И.А. БЕДАРЕВ
Н.Н. ФЕДОРОВА
И.А. ФЕДОРЧЕНКО

КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ В ЗАДАЧАХ СТРОИТЕЛЬСТВА

НОВОСИБИРСК 2012

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ
АРХИТЕКТУРНО-СТРОИТЕЛЬНЫЙ УНИВЕРСИТЕТ
(СИБСТРИН)

И.А. Бедарев, Н.Н. Федорова, И.А. Федорченко

КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ В ЗАДАЧАХ СТРОИТЕЛЬСТВА

Учебное пособие

НОВОСИБИРСК 2012

УДК 004+69
ББК 38+32.97
Б 38

Бедарев И. А.

Компьютерное моделирование в задачах строительства : учебное пособие / И. А. Бедарев, Н. Н. Федорова, И. А. Федорченко ; Новосиб. гос. архитектур.-строит. ун-т (Сибстрин). – Новосибирск : НГАСУ (Сибстрин), 2012. – 152 с.

ISBN 978-5-7795-0595-6

Учебное пособие предназначено для использования в учебном процессе магистрантами НГАСУ (Сибстрин), обучающимся по направлению 270800.68 «Строительство». Основная цель пособия – дать представление о математических моделях, возникающих в процессе исследовательской и проектной деятельности; научить подбирать и модифицировать методы прикладной математики для решения поставленной задачи и анализировать полученное решение. В пособии приведены сведения из линейной алгебры, функционального анализа, дифференциальных уравнений, уравнений в частных производных и численных методов, которые необходимы для понимания методов прикладной математики и активного их использования при решении задач строительства.

Печатается по решению издательско-библиотечного совета
НГАСУ (Сибстрин)

Рецензенты:
- Ю.Е. Воскобойников, д-р физ.-мат. наук, профессор, зав. кафедрой прикладной математики
- НГАСУ (Сибстрин);
- О.Б. Ковалев, д-р физ.-мат. наук, профессор, зав. лабораторией ИТПМ СО РАН

ISBN 978-5-7795-0595-6

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 8

РАЗДЕЛ 1. СВЕДЕНИЯ ИЗ АЛГЕБРЫ И ФУНКЦИОНАЛЬНОГО АНАЛИЗА

Тема 1. Погрешности приближенного решения
1.1. Источники погрешностей 10
1.2. Погрешности арифметических операций 10
1.3. Погрешности машинной арифметики 12
Тема 2. Основные понятия алгебры и функционального анализа
2.1. Метрика, метрическое пространство 14
2.2. Линейное пространство, базис 15
2.3. Норма, нормированное пространство 18
2.4. Скалярное произведение, гильбертово пространство 19
2.5. Углы между векторами, теорема о разложении 20
2.6. Ортогональный базис. Ортогонализация 21
Тема 3. Линейные операторы, матрицы и их спектр
3.1. Линейные операторы 23
3.2. Матрицы и операции над ними 26
3.3. Преобразование матрицы при переходе к новому базису 28
3.4. Собственные значения и собственные векторы 29
3.5. Приведение матрицы к диагональному виду 33
3.6. Сопряженный оператор. Квадратичная форма 36
Тема 4. Системы линейных алгебраических уравнений 38
4.1. Точные методы решения СЛАУ 40
4.2. Итерационные методы решения СЛАУ 46

РАЗДЕЛ 2. ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

Тема 5. Численное интегрирование и дифференцирование
5.1. Постановка задачи численного интегрирования 52
5.2. Формула прямоугольников 53
5.3. Формула трапеций 57
5.4. Формула Симпсона 59
5.5. Численное дифференцирование 61
5.6. Метод неопределенных коэффициентов 65
Тема 6. Численные методы решения задачи Коши для обыкновенных дифференциальных уравнений
6.1. Постановка задачи 67
6.2. Метод Эйлера 68
6.3. Модифицированный метод Эйлера 70
6.4. Методы Рунге – Кутты 70

6.5. Методы приближенного решения задачи Коши для системы ОДУ и ОДУ высших порядков ... 72
6.6. Жесткие ОДУ ... 76
Тема 7. Краевая задача для ОДУ второго порядка ... 78
7.1. Конечно-разностный метод ... 79
7.2. Метод стрельбы ... 81
7.3. Метод коллокаций ... 82
7.4. Вариационные методы ... 83
7.5. Проекционные методы ... 85
7.6. Метод конечных элементов ... 87
РАЗДЕЛ 3. УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ
Тема 8. Общие сведения об уравнениях в частных производных
8.1. Постановка задачи ... 91
8.2. Характеристики. Типы уравнений ... 92
8.3. Приближенные методы. Аппроксимация и устойчивость ... 93
Тема 9. Уравнения теплопроводности
9.1. Одномерное уравнение теплопроводности ... 95
9.2. Двумерное уравнение теплопроводности ... 106
Тема 10. Уравнение Пуассона ... 110
10.1. Метод установления ... 111
10.2. Итерационные методы ... 111
Тема 11. Уравнения гиперболического типа
11.1. Характеристики ... 114
11.2. Линейное уравнение переноса ... 114
11.3. Разрывные решения ... 121
11.4. Первое дифференциальное приближение ... 122
11.5. Монотонность численного решения ... 125
11.6. Схемы высокого порядка ... 125
11.7. Уравнение Хопфа. Градиентная катастрофа ... 127
11.8. Волновое уравнение ... 135
Тема 12. Современные пакеты прикладных программ для решения задач строительства ... 140
12.1. Базовые сведения ... 141
12.2. Основной интерфейс и запуск Workbench ... 144
12.3. Построение геометрии ... 145
12.4. Нагрузки и закрепления ... 149
12.5. Просмотр и анализ результатов расчетов ... 151
ЗАКЛЮЧЕНИЕ ... 153
БИБЛИОГРАФИЧЕСКИЙ СПИСОК ... 154

ВВЕДЕНИЕ

В настоящее время компьютерное моделирование играет важную роль в проектировании технических изделий. Это связано как с усложнением объектов исследования, так и с ростом возможностей ЭВМ. Распространенное мнение о всемогуществе современных ЭВМ часто порождает впечатление, что математики избавились от всех хлопот, связанных с решением задач, и разработка новых методов для их решения не требуется. В действительности это не так, поскольку все время появляются новые более сложные задачи, начинают решаться комплексные задачи, математическое моделирование проникает в те области, где раньше его не было, например, в биологию, медицину, социологию и другие сферы.

Математическое моделирование – это описание на абстрактном математическом языке различных явлений и процессов. Математическая модель объекта или явления – набор формул, таблиц, уравнений, описывающих поведение этого объекта или явления. **Этапы** математического моделирования:

1) уточнение наиболее существенных фактов, свойств описываемого объекта или явления;
2) построение математической модели – системы уравнений: алгебраических, функциональных, в частных производных;
3) нахождение решений, точно или с помощью приближенных методов, которые специально разрабатывают для решения той или иной задачи;
4) проверка адекватности модели: соответствие полученных решений основным существенным фактам, экспериментам, если необходимо – уточнение модели, т.е. введение в модель новых уравнений или добавление новых слагаемых в существующие уравнения;
5) параметрические исследования на основе построенной модели; получение новых сведений об объекте.

Цель настоящего курса – дать представление о математических основах методов, используемых для научных и инженерных расчетов. Разработкой таким методов занимается прикладная математика, которая тесно связана с другими разделами ма-

тематики, такими как алгебра, математический и функциональный анализ, теория вероятностей, дифференциальные уравнения, теория функций комплексного переменного и уравнения в частных производных. Это значит, что специалист, который использует методы прикладной математики для решения задач в своей области науки или техники, должен быть знаком со всеми этими разделами.

Учебное пособие предназначено для использования в учебном процессе магистрантами НГАСУ (Сибстрин), обучающимися по направлению 270800.68 «Строительство», курс М.1.Б.1.3 «Специальные разделы высшей математики». Основная цель этого курса – дать представление о многообразии математических моделей и методов, возникающих в процессе научно-исследовательской и проектной деятельности в области строительства; научить подбирать и модифицировать методы прикладной математики для решения поставленной задачи из предметной области и анализировать полученное решение. Данный курс является неотъемлемой частью математической подготовки в соответствии с требованиями, отраженными в федеральном государственном стандарте.

В пособии приведены сведения из линейной алгебры, функционального анализа, дифференциальных уравнений, уравнений в частных производных и численных методов, необходимые для понимания и активного освоения методов прикладной математики, которые рассматриваются, в частности, в рамках последующего курса М.2.Б.1.3 «Методы решения задач строительства».

Пособие состоит из двенадцати тем, организованных в три раздела, введения и библиографического списка. Темы, составляющие разделы, пронумерованы по порядку в пределах пособия. Формулы, рисунки и примеры имеют двойную нумерацию с учетом номера раздела.

РАЗДЕЛ 1. СВЕДЕНИЯ ИЗ АЛГЕБРЫ И ФУНКЦИОНАЛЬНОГО АНАЛИЗА

Тема 1. Погрешности приближенного решения

Любой приближенный метод нуждается в оценке погрешности и понимании того, как она влияет на конечный результат. При численном решении прикладных задач неизбежно появление ошибок или **погрешностей**.

1.1. Источники погрешностей

Общая погрешность включает в себя:

1) **П1** – погрешности задачи, связанные с приближенным характером исходной модели, невозможностью учесть все факторы, упрощением исходной математической модели, неточностью измерений. Эти погрешности относятся к неустранимым;
2) **П2** – погрешности метода, связанные со способом решения поставленной математической задачи и появлением в результате подмены исходной нелинейной модели последовательности более простых (линейных) моделей. Эта погрешность является устранимой, поскольку при разработке численных методов она отслеживается и доводится до сколь угодно малого уровня;
3) **П3** – погрешность округлений, обусловленная необходимостью выполнять арифметические операции над числами на ЭВМ, которая требует их усечения до определенного количества разрядов.

Общая погрешность задачи есть сумма всех погрешностей:

П = П1 + П2 + П3.

1.2. Погрешности арифметических операций

Пусть $A, \tilde{A}$ – два близких числа (точное и приближение), тогда $\Delta A = \left|A - \tilde{A}\right|$ – абсолютная погрешность, $\delta A = \dfrac{\Delta A}{|A|}$ – отно-

сительная погрешность, $\Delta_A \geq \Delta A$ и $\delta_A \geq \delta A$ – оценки (границы) погрешностей. Зачастую найти ΔA, δA трудно, но можно оценить Δ_A и δ_A.

Пусть известны Δ_x, Δ_y – абсолютные оценки погрешностей чисел *x, y* соответственно. Можно показать, что при сложении, вычитании, умножении и делении чисел абсолютные погрешности складываются, т.е.

$$\Delta_{x\pm y} = \Delta_x + \Delta_y; \quad \Delta_{x\cdot y} = \Delta_x + \Delta_y; \quad \Delta_{x/y} = \Delta_x + \Delta_y.$$

Оценим относительную погрешность суммы:

$$\delta(x+y) = \frac{\Delta(x+y)}{|x+y|} \leq \frac{\Delta_{x+y}}{|x+y|} = \frac{\Delta_x + \Delta_y}{|x+y|} \leq \frac{x\delta_x + y\delta_y}{|x+y|} \leq \frac{\delta(x+y)}{|x+y|} = \delta^*,$$

где $\delta^* = \max(\delta x, \delta y)$, следовательно, суммарная относительная погрешность не превосходит наибольшей относительной погрешности слагаемых.

С вычитанием дело обстоит хуже:

$$\delta(x-y) \leq \frac{\Delta_{x-y}}{|x-y|},$$

и поскольку *x, y* могут быть близкими, то в знаменателе появляется малое число, что приводит к большой погрешности. Это явление называют потерей точности при вычитании близких чисел.

Часто возникает **обратная задача** теории погрешностей: какую точность должны иметь исходные данные, чтобы на выходе получить результат заданной точности.

При больших количествах однотипных вычислений вступают в силу **статистические законы** формирования погрешностей. Можно показать, что математическое ожидание абсолютной погрешности суммы *n* слагаемых с одинаковым уровнем абсолютных погрешностей при достаточно большом *n* пропорционально $\sqrt{n}$, т.е. арифметическое усреднение результатов измерений увеличивает точность. Прямое применение вероятностно-статистических оценок весьма сложно, и часто их заменяют простыми правилами действий над приближенными числами

(**технический подход**), впервые предложенными русским математиком и механиком А.Н. Крыловым:

1) приближенное число должно записываться так, чтобы в нем все значащие цифры, кроме последней, были верными;
2) при сложении и вычитании приближенных чисел в результате следует сохранять столько десятичных знаков, сколько их было в слагаемом с наименьшим количеством десятичных знаков после запятой;
3) при умножении и делении в результате следует сохранять столько знаков, сколько их было в умножаемом с наименьшим числом значащих цифр;
4) все результаты промежуточных вычислений должны иметь один-два запасных знака.

Таким образом, при техническом подходе к учету погрешности приближенных вычислений предполагается, что в самой записи чисел содержится информация о его точности. Прямая выгода от применения приведенных правил может быть получена лишь при ручном счете, однако их знание помогает интерпретировать компьютерные расчеты, а иногда и правильно их организовать.

1.3. Погрешности машинной арифметики

При проведении расчетов на ЭВМ нужно иметь представление о **погрешностях машинной арифметики**. В основу запоминающего устройства компьютера положены однотипные физические устройства, имеющие r устойчивых состояний. Как правило, r есть степень числа 2: $r = 2, 4, 8, 16$ и т.д. Каждому арифметическому числу ставится в соответствие одинаковое количество таких элементов (машинное слово), и с помощью более простых элементов задается знак числа. Упорядоченные элементы образуют разрядную сетку машинного слова, в каждом r-м элементе можно записать одно из базисных чисел: $0, 1, \ldots, r - 1$.

Для представления чисел в компьютере используется два способа: с фиксированной и плавающей запятой. При записи числа **с фиксированной запятой** кроме упомянутых r базисных чисел и длины машинного слова k указывается также l – количе-

ство разрядов, выделяемых под дробную часть. Таким образом, любое вещественное число отображается с помощью конечной последовательности

$$a_{fix} = \alpha_1\alpha_2..\alpha_{k-i}\cdot\alpha_{k-i-1}..\alpha_k =$$

$$= \alpha_1 r^{k-i-1} + \alpha_2 r^{k-i-2} + ... + \alpha_{k-l-1} r^1 + \alpha_{k-l}\ r^0 + \alpha_{k-l+1} r^{-1} + ... + \alpha_k r^{-l}.$$

Диапазон представляемых таким образом чисел определяется положительным и отрицательным числами с наибольшими цифрами, равными $(r-1)$ во всех $(k-l)$ разрядах до запятой, а абсолютная точность такого представления зависит от порядка первого отброшенного после запятой числа, т.е. $r^{-(l+1)}$.

В основе более распространенного представления вещественного числа **с плавающей запятой** лежит экспоненциальная форма записи:

$$a_{float} = M \cdot r^p,$$

где r – основание, p – порядок, а $r^{-1} < M < 1$ – мантисса. Если под мантиссу выделяется l, а под порядок m элементов длины r, то машинное слово в такой системе имеет следующую структуру:

Знак порядка	Порядок	Знак мантиссы	Мантисса
	m разрядов		l разрядов

Любое вещественное число в этой форме записи представляется как $$a_{float} = \pm(\beta_1 r^{-1} + \beta_2 r^{-2} + ... + \beta_l r^{-l}) \cdot r^{\gamma},$$

где γ – целое число из промежутка $[-r^m, r^m - 1]$. Следовательно, все вещественные числа с плавающей запятой лежат в диапазоне $[-r^{r^m}, r^{r^m}]$. Для положительных чисел можно определить машинный нуль r^{-r^m} и машинную бесконечность r^{r^m-1}. Типичная ошибка, возникающая при расчетах на ЭВМ, – это так называемое переполнение (*overflow*), означающее, что в процессе расчета возникли числа, выходящие за границы машинного нуля или машинной бесконечности.

В современных персональных компьютерах на базе процессоров *Intel* в соответствии со стандартом *IEEE754* предусматривается существование двух двоичных форматов с плавающей

запятой: с одинарной (*single*) и двойной (*double*) точностью. В числах с одинарной точностью под мантиссу выделяется 24 разряда, а с двойной точностью – 53 разряда. Таким образом, для одинарной точности машинный нуль равен $M_0 \approx 10^{-38}$, а машинная бесконечность – $M_\infty \approx 10^{38}$. Для двойной точности эти значения равны 10^{-308} и 10^{308} соответственно.

В разных языках программирования этим представлениям соответствуют разные типы данных. Например, в языке Pascal это *real* и *extended*, в Си – *float* и *double*, в Fortran – *real* и *double precision*.

Следует помнить, что при выполнении на ЭВМ арифметических операций с числами, лежащими на границе точности, можно получить нарушение привычных законов ассоциативности и дистрибутивности. Поэтому надо применять алгоритмы расчета, которые не выводят результаты за границы машинного нуля и машинной бесконечности.

Тема 2. Основные понятия алгебры и функционального анализа

2.1. Метрика, метрическое пространство

Пусть **X** – множество элементов произвольной природы, объединенных по какому-то признаку. Множество **X** назовем **метрическим пространством**, если любой паре его элементов $x, y \in \mathbf{X}$ сопоставляется вещественное число $\rho(x, y)$, называемое расстоянием между x и y, или **метрикой**, которое удовлетворяет следующим трем аксиомам:

1. $\rho(x, y) \geq 0$, причем из того, что $\rho(x, y) = 0$, следует, что $x = y$.
2. $\rho(x, y) = \rho(y, x)$.
3. $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$, где $z \in \mathbf{X}$.

Последняя аксиома называется неравенством треугольника, поскольку отражает свойство сторон треугольника на плоскости: сумма длин любых двух сторон не может быть меньше длины третьей стороны. В одном множестве метрика может быть задана различными способами. Две метрики

$\rho_1(x, y)$ и $\rho_2(x, y)$ называются эквивалентными, если $c_1\rho_1(x, y) \le \rho_2(x, y) \le c_2\rho_1(x, y) \quad \forall x,\ y \in \mathbf{X}$.

Элемент x метрического пространства называется пределом бесконечной последовательности $\{x_n\}_{n\to\infty}$, если $\rho(x_n, x)_{n\to\infty} \to 0$. В этом случае говорят, что последовательность $\{x_n\}$ сходится по метрике пространства **X**.

Примеры метрических пространств:

1) множество вещественных чисел $\mathbf{X} = \mathbf{R}$ с метрикой $\rho(x, y) = |x - y|$;
2) $\mathbf{X} = \mathbf{R}^2$, элементами которого являются пары вещественных чисел $\bar{x} = (x_1, x_2)$, $\bar{y} = (y_1, y_2)$ и т.д.

   Можно определить несколько метрик, например,

   $\rho(\bar{x}, \bar{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$ (евклидова метрика),

   $\rho(\bar{x}, \bar{y}) = \max\left(|x_1 - y_1|, |x_2 - y_2|\right)$,

   $\rho(\bar{x}, \bar{y}) = |x_1 - y_1| + |x_2 - y_2|$.

   Аналогично можно определить конечномерные пространства $\mathbf{R}^3$, $\mathbf{R}^4$, …, $\mathbf{R}^n$, для которых обобщаются все приведенные выше метрики;
3) множество всех вещественных функций, непрерывных на отрезке $[a, b]$, для которого введены:
   - равномерная метрика, порождающая на **X** пространство $C_1[a, b]$: $\rho(x(t), y(t)) = \max\limits_{t\in[a,b]} |x(t) - y(t)|$;
   - среднеквадратичная метрика, порождающая на **X** пространство $L_2[a, b]$: $\rho(x(t), y(t)) = \sqrt{\int\limits_a^b [x(t) - y(t)]^2\, dt}$.

2.2. Линейное пространство, базис

Пусть **X** – множество каких-либо элементов; **R** – множество вещественных чисел. Определим над элементами $\bar{a}, \bar{b} \in \mathbf{X}$, $\alpha \in \mathbf{R}$, операции сложения и умножения на вещественное число:

$$\bar{a} + \bar{b} = \bar{c}, \quad \bar{c} \in \mathbf{X},\ \alpha \cdot \bar{a} = \bar{b} \in \mathbf{X}.$$

Пространство **X** называется **линейным**, а его элементы $\bar{a}, \bar{b}$ называются **векторами**, если $\forall \bar{a}, \bar{b}, \bar{c} \in \mathbf{X}$ справедливы аксиомы:

1) коммутативности сложения $\bar{a} + \bar{b} = \bar{b} + \bar{a}$;
2) ассоциативности сложения $\bar{a} + \left(\bar{b} + \bar{c}\right) = \left(\bar{a} + \bar{b}\right) + \bar{c}$;
3) существования нулевого элемента $\bar{0} \in \mathbf{X}$, такого, что $\forall \bar{a} \in \mathbf{X}$ справедливо $\bar{0} + \bar{a} = \bar{a}$;
4) существования обратного элемента $-\bar{a} \in \mathbf{X}$, такого, что $-\bar{a} + \bar{a} = \bar{0}$;
5) ассоциативности умножения на скаляр $\left(\alpha \cdot \beta\right) \cdot \bar{a} = \alpha \cdot \left(\beta \cdot \bar{a}\right)$;
6) дистрибутивности $\alpha \cdot \left(\bar{a} + \bar{b}\right) = \alpha \cdot \bar{a} + \alpha \cdot \bar{b}$, $\left(\alpha + \beta\right) \cdot \bar{a} = \alpha \cdot \bar{a} + \beta \cdot \bar{a}$.

Подмножество **Y** линейного пространства **X** называется **линейным многообразием**. Будем говорить, что многообразие **Y** замкнуто, если

$\forall \bar{a}, \bar{b} \in \mathbf{Y}, \quad \bar{c} = \bar{a} + \bar{b} \in \mathbf{Y}$ и $\forall \alpha \in \mathbf{R}, \quad \bar{b} = \alpha \cdot \bar{a} \in \mathbf{Y}$.

Замкнутое линейное многообразие называется **подпространством**. Примерами подпространств являются плоскость в трехмерном пространстве, линия на плоскости и т.д.

Система векторов $\bar{a}_1, \bar{a}_2, ..., \bar{a}_n \in \mathbf{X}$ называется **линейно независимой**, если из того, что

$$\alpha_1 \bar{a}_1 + \alpha_2 \bar{a}_2 + ... + \alpha_n \bar{a}_n = \bar{0}$$

следует, что $\alpha_1 = \alpha_2 = ... = \alpha_n = 0$. В противном случае векторы называются **линейно зависимыми**, и один из них может быть представлен в виде линейной комбинации остальных. Действительно, пусть $\alpha_1 \neq 0$, тогда вектор $\bar{a}_1$ можно представить в виде линейной комбинации $\bar{a}_2, ..., \bar{a}_n$:

$$\bar{a}_1 = \left(-\frac{\alpha_2}{\alpha_1}\right)\bar{a}_2 + \left(-\frac{\alpha_3}{\alpha_1}\right)\bar{a}_3 + ... + \left(-\frac{\alpha_n}{\alpha_1}\right)\bar{a}_n = \sum_{i=2}^{n} \lambda_i \bar{a}_i \, .$$

Базисом линейного многообразия **X** называется множество линейно независимых векторов $\bar{e}_1, \bar{e}_2, ..., \bar{e}_n \in \mathbf{X}$, таких, что каждый вектор $\bar{x} \in \mathbf{X}$ может быть выражен в виде линейной комбинации векторов базиса $\bar{x} = \alpha_1 \bar{e}_1 + \alpha_2 \bar{e}_2 + ... + \alpha_n \bar{e}_n$. Вещественные числа $\alpha_1, \alpha_2, ..., \alpha_n$ называют координатами вектора $\bar{x}$ в базисе $\bar{e}_1, \bar{e}_2, ..., \bar{e}_n$.

Говорят, что многообразие **X** порождено векторами базиса $\bar{e}_1, \bar{e}_2, ..., \bar{e}_n$, n – размерность многообразия. В конечномерном линейном многообразии или векторном пространстве, порожденном n базисными векторами:

- каждое множество n линейно независимых векторов является базисом;
- всякое множество из $m < n$ векторов не является базисом;
- каждое множество из $m > n$ векторов линейно зависимо.

Примеры базисов в пространстве $\mathbf{R}^2$:

- $\bar{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \bar{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

  Проверим, что векторы линейно независимы:

  $$\alpha_1 \bar{e}_1 + \alpha_2 \bar{e}_2 = \bar{0} \Rightarrow \alpha_1 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \alpha_2 \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

  $$\begin{cases} 0 \cdot \alpha_1 + 1 \cdot \alpha_2 = 0 \\ 1 \cdot \alpha_1 + 0 \cdot \alpha_2 = 0 \end{cases} \Rightarrow \alpha_1 = \alpha_2 = 0.$$

- $\bar{g}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \bar{g}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Проверка линейной независимости:

  $$\alpha_1 \bar{g}_1 + \alpha_2 \bar{g}_2 = \bar{0} \Rightarrow \alpha_1 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \alpha_2 \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

  $$\begin{cases} 1 \cdot \alpha_1 + 0 \cdot \alpha_2 = 0 \\ 1 \cdot \alpha_1 + 1 \cdot \alpha_2 = 0 \end{cases} \Rightarrow \alpha_1 = \alpha_2 = 0.$$

2.3. Норма, нормированное пространство

Линейное пространство **X** называется **нормированным**, если любому его вектору $\bar{x}$ поставлено в соответствие вещественное число $\|\bar{x}\|$, называемое **нормой**, которое удовлетворяет следующим свойствам:

1) $\|\bar{x}\| \geq 0 \; \forall \bar{x} \in \mathbf{X}$, и если $\|\bar{x}\| = 0$, то $\bar{x} = \bar{0}$;
2) $\|\alpha \cdot \bar{x}\| = |\alpha| \cdot \|\bar{x}\| \; \forall \alpha \in \mathbf{R}$;
3) $\|\bar{x} + \bar{y}\| \leq \|\bar{x}\| + \|\bar{y}\| \quad \forall \bar{x}, \bar{y} \in \mathbf{X}$.

