partial^\beta u \in C(\bar{Q}), \quad \text{где} \quad \beta = (\beta_1, \beta_2, \beta_3), \tag{7.69}$$

$$\beta_1 + \beta_2 + 2\beta_3 \le 6 \quad \text{и} \quad \alpha \in (0,1).$$

Справедлива следующая [53]

Теорема 60. *Пусть для задачи (7.67), (7.68) выполнены условия (7.69). Тогда для решения и имеет место разложение*

$$u^\tau = u + h^2 v_1 + \tau v_2 + (h^4 + \tau^2)\eta^\tau, \quad x \in \quad \omega_{h\tau}, \tag{7.70}$$

где функции v_1, v_2 не зависят от h,τ, а сеточная функция η^τ ограничена

$$\partial_i^2 v_l, \partial_t v_l \in C(\bar{Q}), \quad i,l = 1,2, \tag{7.71}$$
$$||\eta^\tau||_{\infty,\omega_{h\tau}} \le c.$$

□

Докажем, что ширина коридора двустороннего решения **u** имеет такой же порядок малости, как погрешность исходного приближенного решения u^τ.

Теорема 61. *Пусть выполнено условие (7.69) и τ, h таковы, что $L_1 s_1 \ge c > 0$ на $\bar{\Omega}$. Тогда ширина wid**s** двустороннего решения задачи (7.67), (7.68) является величиной $O(h^2 + \tau)$ равномерно на $\bar{Q}$.*

Доказательство. Структура доказательства совпадает с той, что использована в теореме 59. Поэтому мы изложим его схематично, указав только основные этапы. Представим разность $u - s$ в виде суммы $u - s = (u - s_2) + (s_2 - s_3) + (s_3 - s)$, где s_2, s_3 — сплайны той же структуры, что и s, т.е. линейные по t и эрмитовы степени 3 по x_1, x_2. Они строятся по следующим данным на $\Omega_h \times \bar{\omega}_\tau$:

$$s_2 = u, \partial_1 s_2 = \partial_1 u, \partial_2 s_2 = \partial_2 u, \partial_{12} s_2 = \partial_{12} u;$$

$$s_3 = u, \partial_1 s_3 = \partial_1^h u, \partial_2 s_3 = \partial_2^h u, \partial_{12} s_3 = d_1^h d_2^h u.$$

В узлах $x \in \Omega_h$, выходящих за пределы Ω, используется гладкое продолжение функции u. На основании [36] для разности $u - s_2$ справедлива оценка

$$||\partial^\beta (u - s_2)||_{\infty,Q} = O(h^2 + \tau), \quad \beta = (\beta_1, \beta_2, \beta_3),$$
$$\beta_1 + \beta_2 + 2\beta_3 \le 2.$$

Для этих же β по аналогии с разделом (7.1) получается оценка

$$||\partial^\beta (s_2 - s_3)||_{\infty,Q} = O(h^2 + \tau).$$

И наконец, на основании теоремы 60, доказывается оценка

$$||\partial^{\beta}(s_3 - s)||_{\infty,Q} = O(h^2 + \tau).$$

Используя три последние оценки, приходим к выводу, что

$$||\varphi||_{\infty,Q} = ||L_1(s - u)||_{\infty,Q} = O(h^2 + \tau). \qquad (7.72)$$

Согласно интерполирующим свойствам кусочно-линейных функций $|\underline{\beta}|, |\bar{\beta}| = O(\tau^2)$. В силу (7.66), (7.72) $|\underline{\alpha}|, |\bar{\alpha}| = O(h^2 + \tau)$. Поэтому для ширины двустороннего решения выполнено равенство

$$\mathrm{wid}\mathbf{u} = \rho(x,t) = \mathrm{wid}\alpha \cdot s_1(x,t) + \mathrm{wid}\beta = O(h^2 + \tau).\square$$

Таким образом, мы рассмотрели построение интервальных решений для уравнений в частных производных. Для нахождения интервальных решений мы использовали апостериорные оценки погрешности. Порядок сходимости полученных решений совпадает с порядком сходимости численных решений.

Список литературы

[1] Агеев М. П., Алик В. П., Марков Ю. И. Алгоритм 61б. Процедуры интервальной математики // Библиотека алгоритмов 51б-100б. — М.: Сов. радио, 1976. — С. 21-26.