Как и метрика, норма в каждом пространстве не единственна. Можно доказать, что всякое нормированное пространство является метрическим. Действительно, пусть $\rho(\bar{x}, \bar{y}) = \|\bar{x} - \bar{y}\|$. Покажем, что выполнены все аксиомы метрики:

1) $\rho(\bar{x}, \bar{y}) = \|\bar{x} - \bar{y}\| \geq 0$ по свойству нормы.

Если $\rho(\bar{x}, \bar{y}) = 0$, то $\|\bar{x} - \bar{y}\| = 0 \Rightarrow \bar{x} - \bar{y} = \bar{0} \Rightarrow \bar{x} = \bar{y}$;

2) $\rho(\bar{x}, \bar{y}) = \|\bar{x} - \bar{y}\| = \|(-1) \cdot (\bar{y} - \bar{x})\| =$

$= |-1| \cdot \|\bar{y} - \bar{x}\| = \|\bar{y} - \bar{x}\| = \rho(\bar{y}, \bar{x})$;

3) $\rho(\bar{x}, \bar{y}) = \|\bar{x} - \bar{z} + \bar{z} - \bar{y}\| \leq \|\bar{x} - \bar{z}\| + \|\bar{z} - \bar{y}\| = \rho(\bar{x}, \bar{z}) + \rho(\bar{z}, \bar{y})$.

Если метрическое пространство с метрикой ρ линейно, то в нем можно ввести норму равенством $\|\bar{x}\| = \|\bar{x} - \bar{0}\| = \rho(\bar{x}, \bar{0})$, т.е. норма вектора – это расстояние от него до нулевого вектора, которое обязательно присутствует в векторном пространстве.

Примеры нормированных пространств:

1. Пусть $\mathbf{R}^n$ – множество n-мерных векторов с вещественными координатами $\bar{x} = (x_1, x_2, \ldots, x_n)$, для которых норму можно определить как:
   - $\|\bar{x}\|_\infty = \max\limits_{0 \leq i \leq n} |x_i|$;
   - $\|\bar{x}\|_2 = \sqrt{\sum\limits_{i=1}^{n} x_i^{\,2}}$;

- $\|\bar{x}\|_1 = \sum_{i=1}^{n} |x_i|$;
- $\|\bar{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$.

2. В пространстве функций $x(t)$, заданных на отрезке $[a, b]$, норму можно определить как
   - $\|x(t)\|_{C_1} = \max_{t \in [a,b]} |x(t)|$;
   - $\|x(t)\|_{L_2} = \sqrt{\int_a^b (x(t))^2 dt}$.

2.4. Скалярное произведение, гильбертово пространство

Линейное пространство **X** называется предгильбертовым, или пространством со **скалярным произведением**, если любой паре его элементов $\bar{x}, \bar{y}$ можно поставить в соответствие вещественное число, называемое скалярным произведением и обозначаемое $(\bar{x}, \bar{y})$, которое удовлетворяет следующим свойствам:

1) $(\bar{x}, \bar{y}) = (\bar{y}, \bar{x})$;
2) $(\bar{x} + \bar{y}, \bar{z}) = (\bar{x}, \bar{z}) + (\bar{y}, \bar{z}) \quad \forall \bar{z} \in \mathbf{X}$;
3) $(\lambda \cdot \bar{x}, \bar{y}) = \lambda \cdot (\bar{x}, \bar{y}) \quad \forall \lambda \in \mathbf{R}$;
4) $(\bar{x}, \bar{x}) \geq 0$, если $(\bar{x}, \bar{x}) = 0$, то $\bar{x} = \bar{0}$.

Можно определить норму вектора как $\|\bar{x}\| = \sqrt{(\bar{x}, \bar{x})}$ (самостоятельно доказать, что выполняются все аксиомы нормы). В таком случае говорят, что норма *порождена* скалярным произведением. В пространстве со скалярным произведением любые два элемента связаны неравенством **Коши – Буняковского**:

$$|(\bar{x}, \bar{y})|^2 = (\bar{x}, \bar{x}) \cdot (\bar{y}, \bar{y}),$$

из которого следует неравенство **Коши – Шварца**:

$$(\bar{x}, \bar{y}) \leq \|\bar{x}\| \cdot \|\bar{y}\|.$$

В пространствах со скалярным произведением можно определить **сильную** и **слабую** сходимость. Будем говорить, что последовательность $\{\bar{x}_n\}$ пространства **X** сходится к $\bar{x}$ в сильном смысле: $\bar{x}_n \underset{n\to\infty}{\to} \bar{x}$, если она сходится по норме, т.е. $\|\bar{x}_n - \bar{x}\| \underset{n\to\infty}{\to} 0$. Последовательность векторов $\{\bar{x}_n\}$ пространства **X** сходится к $\bar{x}$ слабо: $\bar{x}_n \underset{n\to\infty}{\to} \bar{x}$, если $\forall y \in \mathbf{X}$ $(\bar{x}_n, \bar{y}) \underset{n\to\infty}{\to} (\bar{x}, \bar{y})$. Предгильбертово пространство называется гильбертовым пространством **H**, если предел любой последовательности, сходящейся в сильном смысле, принадлежит **H**, т.е. пространство **H** – полное.

Примеры гильбертовых пространств.

- Евклидово пространство *n*-мерных векторов с вещественными координатами

  $$\bar{x} = (x_1, x_2, \ldots, x_n), \bar{y} = (y_1, y_2, \ldots, y_n), \quad (\bar{x}, \bar{y}) = \sum_{i=1}^{n} x_i y_i.$$

- Пространство бесконечных вещественных последовательностей $\bar{x} = \{x_n\}_{n=1,\ldots,\infty}$, для которых существует $\lim\limits_{n\to\infty} (x_n)^2$. Тогда скалярное произведение можно определить как $(\bar{x}, \bar{y}) = \sum\limits_{i=1}^{\infty} x_i y_i$.

- Пространства вещественных функций $x(t)$, определенных на отрезке [*a*, *b*], для которых существует $\int\limits_a^b (x(t))^2 dt$. Тогда $(x(t), y(t)) = \int\limits_a^b x(t) \cdot y(t)\, dt$.

2.5. Углы между векторами, теорема о разложении

Наличие скалярного произведения позволяет измерять не только расстояния, но и углы между элементами гильбертова

пространства (векторами) и переносить на абстрактные пространства многие геометрические свойства двух- и трехмерных пространств. Определим «косинус» угла между векторами:

$$\cos(\bar{x},\bar{y}) = \frac{(\bar{x},\bar{y})}{\|\bar{x}\|\cdot\|\bar{y}\|} \in [-1,1],$$

и на основе этого введем понятие параллельных и ортогональных векторов. Параллельными будут векторы $\bar{x},\bar{y}$, для которых $\cos(\bar{x},\bar{y}) = \pm 1$. Два вектора $\bar{x},\bar{y} \in \mathbf{H}$ называются ортогональными, т.е. $\bar{x} \perp \bar{y}$, если $(\bar{x},\bar{y}) = 0$. Элемент $\bar{x} \in \mathbf{H}$ ортогонален множеству $\mathbf{M} \subset \mathbf{H}$, т.е. $\bar{x} \perp \mathbf{M}$, если он ортогонален каждому элементу $\bar{y} \in \mathbf{M}$. Тождество

$$\|\bar{x}+\bar{y}\|^2 + \|\bar{x}-\bar{y}\|^2 = 2\|\bar{x}\|^2 + 2\|\bar{y}\|^2$$

называется равенством параллелограмма, поскольку оно аналогично формуле, связывающей сумму длин диагоналей параллелограмма и его периметр.

Теорема о разложении. Пусть $\bar{x} \in \mathbf{H}$ – элемент гильбертово пространства, **M** – подпространство **H**; тогда существует единственная пара векторов $\bar{y}$, $\bar{z}$, где $\bar{y} \in \mathbf{M}$, $\bar{z} \perp \mathbf{M}$, таких, что $\bar{x} = \bar{y} + \bar{z}$ (без доказательства).

Элемент $\bar{y}$ называется проекцией элемента $\bar{x}$ на подпространство **M**, вектор $\bar{z} \in \mathbf{H}$, $\bar{z} \notin \mathbf{M}$ – расстоянием от элемента $\bar{x}$ до множества **M**, множество элементов $\bar{z} \perp \mathbf{M}$ – ортогональным дополнением **M**. Следствием теоремы о разложении является аналог теоремы Пифагора:

$$\|\bar{x}\|^2 = \|\bar{y}\|^2 + \|\bar{z}\|^2.$$

2.6. Ортогональный базис. Ортогонализация

Пусть в гильбертовом пространстве **H** размерности *n* задан базис $\{\bar{e}^i\}_{i=1,\dots,n}$. Любой элемент **H** можно разложить по этому базису:

$$\bar{a},\bar{b}\in\mathbf{H},\quad \bar{a}=\sum_{i=1}^{n}\alpha_i\bar{e}^{\,i},\quad \bar{b}=\sum_{i=1}^{n}\beta_i\bar{e}^{\,i},$$

α_i,β_i – координаты векторов $\bar{a},\bar{b}$ в базисе $\{\bar{e}^{\,i}\}$. Вычислим скалярное произведение векторов: $(\bar{a},\bar{b})=\sum_{j=1}^{n}\alpha_j\sum_{i=1}^{n}\beta_i\,g_{ij}$, где $G=\{g_{ij}\}$ матрица, содержащая попарные скалярные произведения векторов базиса $g_{ij}=(\bar{e}^{\,i},\bar{e}^{\,j})$. Система базисных векторов $\bar{e}^{\,1},\bar{e}^{\,2},\ldots,\bar{e}^{\,n}$ называется **ортонормированной**, если

$$(\bar{e}^{\,i},\bar{e}^{\,j})=\delta_{ij}=\begin{cases}1, i=j\\0, i\neq j\end{cases}.$$

Каждая система попарно ортогональных ненулевых элементов линейно независима. Наибольшее число векторов в такой системе не может превышать размерность пространства, поэтому ортонормированная система может служить базисом пространства. Векторы $\bar{a},\bar{b}\in\mathbf{H}$ могут быть представлены в виде линейной комбинации векторов ортонормированного базиса:

$$\bar{a}=\sum_{j=1}^{n}\alpha_j\bar{e}^{\,j},\ \bar{b}=\sum_{j=1}^{n}\beta_j\bar{e}^{\,j},$$

где $\alpha_j=\left(\bar{a},\bar{e}^{\,j}\right),\beta_j=\left(\bar{b},\bar{e}^{\,j}\right),\|\bar{a}\|^2=\sum_{j=1}^{n}\alpha_j{}^2,\left(\bar{a},\bar{b}\right)=\sum_{j=1}^{n}\alpha_j\beta_j$.

Простота вычисления нормы, координат и скалярного произведения векторов делает ортонормированный базис наиболее удобным для использования. Любой базис можно сделать ортонормированным с помощью специального процесса **ортогонализации** векторов.

Пусть $\bar{x}^1,\bar{x}^2,\bar{x}^3,\ldots,\bar{x}^m$ – система линейно независимых векторов в пространстве $\mathbf{R}^m$. Построим ортонормированную систему векторов $\bar{z}^1,\bar{z}^2,\bar{z}^3,\ldots,\bar{z}^m$, которая также будет базисом в

пространстве $\mathbf{R}^m$. Векторы $\bar{z}^k$ можно построить по следующей схеме:

$$\bar{y}^1 = \bar{x}^1, \quad \bar{z}^1 = \frac{\bar{y}^1}{\left\|\bar{y}^1\right\|}, \quad \bar{y}^2 = \bar{x}^2 - \left(\bar{x}^2, \bar{z}^1\right)\cdot \bar{z}^1, \quad \bar{z}^2 = \frac{\bar{y}^2}{\left\|\bar{y}^2\right\|}.$$

Проверим, что векторы ортогональны, т.е. $\left(\bar{y}^2, \bar{z}^1\right) = 0$:

$$\left(\bar{y}^2, \bar{z}^1\right) = \left(\bar{x}^2 - \left(\bar{x}^2, \bar{z}^1\right)\cdot \bar{z}^1, \bar{z}^1\right) = \left(\bar{x}^2, \bar{z}^1\right) - \left(\bar{x}^2, \bar{z}^1\right)\cdot\left(\bar{z}^1, \bar{z}^1\right) = 0,$$

так как $\left(\bar{z}^1, \bar{z}^1\right) = \left\|\bar{z}^1\right\|^2 = 1$. Далее строим $\bar{z}^3$:

$$\bar{y}^3 = \bar{x}^3 - \left(\bar{x}^3, \bar{z}^1\right)\cdot \bar{z}^1 - \left(\bar{x}^3, \bar{z}^2\right)\cdot \bar{z}^2, \quad \bar{z}^3 = \frac{\bar{y}^3}{\left\|\bar{y}^3\right\|}$$

и проверяем, что $\bar{z}^3$ ортогонален $\bar{z}^2, \bar{z}^1$ и т.д. Все вновь построенные векторы $\bar{z}^k$ линейно выражаются через $\bar{x}^k$, и, наоборот, $\bar{x}^k$ выражается через $\bar{z}^k$. Ни один из построенных векторов $\bar{z}^k$ не может обратиться в 0. Действительно, пусть на некотором *k*-м шаге получен нулевой вектор $\bar{z}^k$. Поскольку $\bar{z}^k$ линейно выражается через $\bar{x}^1, \bar{x}^2, \ldots, \bar{x}^k$, это означает линейную зависимость системы векторов $\bar{x}^1, \bar{x}^2, \ldots, \bar{x}^k$, что невозможно, т.к. эти векторы входят в базис.

Тема 3. Линейные операторы, матрицы и их спектр

3.1. Линейные операторы

Оператором, действующим на векторном пространстве **X**, называется отображение, т.е. правило, по которому элементу $\bar{x} \in \mathbf{X}$ ставится в соответствие элемент $\bar{y}$, принадлежащий, вообще говоря, другому пространству **Y**: $\bar{y} = A\,\bar{x}$, или $A: \mathbf{X} \to \mathbf{Y}$.

Пусть **X**, **Y** – линейные пространства; $\bar{x}_1, \bar{x}_2 \in \mathbf{X}$. Оператор $A: \bar{x} \to \bar{y}$ называется **аддитивным**, если

$$A(\bar{x}_1 + \bar{x}_2) = A\bar{x}_1 + A\bar{x}_2 \quad \forall \bar{x}_1, \bar{x}_2 \in \mathbf{X},$$

и **однородным**, если

$$A(\lambda\bar{x}) = \lambda A\bar{x} \quad \forall\bar{x} \in \mathbf{X} \quad \forall\lambda \in \mathbf{R}.$$

Оператор **A** называется линейным, если он аддитивен и однороден, другими словами, если он дистрибутивен:

$$A(\lambda_1\bar{x}_1 + \lambda_2\bar{x}_2) = \lambda_1 A\bar{x}_1 + \lambda_2 A\bar{x}_2.$$

Если **X**, **Y** – нормированные пространства, оператор $A: \mathbf{X} \to \mathbf{Y}$ называется ограниченным, если существует константа C, такая, что $\|A\bar{x}\| \le C \cdot \|\bar{x}\|$. Наименьшая из всех таких констант *C* называется нормой оператора **A**. Таким образом, имеет место неравенство

$$\|A\bar{x}\| \le \|A\| \cdot \|\bar{x}\|,$$

которое называют условием согласования норм.

Оператор $A: \mathbf{X} \to \mathbf{Y}$ называется **непрерывным** в точке $\bar{x} \in \mathbf{X}$, если для любой последовательности $\bar{x}_n \in \mathbf{X}$, таких, что $\bar{x}_n \underset{n\to\infty}{\to} \bar{x}$, справедливо $A\bar{x}_n \underset{n\to\infty}{\to} A\bar{x}$, т.е. $\|A\bar{x}_n - A\bar{x}\| \underset{n\to\infty}{\to} 0$ при $\|\bar{x}_n - \bar{x}\| \underset{n\to\infty}{\to} 0$. Для линейных операторов непрерывность в одной точке влечет непрерывность во всей области; и непрерывность равносильна ограниченности.

Над линейными операторами можно определить операции сложения: $U = A + B$, если $U\bar{x} = A\bar{x} + B\bar{x} \quad \forall\bar{x} \in \mathbf{X}$, и умножения на константу: $V = \lambda A$, если $V\bar{x} = \lambda A\bar{x} \quad \forall\bar{x} \in \mathbf{X}$. Множество линейных непрерывных операторов с определенными таким образом операциями само образует линейное пространство.

Определим произведение линейных операторов. Пусть $A: \mathbf{Y} \to \mathbf{Z}$, $B: \mathbf{X} \to \mathbf{Y}$. Оператор $C: \mathbf{X} \to \mathbf{Z}$ является произведением *A* и *B*, если $C\bar{x} = (A \cdot B)\bar{x} = A(B\bar{x})$. Если *A, B* – линейные непрерывные операторы, то оператор *C* = *A*·*B* тоже линеен и непрерывен, причем $\|AB\| \le \|A\| \cdot \|B\|$.

Пусть **X** – нормированное пространство, определим *I* – тождественный оператор в **X**, отображающий любой элемент в себя: $I\bar{x} = \bar{x} \quad \forall\bar{x} \in \mathbf{X}$. Если $A, B: \mathbf{X} \to \mathbf{X}$, $A \cdot B: \mathbf{X} \to \mathbf{X}$, и мож-

но определить степени оператора: $A^2 = A \cdot A$, $A^3 = A \cdot A \cdot A$ и т.д., причем $\|A^n\| \le \|A\|^n \ \forall n$.

Пусть $A: \mathbf{X} \to \mathbf{X}$, $A\bar{x} = \bar{y}$. Обратным оператором $A^{-1}: \mathbf{X} \to \mathbf{X}$ называется такой оператор, что $A \cdot A^{-1} = A^{-1} \cdot A = I$. Наличие ограниченного обратного оператора A^{-1} для оператора $A: \mathbf{X} \to \mathbf{X}$ означает, что существует константа $\delta > 0$, такая, что $\|A\bar{x}\| \ge \delta \cdot \|\bar{x}\|$, т.е. оператор *A* невырожден. Для обратного оператора выполняется $\|A^{-1}\| \le 1/\delta$. Если оператор *A* имеет обратный, то уравнение $A\bar{x} = \bar{y}$ имеет решение $\bar{x} = A^{-1}\bar{y}$.

Теорема Банаха. Пусть $\mathbf{X}$ – линейное пространство; $A: \mathbf{X} \to \mathbf{X}$ – линейный оператор. Если $\|A\| \le q < 1$, то оператор $(I - A): \mathbf{X} \to \mathbf{X}$ имеет непрерывный обратный оператор, который можно представить в виде бесконечного ряда:

$$(I - A)^{-1} = I + A + A^2 + \ldots + A^k + \ldots,$$

причем $\|(I - A)^{-1}\| \le \dfrac{1}{1 - q}$.

Ядром линейного оператора $A: \mathbf{X} \to \mathbf{Y}$ является множество векторов $\bar{x} \in \mathbf{X}$, таких, что $A\bar{x} = \bar{0}$. Если ядро оператора *A* состоит из одного нулевого элемента, то он имеет обратный. Размерность ядра, т.е. число линейно независимых векторов, входящих в ядро, называется **дефектом** линейного оператора. **Образом** линейного оператора $A: \mathbf{X} \to \mathbf{Y}$ называется множество всех элементов $\bar{y} = A\bar{x}$. **Рангом** линейного оператора называется размерность образа линейного оператора.

Ранг невырожденного оператора $A: \mathbf{R}^n \to \mathbf{R}^n$ равен *n*, а дефект равен 0. Пусть $A: \mathbf{R}^n \to \mathbf{R}^m$ и размерность ядра оператора равна *l*. Можно показать, что $m + l = n$.

3.2. Матрицы и операции над ними

Выберем в пространстве базис $\bar{e}^1, \bar{e}^2, ..., \bar{e}^n$. Как известно, любой вектор может быть представлен в виде линейной комбинации векторов базиса $\bar{x} = \sum_{j=1}^{n} x_j \bar{e}^j$. Подействуем на вектор базиса $\bar{e}^j$ линейным преобразованием $A: \mathbf{R}^n \to \mathbf{R}^n$. Так как результатом будет также вектор из $\mathbf{R}^n$, разложим его по базису $\bar{e}^1, \bar{e}^2, ..., \bar{e}^n$:

$$A\bar{e}^j = \sum_{k=1}^{n} a_{kj} \bar{e}^k, \quad j = 1, \ldots, n.$$

В силу линейности оператора A имеем

$$A\bar{x} = A\sum_{j=1}^{n} x_j \bar{e}^j = \sum_{j=1}^{n} x_j A\,\bar{e}^j = \sum_{j=1}^{n} x_j \sum_{k=1}^{n} a_{kj}\ \bar{e}^k =$$

$$= x_1(a_{11}\bar{e}^1 + a_{21}\bar{e}^2 + \ldots + a_{n1}\bar{e}^n) + x_2(a_{12}\bar{e}^1 + a_{22}\bar{e}^2 + \ldots + a_{n2}\bar{e}^n) +$$

$$+ \ldots + x_n(a_{1n}\bar{e}^1 + a_{2n}\bar{e}^2 + \ldots + a_{nn}\bar{e}^n) =$$

$$+ \bar{e}^1(x_1a_{11} + x_2a_{12} + \ldots + x_na_{1n}) + \bar{e}^2(x_1a_{21} + x_2a_{22} + \ldots + x_na_{2n}) +$$

$$+ \ldots + \bar{e}^n(x_1a_{n1} + x_2a_{n2} + \ldots + x_na_{nn}) = \sum_{j=1}^{n} \bar{e}^j \sum_{k=1}^{n} x_k a_{jk}.$$

Пусть $y_1, y_2, ..., y_n$ – координаты вектора $\bar{y} = A\,\bar{x}$ в том же базисе $\bar{e}^1, \bar{e}^2, ..., \bar{e}^n$, т.е. $\bar{y} = \sum_{j=1}^{n} y_j \bar{e}^j$. В силу единственности разложения по базису имеем $y_j = \sum_{k=1}^{n} x_k a_{jk}$, или

$$\begin{aligned}
y_1 &= x_1a_{11} + x_2a_{12} + \ldots + x_na_{1n} \\
y_2 &= x_1a_{21} + x_2a_{22} + \ldots + x_na_{2n} \\
&\ldots \\
y_n &= x_1a_{n1} + x_2a_{n2} + \ldots + x_na_{nn}
\end{aligned}$$

Таким образом, линейному оператору A в базисе $\bar{e}^1, \bar{e}^2, \ldots, \bar{e}^n$ соответствует квадратная матрица (число строк равно числу столбцов), которой присвоим такое же имя:

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix}.$$

Важным примером линейного оператора является преобразование $\bar{y} = A\bar{x}$ вектор-столбцов $\bar{x} \in \mathbf{R}^n$ в вектор-столбцы $\bar{y} \in \mathbf{R}^m$, которое однозначно определяется матрицей из m строк и n столбцов (в общем случае $m \neq n$).

Пусть $A = \{a_{ij}\}_{i=1,\ldots,m}^{j=1,\ldots,n}$, $B = \{b_{ij}\}_{i=1,\ldots,m}^{j=1,\ldots,n}$ – матрицы одинаковой размерности. Для них можно определить операцию сложения:

$C = A + B$, $C = \{c_{ij}\}_{i=1,\ldots,m}^{j=1,\ldots,n}$, $c_{ij} = a_{ij} + b_{ij}$, $i = 1,\ldots,m,\ j = 1,\ldots,n$,

умножения на числовую константу:

$C = \lambda \cdot A$, $C = \{c_{ij}\}_{i=1,\ldots,m}^{j=1,\ldots,n}$, $c_{ij} = \lambda \cdot a_{ij}$, $i = 1,\ldots,m,\ j = 1,\ldots,n$.

Операцию умножения матриц можно определить для матриц $A = \{a_{ij}\}_{i=1,\ldots,m}^{j=1,\ldots,n}, B = \{b_{ij}\}_{i=1,\ldots,n}^{j=1,\ldots,l}$, тогда

$$C = A \cdot B,\ C = \{c_{ij}\}_{i=1,\ldots,m}^{j=1,\ldots,l},\ c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

Можно показать, что для матричных операций справедливо:

- $A + B = B + A$
- $(A + B) + C = A + (B + C)$
- $(A \cdot B) \cdot C = A \cdot (B \cdot C)$
- существование нулевой матрицы O: $A + O = A$
- существование обратной по отношению к операции сложения матрицы $-A$, такой, что $A + (-A) = O$

- существование единичной матрицы I, $A \cdot I = I \cdot A = A$,

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

- $A \cdot B \neq B \cdot A$ (нет коммутативности операции умножения)
- существование обратной A^{-1} для невырожденных матриц $A \cdot A^{-1} = A^{-1} \cdot A = I$

Невырожденность матрицы означает, что ее определитель отличен от 0: $\det(A) \neq 0$.

Для матриц можно ввести норму, согласованную с евклидовой нормой векторного пространства:

$$\|A\| = \sup_{\|\bar{x}\|=1} \|A\bar{x}\|, \ \|A\| = \max_j \sum_{i=1}^{n} |a_{ij}|, \quad \|\bar{x}\| = \max_j |x_j| .$$

3.3. Преобразование матрицы при переходе к новому базису

Пусть линейный оператор, отображающий $\mathbf{R}^n \to \mathbf{R}^m$ в базисе $\bar{e}^1, \bar{e}^2, \dots, \bar{e}^n$, имеет матрицу $A = \{a_{ij}\}_{i=1,\dots,n}^{j=1,\dots,m}$, а в базисе $\bar{g}^1, \bar{g}^2, \dots, \bar{g}^n$ – матрицу $A' = \{a'_{ij}\}_{i=1,\dots,n}^{j=1,\dots,m}$. Найдем, как связаны между собой матрицы A, A'.

Обозначим через $C = \{c_{ij}\}_{i=1,\dots,n}^{j=1,\dots,n}$ матрицу перехода от базиса $\bar{e}^1, \bar{e}^2, \dots, \bar{e}^n$ к базису $\bar{g}^1, \bar{g}^2, \dots, \bar{g}^n$, т.е. $\bar{g}^i = C\bar{e}^i$. Матрицу C можно связать с некоторым линейным преобразованием $\mathbf{R}^n \to \mathbf{R}^n$. Поскольку это преобразование переводит базис в базис, дефект этого оператора равен 0. Следовательно, оператор C невырожден, и для него существует обратный C^{-1}, $\bar{e}^i = C^{-1}\bar{g}^i, i = 1, \dots, n$.

Поскольку A' – матрица оператора в базисе $\bar{g}^1, \bar{g}^2, \dots, \bar{g}^n$, то

$$A\bar{g}^i = \sum_{j=1}^{n} a'_{ij}\bar{g}^j .$$

Применяя к обеим частям равенства оператор C^{-1}, получим

$$C^{-1}A\bar{g}^i = C^{-1}\sum_{j=1}^{n} a'_{ij}\bar{g}^j = \sum_{j=1}^{n} a'_{ij}C^{-1}\bar{g}^j = \sum_{j=1}^{n} a'_{ij}\bar{e}^j .$$

Подставляя в левую часть равенство $\bar{g}^i = C\bar{e}^i$, получим

$$C^{-1}AC\bar{e}^i = \sum_{j=1}^{n} a'_{ij}\bar{e}^j ,$$

а это означает, что матрицей оператора $C^{-1}A\,C$ в базисе $\bar{e}^1, \bar{e}^2, ..., \bar{e}^n$ является матрица A', т.е.

$$C^{-1}A\,C = A'. \qquad (1.1)$$

Применим к обеим частям последнего равенства матрицу C:

$$A\,C = CA'. \qquad (1.2)$$

3.4. Собственные значения и собственные векторы

Как было отмечено выше, вид матрицы линейного преобразования, действующего в пространстве, зависит от базиса. Для каждой матрицы существует некоторый собственный базис, в котором матрица имеет наиболее простую структуру, например, диагональный вид. Для того, чтобы найти такой вид матрицы и соответствующую ей систему координат, введем понятие собственного числа и собственного вектора матрицы.

Пусть A – матрица линейного оператора, отображающего $\mathbf{R}^n \to \mathbf{R}^n$. Числовая константа λ называется собственным значением, а ненулевой вектор $\bar{x}$ – соответствующим λ собственным вектором матрицы A, если

$$A\bar{x} = \lambda\bar{x}. \qquad (1.3)$$

Равенство (1.3) можно записать также в виде

$$B\bar{x} = (A - \lambda I)\bar{x} = \bar{0},$$

где I – единичный (тождественный оператор). Последнее равенство означает, что вектор $\bar{x}$ принадлежит ядру линейного опе-

ратора $B = A - \lambda I$. Чтобы найти ненулевое решение, необходимо, чтобы определитель матрицы был равен нулю:

$$\det(A - \lambda I) = 0 . \qquad (1.4)$$

Следовательно, чтобы найти собственные значения и собственные векторы матрицы, необходимо решить **характеристическое (вековое) уравнение** (1.4), а затем подставить полученные значения λ в уравнение (1.3) и, решив систему линейных алгебраических уравнений (СЛАУ), найти компоненты собственного вектора. Поскольку матрица СЛАУ вырождена, ее собственный вектор находится с точностью до константы.

Множество всех собственных значений матрицы называется **спектром**, максимальное по модулю собственное значение $|\lambda|_{\max}$ – **спектральным радиусом**. На основе спектра может быть введена норма матрицы $\|A\| = \sqrt{|\lambda|_{\max}}$, которая согласована с евклидовой нормой вектора $\|\bar{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$. С помощью собственных значений вычисляется важная характеристика, которая называется обусловленностью матрицы:

$$cond(A) = \|A\| \cdot \|A^{-1}\| = \frac{|\lambda|_{\max}}{|\lambda|_{\min}} .$$

При нахождении собственных значений могут возникнуть различные ситуации, например, вековое уравнение не имеет вещественных корней, или некоторые из корней кратные. Кратным собственным значениям может соответствовать один собственный вектор, два или больше линейно независимых собственных вектора. Рассмотрим несколько примеров.

Пример 1.1. Найдем собственные значения матрицы

$$A = \begin{pmatrix} 5 & 2 \\ 3 & 6 \end{pmatrix}.$$

Запишем вековое уравнение:

$$\det(A-\lambda I)=\begin{vmatrix}5-\lambda & 2\\ 3 & 6-\lambda\end{vmatrix}=(5-\lambda)\cdot(6-\lambda)-6=-\lambda^2+11\lambda+24=0$$

Решая квадратное уравнение, получим

$$\lambda_{1,2}=\frac{11\pm\sqrt{121-4\cdot 24}}{2}=\frac{11\pm 5}{2};\quad \lambda_1=8,\quad \lambda_2=3.$$

Подставляем λ_1 в (1.3), получим СЛАУ

$$\begin{pmatrix}5-8 & 2\\ 3 & 6-8\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}\Rightarrow\begin{cases}-3x_1+2x_2=0\\ 3x_1-2x_2=0\end{cases},$$

решением которой будут векторы, у которых $3x_1=2x_2$, например, $\bar{x}=(2,3)$. Найдем второй собственный вектор $\bar{y}$, отвечающий собственному значению $\lambda = 3$:

$$\begin{pmatrix}5-3 & 2\\ 3 & 6-3\end{pmatrix}\begin{pmatrix}y_1\\ y_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}\Rightarrow\begin{cases}2y_1+2y_2=0\\ 3y_1+3y_2=0\end{cases}\Rightarrow y_1=-y_2.$$

Вторым собственным вектором будет $\bar{y}=(1,-1)$. Проверим, что векторы $\bar{x},\bar{y}$ линейно независимы. Пусть $\alpha_1\bar{x}+\alpha_2\bar{y}=\bar{0}$. После подстановки в равенство векторов $\bar{x}=(2,3)$, $\bar{y}=(1,-1)$ получим однородную систему с ненулевым определителем: $\begin{cases}2\alpha_1+\alpha_2=0\\ 3\alpha_1-\alpha_2=0\end{cases}\Rightarrow\alpha_1=\alpha_2=0.$

Пример 1.2. Пусть $A=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$. Вековое уравнение $(1-\lambda)^2=0,$ значит, $\lambda_1=\lambda_2=1$. Ищем собственный вектор:

$$\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix},$$

откуда $x_2=0, x_1$ – произвольно, например, $x_1=1$. Второго собственного вектора нет.

Пример 1.3. Найти собственные значения и собственные векторы матрицы $A=\begin{pmatrix}1 & -3 & 3\\ 3 & -5 & 3\\ 6 & -6 & 4\end{pmatrix}$.

$$\det(A-\lambda I)=0\Rightarrow\begin{vmatrix}1-\lambda & -3 & 3\\ 3 & -5-\lambda & 3\\ 6 & -6 & 4-\lambda\end{vmatrix}=0.$$

Раскладываем определитель по первой строке:

$$(1-\lambda)\cdot\begin{vmatrix}-5-\lambda & 3\\ -6 & 4-\lambda\end{vmatrix}-(-3)\cdot\begin{vmatrix}3 & 3\\ 6 & 4-\lambda\end{vmatrix}+3\cdot\begin{vmatrix}3 & -5-\lambda\\ 6 & -6\end{vmatrix}=$$

$$=(1-\lambda)\left[(-5-\lambda)(4-\lambda)+18\right]+3\left[3(4-\lambda)-18\right]+3\left[6(5+\lambda)-18\right]=$$

$$=(1-\lambda)\cdot\left[-2+\lambda-\lambda^2\right]+3\cdot\left[-6-3\lambda\right]+3\cdot\left[12+6\lambda\right]=$$

$$=-2+2\lambda+\lambda-\lambda^2+\lambda^2-\lambda^3-18-9\lambda+36+18\lambda=$$

$$=16+12\lambda-\lambda^3=0$$

Ищем целочисленные решения кубического уравнения, проверяя значения $\lambda=\pm1,\pm2,\ldots$. В результате подстановки находим первый корень $\lambda_1=-2$. Выделяя множитель $(\lambda+2)$ из кубического полинома, получим квадратное уравнение $\lambda^2-2\lambda-8=0$ с действительными корнями $\lambda_2=-2,\ \lambda_3=4.$ Обозначим собственные векторы, отвечающие найденным собственным значениям $\lambda_1=\lambda_2=-2$, $\lambda_3=4$, как $\bar{x},\bar{y},\bar{z}$ соответственно.

Подставляя $\lambda_3=4$ в (1.3), получим систему для определения собственного вектора $\bar{z}$:

$$\begin{cases}-3z_1-3z_2+3z_3=0\\ 3z_1-9z_2+3z_3=0\\ 6z_1-6z_2=0\end{cases},$$

в которой после несложных преобразований останется два независимых уравнения:

$$\begin{cases} z_1 - z_2 = 0 \\ -2z_2 + z_3 = 0 \end{cases}.$$

Полагая $z_2 = 1$, получим собственный вектор $\bar{z}$ = (1, 1, 2).

Ищем собственные векторы, соответствующие кратному собственному значению $\lambda_1 = \lambda_2 = -2$. После преобразований остается одно независимое уравнение $x_1 - x_2 + x_3 = 0$, которое задает целую плоскость собственных векторов. Полагая $x_2 = 1, x_3 = 0$, $y_2 = 0, y_3 = 1$, получим два линейно независимых собственных вектора $\bar{x} = (1,1,0)$ и $\bar{y} = (-1,0,1)$. Проверяем, что полученная система собственных векторов линейно независима. Пусть $\alpha_1\bar{x} + \alpha_2\bar{y} + \alpha_3\bar{z} = \bar{0}$. Тогда

$$\begin{cases} \alpha_1 - \alpha_2 + \alpha_3 = 0 \\ \alpha_1 \qquad + \alpha_3 = 0 \\ \qquad \alpha_2 + 2\alpha_3 = 0 \end{cases}.$$

Определитель однородной системы равен 4, следовательно, она имеет единственное решение $\alpha_1 = \alpha_2 = \alpha_3 = 0$, откуда следует линейная независимость векторов $\bar{x}, \bar{y}, \bar{z}$.

3.5. Приведение матрицы к диагональному виду

В предыдущих примерах было показано, что система собственных векторов матрицы линейного оператора, отображающего $\mathbf{R}^n \to \mathbf{R}^n$, соответствующих разным собственным значениям, линейно независима. Если матрица имеет полную систему собственных векторов, то они образуют базис в пространстве $\mathbf{R}^n$, в котором исходная матрица имеет наиболее простой вид. Матрицы, которые имеют базис, составленный из собственных векторов, называются матрицами простой структуры. Они могут быть приведены к диагональному виду, причем на диагонали будут стоять собственные значения матрицы. Это означает, что ли-

нейное преобразование, соответствующее данной матрице, является преобразованием растяжения по характерному направлению, задаваемому собственным вектором.

Для приведения матрицы к диагональному виду используют линейное преобразование, матрица которого составлена из столбцов, являющихся собственными векторами. Не всякая матрица имеет простую структуру, следовательно, не каждая может быть приведена к диагональному виду.

Пример 1.4. Приведем к диагональному виду матрицу из Примера 1.1. Запишем найденные собственные векторы в матрицу преобразования $C = \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix}$ и найдем обратную

$$C^{-1} = \frac{1}{5}\begin{pmatrix} 1 & 1 \\ 3 & -2 \end{pmatrix}.$$

Проверка: $C \cdot C^{-1} = \frac{1}{5}\begin{pmatrix} 2+3 & 2-2 \\ 3-3 & 3+2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$.

Применяя (1.1), получим

$$A' = C^{-1} \cdot A \cdot C = \frac{1}{5}\begin{pmatrix} 2 & 2 \\ 3 & -4 \end{pmatrix} \cdot \begin{pmatrix} 5 & 2 \\ 3 & 6 \end{pmatrix} \cdot \begin{pmatrix} 2 & 1 \\ 3 & -1 \end{pmatrix} = \begin{pmatrix} 8 & 0 \\ 0 & 3 \end{pmatrix},$$

т.е. в результате перехода к новому базису получена диагональная матрица, на главной диагонали которой стоят собственные значения. В дальнейшем полезно иметь не только ортогональный, но и ортонормированный базис. Вычислим евклидову норму векторов:

$\|\bar{x}\| = \sqrt{2^2 + 3^2} = \sqrt{13}, \quad \|\bar{y}\| = \sqrt{1^2 + (-1)^2} = \sqrt{2}$,

следовательно,

$$\bar{e}^1 = \frac{\bar{x}}{\|\bar{x}\|} = \left(\frac{2}{\sqrt{13}}, \frac{3}{\sqrt{13}}\right)^T, \; \bar{e}^2 = \frac{\bar{y}}{\|\bar{y}\|} = \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right)^T.$$

Пример 1.5.

Для матрицы из Примера 1.3 найдены собственные векторы $\bar{x} = (1,1,0)$, $\bar{y} = (-1,0,1)$, $\bar{z} = (1, 1, 2)$, из которых образуем

$$C = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \; C^{-1} = \frac{1}{2}\begin{pmatrix} -1 & 3 & -1 \\ -2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix}.$$

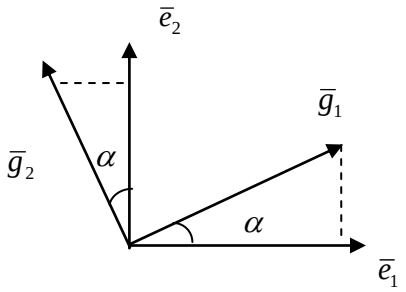
(Проверить, что $C \cdot C^{-1} = C^{-1} \cdot C = I$.)

Приводим к диагональному виду: $A' = C^{-1} \cdot A \cdot C =$

$$= \frac{1}{2}\begin{pmatrix} -1 & 3 & -1 \\ -2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & -3 & 3 \\ 3 & -5 & 3 \\ 6 & -6 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & -3 & 1 \\ 2 & -2 & 0 \\ 2 & -2 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

Пример 1.6. Покажем пример линейного преобразования, матрица которого не может быть приведена к диагональному виду. Рассмотрим оператор поворота системы координат $(\bar{e}_1, \bar{e}_2)$ на плоскости на заданный угол α (см. рисунок).



Преобразование поворота системы координат

Базисные векторы новой системы координат имеют вид:

$$\bar{g}_1 = (\cos\alpha, \sin\alpha);$$
$$\bar{g}_2 = (-\sin\alpha, \cos\alpha).$$

Следовательно, матрицей преобразования будет

$$A = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

Характеристическое уравнение для нахождения собственных значений имеет вид

$$(\cos\alpha - \lambda)^2 + \sin^2\alpha = 0,$$

$$\cos^2\alpha - 2\lambda\cos\alpha + \lambda^2 + \sin^2\alpha = \lambda^2 - 2\lambda\cos\alpha + 1 = 0$$

$$\lambda_{1,2} = \cos\alpha \pm \sqrt{\cos^2\alpha - 1}$$

Уравнение имеет вещественный корень $\lambda = 1$ в случае

$$D = \cos^2\alpha - 1 = 0 \Leftrightarrow \cos\alpha = 1 \Leftrightarrow \alpha = \pi k, \quad k = 1, 2, \ldots,$$

что соответствует тривиальному решению поворота на угол, кратный π. Во всех остальных случаях матрица преобразований не имеет вещественных собственных значений.

3.6. Сопряженный оператор. Квадратичная форма

Пусть A – линейный оператор в гильбертовом пространстве **H**. Если существует такой оператор B, что $\forall \bar{x}, \bar{y} \in \mathbf{H}$ справедливо $(A\bar{x}, \bar{y}) = (\bar{x}, B\bar{y})$, то оператор B называется **сопряженным** к A и обозначается A^*. Сопряженным к оператору поворота системы координат на угол α (см. Пример 1.6) является оператор обратного поворота, т.е. поворот на угол $-\alpha$.

Матрицей сопряженного оператора будет транспонированная матрица: $A^* = A^T$. Можно доказать следующие свойства сопряженных операторов:

- $(A+B)^* = A^* + B^*$;
- $(\alpha \cdot A)^* = \alpha \cdot A^*$, α – вещественная константа;
- $(A \cdot B)^* = B^* \cdot A^*$;
- A, A^* имеют одинаковые характеристические уравнения и, следовательно, одинаковые собственные значения.

Оператор называется **самосопряженным**, если $\forall \bar{x}, \bar{y} \in \mathbf{H}$ $(A\bar{x}, \bar{y}) = (\bar{x}, A\bar{y})$, т.е. $A = A^*$. Легко понять, что матрица самосопряженного оператора симметрична, т.е. $A = A^T$. Важным свойством самосопряженных операторов является то, что они имеют простую структуру, т.е. обладают полным набором веще-

ственных собственных значений и соответствующих им собственных векторов.

С каждой симметричной матрицей связана так называемая **квадратичная форма** $(\bar{x}, A\bar{x})$, представляющая собой уравнение кривой второго порядка. Действительно, пусть $n = 2$:

$$(\bar{x}, A\bar{x}) = (x_1, x_2) \cdot \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = a_{11}x_1^{\,2} + 2a_{12}x_1x_2 + a_{22}x_2^{\,2}.$$

Верно и обратное, т.е. каждой квадратичной форме соответствует своя симметричная матрица. Поскольку симметричная матрица имеет простую структуру, то она может быть приведена к диагональному виду. Для квадратичной формы это означает отсутствие смешанных произведений вида $x_i \cdot x_j, i \neq j$, или приведение к главным осям.

Квадратичная форма и соответствующая ей матрица положительно определена, если $(\bar{x}, A\bar{x}) \geq 0 \quad \forall \bar{x} \in \mathbf{H}$. Можно показать, что все собственные значения симметричной положительно определенной матрицы положительны.

Пример 1.7.

Привести к каноническому виду квадратичную форму

$x_1^{\,2} + 5x_2^{\,2} + x_3^{\,2} + 2x_1x_2 + 6x_1x_3 + 2x_2x_3$.

Составим матрицу, определяющую квадратичную форму:

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 1 & 5 & 1 \\ 3 & 1 & 1 \end{pmatrix}.$$

Поскольку матрица симметричная, она имеет полный набор вещественных собственных значений и собственных векторов, следовательно, ее можно привести к диагональному виду. Находим собственные значения

$\det(A - \lambda I) = 0 \Leftrightarrow \lambda^3 - 7\lambda^2 + 36 = 0$, $\lambda_1 = -2$, $\lambda_2 = 3$, $\lambda_3 = 6$

и соответствующие им собственные векторы

$$\bar{x} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \|\bar{x}\| = \sqrt{2},\ \bar{y} = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \quad \|\bar{y}\| = \sqrt{3},\ \bar{z} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \|\bar{y}\| = \sqrt{6}\ .$$

Следовательно, канонический вид формы имеет вид

$$-2y_1^{\,2} + 3y_2^{\,2} + 6y_3^{\,2},$$

а матрица преобразования координат $\bar{y} = C\bar{x}$ составлена из нормированных собственных векторов:

$$C = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{3} & 1/\sqrt{6} \\ 0 & -1/\sqrt{3} & 2/\sqrt{6} \\ -1/\sqrt{2} & 1/\sqrt{3} & 1/\sqrt{6} \end{pmatrix}.$$

Тема 4. Системы линейных алгебраических уравнений

С алгеброй матриц связаны следующие задачи:

- решение СЛАУ;
- вычисление определителей матриц;
- нахождение обратных матриц;
- отыскание спектра и собственных векторов.

Все методы решения СЛАУ можно разбить на два класса: **прямые** (точные) и **итерационные** (приближенные).

Прямые методы позволяют получить решение за конечное число арифметических операций. Если операции реализуются точно, то и решение будет точным (поэтому класс прямых методов еще называют точными методами). В итерационных методах решением является предел некоторой бесконечной последовательности единообразных действий.

В данном пособии описаны методы решения СЛАУ, у которых число уравнений совпадает с числом неизвестных.

Задана СЛАУ размерности m, которую можно записать в скалярном:

$$\begin{cases} a_{11}\cdot x_1 + a_{12}\cdot x_2 + \ldots + a_{1m}\cdot x_m = f_1 \\ a_{21}\cdot x_1 + a_{22}\cdot x_2 + \ldots + a_{2m}\cdot x_m = f_2 \\ \ldots \\ a_{m1}\cdot x_1 + a_{m2}\cdot x_2 + \ldots + a_{mm}\cdot x_m = f_m \end{cases} \qquad (1.5)$$

векторном:

$$\begin{pmatrix} a_{11} \\ a_{21} \\ \ldots \\ a_{m1} \end{pmatrix} x_1 + \begin{pmatrix} a_{12} \\ a_{22} \\ \ldots \\ a_{m2} \end{pmatrix} x_2 + \ldots \begin{pmatrix} a_{1m} \\ a_{2m} \\ \ldots \\ a_{mm} \end{pmatrix} x_m = \begin{pmatrix} f_1 \\ f_2 \\ \ldots \\ f_m \end{pmatrix} \qquad (1.5')$$

или матричном виде:

$$A\vec{x} = \vec{f}\,, \qquad (1.5'')$$

где $$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ & & & \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{pmatrix},\ \vec{f} = \begin{pmatrix} f_1 \\ f_2 \\ \ldots \\ f_{m-1} \\ f_m \end{pmatrix},\ \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_{m-1} \\ x_m \end{pmatrix},$$

A – матрица системы, $\vec{f}$ – вектор правых частей, $\vec{x}$ – вектор неизвестных. Назовем расширенной матрицей системы матрицу A', дополненную столбцом вектора правых частей:

$$A' = \begin{pmatrix} a_{11} & \ldots & a_{1m} & f_1 \\ a_{21} & \ldots & a_{2m} & f_2 \\ & & & \ldots \\ a_{m1} & \ldots & a_{mm} & f_m \end{pmatrix}.$$

Система имеет решение, если $\det A \neq 0$. По определению решения, подставив вектор $\vec{x}$ в СЛАУ, получим m тождественных уравнений.

Эффективность способов решения системы (1.5) во многом зависит от структуры и свойств матрицы A: размерности, обусловленности, симметричности, заполненности (т.е. соотношения между числом ненулевых и нулевых элементов) и др.

4.1. Точные методы решения СЛАУ

При небольшой размерности системы m (m = 2, 3) на практике часто используют формулы Крамера решения СЛАУ:

$$x_i = \frac{\det A_i}{\det A} \quad (i = 1, 2, \ldots, m).$$

Эти формулы позволяют находить неизвестные в виде дробей, знаменателем которых является определитель матрицы системы, а числителем – определители матриц A_i, полученных из A заменой столбца коэффициентов при вычисляемом неизвестном столбцом вектора правых частей.

Размерность системы (т.е. число m) является главным фактором, из-за которого **формулы Крамера** не могут быть использованы для численного решения СЛАУ большого порядка. При непосредственном раскрытии определителей решение системы с m неизвестными требует порядка $m!m$ арифметических операций. Таким образом, для решения системы, например, из m = 100 уравнений потребуется совершить 10^{158} операций, что не под силу даже самым мощным современным ЭВМ.

Для решения небольших систем используют метод **обратной матрицы**. Если $\det A \neq 0$, то существует обратная матрица A^{-1}. По определению обратной матрицы: $A\,A^{-1} = A^{-1}\,A = I$, где I – единичная матрица. Если обратная матрица известна, то, умножая на нее СЛАУ слева, получим:

$$A^{-1}A\vec{x} = A^{-1}\vec{f}, \quad I\,\vec{x} = A^{-1}\vec{f}, \quad \vec{x} = A^{-1}\vec{f}.$$

Следовательно, решение СЛАУ свелось к умножению известной обратной матрицы на вектор правых частей. Таким образом, задача решения СЛАУ и задача нахождения обратной матрицы связаны между собой, поэтому часто решение СЛАУ называют задачей обращения матрицы. Проблемы применения этого метода те же, что и при использовании метода Крамера: нахождение обратной матрицы – трудоемкая операция.

Наиболее популярным точным способом решения линейных систем вида (1.5) является **метод Гаусса**, или последовательного исключения неизвестных. Метод состоит из двух этапов: прямого и обратного. На первом этапе исходная система с

помощью эквивалентных преобразований сводится к системе с треугольной матрицей, на втором этапе решается система с треугольной матрицей. Эквивалентными преобразованиями будут следующие преобразования расширенной матрицы A':

- перестановка строк;
- умножение строк на ненулевую константу;
- сложение строк.