[2] Алексеев А.В., Берсенев С.М., Добронец Б.С., Щайдуров В.В. Решение двумерных краевых задач для эллиптических уравнений второго порядка // Алгоритмы и программы. 1985. — №3.

[3] Алефельд Г., Херцбергер Ю. Введение в интервальные вычисления. — М.:Мир, 1987. — 360 с.

[4] Альберг Дж., Нильсон Э., Уолш Дж. Теория сплайнов и ее приложения. — М.: Мир, 1972.

[5] Арнольд В.И. Обыкновенные дифференциальные уравнения. — М., Наука, 1975. — 240 с.

[6] Бабушка И., Витасек Э., Прагер М. Численные процессы решения дифференциальных уравнений. — М.: Мир, 1969.

[7] Бахвалов Н. С. Численные методы. — М.: Наука, 1973. — Т. 1.

[8] Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы. — М.: Наука, 1987.

[9] Беклемишев Д.В. Дополнительные главы линейной алгебры. — М.: Наука, 1983. — 336 с.

[10] Быков В.И., Добронец Б.С. К интервальному анализу уравнений химической кинетики // Математические проблемы химической кинетики. — Новосибирск, 1989. — С.226–232.

[11] Варга Р. Функциональный анализ и теория аппроксимации в численном анализе. — М.: Мир, 1974.

[12] Васильев Ф.П. Численные методы решения экстремальных задач. — М.: Наука, 1980.

[13] Вербицкий В.И., Горбань А.Н., Утюбаев Г.Ш., Шокин Ю.И. Эффект Мура в интервальных пространствах // ДАН СССР. — 1989. Т.304, №1. — С. 17-22.

[14] Воеводин В. В. Об асимптотическом распределении ошибок округления при линейных преобразованиях // Журн. вычисл. математики и мат. физики. — 1967. Т. 1, №5. — С. 965-977.

[15] Воеводин В. В. Вычислительные основы линейной алгебры. — М.: Наука, 1977.

[16] Воеводин В.В., Кузнецов Ю.А. Матрицы и вычисления. — М.: Наука, 1984.

[17] Волков Е.А. Эффективные оценки погрешности решений методом сеток краевых задач для уравнений Лапласа и Пуассона на прямоугольнике и некоторых треугольниках // Тр. МИАН СССР. — 1967. Т. 74. — С. 55-86.

[18] Волков Е. А. Поточечные оценки погрешности разностного решения краевой задачи для обыкновенного дифференциального уравнения // Дифференц. уравнения. — 1973. Т. 9, №4. — С. 717-726.

[19] Волков Е. А. Об асимптотике апостериорной оценки погрешности разностного решения обыкновенного дифференциального уравнения // Дифференц. уравнения. — 1974. Т. 10, №12. — С. 2263-2266.

[20] Волков Е. А. О поиске максимума функции и приближенном глобальном решении системы нелинейных уравнений // Тр. МИАН СССР. — 1974. Т. 131. — С. 64-80.

[21] Волков Е. А. Апостериорная оценка погрешности разностных уравнений Лапласа и Пуассона // Тр. МИАН СССР. — 1975. Т. 134. — С. 47-62.

[22] Ганшин Г. С. Методы оптимизации и решение уравнений. — М.: Наука, 1987.

[23] Герасимов В.А., Добронец Б.С., Шустров М.Ю. Численные операции гистограммной арифметики и их применения // АиТ. — 1991, №2. — С. 83-88.

[24] Гилбарг Д., Трудингер Н. Эллиптические дифференциальные уравнения с частными производными второго порядка. — М.: Наука, 1989. — 463 с.

[25] Годунов С.К., Рябенький В. С. Разностные схемы. — М.: Наука, 1977.

[26] Горбунов А.Д., Шахов Ю. А. О приближенном решении задачи Коши для обыкновенных дифференциальных уравнений с наперед заданным числом верных знаков. 1, IV // Журн. вычисл. математики и мат. физики. — 1963. — Т. 3, №2. — С. 239-259; Т. 4, №3. — С. 426-433.

[27] Годунов С.К., Антонов А.Г., Кирилюк О.П., Костин В.И. Гарантированная точность решения систем линейных алгебраических уравнений в евклидовых пространствах. — Новосибирск: Наука. Сиб. от-ние, 1988. — 456 с.

[28] Давиденко Д.Ф. К вопросу об оценке погрешности при решении методом сеток задачи Дирихле для уравнения Лапласа // Докл. АН СССР. — 1961. — Т. 138, №2. — С. 267-270.