Используя эти преобразования, перепишем исходную систему так, чтобы один из коэффициентов столбца был равен 1, а все коэффициенты, стоящие ниже, были равны 0.

Пусть в исходной системе уравнений

$$\begin{cases} a_{11}^{(0)} \cdot x_1 + a_{12}^{(0)} \cdot x_2 + \ldots + a_{1m}^{(0)} \cdot x_m = f_1^{(0)} \\ a_{21}^{(0)} \cdot x_1 + a_{22}^{(0)} \cdot x_2 + \ldots + a_{2m}^{(0)} \cdot x_m = f_2^{(0)} \\ \ldots \\ a_{m1}^{(0)} \cdot x_1 + a_{m2}^{(0)} \cdot x_2 + \ldots + a_{mm}^{(0)} \cdot x_m = f_m^{(0)} \end{cases}$$

первый элемент $a_{11}^{(0)} \neq 0$. Назовем его ведущим элементом первой строки. Поделим все элементы этой строки на $a_{11}^{(0)}$ и исключим x_1 из всех последующих строк, начиная со второй, путем вычитания первой (преобразованной), умноженной на коэффициент при x_1 в соответствующей строке. Получим

$$\begin{cases} x_1 + a_{12}^{(1)} \cdot x_2 + a_{13}^{(1)} \cdot x_3 + \ldots + a_{1m}^{(1)} \cdot x_m = f_1^{(1)} \\ \quad a_{22}^{(1)} \cdot x_2 + a_{23}^{(1)} \cdot x_3 + \ldots + a_{2m}^{(1)} \cdot x_m = f_2^{(1)} \\ \ldots \\ \quad a_{m2}^{(1)} \cdot x_2 + a_{m3}^{(1)} \cdot x_3 + \ldots + a_{mm}^{(1)} \cdot x_m = f_m^{(1)} \end{cases}$$

Если $a_{22}^{(1)} \neq 0$, то, продолжая аналогичное исключение, приходим к системе уравнений с верхней треугольной матрицей

$$\begin{cases} x_1 + a_{12}^{(1)} \cdot x_2 + a_{13}^{(1)} \cdot x_3 + \ldots + a_{1m}^{(1)} \cdot x_m = f_1^{(1)} \\ \qquad x_2 + a_{23}^{(2)} \cdot x_3 + \ldots + a_{2m}^{(2)} \cdot x_m = f_2^{(2)} \\ \qquad\qquad x_3 + \ldots + a_{3m}^{(3)} \cdot x_m = f_3^{(3)} \\ \qquad\qquad \ldots \quad \ldots \quad \ldots \\ \qquad\qquad\qquad\qquad x_m = f_m^{(m)} \end{cases}$$

Из нее в обратном порядке находим все значения x_i:

$$\begin{cases} x_m = f_m^{(m)} \\ x_{m-1} = f_{m-1}^{(m-1)} - a_{m-1m}^{(m-1)} \cdot x_m \\ \ldots \quad \ldots \quad \ldots \\ x_1 = f_1^{(1)} - a_{12}^{(1)} \cdot x_2 - a_{13}^{(1)} \cdot x_3 - \ldots - a_{1m}^{(1)} \cdot x_m \end{cases}$$

Процесс приведения к системе с треугольной матрицей называется прямым ходом, а нахождения неизвестных – обратным. В случае, если один из ведущих элементов равен нулю, изложенный алгоритм метода Гаусса неприменим. Кроме того, если какие-либо ведущие элементы малы, то это приводит к увеличению ошибок округления и ухудшению точности счета. Поэтому обычно используется другой вариант метода Гаусса – схема Гаусса с выбором главного элемента. Путем перестановки строк, а также столбцов с соответствующей перенумерацией коэффициентов и неизвестных добиваются выполнения условия:

$$\left|a_{ii}^{(0)}\right| \geq \left|a_{ij}^{(0)}\right|, \qquad i, j = 1, 2, \ldots, m,$$

т.е. осуществляется выбор первого главного элемента. Разделив первую строку на главный элемент, как и прежде, исключают x_1 из остальных уравнений. Затем для оставшихся столбцов и строк выбирают второй главный элемент и т.д.

Метод Гаусса – Жордано использует аналогичный прием с исключением элементов в столбцах матрицы таким образом, чтобы перевести исходную систему $A\vec{x} = \vec{f}$ к $A'\vec{x} = \vec{f}'$, где A' – диагональная матрица. Этот метод можно использовать для нахождения обратной матрицы, для чего в расширенную матрицу

надо включить единичную матрицу. Преобразования в строках, приводящие исходную матрицу к диагональному виду, будут переводить единичную матрицу в обратную, а вектор правых частей – в решение СЛАУ.

Пример 1.8. Решить СЛАУ

$$\begin{cases} x_1+x_2-x_3=2 \\ -2x_1+x_2+x_3=3 \\ x_1+x_2+x_3=6 \end{cases}$$

методом Гаусса – Жордано и найти обратную матрицу. Записываем расширенную матрицу и с помощью эквивалентных преобразований переводим матрицу системы в единичную:

A			I			f
1	1	−1	1	0	0	2
−2	1	1	0	1	0	3
1	1	1	0	0	1	6

Прибавляем ко второй строке первую, умноженную на 2, а из третьей вычитаем первую:

1	1	−1	1	0	0	2
0	3	−1	2	1	0	7
0	0	2	−1	0	1	4

Делим вторую строку на 3 и вычитаем ее из первой:

1	0	$-\frac{1}{3}$	$\frac{1}{3}$	$-\frac{1}{3}$	0	$-\frac{1}{3}$
0	1	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	0	$\frac{7}{3}$
0	0	2	−1	0	1	4

Делим последнюю строку на 2 и прибавляем ее с коэффициентом $\frac{1}{3}$ к первой строке и с коэффициентом $-\frac{1}{3}$ – ко второй:

1	0	0	0	$-1/3$	$1/3$	1
0	1	0	$1/2$	$1/3$	$1/6$	3
0	0	1	$-1/2$	0	$1/2$	2
	I			A^{-1}		$\bar{x}$

Выполняя проверку $A \cdot A^{-1} = I, A\bar{x} = \bar{f}$, убеждаемся, что решение найдено правильно.

Часто возникает необходимость в решении СЛАУ, матрицы которых являются слабо заполненными, т.е. содержат много нулевых элементов. В то же время эти матрицы имеют определенную структуру. Среди таких систем выделим системы с матрицами ленточной структуры, в которых ненулевые элементы располагаются на главной диагонали и на нескольких побочных диагоналях. Для решения систем с ленточными матрицами коэффициентов вместо метода Гаусса можно использовать более эффективные методы, например, **метод прогонки.**

Рассмотрим наиболее простой случай: систему с трехдиагональной матрицей коэффициентов, к которой сводится решение ряда численных задач (сплайн-интерполяция таблично заданной функции, дискретизация краевых задач для дифференциальных уравнений методами конечных разностей и др.). В этом случае СЛАУ имеет вид:

$$\begin{cases} -c_1 x_1 + b_1 x_2 = f_1 & (1.6) \\ a_i x_{i-1} - c_i x_i + b_i x_{i+1} = f_i, \quad i = 2, 3, \ldots, m-1 & (1.7) \\ a_m x_{m-1} - c_m x_m = f_m & (1.8) \end{cases}$$

Матрица системы (1.6) – (1.8) имеет трехдиагональную структуру, что хорошо видно из следующего эквивалентного векторно-матричного представления:

$$\begin{bmatrix} -c_1 & b_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ a_2 & -c_2 & b_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & a_3 & -c_3 & b_3 & 0 & \dots & 0 & 0 \\ 0 & 0 & a_4 & -c_4 & b_4 & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & a_{m-1} & -c_{m-1} & b_{m-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & a_m & -c_m \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{m-1} \\ x_m \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{m-1} \\ f_m \end{bmatrix}.$$

Если при этом выполняется условие $|c_i| \geq |b_i| + |a_i|$, то говорят, что матрица данной системы имеет диагональное преобладание.

Предположим, что существуют такие наборы чисел α_i и β_i, $i = 2, 3, \dots, m$, при которых

$$x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1}. \tag{1.9}$$

Записав уравнение (1.6) в виде (1.9):

$$x_1 = \frac{b_1}{c_1} x_2 - \frac{f_1}{c_1},$$

получим формулы для определения α_2, β_2:

$$\alpha_2 = \frac{b_1}{c_1}, \ \beta_2 = -\frac{f_1}{c_1}. \tag{1.10}$$

Уменьшим в (1.9) индекс на единицу: $x_{i-1} = \alpha_i x_i + \beta_i$, и подставим полученное выражение в (1.7):

$$a_i \alpha_i x_i + a_i \beta_i - c_i x_i + b_i x_{i+1} = f_i,$$

откуда

$$x_i = \frac{b_i}{c_i - a_i \alpha_i} x_{i+1} + \frac{a_i \beta_i - f_i}{c_i - a_i \alpha_i}.$$

Данное равенство совпадает с (1.9), если при всех $i = 1, 2, \dots, m-1$ выполняются рекуррентные соотношения

$$\alpha_{i+1} = \frac{b_i}{c_i - a_i \alpha_i}, \ \beta_{i+1} = \frac{a_i \beta_i - f_i}{c_i - a_i \alpha_i}. \tag{1.11}$$

Они позволяют получить все остальные коэффициенты α_i, β_i.

При $i = m - 1$ из (1.9) получим $x_{m-1} = \alpha_m x_m + \beta_m$.

Подставляя это выражение в (1.8) и разрешая полученное выражение относительно x_m, записываем:

$$x_m = \frac{a_m\beta_m - f_m}{c_m - a_m\alpha_m}, \qquad (1.12)$$

где α_m и β_m известны. Далее по формулам (1.9) последовательно находятся $x_{m-1}, x_{m-2}, \ldots, x_1$.

Для успешного применения метода прогонки нужно, чтобы в процессе вычислений не возникало ситуаций с делением на нуль, а при больших размерностях систем не должно быть быстрого роста погрешностей округления.

Будем называть прогонку корректной, если знаменатели прогоночных коэффициентов в формуле (1.11) не обращаются в нуль, и устойчивой, если $|\alpha_i| < 1$ при $i = 1, 2, 3, \ldots, m$.

Теорема. Пусть коэффициенты a_i, b_i уравнения (1.7) отличны от нуля и пусть $|c_i| \geq |b_i| + |a_i|$ при $i = 1, 2, 3, \ldots, m$. Тогда прогонка корректна и устойчива.

Условия этой теоремы, которые во многих приложениях выполняются автоматически, являются достаточными условиями корректности и устойчивости прогонки. Если эти условия не выполняются, то можно организовать выбор главного элемента аналогично схеме Гаусса.

4.2. Итерационные методы решения СЛАУ

Рассмотрим систему линейных алгебраических уравнений (1.5). Итерационные методы, или методы последовательных приближений, дают возможность построить последовательность векторов $\vec{x}^{(0)}, \vec{x}^{(1)}, \vec{x}^{(2)}, \ldots, \vec{x}^{(k)}, \ldots$, пределом которой должно быть точное решение $\vec{x}^* = \lim\limits_{k\to\infty} \vec{x}^{(k)}$.

На практике построение последовательности обрывается, как только достигается желаемая точность. Чаще всего для достаточно малого значения $\varepsilon > 0$ контролируется выполнение оценки $\left|\vec{x}^* - \vec{x}^{(k)}\right| < \varepsilon$.

Метод последовательных приближений может быть построен по следующей схеме. Эквивалентными преобразованиями приведем систему (1.5) к виду

$$\vec{x} = C\,\vec{x} + \vec{d}\,, \qquad (1.13)$$

где $\vec{x}$ – тот же самый вектор, а C и $\vec{d}$ – некоторые новые матрица и вектор соответственно.

При решении методом последовательных приближений необходимо выбрать начальное (нулевое) приближение. За нулевое приближение можно принять столбцы правых частей $\vec{f}$, $\vec{d}$ системы (1.13) или нулевой вектор. Следующее приближение $\vec{x}^{(1)}$ определяется рекуррентным равенством

$$\vec{x}^{(1)} = C\,\vec{x}^{(0)} + \vec{d}\,.$$

Далее находим $\vec{x}^{(2)}$:

$$\vec{x}^{(2)} = C\,\vec{x}^{(1)} + \vec{d}\,,$$

и т.д. Для *k*-й итерации получаем

$$\vec{x}^{(k+1)} = C\,\vec{x}^{(k)} + \vec{d}, \quad k = 0, 1, 2, \ldots \qquad (1.14)$$

Такой итерационный процесс будем называть **одношаговым итерационным методом.**

Изучим вопрос о сходимости итерационного процесса, т.е. определим, какие нужно предъявить требования к виду матрицы C, чтобы последовательность $\left\{\vec{x}^{(k)}\right\}$ при $k \to \infty$ имела пределом $\vec{x}^{*}$, т.е. была решением системы (1.13), эквивалентной исходной системе (1.5): $\lim\limits_{k\to\infty} \vec{x}^{(k)} = \vec{x}^{*}$.

Достаточным условием сходимости итерационного метода (1.13) к решению системы (1.5) при любом начальном векторе $\vec{x}^{(0)}$ является требование $\|C\| < 1$, где $\|C\|$ – норма матрицы *C*.

В общем виде одношаговый итерационный метод можно записать в так называемой канонической форме:

$$B^{(k+1)}\,\frac{\vec{x}^{(k+1)} - \vec{x}^{(k)}}{\tau^{(k+1)}} + A\vec{x}^{(k)} = \vec{f}\,.$$

Здесь $B^{(k+1)}$ – матрица, задающая итерационный метод, τ^{k+1} – итерационный параметр. Если $B^{(k+1)} = E$ (где E – единичная матрица), то метод называют **явным**, а в противном случае – **неявным**. Если матрица $B^{(k+1)}$ и итерационный параметр $\tau^{(k+1)}$ не зависят от номера итерации ($B^{(k+1)} = B$, $\tau^{(k+1)} = \tau$), то метод называют **стационарным**, и **нестационарным** – в противном случае.

Использование неявных методов сопровождается обращением матрицы B^{k+1}, поэтому для сохранения эффективности алгоритма эта матрица должна быть легко обратима (например, B^{k+1} – диагональная, треугольная, трехдиагональная или ортогональная).

Приведем некоторые **примеры**. Для этого представим матрицу A в виде $A = A_1 + D + A_2$, где D – диагональная матрица, A_1 – левая треугольная, A_2 – правая треугольная. A_1 и A_2 имеют нулевую главную диагональ. Ненулевые элементы всех трех матриц совпадают с соответствующими элементами матрицы A .

- *Метод релаксации* (*простой итерации*). Здесь $B^{(k+1)} = E$, а $\tau^{(k+1)} = \tau$.
- *Итерационный метод Ричардсона*. Здесь $B^{(k+1)} = E$, а $\tau^{(k+1)}$ – переменный параметр.
- *Метод Якоби*. $B^{(k+1)} = D$, а $\tau^{(k+1)} = 1$.
- *Метод верхней релаксации*. $B^{(k+1)} = D + \omega A_1$, $\tau^{k+1} = \omega$, $0 < \omega < 2$, где ω – заданный числовой параметр. Для $\omega = 1$ как частный случай получается *метод Гаусса – Зейделя*. Для симметричных положительно определенных матриц A условие $0 < \omega < 2$ является условием сходимости метода.

Для некоторых из вышеназванных методов рассмотрим более детально способы получения матрицы C.

Предположим, что диагональные элементы матрицы A исходной системы (1.5) не равны 0 ($a_{ii} \neq 0,\ i = 1, 2, \ldots, m$). Разрешим первое уравнение системы (1.5) относительно x_1, второе относительно x_2 и т.д. Получим следующую эквивалентную систему, записанную в скалярном виде:

$$\begin{cases} x_1 = C_{12}x_2 + C_{13}x_3 + \ldots + C_{1n}x_n + d_1 \\ x_2 = C_{21}x_1 + C_{23}x_3 + \ldots + C_{2n}x_n + d_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = C_{n1}x_1 + C_{n2}x_2 + \ldots + C_{n,n-1}x_{n-1} + d_n \end{cases},$$

что совпадает с формулой (1.13), в которой матрица C и вектор $\vec{d}$ определены по следующим формулам:

$$C_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}}, i \neq j \\ 0, \quad i = j \end{cases}, \quad d_i = \frac{f_i}{a_{ii}}, \quad i = 1, 2, \ldots, n. \qquad (1.15)$$

Метод, основанный на таком приведении системы (1.5) к виду (1.13), называют **методом Якоби**. Теперь, задав нулевое приближение, по рекуррентным соотношениям (1.14) можем выполнять итерационный процесс.

Условие сходимости $\|C\| < 1$ в методе Якоби равносильно условию диагонального преобладания для исходной матрицы A:

$$|a_{ii}| > \sum_{\substack{j=1 \\ i \neq j}}^{m} |a_{ij}|, \quad i = 1, 2, \ldots, m.$$

Действительно, пусть для матрицы A выполняется условие диагонального преобладания. Разделим обе части данного неравенства на $|a_{ii}|$, получим неравенство

$$1 > \sum_{\substack{j=1 \\ i \neq j}}^{m} \frac{|a_{ij}|}{|a_{ii}|}, \quad i = 1, 2, \ldots, m.$$

С учетом (1.15) можно перейти к следующему неравенству:

$$1 > \sum_{j=1}^{m} |C_{ij}|, \quad i = 1, 2, \ldots, m.$$

То есть сумма элементов любой строки матрицы C меньше 1. Выполнение такого условия равносильно выполнению условия сходимости метода $\|C\| = \max\limits_{1\le i\le m} \sum\limits_{j=1}^{m} |C_{ij}| < 1$.

Под методом **Гаусса – Зейделя** обычно понимается такое видоизменение одношагового итерационного метода (1.15) решения СЛАУ, в котором для подсчета i-й компоненты $(k+1)$-го приближения $x_i^{(k+1)}$ к искомому вектору $\vec{x}^*$ используются уже вычисленные на этом, т.е. $(k+1)$-м шаге, значения первых $i-1$ компонент $x_{i-1}^{(k+1)}$. Это означает, что если система (1.5) тем или иным способом сведена (например, с помощью метода Якоби) к системе (1.13) с матрицей коэффициентов C и вектором свободных членов $\vec{d}$, то ее приближение к решению по методу Зейделя определяется системой равенств

$$
\begin{aligned}
x_1^{(k+1)} &= C_{11}x_1^{(k)} + C_{12}x_2^{(k)} + C_{13}x_3^{(k)} + \ldots + C_{1m}x_m^{(k)} + d_1 \\
x_2^{(k+1)} &= C_{21}x_1^{(k+1)} + C_{22}x_2^{(k)} + C_{23}x_3^{(k)} + \ldots + C_{2m}x_m^{(k)} + d_2 \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
x_m^{(k+1)} &= C_{m1}x_1^{(k+1)} + C_{m2}x_2^{(k+1)} + \ldots + C_{m,m-1}x_{m-1}^{(k+1)} + C_{m,m}x_m^{(k)} + d_3
\end{aligned}
$$

С точки зрения компьютерной реализации одношагового итерационного метода использование метода Гаусса – Зейделя означает, что элементы массива $\vec{x}$ будут постепенно замещаться новыми элементами. В связи с такой интерпретацией метод Гаусса – Зейделя иногда называют методом последовательных смещений.

Иногда исходную систему (1.5) не удается привести к виду (1.13), выполнив при этом условие сходимости метода. В этом случае можно воспользоваться **методом релаксации**. Этот метод основывается на соотношении

$$\frac{\vec{x}^{(k+1)} - \vec{x}^{(k)}}{\tau} = -A\vec{x}^{(k)} + \vec{f},$$

откуда $\vec{x}^{(k+1)} = \vec{x}^{(k)} - \tau\left(A\vec{x}^{(k)} - \vec{f}\right)$, где τ – итерационный параметр. Скалярные формулы метода релаксации имеют следующий вид:

$$
\begin{aligned}
x_1^{(k+1)} &= x_1^{(k)} - \tau\left(a_{11}x_1^{(k)} + a_{12}x_2^{(k)} + \ldots + a_{1n}x_n^{(k)} - f_1\right)\\
x_2^{(k+1)} &= x_2^{(k)} - \tau\left(a_{21}x_1^{(k)} + a_{22}x_2^{(k)} + \ldots + a_{2n}x_n^{(k)} - f_2\right)\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
x_n^{(k+1)} &= x_1^{(k)} - \tau\left(a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \ldots + a_{nn}x_n^{(k)} - f_n\right)
\end{aligned}
\quad (1.16)
$$

Раскрыв скобки, можно привести (1.16) к виду (1.14), где коэффициенты матрицы C и вектор свободных членов $\vec{d}$ будут иметь вид: $C_{ij} = \begin{cases} 1-\tau\ a_{ij}, i = j \\ -\tau\ a_{ij},\ i \neq j \end{cases}$, $d_i = \tau\ f_i$, $i = 1, 2, \ldots, n$. Подбором параметра τ можно добиться сходимости метода релаксации.

При использовании итерационных методов мы можем найти решение исходной СЛАУ (1.5) лишь приближенно с заданной точностью. Поэтому важной проблемой является вопрос о способе остановки итерационного процесса при достижении точности. Наиболее простой способ – это сравнение между собой соответствующих неизвестных с двух соседних итераций: (k + 1) и (k). Если максимальная из всех разностей становится меньше заданной точности ε, то итерационный процесс останавливается:

$$\max_{1 \le i \le m} \left| x_i^k - x_i^{k+1} \right| < \varepsilon\ .$$

При использовании метода релаксации для остановки итерационного процесса можно применить способ, связанный с вычислением вектора невязки $\vec{r}$:

$$r_i = \sum_{j=1}^{m} a_{ij} x_j^k - f_i\ ,$$

показывающего, насколько полученное приближение $\vec{x}^k$ отличается от точного решения. Затем вычисляется норма вектора невязки

$$\left\| \vec{r} \right\| = \max_{1 \le i \le m} | r_i |\ .$$

Если она мала, т.е. $\left\| \vec{r} \right\| < \varepsilon$, то итерационный процесс останавливается.

РАЗДЕЛ 2. ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

Тема 5. Численное интегрирование и дифференцирование

5.1. Постановка задачи численного интегрирования

Дано: определенный интеграл вида $I = \int_a^b f(x)dx$, где функция $f(x)$ задана на отрезке $[a,b]$. Найти: значение определенного интеграла.

Для некоторых видов функции $f(x)$ значение интеграла можно рассчитать точно, найдя первообразную и вычислив разность значений в верхнем и нижнем пределе. Однако в общем случае, если функция сложная либо задана дискретно, интеграл можно найти только приближенно, используя один из способов численного интегрирования.

Методы численного интегрирования основаны на замене интеграла некоторой суммой $I_n = \sum_{k=0}^{n} c_k f(x_k)$. Такая замена следует из определения интеграла как предела суммы $I = \lim_{n\to\infty} \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1})$, где $x_i \in [a,b]$ и $\xi_i \in (x_{i-1}, x_i)$. Положив значение n равным некоторому конечному числу, мы перейдем от предела к вышеописанной формуле.

Приближенное равенство $I = I_n$ называется **квадратурной формулой**, x_k – **узлами**, а c_k – **коэффициентами** квадратурной формулы. Разность между точным и приближенно вычисленным значениями интеграла $\psi_n = \int_a^b f(x)dx - \sum_{k=0}^{n} c_k f(x_k)$ называется **погрешностью** квадратурной формулы.

Итак, разобьем отрезок $[a,b]$ на n частей точками $a = x_0 < x_1 < x_2 < \ldots < x_n = b$. Набор этих точек называется **вычислительной сеткой**. В случае, если отрезок разбивается на равные части, длина которых равна h, мы получим равномерную сетку: $x_i = a + ih,\ i = 0,1,\ldots,n$. В этом случае

$$I = \int_a^b f(x)\,dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx.$$

Для построения квадратурной формулы на всем отрезке $[a,b]$ достаточно построить квадратурную формулу на отдельном отрезке $[x_{i-1}, x_i]$ и затем полученные значения просуммировать. Рассмотрим несколько основных формул численного интегрирования.

5.2. Формула прямоугольников

На отрезке $[a,b]$ построим сетку, состоящую из n частей и, соответственно, $(n + 1)$ узла. Пусть $f(x) = f(x_{i-1}), x \in [x_{i-1}, x_i]$. Это означает, что на каждом отрезке мы интерполируем функцию $f(x)$ при помощи левой кусочно-постоянной интерполяции. Тогда

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = \int_{x_{i-1}}^{x_i} f(x_{i-1})\,dx = f(x_{i-1}) \int_{x_{i-1}}^{x_i} dx = f(x_{i-1})(x_i - x_{i-1}).$$

Суммируя найденные значения, получим формулу $\int_a^b f(x)\,dx = \sum_{i=1}^{n} f(x_{i-1})(x_i - x_{i-1})$. Она называется формулой **левых прямоугольников**.

В случае, если расстояние (шаг) между соседними узлами x_i постоянный, т.е. сетка равномерная, формула примет более простой вид: $\int_a^b f(x)dx = h\sum_{i=1}^{n} f(x_{i-1})$.

Геометрическая интерпретация метода левых прямоугольников представлена на рис. 2.1, который показывает, что точное значение интеграла (площадь криволинейной области под графиком $f(x)$) заменяется суммой площадей прямоугольников, построенных под кусочно-постоянной интерполирующей функцией.