[29] Даугаве И.К., Самокиш Б. А. Об апостериорной оценке погрешности численного решения дифференциального уравнения // Методы вычислений. — Л.: Изд-во ЛГУ, 1963. — №1. С. 52-57.

[30] Девятко В.И. О двустороннем приближении при численном интегрировании обыкновенных дифференциальных уравнений // Журн. вычисл. математики и мат. физики. — 1963. — Т. 3. №2. — С. 254-265.

[31] Добронец Б.С., Рощина Е.Л. Приложения интервального анализа чувствительности // Вычислительные технологии. — Т. 7, №1. — 2002. С. 75–82.

[32] Добронец Б.С., Рощина С.Л. Специальные приближения множеств решений систем ОДУ с интервальными параметрами // Вопросы математического анализа. Вып. 5. — Красноярск: Из-во КГТУ, 2002. — С. 12–17.

[33] Добронец Б.С., Сенашов В.И. Об интервальных расширениях некоторых классов функций // Интервальные вычисления. — 1991. — №1. — С.54-59.

[34] Добронец Б.С., Шайдуров В.В. Двусторонние численные методы. — Новосибирск: Наука, 1990. — 208 с.

[35] Дулан Э., Миллер Д., Шилдерс У. Равномерные численные методы решения задач с пограничным слоем. — М.: Мир, 1983.

[36] Завьялов Ю.С., Квасов Б.И., Мирошниченко В.Л. Методы сплайн-функций. — М.: Наука,1980.

[37] Загаллер В.А. Теория огибающих. — М.: Наука. 1975. — 104 с.

[38] Ильин А.М. Разностная схема для дифференциального уравнения с малым параметром при старшей производной // Мат. заметки. — 1969. — Т. 6, вып. 2. — С. 237-248.

[39] Калмыков С.А., Шокин Ю.И., Юлдашев З.Х. Методы интервального анализа. — Новосибирск: Наука, 1986. — 224 с.

[40] Канторович Л. В., Акилов Г. П. Функциональный анализ. — М.: Наука, 1984.

[41] Клатте Р., Кулиш У., Неага М., Рац Д., Уллрих Х. PASCAL–XSC. Язык численного программирования. — М.: ДМК Пресс, 2000. — 352 с.

[42] Коллатц Л. Функциональный анализ и вычислительная математика. — М.: Мир, 1969

[43] Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. — М.: Наука, 1976.

[44] Курпель Н.С., Шувар Б.А. Двусторонние неравенства и их приложения. — Киев: Наук. думка. 1980. — 268 с.

[45] Ладыженская О.А., Уральцева Н.Н. Линейные и квазилинейные уравнения эллиптического типа. — М.: Наука, 1973.

[46] Ладыженская О.А., Солонников В.А., Уральцева Н.Н. Линейные и квазилинейные уравнения параболического типа. — М.: Наука, 1967.

[47] Лисейкин В. Д. О численном решении обыкновенного дифференциального уравнения второго порядка с малым параметром при старшей производной // Числ. методы механики сплошн. среды. — 1982. — Т. 13, №3. — С. 98-106.

[48] Лозинский С. М. Оценка погрешности приближенного решения системы обыкновенных дифференциальных уравнений // Докл. АН СССР. — 1953. — Т. 92, №2. — С. 225-228.

[49] Лозинский С. М. О приближенном решении систем обыкновенных дифференциальных уравнений // Докл. АН СССР. — 1954. — Т. 97, №1. — С. 29-32.

[50] Лозинский С. М. Недостаточные и избыточные методы численного интегрирования обыкновенных дифференциальных уравнений // Вестн. Ленингр. ун-та. — 1967. — №7. — С. 74-86.

[51] Марчук Г.И. Методы вычислительной математики. — М.: Наука, 1980.

[52] Марчук Г.И., Агошков В.И. Введение в проекционно-сеточные методы. — М.: Наука, 1981.

[53] Марчук Г.И., Шайдуров В.В. Повышение точности решений разностных схем. — М.: Наука, 1980.

[54] Назаренко Т. И., Марченко Л. В. Введение в интервальные методы вычислительной математики. — Иркутск: Изд-во Иркут. ун-та, 1982.

[55] Оганесян Л. А., Руховец Л. А. Вариационно-разностные методы решения эллиптических уравнений. — Ереван: Изд-во АН Арм-ССР, 1979.

[56] Ремез Е. Я. Деякі способи чисельноі интеграціі дифференціальних рівняиь з оціпкою границ допущеноі похибки // Зап. Природничо-Технічного відділу АН УРСР. — 1931. — №1. С.1-38.

[57] Ремез Е. Я. Некоторые вопросы структуры формул механических квадратур, могущих служить для двусторонней численной оценки решений дифференциальных уравнений // Укр. мат. журн. — 1958. — Т. 10, №4. — С. 413-418.

[58] Салихов Н. П. О полярных методах решения задачи Коши для систем обыкновенных дифференциальных уравнений // Журн. вычисл. математики и мат. физики. — 1962. — Т. 2, №4. — С. 515-528.

[59] Самарский А.А. Теория разностных схем. — М.: Наука, 1977.

[60] Самарский А.А., Андреев В.Б. Разностные методы для эллиптических уравнений. — М.: Наука, 1976.

[61] Самарский А. А., Лазаров Р.Д., Макаров В.Л. Разностные схемы для дифференциальных уравнений с обобщенными решениями: Учеб. пособие для ун-тов. — М.: Высш. шк., 1987. — 296 с.

[62] Самарский А. А., Николаев Е. С. Методы решения сеточных уравнении. — М.: Наука. 1978.

[63] Стренг Г., Фикс Дж. Теория метода конечных элементов. — М.: Мир. 1977.

[64] Чаплыгин С. А. Новый метод приближенного интегрирования дифференциальных уравнений. — М.; Л.: Гостехиздат, 1950.

[65] Черноусько Ф.Л. Оценивание фазового состояния динамических систем. Метод эллипсоидов. — М.: Наука, 1988. — 320 с.

[66] Шайдуров В. В. Многосеточные методы конечных элементов. — М.: Наука, 1989.

[67] Шокин Ю.И. Интервальный анализ. — Новосибирск: Наука, 1981.

[68] Adams,
E., Spreuer, H. Konvergente numerische Schrankenkonstruktion mit Spline-funktionen für nichtlinear gewönliche bzv. Lineare Parabolische Randwertaufgaben // Lecture Notes in Computer Science. 1975. V.29. S. 118-126.

[69] Applet, W. Fehlereinschliessung für die Lösung einer Klasse elliptischer Randwertaufgaben // Z. Angew. Math. Mech. 1974. Bd 54 S. T207-T208.

[70] Barth, W., Nuding, E. Optimale Lösung von Intervallgleichugssystemen // Computting. 1974. V.12. P. 117-125.

[71] Bauch, H. Zur Lösungseinschliessung bei Anfangwertaufraben gewöhnlicher Differentialgleichungen nach Defektmethode //Z. Angew. Math. Mech. 1977. Bd 57. S. 387-396.

[72] Beeck H. Über die Struktur und Abschätzungen der Lösungsmenge von linearen Gleichungssystemen mit Intervallkoeffizienten // Computing. 1972. Vol. 10. P.231–244.

[73] Beierbaum, F., Schwierz, K.P. A bibliography on interval mathematics // J. Comput. Appl. Math. V. 4, N 1. P.59-86.

[74] Brouwer L.E.J. Über die Abbidildung von Mannigfatigkeiten, // Math. Ann., 1912, 71, 97-115.

[75] Caprani, O., Madsen, K. Mean value forms in interval analysis // Computing. 1980. V.25. P.147-154.

[76] Ciarlet P.G., Schultz M.N., Varga R.S. Numerical methods of high-oder accuracy for nonlinear boundary value problems. V. Monotone operator theory. // Numer. Math., 1969, 13, 51-77

[77] Dobronets B.S. Two-sided solution of ODE's via a posteriori error estimates // J. of Comp. and Appl. Math. v.23 1988 p.53–61.

[78] Dobronets B.S. Interval Methods based on a posteriori error estimates. // Interval Computations №3(5) 1992, C.50–55.

[79] Dobronets B.S. Numerical Methods using Defects. Reliable Computing 1(4), (1995), 383-391.

[80] B.S.Dobronets, R.B.Kearfott, L.V.Kuprianova, A.G.Yakovlev, V.S.Zyzin Bibliography of works on interval computations published in russion // Interval Computations, Suppl. 1, 1994

[81] Faas, E. Belienbig genaue numersche Schranken für die Lösung Parabolischer Randwertaufgaben // Ibid. P.178-183.