Аналогично может быть получена формула **правых прямоугольников**. В этом случае используется правая кусочно-постоянная интерполяция и за постоянное берется значение функции на правом конце отрезка (рис. 2.2). То есть при $x \in [x_{i-1}, x_i]$ $f(x) = f(x_i)$. В результате получим: $\int_a^b f(x)\,dx = \sum_{i=1}^{n} f(x_i)(x_i - x_{i-1})$ или в случае постоянного шага $\int_a^b f(x)\,dx = h\sum_{i=1}^{n} f(x_i)$.

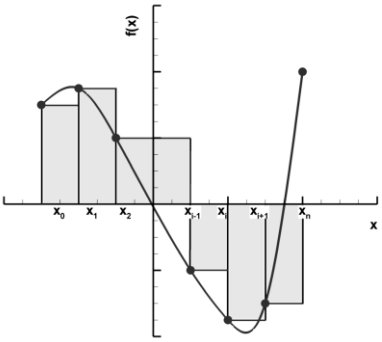


Рис. 2.1. Метод левых прямоугольников

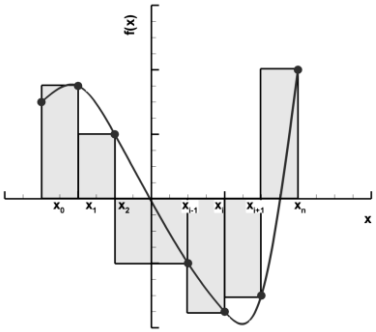


Рис. 2.2. Метод правых прямоугольников

Оценим погрешность формулы левых прямоугольников в случае постоянного шага сетки:

$$\Psi_n = \int_a^b f(x)dx - h\sum_{i=1}^{n} f(x_{i-1}) = \sum_{i=1}^{n}\left(\int_{x_{i-1}}^{x_i} f(x)dx - hf(x_{i-1})\right) = \sum_{i=1}^{n}\varphi_i,$$

$$\varphi_i = \int_{x_{i-1}}^{x_i} f(x)dx - hf(x_{i-1}) = \int_{x_{i-1}}^{x_i}(f(x) - f(x_{i-1}))dx.$$

Воспользуемся разложением в ряд Тейлора в окрестности точки x_{i-1}:

$$f(x) = f(x_{i-1}) + f'(\xi_i)(x - x_{i-1}), \xi_i \in [x_{i-1}, x_i].$$

Тогда

$$\varphi_i = \int_{x_{i-1}}^{x_i}[f(x_{i-1}) + f'(\xi_i)(x - x_{i-1}) - f(x_{i-1})]dx =$$

$$= \int_{x_{i-1}}^{x_i} f'(\xi_i)(x - x_{i-1})dx = \frac{h^2}{2} f'(\xi_i).$$

Пусть $M = \max\limits_{x\in[a,b]} |f'(x)|$, тогда

$$|\Psi_n| = \sum_{i=1}^{n}\frac{h^2}{2}|f'(\xi_i)| \le \frac{h^2}{2}M\sum_{i=1}^{n}1 = \frac{h^2}{2}Mn = \frac{h}{2}Mnh = \frac{Mh}{2}(b-a).$$

Отсюда видно, что погрешность зависит от шага по пространству в первой степени, т.е. при уменьшении шага в *k* раз погрешность также уменьшится в *k* раз. В этом случае говорят, что формула левых прямоугольников имеет **первый по** h **порядок точности**. Аналогичную оценку можно получить для формулы правых прямоугольников.

Если на каждом отрезке $[x_{i-1}, x_i]$ заменить значение функции *f*(*x*) на ее значение в середине отрезка, т.е. $f(x) = f\left(x_{i-1/2}\right), x \in [x_{i-1}, x_i]$, получим формулу **средних прямоугольников**:

$$\int_a^b f(x)dx = \sum_{i=1}^{n} f\left(x_{i-1/2}\right)\left(x_i - x_{i-1}\right)$$

или $\int_a^b f(x)dx = h\sum_{i=1}^{n} f\left(x_{i-1/2}\right)$ при постоянном шаге.

Если функция $f(x)$ задана таблично, среднее значение на локальном отрезке можно вычислить с помощью линейной интерполяции $x_{i-1/2} = \frac{x_{i-1} + x_i}{2}$, и тогда метод средних прямоугольников имеет вид: $\int_a^b f(x)dx = h\sum_{i=1}^{n} f\left(\frac{x_{i-1} + x_i}{2}\right)$.

Проведем аналогичные выкладки для оценки погрешности метода средних прямоугольников:

$$\Psi_n = \int_a^b f(x)dx - h\sum_{i=1}^{n} f\left(x_{i-1/2}\right) = \sum_{i=1}^{n}\left(\int_{x_{i-1}}^{x_i} f(x)dx - hf\left(x_{i-1/2}\right)\right) = \sum_{i=1}^{n}\varphi_i,$$

$$\varphi_i = \int_{x_{i-1}}^{x_i} f(x)dx - f\left(x_{i-1/2}\right)h = \int_{x_{i-1}}^{x_i}\left(f(x) - f\left(x_{i-1/2}\right)\right)dx.$$

Воспользуемся формулой Тейлора:

$$f(x) = f\left(x_{i-1/2}\right) + f'(\xi_i)\left(x - x_{i-1/2}\right) + \frac{1}{2}f''(\xi_i)\left(x - x_{i-1/2}\right)^2,$$

$\xi_i \in [x_{i-1}, x_i]$.

Тогда

$$\varphi_i = \int_{x_{i-1}}^{x_i}\left[\begin{array}{l} f\left(x_{i-1/2}\right) + f'(\xi_i)\left(x - x_{i-1/2}\right) + \frac{1}{2}f''(\xi_i)\left(x - x_{i-1/2}\right)^2 \\ - f\left(x_{i-1/2}\right)\end{array}\right]dx =$$

$$= \int_{x_{i-1}}^{x_i}\left[f'(\xi_i)\left(x - x_{i-1/2}\right) + \frac{1}{2}f''(\xi_i)\left(x - x_{i-1/2}\right)^2\right]dx =$$

$$= \frac{f'(\xi_i)}{2}\left(x - x_{i-1/2}\right)^2\Big|_{x_{i-1}}^{x_i} + \frac{1}{6}f''(\xi_i)\left(x - x_{i-1/2}\right)^3\Big|_{x_{i-1}}^{x_i} =$$

$$= \frac{f'(\xi_i)}{2}\left[\left(x_i - x_{i-1/2}\right)^2 - \left(x_{i-1} - x_{i-1/2}\right)^2\right] +$$

$$+ \frac{1}{6}f''(\xi_i)\left[\left(x_i - x_{i-1/2}\right)^3 - \left(x_{i-1} - x_{i-1/2}\right)^3\right].$$

Поскольку $x_i - x_{i-1/2} = \frac{h}{2}$ и $x_{i-1} - x_{i-1/2} = -\frac{h}{2}$, то

$$\varphi_i = \frac{1}{6}f''(\xi_i)\left[\frac{h^3}{8} + \frac{h^3}{8}\right] = \frac{h^3}{24}f''(\xi_i).$$

Пусть $M = \max\limits_{x\in[a,b]}\left|f''(x)\right|$, тогда

$$\left|\Psi_n\right| = \sum_{i=1}^{n}\frac{h^3}{24}\left|f''(\xi_i)\right| \le \frac{h^3}{24}M\sum_{i=1}^{n}1 = \frac{h^3}{24}Mn = \frac{Mh^2}{24}(b-a),$$

т.е. формула средних прямоугольников имеет **второй порядок точности**. При уменьшении шага в k раз погрешность уменьшится пропорционально квадрату шага, т.е. в k^2 раз.

5.3. Формула трапеций

Во всех рассмотренных формулах площадь криволинейной трапеции под функцией заменялась площадью прямоугольников.

В методе трапеций криволинейная трапеция заменяется на прямоугольную (рис. 2.3) путем интерполяции функции при помощи кусочно-линейной зависимости. Площадь полученной прямоугольной трапеции вычисляется по известной формуле: $\int\limits_{x_{i-1}}^{x_i} f(x)dx \approx \frac{f(x_{i-1}) + f(x_i)}{2}(x_i - x_{i-1})$. Тогда интеграл находится по формуле $I_n = \sum\limits_{i=1}^{n}\frac{f(x_{i-1}) + f(x_i)}{2}(x_i - x_{i-1})$.

Формула трапеций может быть также получена путем замены подынтегральной функции интерполяционным полиномом первой степени:

$$L_{1,i}(x)=\frac{1}{h}\left[(x-x_{i-1})f(x_i)-(x-x_i)f(x_{i-1})\right].$$

Предположим, что шаг по пространству постоянен и равен h. Тогда

$$\int_{x_{i-1}}^{x_i} L_{1,i}(x)dx=\frac{1}{h}\int_{x_{i-1}}^{x_i}(x-x_{i-1})f(x_i)dx-\frac{1}{h}\int_{x_{i-1}}^{x_i}(x-x_i)f(x_{i-1})dx=$$

$$=\frac{1}{2h}f(x_i)h^2-\frac{1}{2h}f(x_{i-1})\left(-h^2\right)=\frac{f(x_i)+f(x_{i-1})}{2}h.$$

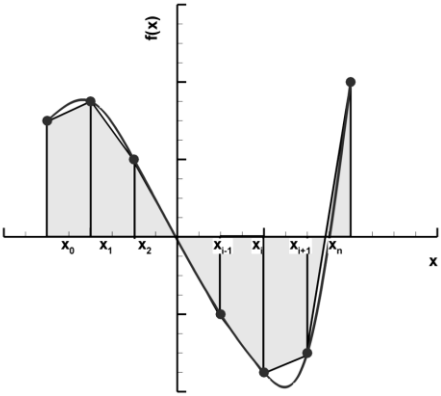


Рис. 2.3. Метод трапеций

Приближенное значение интеграла на отрезке $[a,b]$ будет равно:

$$\int_a^b f(x)dx=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} f(x)dx=\sum_{i=1}^{n}\frac{f(x_i)+f(x_{i-1})}{2}h\,.$$

Можно показать, что в случае таблично заданной (дискретной) функции формула трапеций совпадает с формулой средних прямоугольников и также имеет второй порядок точности.

Формулу трапеций для случая постоянного шага можно также переписать в виде:

$$\int_a^b f(x)dx = h\left(\frac{f_0 + f_n}{2} + \sum_{i=1}^{n-1} f(x_i)\right).$$

5.4. Формула Симпсона

При вычислении интеграла $\int_{x_{i-1}}^{x_i} f(x)dx$ с помощью метода Симпсона (метода парабол) функцию $f(x)$ на локальном отрезке $[x_{i-1}, x_i]$ интерполируют при помощи кусочно-параболической интерполяции. В этом случае требуется третья точка для построения параболы, и в качестве нее выбирают середину отрезка $[x_{i-1}, x_i]$. Таким образом, парабола проходит через точки $(x_{i-1}, f(x_{i-1}))$, $\left(x_{i-1/2}, f\left(x_{i-1/2}\right)\right)$, $(x_i, f(x_i))$, где $x_{i-1/2} = \frac{x_{i-1} + x_i}{2}$. Для этих трех точек построим полином Лагранжа, который будет иметь вторую степень:

$$f(x) \approx L_{2,i}(x), x \in [x_{i-1}, x_i],$$

$$L_{2,i}(x) = \frac{2}{h^2}\left[\left(x - x_{i-1/2}\right)(x - x_i) f_{i-1} - 2(x - x_{i-1})(x - x_i) f_{i-1/2} + \right.$$

$$\left. + (x - x_i)\left(x - x_{i-1/2}\right) f_i \right].$$

Здесь $f_i = f(x_i)$, $f_{i-1/2} = f\left(x_{i-1/2}\right)$, $f_{i-1} = f(x_{i-1})$.

Тогда

$$\int_{x_{i-1}}^{x_i} f(x)dx = \int_{x_{i-1}}^{x_i} L_{2,i}(x)dx = \frac{2}{h^2}\left[f_{i-1}\int_{x_{i-1}}^{x_i}\left(x - x_{i-1/2}\right)(x-x_i)dx - \right.$$

$$\left. - 2f_{i-1/2}\int_{x_{i-1}}^{x_i}(x-x_{i-1})(x-x_i)dx + f_i\int_{x_{i-1}}^{x_i}(x-x_i)\left(x-x_{i-1/2}\right)dx\right] =$$

$$= \frac{2}{h^2}\left[f_{i-1}\int_{x_{i-1}}^{x_i}\left(x - x_i + \frac{h}{2}\right)(x-x_i)dx - \right.$$

$$- 2f_{i-1/2}\int_{x_{i-1}}^{x_i}(x-x_{i-1})(x-x_{i-1}-h)dx +$$

$$\left. + f_i\int_{x_{i-1}}^{x_i}(x-x_i)\left(x-x_{i-1}-\frac{h}{2}\right)dx\right] =$$

$$= \frac{2}{h^2}\left[f_{i-1}\left(\frac{h^3}{3}-\frac{h^3}{4}\right) - 2f_{i-1/2}\left(\frac{h^3}{3}-\frac{h^3}{2}\right) + f_i\left(\frac{h^3}{3}-\frac{h^3}{4}\right)\right] =$$

$$= \frac{h}{6}\left[f_{i-1} + 4f_{i-1/2} + f_i\right].$$

Таким образом, мы получаем **формулу Симпсона**

$$\int_a^b f(x)dx = \frac{h}{6}\sum_{i=1}^{n}\left[f(x_{i-1}) + 4f\left(\frac{x_{i-1}+x_i}{2}\right) + f(x_i)\right].$$

Можно показать, что формула Симпсона имеет **четвертый порядок точности**.

Пример 2.1. Вычислить интеграл $J = \int_{-1}^{2}\left(3-x^2\right)(1-x)dx$.

Найдем значение определенного интеграла точно:

$$J = \frac{x^4}{4} - \frac{x^3}{4} - 3\frac{x^2}{2} + 3x\Big|_a^b = 5.25.$$

Разобьем отрезок интегрирования $[-1, 2]$ на 10 частей, т.е. $h = \frac{2-(-1)}{10} = 0.3$. Проведем интегрирование рассмотренными численными методами. В результате получим следующие значения:

Название метода	Приближенное значение	Погрешность
Формула левых прямоугольников	5.7225	0.4725
Формула правых прямоугольников	4.8225	0.4275
Формула средних прямоугольников	5.23875	0.01125
Формула трапеций	5.2725	0.0225
Формула Симпсона	5.25	0

Можно заметить, что метод Симпсона дал абсолютно точное значение интеграла. Это связано с тем, что первообразная функция в данном примере является полиномом четвертого порядка, для которого метод Симпсона дает точное значение.

5.5. Численное дифференцирование

Численное дифференцирование, т.е. нахождение значений производных заданной функции $y = f(x)$ в заданных точках x, в отличие от численного интегрирования, можно считать не столь актуальной проблемой в связи с отсутствием принципиальных трудностей с аналитическим нахождением производных. Однако имеется ряд важных задач, для которых численное дифференцирование является единственным способом нахождения производной. Это, например, поиск производной таблично заданной функции или дифференцирование функции в процессе численного решения, когда значения этой функции известны только в узлах сетки. Кроме того, если при аналитическом диф-

ференцировании получаются громоздкие выражения, использование численного подхода упрощает задачу.

Существует несколько способов для получения формул численного дифференцирования, которые в конечном счете могут привести к одним и тем же формулам. Во-первых, можно аппроксимировать таблично заданную функцию каким-либо способом (линейная интерполяция, многочлен Лагранжа, сплайн-функции и т.д.) и дифференцировать полученную непрерывную функцию, приближающую исходную. Во-вторых, для вывода формул численного дифференцирования можно воспользоваться понятием **конечных разностей.**

Пусть узлы таблицы x_i, $i = 0, 1, \ldots, N$, расположены на равных расстояниях: $x_i = x_0 + ih$, f_i – соответствующие значения функции; величину h называют шагом таблицы. Разности значений функции в соседних узлах называют разностями первого порядка. В каждом внутреннем узле x_i, $i = 1, \ldots, N-1$, можно составить три разности первого порядка: разность вперед

$$\Delta_+ f_i = f_{i+1} - f_i ,$$

разность назад:

$$\Delta_- f_i = f_i - f_{i-1} = \Delta_+ f_{i-1}$$

и центральную разность

$$\Delta_\pm f_i = f_{i+1} - f_{i-1}.$$

Разности высших порядков образуют при помощи рекуррентных соотношений

$$\Delta^m f_i = \Delta\left(\Delta^{m-1} f_i\right) = \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i .$$

Используя эти формулы, первую производную можно определить тремя разными способами:

$$f_+'(x_i) = \frac{f_{i+1} - f_i}{h} , \qquad (2.1)$$

$$f_-'(x_i) = \frac{f_i - f_{i-1}}{h} , \qquad (2.2)$$

$$f_\pm'(x_i) = \frac{f_{i+1} - f_{i-1}}{2h} . \qquad (2.3)$$

Геометрически вычисление производной по трем этим формулам эквивалентно замене касательной в точке B прямыми

BC, *AB* и *AC* соответственно и поиску тангенса угла наклона этих прямых вместо тангенса угла наклона касательной (рис. 2.4).

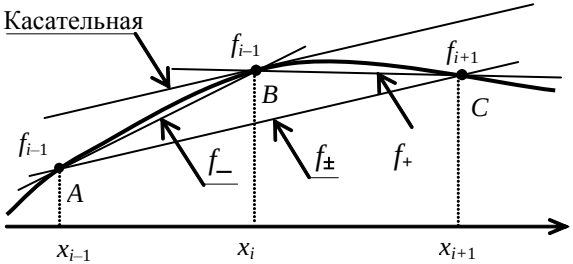


Рис. 2.4. Геометрическая иллюстрация разностного дифференцирования

Рисунок 2.4 показывает, что наиболее точно поведение производной (касательной) в точке x_i передает прямая *AC*, угол наклона которой определяет формула центральной разности $f_{\pm}$. Изучим вопрос о порядке точности (аппроксимации) этих формул. Разложим $f(x)$ в ряд Тейлора в окрестности точки x_i:

$$f_{i+1} = f_i + hf_i' + \frac{h^2}{2} f_i'' + \frac{h^3}{6} f_i''' + \dots ,$$

$$f_{i-1} = f_i - hf_i' + \frac{h^2}{2} f_i'' - \frac{h^3}{6} f_i''' + \dots .$$

Подставив эти разложения в (2.1), получаем

$$f_+'(x_i) = \frac{f_i + hf_i' + \frac{h^2}{2} f_i'' + \frac{h^3}{6} f_i''' + \dots - f_i}{h} = f_i' + \frac{h}{2} f_i'' + \frac{h^2}{6} f_i''' + \dots$$

Здесь f_i' – первая производная, которую необходимо найти, а $\frac{h}{2} f_i'' + \frac{h^2}{6} f_i''' + \dots$ – погрешность, с которой вычисляется производная. Видим, что первый, самый большой член погрешности имеет порядок h, значит, при измельчении шага сетки погрешность будет уменьшаться пропорционально h в степени 1. Поэтому говорят, что формула имеет первый порядок точности.

Нетрудно показать, что формула (2.2) также имеет первый порядок аппроксимации.

Покажем, что формула центральной разности имеет второй порядок точности. Подставим в (2.3) разложения для f_{i+1} и f_{i-1}:

$$f'_{\pm}(x_i) = \frac{f_i + hf'_i + \frac{h^2}{2} f''_i + \frac{h^3}{6} f'''_i + \ldots - f_i + hf'_i - \frac{h^2}{2} f''_i + \frac{h^3}{6} f'''_i + \ldots}{2h} = f'_i + \frac{h^2}{6} f'''_i + \ldots$$

Погрешность вычисления производной пропорциональна h^2, значит, формула (2.3) имеет второй порядок аппроксимации. Таким образом, мы подтвердили вывод, который сделан на основании рис. 2.4.

Используя понятие конечных разностей, можно получить формулы для аппроксимации производных высших порядков. Покажем это на примере формулы для аппроксимации второй производной:

$$f''(x_i) \approx \frac{(f')_i - (f')_{i-1}}{h} = \frac{\frac{f_{i+1} - f_i}{h} - \frac{f_i + f_{i-1}}{h}}{h} = \frac{f_{i+1} - 2f_i + f_{i-1}}{h^2}.$$

Можно доказать, что эта формула имеет второй порядок точности (доказать самостоятельно).

Пример 2.2. Вычислить приближенные значения производных от функции $f(x) = \sin(x)$, заданной на отрезке $[0, \pi/2]$ с шагом $\pi/6$ с разным порядком точности.

Построим таблицу значений функции и вычислим ее производные точно: $f'(x) = \cos(x)$ и с использованием формул (2.1) – (2.3). Шаг аргумента $h = \pi/6 \approx 0.524$, $\sin(0) = \cos(\pi/2) = 0$, $\sin(\pi/6) = \cos(\pi/3) = 0.5$, $\sin(\pi/3) = \cos(\pi/6) \approx 0.866$, $\sin(\pi/2) = \cos(0) = 1$.

x	$f(x)$	$f'(x)$ точно	$f'_+(x)$	$f'_-(x)$	$f'_{\pm}(x)$
0	0	1	0.954	–	–
0.524	0.5	0.866	0.698	0.954	0.827
1.047	0.866	0.5	0.256	0.698	0.478
1.571	1	0	–	0.256	–

Анализ результатов, содержащихся в таблице, показывает, что формулы «разность назад» и «разность вперед» дают значительные погрешности в вычислении производных, что можно объяснить довольно грубой сеткой, т.е. большим шагом h. В то же время центральная разность, несмотря на грубый шаг, позволяет получить значение производной с хорошей точностью. Однако эту формулу нельзя применять в крайних точках таблицы.

5.6. Метод неопределенных коэффициентов

Для того, чтобы получить формулы численного дифференцирования в любой точке сетки с любым порядком точности, используют метод неопределенных коэффициентов. Покажем, как работает этот метод на примере вывода формулы для первой производной в крайней левой точке таблицы. Представим приближенно в точке $x = x_0$ первую производную таблично заданной функции в виде линейной комбинации ее значений в узлах:

$$f'(x_0) \approx \frac{\alpha f_0 + \beta f_1 + \gamma f_2}{h}, \qquad (2.4)$$

где α, β, γ – неопределенные коэффициенты, которые выберем из условия, чтобы эта формула имела второй порядок аппроксимации, т.е. главный член погрешности был равен $c \cdot h^2$.

Разложим f_1, f_2 в ряд Тейлора и подставим эти выражения в формулу (2.4):

$$f_1 = f_0 + hf_0' + \frac{h^2}{2} f_0'' + \frac{h^3}{6} f_0''' + \dots,$$

$$f_2 = f_0 + 2hf_0' + 2h^2 f_0'' + \frac{4h^3}{3} f_0''' + \dots.$$

$$f'(x_0) + c \cdot h^2 \approx \frac{1}{h}\Big[\ \alpha f_0 + \beta\left(f_0 + hf_0' + \frac{h^2}{2} f_0'' + \frac{h^3}{6} f_0''' \right) +$$

$$+ \gamma\left(f_0 + 2hf_0' + 2h^2 f_0'' + \frac{4h^3}{3} f''' \right)\ \Big] =$$

$$= \frac{f_0}{h}(\alpha + \beta + \gamma) + f_0'(\beta + 2\gamma) + hf_0''\left(\frac{\beta}{2} + 2\gamma\right) + h^2 f'''\left(\frac{\beta}{6} + \frac{4\gamma}{3}\right).$$

Приравняв коэффициенты, стоящие перед соответствующими степенями h, получим систему линейных уравнений:

$$\begin{cases} \alpha + \beta + \gamma = 0 \\ \beta + 2\gamma = 1 \\ \dfrac{\beta}{2} + 2\gamma = 0 \end{cases},$$

из которой найдем искомые коэффициенты:

$$\alpha = -\frac{3}{2},\ \beta = 2,\ \ \gamma = -\frac{1}{2}.$$

Подставив коэффициенты в (2.4), получим формулу:

$$f'(x_0) + ch^2 \approx \frac{-3f_0 + 4f_1 - f_2}{2h}. \qquad (2.5)$$

Аналогичные выкладки позволяют получить формулу для вычисления первой производной в последней, крайней правой точке таблицы:

$$f'(x_N) + ch^2 \approx \frac{3f_N - 4f_{N-1} + f_{N-2}}{2h}. \qquad (2.6)$$

Пример 2.3. Вычислить приближенные значения производных от функции $f(x) = \sin(x)$, заданной на отрезке $[0, \pi/2]$ с шагом $\pi/6$ при $x = 0$ и $x = \pi/2$ со вторым порядком точности.

Для вычисления производной используем таблицу значений функции из Примера 2.2. Точные значения производной: $f'(x_0) = \cos(x_0) = 1$, $f'(x_N) = \cos(x_N) = 0$. Производная, вычисленная по формуле первого порядка точности, дает большую погрешность в последней точке. Вычислим производные по формулам (2.5) и (2.6):

$$f'(x_0) = \frac{-3f_0 + 4f_1 - f_2}{2h} = \frac{-3\cdot 0 + 4\cdot 0.5 - 0.866}{1.047} = 1.083.$$

$$f'(x_N) \approx \frac{3f_N - 4f_{N-1} + f_{N-2}}{2h} = \frac{3\cdot 1 - 4\cdot 0.866 + 0.5}{1.047} = 0.034.$$

Видно, что формулы второго порядка позволяют достаточно точно вычислить значения первой производной.