[82] Gay D. Solving Interwal Linear Equations, SIAM J. Numer. Anal., 19, 858 - 870,(1980).

[83] Hansen E. On linear algebraic equations with interval coefficient // Topics in interval analysis. Oxford: Clarendon Press, 1969. pp. 35-46.

[84] Hansen E. On solving two-point boundary value problems using interval arithmetic // Topics in interval analysis. - Oxford: Clarendon Press, 1969. pp.74-90.

[85] Krawczyk R. Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. Computing, 1969, V. 4, p. 187–201.

[86] Krückberg, F. Partial differential equations // Topics in interval analysis. Oxford: Clarendon Press, 1969. P.98-101.

[87] Lohner R. Enclosing the Solutions of Ordinary Initial and Boundary Value Problems // E. Kausher, U. Kulisch, Ch. Ullrich (eds.): Computer Arithmetic, B.G. Teubner, 1987.

[88] Moore R.E.Automatic local coordinate transformations to reduce the growth of error bounds in interval computation of solutions of ordinary differential equations // Rall L.E. (ed), Error in Digital Computation. Vol.II,Wiley & Sons Inc.,New York 1965, pp.103-140.

[89] Moore R.E. Interval analysis. Englewood Cliffs. N.J.:Prentice-Hall, 1966.

[90] Moore R.E. Methods and Applications of Interval Analysis, SIAM, Philadelrhia 1979.

[91] Moore R.E., Kioustelidis J.B. A simple test for accyracy of approximate solutions to nonlinear (or linear) systems, SIAM J. Numer. Anal., 17, 521 - 529, 1980.

[92] Neumaier, A. Interval methods for systems of equations. Cambidge: Cambidge University Press, 1990.

[93] Nickel K. Bounds for the Set of solutions of Functional-Differential Equations, MRC Techn. Summary Report N 1782, Univ. of Wisconsin, Madison 1977.

[94] Nickel K.L.E. Using Interval Methods for the Numerical Solution of ODE's // ZAMM, 1986, Vol. 66, N 11, 513-523.

[95] Oettli W., Prager W. Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides // Numerische Mathematik. 1964. Vol. 6. P.405–409.

[96] Oliveria, F.A. Interval analysis and two-point boundary value problems //SIAM J. Numer. Anal. -1974. -V.11. -P.382-391.

[97] Ratschek H. Schröder G. Über die Adleitung von Intervallwertigen funktionen. Computing 7, 172-187 (1971)

[98] Rogalev A. N., Solving Systems of Ordinary Differential Equations with Interval Data: Rigorous and Optimal Bounds, IMACS/GAMM International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics, Sept. 26-29, 1995.- Bergische Universität Gesamthochschule Wuppertal , Fachbereich Mathematik und Institut für Angewandte Informatic (Germany) . - p.113-114.

[99] Schröder G. Differention of interval functions. Proc. Amer. Math. Soc. 36, 485-490 (1972)

[100] Schröder, J. Upper and lower bounds for solutions of generalized two-point boundary value problem // Numer. Math. 1975. V.23. P.433-457.

[101] Shary S. P. On optimal solution of interval linear equations // SIAM J. Numer. Anal., 32 (1995), pp. 610–630.

[102] Streng G., Fix G. J. An analysis of the finite element method. – Englewood Cliffs: Prentice-Hall, 1973.

[103] Ström, T. Trict estimation of the maximum of a function of one variable //BIT. 1971. V.11. P.199-211.

[104] Tost, R. Zur numerischen Lösungen von Randwertaufgabe mit Gesicherter Fehlereinschliessung bei Partiellen Differentialgleichungen // Z. Angew. Math. Mech. 1971. Bd 51. S. T74-T75.

[105] Walzel A. Fehlerabschatzung bei Anfangswertaufgaben fur Systeme von gewohnlichen Differentialgleichungen. Diss., Univ. Koln 1969.

Предметный указатель

Оглавление

Учебное издание
Добронец Борис Станиславович
Интервальная математика

Редактор И. А. Вейсиг
Корректор Т. Е. Бастрыгина

Подписано в печать 30.06.2004 г. Формат 60×84/16
Бумага тип. Печать офсетная Усл. печ. л. 12.5
Тираж экз. Заказ

Издательский центр Красноярского государственного университета
660041 г. Красноярск, пр. Свободный, 79