Тема 6. Численные методы решения задачи Коши для обыкновенных дифференциальных уравнений

Решение обыкновенных дифференциальных уравнений (ОДУ) занимает важное место среди прикладных задач механики, физики, химии и техники. ОДУ описывают движение системы взаимодействующих материальных точек, химической кинетики, электрических цепей, моделируют статический прогиб упругого стержня (сопротивление материалов) и многие другие процессы. Ряд важных задач для уравнений в частных производных также сводится к задачам для ОДУ. Так бывает, если многомерная задача допускает разделение переменных (например, задачи на нахождение собственных колебаний упругих балок и мембран простейшей формы).

6.1. Постановка задачи

Требуется найти решение ОДУ первого порядка

$$\frac{dy}{dx} = f(x, y). \tag{2.7}$$

Известно, что общее решение (2.7) содержит произвольную константу C, т.е. является однопараметрическим семейством интегральных кривых

$$y(x) = \int f(x, y)\, dx + C.$$

Для выбора конкретной интегральной кривой следует определить значение константы C, для чего достаточно задать при каком-либо значении $x = x_0$ значение

$$y(x_0) = y_0. \tag{2.8}$$

Поэтому задача Коши, или задача с начальными данными, позволяющая получить единственное решение уравнения (2.7), формулируется так: найти $y(x)$ – решение уравнения (2.7) с начальным условием (2.8).

В случае точного решения ОДУ мы получаем аналитическое представление искомой функции. Несмотря на внешнюю простоту уравнения (2.7), решить его аналитически, т.е. найти общее решение $y = y(x, C)$ с тем, чтобы потом выделить из не-

го интегральную кривую $y = y(x)$, проходящую через точку (x_0, y_0), удается лишь для некоторых специальных типов уравнений. Поэтому большое значение приобретают приближенные способы решения начальных задач ОДУ. При численном же решении мы будем искать приближенное решение в узлах расчетной сетки $x_i = x_0 + ih, \quad i = 0,1,\ldots,n,$ с шагом $h = \dfrac{x_n - x_0}{n}$. То есть вместо непрерывной зависимости $y(x)$ мы найдем приближенные значения в узлах сетки $y_i = y(x_i)$.

Для построения численных методов решения ОДУ проинтегрируем уравнение на отрезке $[x_i, x_{i+1}]$, получим

$$y_{i+1} - y_i = \int_{x_i}^{x_{i+1}} f(x, y)\,dx. \qquad (2.9)$$

Чтобы найти все значения y_i, нужно каким-то образом вычислить интеграл, стоящий в правой части (2.9). Применяя различные квадратурные формулы, будем получать методы решения задачи (2.7), (2.8) разного порядка точности.

6.2. Метод Эйлера

Если для вычисления интеграла в (2.9) воспользоваться простейшей формулой левых прямоугольников первого порядка

$$\int_{x_i}^{x_{i+1}} f(x, y)\,dx = hf(x_i, y_i),$$

то получается **явная формула Эйлера**:

$$y_{i+1} = y_i + hf(x_i, y_i),\ i = 0,1,\ldots,n-1, \qquad (2.10)$$

имеющая первый порядок аппроксимации.

Реализация метода. Поскольку $x_0, y_0, f(x_0, y_0)$ известны, последовательно применяя (2.10), определим все y_i: $y_1 = y_0 + hf(x_0, y_0),\ y_2 = y_1 + hf(x_1, y_1), \ldots$.

Геометрическая интерпретация метода Эйлера приведена на рис. 2.5. Пользуясь тем, что в точке x_0 известно решение

$y(x_0)=y_0$ и значение его производной $y'(x_0)=\left.\frac{dy}{dx}\right|_{x=x_0}=$ $=f(x_0,y_0)$, можно записать уравнение касательной к графику искомой функции $y=y(x)$ в точке $f(x_0,y_0)$: $y=y_0+f(x_0,y_0)(x-x_0)$. При достаточно малом шаге h ордината $y_1=y_0+hf(x_0,y_0)$ этой касательной, полученная подстановкой в правую часть значения $x_1=x_0+h$, должна мало отличаться от ординаты $y(x_1)$ решения $y(x)$ задачи Коши. Следовательно, точка (x_1,y_1) пересечения касательной с прямой $x=x_1$ может быть приближенно принята за новую начальную точку. Через эту точку снова проведем прямую $y=y_1+f(x_1,y_1)(x-x_1)$, которая приближенно отражает поведение касательной к $y(x)$ в точке $(x_1,y(x_1))$. Подставляя сюда $x_2=x_1+h$ (т.е. пересечение с прямой $x=x_2$), получим приближенное значение $y(x)$ в точке x_2: $y_2=y_1+hf(x_1,y_1)$ и т.д. В итоге для i-й точки получим формулу Эйлера.

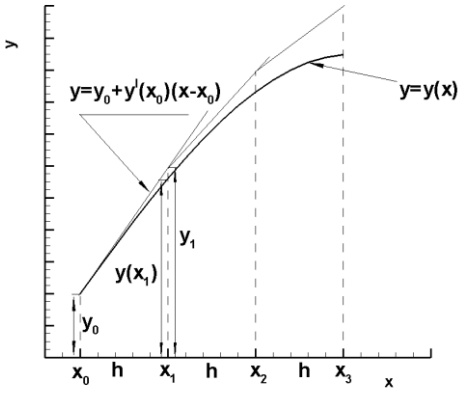


Рис. 2.5. Геометрическая интерпретация метода Эйлера

Если в (2.9) использовать формулу правых прямоугольников, то получим неявный метод Эйлера:

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}),\ i = 0,1,\ldots,n-1. \qquad (2.11)$$

Этот метод называют неявным, поскольку для вычисления неизвестного значения $y_{i+1} \approx y(x_{i+1})$ по известному $y_i \approx y(x_i)$ требуется решать в общем случае нелинейное уравнение. Неявный метод Эйлера также имеет первый порядок аппроксимации.

6.3. Модифицированный метод Эйлера

В данном методе вычисление y_{i+1} состоит из двух этапов:

$$\tilde{y}_{i+1} = y_i + hf(x_i, y_i),$$

$$y_{i+1} = y_i + \frac{h}{2}\left[f(x_i, y_i) + f(x_{i+1}, \tilde{y}_{i+1})\right]. \qquad (2.12)$$

Эта схема называется также методом предиктор – корректор (от англ. предсказать – исправить). Действительно, на первом этапе, совпадающем с формулой Эйлера, приближенное значение предсказывается с первым порядком точности, а на втором этапе это значение корректируется, так что в результате схема имеет второй порядок точности.

6.4. Методы Рунге – Кутты

Идея построения явных методов Рунге – Кутты *p*-го порядка заключается в получении приближений к значениям $y(x_{i+1})$ по формуле вида $y_{i+1} = y_i + h\varphi(x_i, y_i, h)$, где

$$\varphi(x_i, y_i, h) = \sum_{n=1}^{q} c_n k_n^i(h),$$

$$k_1^i(h) = f(x_i, y_i),$$

$$k_2^i(h) = f\left(x_i + \alpha_2 h, y_i + \beta_{21} k_1^i(h)\right),$$

$$k_3^i(h) = f\left(x_i + \alpha_3 h, y_i + \beta_{31} k_1^i(h) + \beta_{32} k_2^i(h)\right),$$

$$\ldots$$

$$k_q^i(h) = f\left(x_i + \alpha_q h, y_i + \beta_{q1} k_1^i(h) + \ldots + \beta_{q,q-1} k_{q-1}^i(h)\right).$$

Здесь α_n, β_{nj}, $0 < j < n \le q$ – некоторые фиксированные числа (параметры), которые подбирают таким образом, чтобы получить нужный порядок аппроксимации *p*. Как правило, для каждого p существует не одна схема Рунге – Кутты порядка *p*, а целое параметрическое семейство. Так, схемы Рунге – Кутты **второго порядка точности** образуют однопараметрическое семейство

$$k_1^i = f(x_i, y_i), \quad k_2^i = f\left(x_i + \frac{h}{2a}, y_i + \frac{h}{2a}k_1^i\right), \qquad (2.13)$$
$$y_{i+1} = y_i + h\left[(1-a)k_1^i + ak_2^i\right].$$

Выделим из семейства методов (2.13) два наиболее простых и часто используемых частных случая. При $a = \frac{1}{2}$ получаем формулы

$$k_1^i = f(x_i, y_i), k_2^i = f\left(x_i + h, y_i + hk_1^i\right),$$
$$y_{i+1} = y_i + \frac{h}{2}\left[k_1^i + k_2^i\right], \quad i = 0, 1, 2, \ldots, \qquad (2.14)$$

которые совпадают с формулами модифицированного метода Эйлера (2.12). При *a* = 1 выводим новый простой метод:

$$k_1^i = f(x_i, y_i), k_2^i = f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}k_1^i\right),$$
$$y_{i+1} = y_i + hk_2^i, \quad i = 0, 1, 2, \ldots,$$

который называется методом средней точки.

Схема Рунге – Кутты четвертого порядка точности. При *p* = 4 можно получить один из вариантов метода:

$$k_1 = f(x_i, y_i), \qquad k_2 = f\left(x_i + \frac{h}{2}, y_i + \frac{hk_1}{2}\right),$$
$$k_3 = f\left(x_i + \frac{h}{2}, y_i + \frac{hk_2}{2}\right), \quad k_4 = f(x_i + h, y_i + hk_3), \qquad (2.15)$$
$$y_{i+1} = y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4), \; i = 0, 1, 2, \ldots$$

6.5. Методы приближенного решения задачи Коши для системы ОДУ и ОДУ высших порядков

Рассмотренные выше методы решения задачи Коши для одного уравнения могут быть использованы также для решения систем дифференциальных уравнений первого порядка и уравнений высших порядков.

Пусть задана задача Коши для системы двух уравнений первого порядка:

$$\begin{cases} \dfrac{dy}{dx} = \varphi(x, y, z) \\ \dfrac{dz}{dx} = \psi(x, y, z) \end{cases} \tag{2.16}$$

с начальными условиями $y(x_0) = y_0, \quad z(x_0) = z_0$. Обобщим формулы явного метода Эйлера (2.4) для этой системы, записав схему для каждого уравнения (2.16):

$$y_{i+1} = y_i + h\varphi(x_i, y_i, z_i),$$
$$z_{i+1} = z_i + h\psi(x_i, y_i, z_i).$$

Модифицированный метод Эйлера (2.12) примет вид:

$$\tilde{y}_{i+1} = y_i + h\varphi(x_i, y_i, z_i),$$
$$\tilde{z}_{i+1} = z_i + h\psi(x_i, y_i, z_i),$$
$$y_{i+1} = y_i + \frac{h}{2}\left[\varphi(x_i, y_i, z_i) + \varphi(x_i, \tilde{y}_{i+1}, \tilde{z}_{i+1})\right],$$
$$z_{i+1} = z_i + \frac{h}{2}\left[\psi(x_i, y_i, z_i) + \psi(x_i, \tilde{y}_{i+1}, \tilde{z}_{+1i})\right],$$

а схема Рунге – Кутты четвертого порядка точности (2.15):

$$y_{i+1} = y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$
$$z_{i+1} = z_i + \frac{h}{6}(l_1 + 2l_2 + 2l_3 + l_4),$$
$$k_1 = \varphi(x_i, y_i, z_i), l_1 = \psi(x_i, y_i, z_i),$$
$$k_2 = \varphi\left(x_i + \frac{h}{2}, y_i + \frac{hk_1}{2}, z_i + \frac{hl_1}{2}\right), l_2 = \psi\left(x_i + \frac{h}{2}, y_i + \frac{hk_1}{2}, z_i + \frac{hl_1}{2}\right),$$

$$k_3 = \varphi\left(x_i + \frac{h}{2}, y_i + \frac{hk_2}{2}, z_i + \frac{hl_2}{2}\right), l_3 = \psi\left(x_i + \frac{h}{2}, y_i + \frac{hk_2}{2}, z_i + \frac{hl_2}{2}\right),$$

$$k_4 = \varphi(x_i + h, y_i + hk_3, z_i + hl_3), l_4 = \psi(x_i + h, y_i + hk_3, z_i + hl_3).$$

Приближенное решение вычисляется путем последовательного применения этих формул для каждого узла расчетной сетки.

Также аналогичным способом можно решить ОДУ высокого порядка. Например, рассмотрим задачу Коши для уравнения второго порядка

$$\frac{d^2y}{dx^2} = f\left(x, y, \frac{dy}{dx}\right), y(x_0) = y_0, \frac{dy}{dx}(x_0) = z_0.$$

Введем обозначение $z(x) = \frac{dy}{dx}$. Тогда исходная задача Коши для уравнения второго порядка сводится к следующей задаче для системы двух ОДУ первого порядка:

$$\begin{cases} \frac{dy}{dx} = z, \\ \frac{dz}{dx} = f(x, y, z), \end{cases}$$

$$y(x_0) = y_0, z(x_0) = z_0.$$

Можно заметить, что эта запись эквивалентна системе (2.16) при $\varphi(x, y, z) = z$ и $\psi(x, y, z) = f(x, y, z)$. Таким образом, полученная система решается вышеописанным способом при помощи одного из методов решения задачи Коши.

Пример 2.4. Найти решение задачи Коши

$\frac{d^2y}{dx^2} + 2\frac{dy}{dx} + y(x) = x, y(0) = 1, \frac{dy}{dx}(0) = 0$ на отрезке $[0,1]$.

Пусть известно точное решение данного ОДУ:

$y(x) = 3e^{-x} + 2xe^{-x} + x - 2$.

Проверим, что точное решение удовлетворяет уравнению:

$$\frac{dy}{dx} = -e^{-x} - 2xe^{-x} + 1, \quad \frac{d^2y}{dx^2} = -e^{-x} + 2xe^{-x},$$

$$\frac{d^2y}{dx^2}+2\frac{dy}{dx}+y(x)=$$
$$=-e^{-x}+2xe^{-x}+2\left(-e^{-x}-2xe^{-x}+1\right)+3e^{-x}+2xe^{-x}+x-2=x,$$
$$y(0)=3e^0+0+0-2=1,\frac{dy}{dx}(0)=-e^0+1=0.$$

Введем функцию $z(x)=\frac{dy}{dx}$ и получим следующую задачу Коши для системы двух ОДУ первого порядка:

$$\frac{dy}{dx}=z,\quad \frac{dz}{dx}=-2z-y+x,\quad y(0)=1,\quad z(0)=0.$$

Используем формулы явного метода Эйлера:

$y_{i+1}=y_i+hz_i$,

$z_{i+1}=z_i+h(-2z_i-y_i+x_i)$,

$y_0=1,\quad z_0=0$,

модифицированного метода Эйлера:

$\tilde{y}_{i+1}=y_i+hz_i$,

$\tilde{z}_{i+1}=z_i+h(-2z_i-y_i+x_i)$,

$$y_{i+1}=y_i+\frac{h}{2}[z_i+\tilde{z}_{i+1}],$$

$$z_{i+1}=z_i+\frac{h}{2}[(-2z_i-y_i+x_i)+(-2\tilde{z}_{i+1}-\tilde{y}_{i+1}+x_{i+1})],$$

$x_{i+1}=x_i+h$,

и четырехэтапного метода Рунге – Кутты:

$$y_{i+1}=y_i+\frac{h}{6}(k_1+2k_2+2k_3+k_4),$$

$$z_{i+1}=z_i+\frac{h}{6}(l_1+2l_2+2l_3+l_4),$$

$k_1=z_i, l_1=-2z_i-y_i+x_i$,

$$k_2=z_i+\frac{hl_1}{2}, l_2=-2\left(z_i+\frac{hl_1}{2}\right)-\left(y_i+\frac{hk_1}{2}\right)+x_i+\frac{h}{2},$$

$$k_3 = z_i + \frac{hl_2}{2}, l_3 = -2\left(z_i + \frac{hl_2}{2}\right) - \left(y_i + \frac{hk_2}{2}\right) + x_i + \frac{h}{2},$$

$$k_4 = z_i + hl_3, l_4 = -2(z_i + hl_3) - (y_i + hk_3) + x_i + h.$$

Решение удобно оформить в виде таблиц.

Схема Эйлера:

k	x_i	y_i	z_i	Точное решение	Погреш-ность
0	0	1	0	1	0
1	0.2	1	−0.2	0.983685	0.016315
2	0.4	0.96	−0.28	0.947216	0.012784
3	0.6	0.904	−0.28	0.905009	0.001009
4	0.8	0.848	−0.2288	0.866913	0.018913
5	1	0.80224	−0.14688	0.839397	0.037157

Модифицированный метод Эйлера:

k	x_i	y_i	z_i	Погрешность
0	0	1	0	0
1	0.2	1	−0.18	0.016315
2	0.4	0.962	−0.244	0.014784
3	0.6	0.9108	−0.2342	0.005791
4	0.8	0.8615	−0.178	0.005413
5	1	0.823432	−0.09441	0.015965

Схема Рунге – Кутты:

x_i	y_i	z_i	Погрешность
0	1	0	0
0.2	0.9837	−0.146	1.79E−05
0.4	0.9472	−0.207	2.76E−05
0.6	0.905	−0.207	3.18E−05
0.8	0.8669	−0.168	3.25E−05
1	0.8394	−0.104	3.09E−05

Как можно видеть, максимальная погрешность, определяемая как разность между точным и рассчитанным значением функции y, уменьшается с увеличением порядка точности метода и для четырехэтапной схемы Рунге – Кутты не превышает $3.25 \cdot 10^{-5}$.

6.6. Жесткие ОДУ

До сих пор мы имели дело с ОДУ, которые надежно решались явными численными методами Рунге – Кутты. Однако имеется класс так называемых жестких (*stiff*) систем ОДУ, для которых явные методы практически неприменимы, поскольку их решение требует исключительно малого значения шага численного метода. Рассмотрим пример такой жесткой задачи.

Пример 2.5. Решить задачу Коши

$$y' = -100y + 100,\ y(0) = 2.$$

Точным решением задачи является функция $y = 1+e^{-100x}$, имеющая очень большой градиент вблизи точки $x = 0$. Действительно, $y = 2$ при $x = 0$ (в силу начальных данных), однако уже при малых положительных значениях x решение близко к своему асимптотическому значению $y = 1$. Получим численное решение этой задачи явным методом Эйлера (2.10) с шагом $h = 0.02$.

$$y_{i+1} = (1-100h)y_i + 100h = 2 - y_i,\ i = 0, 1, \ldots, n-1,\ y_0 = 2.$$

Решение будет представлять собой последовательность

$$y(0) = 2,\ y(0.02) = 0,\ y(0.04) = 2,\ y(0.06) = 0, \ldots$$

что не соответствует точному решению. При $h = 0.01$ первая же вычисленная точка $y_1 = 1$ попадает на асимптоту решения, и последующие вычисления не изменяют значения приближенного решения. Существенно более мелкий шаг, например $h = 0.001$, позволит получить вполне удовлетворительное соответствие между приближенным и точным решением. Однако вычисления с таким мелким шагом потребуют больших вычислительных затрат.

$$y(0) = y_0 = 2,\ y(0.001) \approx y_1 = 1.9,\ y(0.002) \approx y_2 = 1.81, \ldots$$

Воспользуемся неявным методом Эйлера:

$$y_{i+1} = \frac{100h + y_i}{1 + 100h},\ i = 0, 1, \ldots, n-1,\ y_0 = 2$$

с шагом $h = 0.1$. Получим последовательность

$$y(0) = 2,\ y(0.1) = 1.091,\ y(0.2) = 1.008,\ y(0.3) = 1.0007, \ldots$$

Даже при очень крупном шаге $h = 0.99$ приближенное решение, полученное неявным методом Эйлера, оказывается качественно правильным. $y_0 = 2,\ y_1 = 1.01,\ y_2 = 1.0001, \ldots$

Данный пример показывает, что получить приближенное решение данной задачи гораздо рациональнее с помощью неявного метода Эйлера.

В приведенном выше примере коэффициенты уравнения различаются на порядки, причем коэффициент при старшей производной меньше остальных. Рассмотрим уравнение

$$\frac{dy}{dx} = \alpha y,\ \alpha < 0,\ |\alpha| >> 1, x \geq 0,\ y(0) = y_0 \qquad (2.17)$$

с точным решением $y = y_0 e^{\alpha x}$. Поскольку при $\alpha < 0$ точное решение является убывающим, для численного решения должна выполнятся цепочка неравенств

$$\|y_{i+1}\| \leq \|y_i\| \leq \|y_{i-1}\| \leq \ldots \leq \|y_0\|,$$

известных из теории разностных схем как принцип максимума. Методы, решения которых удовлетворяют этим условиям, называются *A*-устойчивыми методами.

Запишем для уравнения (2.17) явный метод Эйлера и двухэтапный метод Рунге – Кутты.

$$y_{i+1} = y_i + h\alpha y_i = (1 + \alpha h) y_i = \lambda_1 y_i,$$

$$y_{i+1} = y_i + h\alpha(y_i + \alpha y_i\, h/2) = (1 + \alpha h + \alpha^2 h^2/2) y_i = \lambda_2 y_i.$$

Используя эти формулы, можно последовательно выразить каждое y_i через предыдущее, тогда

$$y_{i+1} = \lambda_k^{i+1} y_0,\ i = 0, 1, 2, \ldots,\ k = 1, 2.$$

Для выполнения принципа максимума $\|y_{i+1}\| \le \lambda_k^{i+1} \|y_0\|$ необходимо и достаточно, чтобы выполнялось условие $0 \le \lambda_k \le 1$. Отсюда сразу следуют ограничения на шаги интегрирования для явных методов. Например, для явного метода Эйлера $h \le 1/|\alpha|$, для двухэтапного метода Рунге – Кутты $h \le 2/|\alpha|$.

Теперь рассмотрим простейший неявный метод Эйлера для решения уравнения (2.17):

$$y_{i+1} = y_i + h\alpha y_{i+1} = y_i / (1 - \alpha h) = \lambda y_i.$$

Можно видеть, что условие $0 \le \lambda \le 1$ выполняется для любых α, следовательно, имеет место принцип максимума, т.е. неявный метод Эйлера не имеет ограничения по α на шаг интегрирования и является A-устойчивым. Неявный метод Эйлера может быть обобщен на систему жестких ОДУ (2.16).

Тема 7. Краевая задача для ОДУ второго порядка

Дано дифференциальное уравнение второго порядка

$$u'' + p(t)u' + g(t)u = f(t), \quad t \in [a, b]. \qquad (2.18)$$

Здесь $p(t), g(t), f(t)$ – заданные функции коэффициентов.

Уравнение можно свести к системе двух ОДУ первого порядка:

$$\begin{cases} u' = f(t,u,v) \equiv v \\ v' = g(t,u,v) \equiv -p(t)v - g(t)u - f(t) \end{cases}. \qquad (2.19)$$

Для определения единственного решения необходимо задать два дополнительных условия на искомую функцию $u(t)$. Если оба условия заданы в одной точке $t = t_0$, то мы имеем задачу Коши, которая может быть решена методами, описанными в Теме 6. Допустим теперь, что два дополнительных условия поставлены в разных точках: $x = a$ и $x = b$:

$$\begin{cases} k_1 u(a) + k_2 u'(a) = A \\ m_1 u(b) + m_2 u'(b) = B \end{cases}, \qquad (2.20)$$

где A, B, k_1, k_2, l_1, l_2 – заданные константы. Задача (2.18), (2.20) называется **краевой**, для приближенного решения которой используются методы:

- конечных разностей;
- сведения краевой задачи к задаче Коши (стрельбы, дифференциальной прогонки, редукции);
- балансов (конечных объемов);
- коллокации;
- проекционные (Галеркина);
- вариационные (наименьших квадратов, Ритца);
- проекционно-сеточные (конечных элементов).

Ниже рассмотрим некоторые из перечисленных методов.

7.1. Конечно-разностный метод

Введем на отрезке $[a, b]$ разностную сетку $(t_0, t_1, t_2, ..., t_M)$, $t_i = a + \tau \cdot i$, $i = 0, 1, ..., M$, $\tau = (b - a)/M$, M – число точек разностной сетки (параметр задачи). Вместо точного решения $u(t)$ будем отыскивать приближенное решение в узлах разностной сетки: $y_i = y(t_i)$. Используя формулы приближенного дифференцирования: $u''(t_i) \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{\tau^2}$, $u'(t_i) \approx \frac{y_{i+1} - y_{i-1}}{2\tau}$, заменим исходное уравнение и краевые условия разностной схемой:

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{\tau^2} + p(t_i)\frac{y_{i+1} - y_{i-1}}{2\tau} + g(t_i)y_i = f(t_i), i = 1, ..., M,$$

$$k_1 y_0 + k_2 \frac{y_1 - y_0}{\tau} = A,$$

$$m_1 y_M + m_2 \frac{y_M - y_{M-1}}{\tau} = B.$$

В итоге получаем следующую систему $M + 1$ линейных алгебраических уравнений на вектор неизвестных $(y_0, y_1, y_2, ..., y_M)$:

$$\begin{cases} -C_0 y_0 + B_0 y_1 = F_0 \\ A_i y_{i-1} - C_i y_i + B_i y_{i+1} = F_i,\ i = 1, 2, ..., M - 1, \\ A_M y_{M-1} - C_M y_M = F_M \end{cases}$$

где $C_i = 2 - g(t_i)\tau^2$, $A_i = 1 - p(t_i) \cdot \tau/2$, $B_i = 1 + p(t_i) \cdot \tau/2$, $F_i = \tau^2 f(t_i)$, $C_0 = k_2 - \tau k_1$, $B_0 = k_2$, $F_0 = A\,\tau$, $C_M = m_2 + \tau m_1$, $A_M = m_2$, $F_M = -B\tau$. Выпишем систему в матричном виде:

$$\begin{pmatrix} -C_0 & B_0 & 0 & 0 & \dots & 0 \\ A_1 & -C_1 & B_1 & 0 & \dots & 0 \\ 0 & A_2 & -C_2 & B_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & A_{M-1} & -C_{M-1} & B_{M-1} \\ 0 & 0 & \dots & 0 & A_M & -C_M \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \dots \\ y_{M-1} \\ y_M \end{pmatrix} = \begin{pmatrix} F_0 \\ F_1 \\ F_2 \\ \dots \\ F_{M-1} \\ F_M \end{pmatrix}.$$

Поскольку матрица СЛАУ имеет трехдиагональный вид, то система решается методом прогонки.

Метод прогонки является точным методом решения СЛАУ, следовательно, погрешность в приближенное решение была внесена на этапе замены исходных уравнений и краевых условий конечно-разностными соотношениями. Оценку погрешности метода можно провести, если вспомнить, что используемые формулы приближенного дифференцирования (центральная разность для первой производной и симметричная аппроксимация для второй производной) имеют второй порядок точности. В то же время при замене краевых условий (2.20) на разностные соотношения в приближенное решение вносится погрешность порядка τ (именно такую погрешность имеют формулы «разность вперед» и «разность назад»). Следовательно, суммарная погрешность аппроксимации уравнения и краевых условий будет пропорциональна τ. Однако в тех случаях, когда в краевые условия не входит производная, т.е. $k_2 = 0$, $l_2 = 0$, краевые условия для приближенного решения выполняются точно, и тогда метод имеет погрешность порядка τ^2.

7.2. Метод стрельбы

Этот метод основан на сведении краевой задачи к задаче Коши для системы (2.19). Пусть краевые условия имеют вид:

$$u(a) = u_0,\ u(b) = u_1.$$

Для того, чтобы свести исходную краевую задачу к задаче Коши, необходимо в точке $t = a$ задать дополнительное краевое условие $u'(a) = v_0$. Величина v_0 имеет геометрический смысл тангенса α – угла наклона касательной к решению в начальной точке. Графическая иллюстрация метода стрельбы приведена на рис. 2.6.

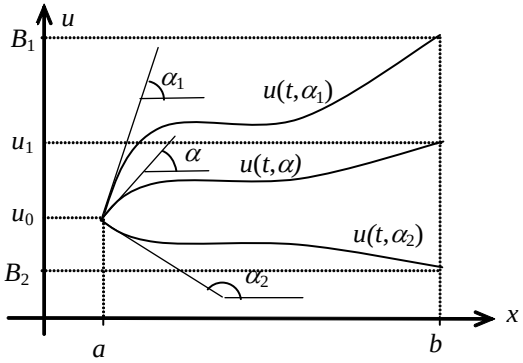


Рис. 2.6. Иллюстрация метода стрельбы

Задача Коши для системы (2.19) может быть решена, например, методом Рунге – Кутты. Поскольку решение задачи Коши зависит от выбора α, можно записать: $u = u(t, \alpha)$. Требуется подобрать значение α, обеспечивающее «попадание», т.е. выполнение условия $u(t, \alpha) = u_1$. Понятно, что при произвольном выборе α полученное решение может не удовлетворять краевому условию на правом конце отрезка $t = b$. Может быть получено, что $u(t, \alpha_1)\big|_{t=b} = B_1 > u_1$ («перелет») $u(t, \alpha_1)\big|_{t=b} = B_2 < u_1$ («недолет») (см. рис. 2.6). Задачу подбора нужного угла α можно рассматривать как решение нелинейного алгебраического уравнения $u(t, \alpha)\big|_{t=b} = u_1$, особенностью которого является то, что $F(x)$, определяющая нелинейное уравнение, не задана явно, а определен только способ нахождения $u(t, \alpha)\big|_{t=b}$.

7.3. Метод коллокаций

Запишем краевую задачу для ОДУ второго порядка в операторном виде:

$$Lu = f(t),\ l_0 u(a) = A,\ l_1 u(b) = B, \qquad (2.21)$$

где $L = \frac{d^2}{dt^2} + p(t)\frac{d}{dt} + q(t)$, $l_0 = k_1 + k_2 \frac{d}{dt}$, $l_1 = m_1 + m_2 \frac{d}{dt}$.

Зададим на $[a, b]$ некоторую систему базисных функций $\varphi_0(t)$, $\varphi_1(t)$, …, $\varphi_n(t)$, таких, что $\varphi_0(t)$ удовлетворяет краевым условиям $l_0\varphi_0(a) = A$, $l_1\varphi_0(b) = B$, а остальные $\varphi_k(t)$ удовлетворяют однородным краевым условиям $l_0\varphi_k(a) = 0$, $l_1\varphi_k(b) = 0$, $k = 1, ..., n$. Представим приближенное решение задачи (2.21) в виде линейной комбинации базисных функций:

$$y_n(t) = \varphi_0(t) + a_1\varphi_1(t) + a_2\varphi_2(t) + ... + a_n\varphi_n(t) \qquad (2.22)$$

с неизвестными пока коэффициентами a_1, a_2, ..., a_n. При этом $y_n(t)$ при любых значениях a_1, a_2, ..., a_n удовлетворяет краевым условиям. Подействуем на (2.22) оператором L. Функция

$$\psi(t, a_1, a_2, ..., a_n) = Ly_n(t) - f(t) = L\varphi_0(t) - f(t) + \sum_{k=1}^{n} a_k L\varphi_k$$

называется **невязкой уравнения**. Если $\psi = 0$, то $y_n(t)$ – точное решение задачи (2.21). Подберем параметры a_1, a_2, ..., a_n, чтобы невязка была минимальной.

Зафиксируем на $[a, b]$ n точек t_1, t_2, ..., t_n, называемых **точками коллокации**, и потребуем, чтобы в этих точках $\psi(t, a_1, a_2, ..., a_n) = 0$. Получается система n линейных алгебраических уравнений

$$\begin{cases} a_1 L\varphi_1(t_1) + a_2 L\varphi_2(t_1) + ... + a_n L\varphi_n(t_1) = f(t_1) - L\varphi_0(t_1) \\ a_1 L\varphi_1(t_2) + a_2 L\varphi_2(t_2) + ... + a_n L\varphi_n(t_2) = f(t_2) - L\varphi_0(t_2) \\ ..... \\ a_1 L\varphi_1(t_n) + a_2 L\varphi_2(t_n) + ... + a_n L\varphi_n(t_n) = f(t_n) - L\varphi_0(t_n) \end{cases},$$

решение которой дает a_1, a_2, ..., a_n. Между точками коллокации $\psi(t, a_1, a_2, ..., a_n) \neq 0$, и поэтому решение будет приближенным. Заметим, что на выбор точек t_j никаких условий не накладывает-

ся и их можно сгущать в предполагаемых местах больших градиентов решения. Это позволяет получить хорошую точность при сравнительно небольшом количестве точек коллокации.

7.4. Вариационные методы

Вариационное исчисление – это раздел математики, который изучает задачи на нахождение экстремумов функционалов. Примером функционала является, например, интеграл

$$I[y(x)] = \int_0^1 y(x)dx,$$

значение которого зависит от того, какая функция в него подставлена: $y(x) = x$, то $I[y(x)] = 1/2$, $y(x) = x^2$, $I[y(x)] = 1/3$ и т.д.

Функционалы и вариационные принципы широко используются в механике и физике (принцип наименьшего действия Гамильтона). Каждому линейному уравнению

$$Lu = f\ , \tag{2.23}$$

где L – положительно определенный оператор, можно поставить в соответствие функционал энергии

$$J(u) = (Lu, u) - 2(f, u)\,. \tag{2.24}$$

Доказано, что если функция u является решением уравнения (2.23), то на ней функционал (2.24) достигает экстремума, и, наоборот, функции, поставляющие экстремум функционала (2.24), являются решениями (2.23). Существуют также другие функционалы, связанные с уравнением (2.23), например,

$$I(u) = (Lu - f, Lu - f)\,,$$

представляющий собой квадрат нормы невязки уравнения.

Таким образом, вместо того, чтобы искать решение уравнения, можно отыскивать функции, на которых тот или другой функционал достигает экстремума.

Рассмотрим задачу о растяжении стержня, находящегося под действием распределенной вдоль оси нагрузки $q(x)$. Дифференциальное уравнение, описывающее этот процесс, имеет вид

$$E \cdot F \frac{d^2u}{dx^2} = -q.$$

Предположим, что модуль упругости E и площадь поперечного сечения стержня F постоянны. Тогда уравнение можно переписать в виде

$$\frac{d^2u}{dx^2} = Q, \tag{2.25}$$

где $Q = -\frac{q(x)}{E \cdot F}$. Будем искать приближенное решение v в виде комбинации n линейно независимых базисных функций $\varphi_i(x)$:

$$v(x) = \sum_{i=1}^{n} c_i \varphi_i(x), \tag{2.26}$$

где c_i – неизвестные коэффициенты, которые должны быть определены из некоторых условий. Введем функционал квадрата нормы невязки и потребуем, чтобы он был минимален:

$$G(v) = \int_a^b \left(\frac{d^2v}{dx^2} - Q \right)^2 dx \to \min. \tag{2.27}$$

Подставим (2.26) в (2.27):

$$G(v) = \int_a^b \left(\frac{d^2 \sum_{i=1}^{n} c_i \varphi_i(x)}{dx^2} - Q \right)^2 dx.$$

Поскольку искомые коэффициенты не зависят от x, можно вынести их из-под знака дифференцирования и переписать выражение в виде

$$G(v) = \int_a^b \left(\sum_{i=1}^{n} c_i \frac{d^2\varphi_i(x)}{dx^2} - Q \right) \left(\sum_{i=1}^{n} c_i \frac{d^2\varphi_i(x)}{dx^2} - Q \right) dx =$$

$$= \int_a^b \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \frac{d^2\varphi_i(x)}{dx^2} \frac{d^2\varphi_j(x)}{dx^2} dx - 2 \int_a^b \sum_{i=1}^{n} c_i \frac{d^2\varphi_i(x)}{dx^2} Q dx + \int_a^b Q^2 dx.$$

Введем обозначения:

$$A_{ij} = \int_a^b \frac{d^2\varphi_i(x)}{dx^2}\frac{d^2\varphi_j(x)}{dx^2}dx\,,\ \alpha_i = \int_a^b \frac{d^2\varphi_i(x)}{dx^2}Qdx\,,\ \beta = \int_a^b Q^2 dx\,.$$

Тогда

$$G(v) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j A_{ij} - 2\sum_{i=1}^{n} c_i\alpha_i + \beta\,.$$

По условию минимума функционала G его производные по коэффициентам c_i должны равняться нулю:

$$\frac{\partial G}{\partial c_i} = \sum_{j=1}^{n} c_i A_{ij} - \sum_{i=1}^{n}\alpha_i = 0\,.$$

Таким образом, мы получили СЛАУ из n уравнений

$$A\vec{c} = \vec{\alpha}\,. \qquad (2.28)$$

Поскольку матрица A симметрична и положительно определена, как следует из способа ее задания, система (2.28) имеет решение. Решим СЛАУ каким-либо из описанных выше методов и, подставляя коэффициенты c_i в (2.26), получим приближенное решение уравнения (2.25).

7.5. Проекционные методы

Пусть задано уравнение вида $Lu = f$, где L – линейный оператор $L: H \to H, u, f \in H, H$ – гильбертово пространство (см. Раздел 1), т.е. пространство со скалярным произведением. Скалярное произведение в пространстве функций $u(t)$, $v(t)$, заданных на отрезке $[0, T]$, можно определить как

$$(u, v) = \int_0^T u(t)v(t)dt = 0\,.$$

Решение уравнения можно представить в виде разложения по базису пространства H:

$$u = \sum_{k=1}^{\infty} c_k\varphi_k\,.$$

Ищем приближенное решение в виде (2.26). Таким образом, $v \in H_n$ – конечномерному подпространству H.

Подставив приближенное решение в исходное уравнение, получим невязку $\psi = f - Lv$. Запишем тождество:

$$Lv + \psi = f .$$

По теореме о разложении (стр. 18) всякий элемент из гильбертова пространства может быть представлен в виде суммы элементов из подпространства H_n и его ортогонального дополнения $H_n^{\perp}$. Поскольку $Lv \in H_n$, $f \in H$, то $\psi \in H_n^{\perp}$. Отсюда следует, что невязка должна быть ортогональна всем базисным функциям φ_j, $j = 1, 2, \ldots, n$:

$$(\psi, \varphi_j) = (Lv - f, \varphi_j) = (L\sum_{k=1}^{n} c_k \varphi_k - f, \varphi_j) = 0,\ j = 1, 2, \ldots, n. \quad (2.29)$$

В силу линейности оператора L из (2.29) получим систему n линейных алгебраических уравнений для определения коэффициентов c_k:

$$\sum_{k=1}^{n} c_k \cdot (L\varphi_k, \varphi_j) = \left(f, \varphi_j\right),\ j = 1, 2, \ldots, n.$$

Как и в ранее описанном вариационном методе Ритца, решение сводится к нахождению решения СЛАУ вида (2.28), где коэффициенты матрицы A и вектора правых частей определяются как

$$a_{ij} = \left(L\varphi_j, \varphi_i\right),\ \alpha_i = \left(f, \varphi_i\right), \quad (2.30)$$

что также совпадает с приведенными выше формулами вариационного метода Ритца.

В качестве базисных функций можно выбрать, например, степенной $\varphi_i = x^{i-1}$ или тригонометрический базис

$$\varphi_i = \sin\left(\frac{x-a}{b-a}\left(i\pi - \frac{\pi}{2}\right)\right),\ i = 1, \ldots, n.$$

Недостатком этих базисов является то, что матрица системы (2.28) является заполненной, и расчетные затраты растут пропорционально n^3, что ограничивает применение методов для больших n.

7.6. Метод конечных элементов

Как отмечено выше, проекционные и вариационные методы решения краевой задачи (2.21) приводят к одинаковой системе линейных алгебраических уравнений с коэффициентами и правыми частями, заданными (2.30). Заполненность матрицы зависит от вида базисных функций. В методе конечных элементов в качестве базисных выбираются финитные функции, отличные от нуля на некотором ограниченном интервале отрезка $[a, b]$.

Метод конечных элементов (МКЭ) широко применяется в современной инженерной практике, в частности, при решении задач строительства. Например, он используется в пакете ANSYS Mechanical, описание которого приведено в Теме 12.

Для решения уравнения (2.21) с нулевыми краевыми условиями $u(a) = u(b) = 0$ разобьем отрезок $[a, b]$ на $n - 1$ часть и запишем базисные финитные функции в виде

$$\varphi_i = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i], \\ -\dfrac{x - x_{i+1}}{x_{i+1} - x_i}, & x \in [x_i, x_{i+1}], \\ 0, & x \notin [x_{i-1}, x_{i+1}]. \end{cases}$$

Предположим, что размер элементов постоянен и равен h:

$$\varphi_i = \begin{cases} \dfrac{x - x_{i-1}}{h}, & x \in [x_{i-1}, x_i], \\ -\dfrac{x - x_{i+1}}{h}, & x \in [x_i, x_{i+1}], \\ 0, & x \notin [x_{i-1}, x_{i+1}]. \end{cases} \tag{2.31}$$

В этом случае производные базисных функций равны

$$\varphi_i' = \begin{cases} 1/h, & x \in [x_{i-1}, x_i], \\ -1/h, & x \in [x_i, x_{i+1}], \\ 0, & x \notin [x_{i-1}, x_{i+1}]. \end{cases} \tag{2.32}$$

График одной базисной функции показан на рис. 2.7. Будем искать приближенное решение в виде (2.26). Найдем правые части системы согласно (2.30):

$$\alpha_i = \int_a^b f(x)\varphi_i(x)\,dx = \sum_{k=1}^{n-1} \int_{x_k}^{x_{k+1}} f(x)\varphi_i(x)\,dx\,.$$

Подставив сюда выражение (2.31), получим

$$\alpha_i = \int_{x_{i-1}}^{x_i} f(x)\frac{(x-x_{i-1})}{h}dx - \int_{x_i}^{x_{i+1}} f(x)\frac{(x-x_{i+1})}{h}dx = \\ = \frac{1}{h}\left\{ \int_{x_{i-1}}^{x_i} f(x)(x-x_{i-1})\,dx - \int_{x_i}^{x_{i+1}} f(x)(x-x_{i+1})\,dx \right\}. \qquad (2.33)$$

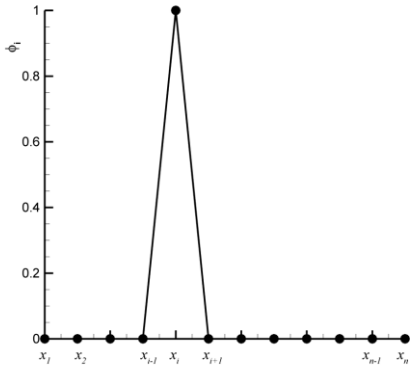


Рис. 2.7. Базисная функция в МКЭ

Зная функцию $f(x)$, выражение (2.33) можно проинтегрировать и получить конкретный вид правых частей СЛАУ.

Теперь получим коэффициенты матрицы СЛАУ

$$a_{ij} = \left(L\varphi_j, \varphi_i\right) = \left(\frac{d^2\varphi_j}{dx^2}, \varphi_i\right) = \int_a^b \frac{d^2\varphi_j}{dx^2}\varphi_i dx\,.$$

Применяя интегрирование по частям, имеем

$$a_{ij} = \varphi_i(x)\varphi_j'(x)\Big|_a^b - \int_a^b \varphi_j'(x)\varphi_i'(x)\,dx =$$

$$= \varphi_i(b)\varphi_j'(b) - \varphi_i(a)\varphi_j'(a) - \int_a^b \varphi_j'(x)\varphi_i'(x)\,dx.$$

Как видно из определения базисных функций, на концах отрезка $[a,b]$ все они обращаются в ноль. Поэтому два первых члена в этом выражении равны нулю и

$$a_{ij} = -\int_a^b \varphi_j'(x)\varphi_i'(x)\,dx = \sum_{k=1}^{n} \int_{x_k}^{x_{k+1}} -\varphi_j'(x)\varphi_i'(x)\,dx.$$

Согласно (2.32) производные базисных функций равны нулю везде, кроме соседних с исследуемой точкой отрезков (элементов). То есть $\varphi_j'(x) = 0$ при $i-1 \le j \le i+1$. Отсюда следует вывод, что полученная нами матрица *A* является трехдиагональной. Элементы главной диагонали (*i* = *j*) определяются как

$$a_{ii} = \sum_{k=1}^{n} \int_{x_k}^{x_{k+1}} -(\varphi_i')^2\,dx = \int_{x_{i-1}}^{x_i} -\frac{1}{h^2}\,dx + \int_{x_i}^{x_{i+1}} -\left(-\frac{1}{h}\right)^2 dx =$$

$$= -\frac{1}{h^2}\left(\int_{x_{i-1}}^{x_i} dx + \int_{x_i}^{x_{i+1}} dx\right) = -\frac{2}{h}.$$

На соседних с главной диагональю

$$a_{i,i+1} = \sum_{k=1}^{n} \int_{x_k}^{x_{k+1}} -\varphi_i'\varphi_{i+1}'\,dx = \int_{x_i}^{x_{i+1}} -\frac{1}{h}\left(-\frac{1}{h}\right)dx\,dx = \frac{1}{h^2}\left(\int_{x_i}^{x_{i+1}} dx\right) = \frac{1}{h},$$

$$a_{i,i-1} = \sum_{k=1}^{n} \int_{x_k}^{x_{k+1}} -\varphi_i'\varphi_{i-1}'\,dx = -\int_{x_{i-1}}^{x_i} -\frac{1}{h}\frac{1}{h}\,dx = \frac{1}{h}.$$

Таким образом, матрицу A можно записать как

$$A = \begin{vmatrix} -2/h & 1/h & 0 & 0 & \dots & 0 \\ 1/h & -2/h & 1/h & 0 & \dots & 0 \\ 0 & 1/h & -2/h & \dots & 0 & 0 \\ 0 & 0 & \dots & 1/h & -2/h & 1/h \\ 0 & 0 & \dots & 0 & 1/h & -2/h \end{vmatrix}.$$

Систему $A\vec{c} = \vec{\alpha}$ можно решить точным экономичным методом прогонки, описанным в Разделе 1. В других случаях возможно применение итерационных методов решения СЛАУ.

Если краевые условия имеют более сложный вид, необходимо ввести замену переменных с тем, чтобы привести их к виду $u(a) = u(b) = 0$.

РАЗДЕЛ 3. УРАВНЕНИЯ В ЧАСТНЫХ ПРОИЗВОДНЫХ

Тема 8. Общие сведения об уравнениях в частных производных

8.1. Постановка задачи

Для описания реального физического процесса (прогиб балки под действием нагрузки, движение газа в некотором объеме, распространение электромагнитных волн и пр.) строятся физико-математические модели, на основе которых можно анализировать процессы качественно и количественно. Задачи описания движения сплошных сред (газа, жидкости, твердых тел), а также задачи теплопроводности, теории упругости, электрических и магнитных полей и многие другие приводят к дифференциальным уравнениям. Независимыми переменными в физических задачах являются время t и пространственные координаты x, y, z. В качестве зависимых переменных в разных задачах используются компоненты скорости частиц среды, плотность, давление, температура, упругие напряжения, деформации и др. характеристики.

Допустим, что требуется найти решение на временном промежутке $[t_0, t_1]$ в некоторой области изменения независимых переменных $G(x, y, z)$. Математическая постановка задачи состоит из дифференциального уравнения (уравнений), а также дополнительных условий, позволяющих выделить единственное решение среди семейства всех решений данного уравнения. Дополнительные условия, заданные при $t = t_0$, называются начальными данными, а условия, заданные на границе области $G(x, y, z)$, – граничными или краевыми условиями. В качестве начальных и краевых условий, как правило, задают значения искомых функций и их производных. Задачи, у которых имеются только начальные условия, называются **задачей Коши**. Задачи с начальными данными и граничными условиями называют **смешанной краевой задачей** или нестационарной краевой задачей.

При исследовании установившихся состояний или стационарных (не зависящих от времени) процессов используются

уравнения, не зависящие от времени. В этом случае решение ищется в области $G(x, y, z)$, на границе которой задаются граничные условия. Такие задачи называются **краевыми.**

Особым вопросом в теории дифференциальных уравнений является **корректность** постановки начальных и смешанных задач. Корректной называется такая постановка дополнительных (начальных и граничных) условий, при которой решение задачи в целом существует, единственно и непрерывно зависит от этих данных и коэффициентов уравнения. Требование непрерывной зависимости необходимо, чтобы небольшие изменения коэффициентов уравнения, начальных данных и краевых условий не приводили к сильным изменениям решения задачи. В механике и физике существуют задачи, решение которых неустойчиво. Изучением таких некорректных задач занимается специальный раздел математики. Здесь мы будем рассматривать только корректные постановки задач, при решении которых не возникает неустойчивости, связанной с исходными уравнениями.

8.2. Характеристики. Типы уравнений

Многие физические задачи приводят к решению уравнений второго порядка, которые достаточно хорошо изучены в теоретическом плане и для которых разработаны стандартные методы приближенного решения. В случае одной пространственной координаты уравнение в частных производных второго порядка можно записать в виде

$$Au_{tt} + Bu_{tx} + Cu_{xx} + Du_t + Eu_x = F. \qquad (3.1)$$

Здесь $u(t, x)$ – искомая функция, t, x – независимые переменные, A, B, C, D, E и F – коэффициенты уравнения, которые, вообще говоря, могут зависеть от t, x и u.

Если все коэффициенты являются константами, то это **линейное уравнение с постоянными коэффициентами**. Если коэффициент F – линейная функция от неизвестной u, а остальные коэффициенты от u не зависят, то такое уравнение называется **линейным с переменными коэффициентами**. Если все коэффициенты зависят от u, то такие уравнения называются **квазилинейными**. Если коэффициенты зависят не только

от искомых функций, но и от ее производных, уравнение будет **нелинейным**.

Если коэффициенты A, B, C – нулевые, а $D \neq 0$ и $E \neq 0$, то уравнение имеет первый порядок и называется **уравнением переноса (адвекции)**. Если хотя бы один из коэффициентов A, B, C отличен от нуля, уравнение имеет второй порядок и может быть классифицировано по типам аналогично кривым второго порядка. Классификация уравнений связана с наличием характеристик – особых направлений, вдоль которых исходное уравнение может быть записано в виде полного дифференциала и, следовательно, может быть проинтегрировано. Уравнение характеристик для (3.1) имеет вид

$$\frac{dx}{dt} = \frac{B \pm \sqrt{B^2 - 4AC}}{2A}. \qquad (3.2)$$

Количество характеристик зависит от знака дискриминанта $B^2 - 4AC$. Если он положителен, то уравнение (3.1) имеет две вещественные характеристики и называется **гиперболическим**. В случае нулевого дискриминанта уравнение имеет одну вещественную характеристику и является **параболическим**. **Эллиптические** уравнения не имеют вещественных характеристик (дискриминант отрицателен).

Физические процессы, описываемые уравнениями перечисленных типов, корректные постановки начально-краевых задач и свойства решений существенно отличаются друг от друга.

8.3. Приближенные методы. Аппроксимация и устойчивость

Как правило, уравнения в частных производных не могут быть решены аналитически (точно), и для нахождения решений используются приближенные методы, которые можно разделить на три основные группы: методы конечных разностей, методы конечных объемов и методы конечных элементов.

В **конечно-разностных методах** приближенное решение ищется в узлах специально построенной разностной сетки, покрывающей область решения исходной дифференциальной задачи. Дифференциальная задача с помощью формул приближенного дифференцирования заменяется системой алгебраиче-

ских уравнений (разностной схемой), которая далее решается с помощью точных или приближенных методов решения СЛАУ. При этом необходимо использовать такие разностные схемы, которые обеспечивают сходимость получаемого решения разностной задачи к решению исходной дифференциальной при уменьшении шага сетки.

Аппроксимация (от англ. *approximation* – приближение) характеризует, насколько хорошо разностная задача приближает исходную дифференциальную, и зависит от точности формул разностного дифференцирования и, конечно же, размеров разностной сетки. Для простых задач аппроксимация может быть исследована аналитически с помощью оценки главного члена погрешности разностной схемы при подстановке в нее точного решения. В более сложных случаях аппроксимацию исследуют экспериментально, с помощью оценки поведения погрешности при измельчении сетки.

Устойчивость характеризует способность решения не накапливать ошибку. Другими словами, устойчивость – это непрерывная зависимость решения от входных данных: коэффициентов, правых частей, начальных данных и краевых условий. При этом следует различать устойчивость решения исходной дифференциальной задачи и устойчивость приближенного метода. Аппроксимируем записанную в общем виде дифференциальную задачу с начальными данными и краевыми условиями:

$$Lu = \varphi, \qquad u^0 = \mathrm{v}_0, \qquad lu = \mu \tag{3.3}$$

с помощью конечно-разностной схемы:

$$L_h u_h = \varphi_h, \qquad u^0{}_h = \mathrm{v}_0, \qquad l_h u_h = \mu_h. \tag{3.4}$$

Рассмотрим также конечно-разностную задачу с «возмущенными» правыми частями, начальными данными и краевыми условиями:

$$L_h \tilde{u}_h = \tilde{\varphi}_h, \qquad \tilde{u}^0 = \tilde{\mathrm{v}}_0, \qquad l_h \tilde{u}_h = \tilde{\mu}_h.$$

Говорят, что решение разностной задачи (3.4) непрерывно зависит от входных данных задачи, если существуют такие константы $M_1 > 0$, $M_2 > 0$, $M_3 > 0$, не зависящие от t и h, что

$$\left\| u_h(t) - \tilde{u}_h(t) \right| \leq \left| M_1 \left| \varphi_h - \tilde{\varphi}_h \right| + M_2 \left| \mu_h - \tilde{\mu}_h \right| \right\| + M_3 \left| \mathrm{v}_0 - \tilde{\mathrm{v}}_0 \right| \Big\|. \tag{3.5}$$

Для линейных разностных схем доказано, что устойчивость по начальным данным эквивалентна устойчивости по правой части, т.е. для линейных операторов L_h, l_h достаточно показать, что

$$\left\|u_h(t)-\tilde{u}_h(t)\right| \quad \le \quad M\left\|\mathrm{v}_0-\tilde{\mathrm{v}}_0\right\|.$$

Если независимых переменных несколько (например, x и t), то вводится понятие условной и безусловной устойчивости. Устойчивость называется *безусловной*, если (3.5) имеет место при произвольном соотношении между шагами разностной сетки τ и h. Схема будет *условно устойчивой*, если для выполнения (3.5) шаги по независимым переменным должны подчиняться дополнительным соотношениям.

В теории разностных схем доказана **теорема Лакса** о том, что если разностная задача аппроксимирует дифференциальную и устойчива, то при измельчении сетки решение разностной задачи сходится к решению дифференциальной.

Подробнее о свойствах аппроксимации и устойчивости будет рассказано ниже при рассмотрении примеров уравнений в частных производных.

Тема 9. Уравнения теплопроводности

9.1. Одномерное уравнение теплопроводности

К параболическим уравнениям приводят задачи теплопроводности, диффузии примеси и некоторые другие. Рассмотрим уравнения этого типа на примере линейного уравнения теплопроводности с одной пространственной переменной и постоянными коэффициентами:

$$\frac{\partial u}{\partial t} = A\frac{\partial^2 u}{\partial x^2} + F(x,t). \qquad (3.6)$$

Уравнение (3.6) описывает распространение тепла в тонком длинном стержне длины L. Здесь $A > 0$ – константа (коэффициент теплопроводности), $u(x, t)$ – искомое решение (температура), $F(x, t)$ – правая часть, с помощью которой можно задать источники или стоки тепла.

Будем искать решение в области $0 \le x \le L$, $0 \le t \le T$. Корректная постановка задачи кроме уравнения (3.6) должна содержать начальные данные:

$$u(x, 0) = u_0(x) \qquad (3.7)$$

и краевые условия. Существует три типа краевых условий, которые называют условиями первого, второго и третьего рода. Условия первого рода означают, что на границах области задана зависимость температуры от времени:

$$u(0, t) = \mu_{11}(t),\ u(L, t) = \mu_{12}(t). \qquad (3.8')$$

Условия второго рода задают тепловые потоки (производные от температуры) через границы области:

$$u_x(0, t) = \mu_{21}(t),\ u_x(L, t) = \mu_{22}(t). \qquad (3.8'')$$

И наконец, условия третьего рода задают на границе линейную комбинацию искомой функции и ее производной:

$$u(0, t) + \alpha_1 u_x(0, t) = \mu_{31}(t),\ u(L, t) + \alpha_2 u_x(L, t) = \mu_{32}(t). \qquad (3.8''')$$

В курсе дифференциальных уравнений доказано, что уравнение (3.6) с начальными данными (3.7) и краевыми условиями (3.8) имеет единственное решение.

Построим в области решения прямоугольную равномерную разностную сетку. Для этого разобьем отрезок $[0, T]$ на N равных частей: $t^i = n \cdot \tau$, а отрезок $[0, L]$ – на M равных частей: $x_j = jh$, $h = L/M$. Область решения и разностная сетка представлены на рис. 3.1. Вместо точного решения $u(x, t)$ будем искать приближенное решение, заданное в узлах сетки $u^i_j = u(x_j, t^i)$. На линиях $t = 0$, $x = 0$ и $x = L$ решение определено начальными данными (3.7) и краевыми условиями (3.8), во всех остальных узлах сетки решение должно быть найдено из разностных аналогов уравнения (3.6).

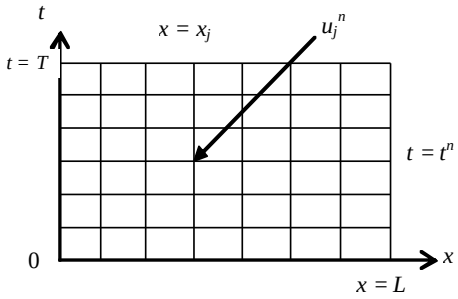


Рис. 3.1. Расчетная область и сетка для уравнения (3.6)

Аппроксимируем (приблизим) исходную дифференциальную задачу конечно-разностной. Для этого заменим все входящие в уравнение (3.6) и краевые условия (3.8′), (3.8″) или (3.8‴) производные их конечно-разностными аналогами:

$$\frac{\partial u}{\partial t}(x_j, t^n) \approx \frac{u(x_j, t^{n+1}) - u(x_j, t^n)}{\tau} = \frac{u_j^{\,n+1} - u_j^{\,n}}{\tau},$$

$$\frac{\partial^2 u}{\partial x^2}(x_j, t^n) \approx \frac{u(x_{j+1}, t^n) - 2u(x_j, t^n) + u(x_{j-1}, t^n)}{h^2} =$$

$$= \frac{u_{j+1}^{\,n} - 2u_j^{\,n} + u_{j-1}^{\,n}}{h^2},$$

$$\frac{\partial u}{\partial x}(x_0, t^i) = \frac{u_1^{\,n} - u_0^{\,n}}{h}, \quad \frac{\partial u}{\partial x}(x_M, t^n) = \frac{u_M^n - u_{M-1}^n}{h}.$$

Подставляя эти выражения в уравнение (3.6), получим разностную схему:

$$\frac{u_j^{n+1} - u_j^n}{\tau} = A\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} + F(x_j, t^n). \qquad (3.9)$$

На первом временном слое решение известно: $u^0{}_j = u(x_j, t^0) = u(x_j, 0) = u_0(x_j)$. Во всех внутренних точках расчетной области оно находится из явных формул, которые легко получаются из схемы (3.9):

$$u_j^{n+1} = u_j^n + \tau A \frac{u_{j+1}^n - 2u_j^{ni} + u_{j-1}^{in}}{h^2} + F(x_j, t^n),$$

$$n = 1,\ 2, \ldots, N-1,\ j = 0,\ 1, \ldots, M-1.$$

Для нахождения решения в крайних точках отрезка $[0, L]$ необходимо использовать краевые условия. Если заданы краевые условия (3.8′), можно сразу определить значения искомых функций:

$$u^n{}_0 = \mu_{11}(t^i),\ u^n{}_M = \mu_{12}(t^i).$$

Для условий (3.8″) получим:

$$u_0^{n+1} = u_1^{n+1} - h\mu_{21}(t^{n+1}),\ u_M^{n+1} = u_{M-1}^{n+1} + h\mu_{22}(t^{n+1}).$$

Для условий (3.8‴) необходимо выразить искомые величины $u(t^{n+1}, x_0)$, $u(t^{n+1}, x_M)$ из линейных конечно-разностных соотношений, аппроксимирующих краевые условия:

$$u_0^{n+1} = \frac{h\mu_{31}(t^{n+1}) - \alpha_1 u_1^{n+1}}{h - \alpha_1},\ u_M^{n+1} = \frac{\alpha_2 u_{M-1}^{n+1} + h\mu_{32}(t^{n+1})}{h - \alpha_2}.$$

Исследуем, насколько численное решение, полученное по схеме (3.9), отличается от точного. Для этого разложим точное решение $u(t^n, x_{j\pm1}) = u^n_{j\pm1}$, $u(t^{n+1}, x_j) = u_j^{n+1}$ в ряд Тейлора в окрестности точки (x_j, t^n):

$$u_{j\pm1}^n = u(t^n, x_j \pm h) = u_j^n \pm h(u_x)_j^{in} + \frac{h^2}{2}(u_{xx})_j^n \pm \frac{h^3}{6}(u_{xxx})_j^n + \frac{h^4}{24}(u_{xxxx})_j^n \ldots,$$

$$u_j^{n+1} = u(t^n + \tau, x_j) = u_j^n + \tau(u_t)_j^n + \frac{\tau^2}{2}(u_{tt})_j^n + \ldots,$$

и подставим эти выражения в разностную схему (3.9):

$$\frac{u_j^{n+1} - u_j^i}{\tau} - A\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} - F(x_j, t^n) =$$

$$= \frac{u_j^{in} + \tau(u_t)_j^n + \dfrac{\tau^2}{2}(u_{tt})_j^{ni} + \ldots - u_j^n}{\tau} -$$

$$-A\left(\frac{u_j^n + h(u_x)_j^n + \frac{h^2}{2}(u_{xx})_j^{in} + \frac{h^3}{6}(u_{xxx})_j^n + \frac{h^4}{24}(u_{xxxx})_j^n - 2u_j^n}{h^2} + \right.$$

$$\left. + \frac{u_j^{in} - h(u_x)_j^n + \frac{h^2}{2}(u_{xx})_j^n - \frac{h^3}{6}(u_{xxx})_j^n + \frac{h^4}{24}(u_{xxxx})_j^n \ldots}{h^2}\right) -$$

$$- F(x_j, t^n) = (u_t)_j^n - A(u_{xx})_j^n - F(x_j, t^n) +$$

$$+ \frac{\tau}{2}(u_{tt})_{jn}^i - A\frac{h^2}{12}(u_{xxxx})_j^n.$$

Первые три члена являются невязкой уравнения (3.6) в точке (t^n, x_j) и равны 0, поскольку $u(x, t)$ – точное решение уравнения. Главный член погрешности схемы равен $\frac{\tau}{2}(u_{tt})_j^n - A\frac{h^2}{12}(u_{xxxx})_j^n$, т.е. схема имеет первый порядок аппроксимации по времени и второй порядок – по пространству.

Исследуем второе важное свойство разностной схемы – устойчивость. Существуют различные методы исследования устойчивости: принцип максимума, метод разделения переменных (Фурье), метод операторных неравенств и др.

Известен точный метод решения уравнений в частных производных – метод разделения переменных, согласно которому решение можно представить в виде ряда Фурье с бесконечным числом слагаемых. Коэффициенты ряда получены путем разложения в ряд Фурье начальных данных. В случае конечно-разностной задачи число членов ряда не бесконечно, а зависит от числа узлов разностной сетки. Для линейных задач можно ограничиться рассмотрением частного решения в виде одной гармоники $u_j^n = \rho^n e^{ij\varphi}$. Коэффициент ρ определяет скорость роста этой гармоники при переходе с n-го временного слоя на

(n + 1)-й слой. Чтобы ошибка не нарастала с течением времени, необходимо, чтобы $|\rho| \le 1$.

Подставим гармонику в разностную схему (3.9):

$$\rho^n e^{ij\varphi} \frac{\rho - 1}{\tau} = A \frac{e^{i\varphi} - 2 + e^{-i\varphi}}{h^2} \rho^n e^{ij\varphi}.$$

Поскольку $e^{i\varphi} = \cos(\varphi) + i \cdot \sin(\varphi)$, то

$$\frac{\rho - 1}{\tau} = \frac{A\left(\cos(\varphi) + i \cdot \sin(\varphi) - 2 + \cos(\varphi) - i \cdot \sin(\varphi)\right)}{h^2} =$$

$$= \frac{2A}{h^2}\left(\cos(\varphi) - 1\right) = -\frac{4A}{h^2} \sin^2\left(\frac{\varphi}{2}\right); \quad \rho = 1 - \frac{4A\tau}{h^2} \sin^2\left(\frac{\varphi}{2}\right).$$

Обозначим $\gamma = \dfrac{A\tau}{h^2}$ – число Куранта. Для того, чтобы приближенное решение на временном слое (n + 1) не превосходило решение на предыдущем слое, для всех φ должно выполняться условие $|\rho| \le 1$. Это означает $\left|1 - 4\gamma \sin^2\left(\dfrac{\varphi}{2}\right)\right| \le 1$, откуда следуют два неравенства:

1) $1 - 4\gamma \sin^2\left(\dfrac{\varphi}{2}\right) \le 1$

$-4\gamma \sin^2\left(\dfrac{\varphi}{2}\right) \le 0$

$4\gamma \sin^2\left(\dfrac{\varphi}{2}\right) \ge 0$

Выполняется для любого φ

2) $1 - 4\gamma \sin^2\left(\dfrac{\varphi}{2}\right) \ge -1$

$-4\gamma \sin^2\left(\dfrac{\varphi}{2}\right) \ge -2$

$\gamma \sin^2\left(\dfrac{\varphi}{2}\right) \le \dfrac{1}{2}$

Выполняется при $\gamma \le \dfrac{1}{2}$

Таким образом, исследование устойчивости явной схемы для уравнения теплопроводности, выполненное на простейших решениях в виде единичной гармоники, показывает, что решения будут устойчивы, если

$$\gamma = A\frac{\tau}{h^2} \le \frac{1}{2}. \qquad (3.10)$$

Преимуществом явной схемы является то, что решение может быть найдено по явным алгебраическим формулам. Однако, как показали расчеты, приближенное решение, полученное с помощью явной схемы, может быть неустойчивым. Неустойчивость приводит к быстрому (экспоненциальному) росту погрешностей, вносимых в численное решение за счет ошибок округления.

Чтобы понять, как неустойчивость проявляется в расчетах, решим численно уравнение (3.6) при $A = 1$, $F(x, t) = 0$ с нулевыми краевыми условиями первого рода

$$u(0,t) = u(1,t) = 0 \qquad (3.11)$$

и с начальными данными в виде гауссоиды, центрированной относительно точки $x = 1/2$:

$$\varphi(x,0) = e^{-20(x-0.5)^2} - e^{-20(x-1.5)^2} - e^{-20(x+0.5)^2}. \qquad (3.12')$$

Задача имеет точное решение

$$u(x,t) = \frac{1}{\sqrt{1+80t}}\left(e^{-\frac{20(x-0.5)^2}{1+80t}} - e^{-\frac{20(x-1.5)^2}{1+80t}} - e^{-\frac{20(x+0.5)^2}{1+80t}} \right), \qquad (3.12'')$$

график которого приведен на рис. 3.2. Как показывает рисунок, точное решение со временем монотонно убывает.

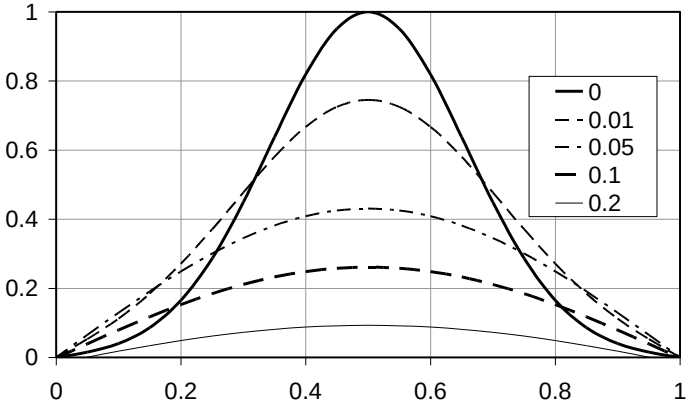


Рис. 3.2. Точное решение (3.12″)
на различные моменты времени

Воспользуемся для решения явной схемой (3.9) на сетке $h = 0.1$, $\tau = 0.02$. Легко проверить, что в этом случае условие устойчивости (3.10) нарушается: $A\dfrac{\tau}{h^2} = \dfrac{0.02}{0.01} = 2 > \dfrac{1}{2}$, и следует ожидать, что решение будет неустойчиво. Действительно, расчеты показывают, что уже через несколько временных шагов численное решение становится немонотонным, и в дальнейшем его график приобретает характерный «пилообразный» вид. Амплитуда «осцилляций» быстро растет, и за несколько временных шагов решение «разваливается», что приводит к характерной ошибке: «Переполнение арифметического устройства». Это означает, что расчет следует вести с маленьким шагом по временной переменной, что существенно ограничивает применение явных схем для решения уравнения теплопроводности. Действительно, пусть $h = 10^{-2}$, $A = 1$, тогда согласно (3.10) для получения устойчивого решения необходимо соблюдать условие $\tau < 5\cdot10^{-5}$. Если решение надо получить на момент времени $T = 1$, то для этого необходимо сделать $N = 2\cdot10^4$ временных шагов. Если же решение надо получить на более подробной сетке по пространственной переменной, например $h = 10^{-3}$, то число временных шагов возрастет до $N = 2\cdot10^6$, и использование явной схемы сделает решение задачи нереализуемым.

Построим для решения задачи (3.6) – (3.8) неявную разностную схему:

$$\frac{u_j^{n+1} - u_j^n}{\tau} = A\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} + F(x_j, t^n). \qquad (3.13)$$

Основное отличие от (3.9) состоит в использовании «неявных», т.е. взятых с верхнего $(n + 1)$-го слоя, значений искомой функции u для аппроксимации второй производной по пространственной переменной. Эта схема также имеет погрешность порядка $\tau^1 + h^2$ и устойчива при любом соотношении шагов τ, h. Такие схемы называют **абсолютно устойчивыми**. На практике это означает, что расчет можно вести с произвольным временным шагом.

Схемы (3.9) и (3.13) являются представителями семейства

$$\frac{u_j^{n+1}-u_j^n}{\tau}=A\left[\begin{array}{l}\sigma\dfrac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{h^2}+\\ (1-\sigma)\dfrac{u_{j+1}^{n}-2u_j^{n}+u_{j-1}^{n}}{h^2}\end{array}\right]+F(x_j,t^n), \quad (3.14)$$

где $1 \geq \sigma \geq 0$ – параметр, который можно подбирать для того, чтобы добиться улучшения аппроксимации или устойчивости схемы. При $\sigma = 0$ (3.14) переходит в явную схему (3.9), а при $\sigma = 1$ – в чисто неявную схему (3.13). При всех других значениях σ в каждом разностном уравнении будут завязаны значения неизвестной функции в шести разных точках, в отличие от явной и неявной схем, в которых завязаны четыре различные точки.

Графическое представление точек расчетной области, входящих в разностное уравнение, называется **шаблоном** конечно-разностной схемы. Шаблоны схем (3.9), (3.13) и (3.14) при $\sigma \neq 0$ представлены на рис. 3.3*а–в* соответственно. В зависимости от того, сколько временных слоев входит в шаблон, схемы бывают *двухслойными* или *трехслойными*. Рисунок 3.3 показывает, что все схемы семейства (3.14) являются двухслойными. Реализация схемы зависит от того, сколько точек находится на верхнем слое шаблона, представляющем искомые величины. Если число точек на верхнем слое меньше или равно двум, решение можно найти с помощью явной процедуры. Схемы, шаблон которых имеет на верхнем слое три точки, реализуются с помощью точного экономичного метода прогонки (см. Раздел 1). Если же на верхнем слое больше трех точек, необходимо применять метод решения СЛАУ с заполненной матрицей, что приводит к существенному увеличению времени расчета.

Для схемы (3.13) на каждом временном шаге необходимо решить СЛАУ с трехдиагональной матрицей:

$$\begin{cases} u^n{}_0 = \mu_{11}(t^n), \\ \gamma u_{j-1}^{n+1} - (1+2\gamma)u_j^{n+1} + \gamma u_{j+1}^{n+1} = u_j^n + \tau\, F(x_j, t^n),\ i = 1, 2, \ldots, N-1. \\ u^n{}_M = \mu_{12}(t^n). \end{cases} \quad (3.15)$$

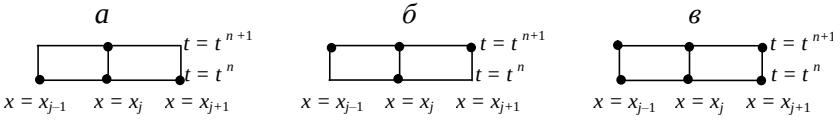


Рис. 3.3. Шаблоны разностных схем (3.9), (3.13) и (3.14)

Решение системы (3.15) находится с помощью прогонки. Для случаев второй и третьей краевой задачи изменяются первое уравнение (3.15), из которого определяются значения первых прогоночных коэффициентов, и последнее уравнение, из которого на обратном этапе прогонки определяется решение в последнем узле сетки. В случае использования более общей схемы (3.14) изменится правая часть уравнений (3.15), однако СЛАУ останется трехдиагональной.

Как было отмечено выше, за счет выбора параметра σ можно добиться, чтобы схема имела более высокий порядок аппроксимации. В частности, легко показать, что симметричная схема ($\sigma = 0.5$) будет иметь порядок аппроксимации $\tau^2 + h^2$. Кроме того, при специальном выборе весового параметра $\sigma = \frac{1}{2} - \frac{h^2}{12A\tau}$ можно добиться, чтобы схема имела порядок аппроксимации $\tau^2 + h^4$.

Схема (3.14) устойчива при выполнении условия $\sigma \geq \frac{1}{2} - \frac{h^2}{4A^2\tau}$, откуда, в частности, легко показать, что неявная ($\sigma = 1$), симметричная ($\sigma = 0.5$) схемы и схема повышенного порядка аппроксимации $\left(\sigma = \frac{1}{2} - \frac{h^2}{12A\tau}\right)$ абсолютно устойчивы.

Нахождение решения разностной схемы (3.14) при $\sigma \neq 0$ аналогично случаю чисто неявной схемы. Система трехточечных уравнений, связывающих решение в точках верхнего ($n + 1$)-го слоя, имеет вид:

$$\sigma\gamma u_{j-1}^{n+1} - (1+2\sigma\gamma)u_j^{n+1} + \gamma\sigma u_{j+1}^{n+1} = u_j^n + \tau(1-\sigma)\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^m}{h^2} +$$

$$+ \tau F(x_j, t^m), \quad n = 0,1,\ldots,M-1,\ j = 1,2,\ldots,N-1,$$

который отличается от (3.15) только правой частью и, следовательно, также решается методом прогонки.

По аналогии с двухслойными схемами (3.15), для уравнения (3.6) можно построить семейство трехслойных схем. Обозначим

$$\Lambda u_j^n = A\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = u_{xx} + \frac{h^2}{12}u^{\mathrm{IV}} + O(h^4).$$

Тогда семейство трехслойных схем можно записать в виде

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\tau} = \Lambda\left(\sigma u_j^{n+1} + (1-2\sigma)u_j^n + \sigma u_j^{n-1}\right). \qquad (3.16)$$

При любых σ схема (3.16) аппроксимирует уравнение (3.6) с порядком $O(\tau^2 + h^2)$, а устойчивость имеет место при $\sigma > 0.25$. При любом $\sigma \neq 0$ шаблон схемы будет иметь три точки на верхнем (неявном) слое, что обусловливает необходимость использования метода прогонки.

Чтобы начать расчеты по трехслойной схеме, нужно знать решение на первых двух временных слоях: t^0, t^1. Однако из начальных данных известно решение только на слое t^0. Есть два способа реализации схемы:

1. Решение на первом временном слое находится из разложения в ряд Тейлора с учетом исходного уравнения.

$$u_j^1 = u_j^0 + \tau u_t + \frac{\tau^2}{2}u_{tt} + O(\tau^3),$$

$$u_t = A^2 u_{xx} \equiv \Lambda u, \quad u_{tt} = (\Lambda u)_t = \Lambda(u_t) = \Lambda^2 u,$$

$$u_j^1 = u_j^0 + \tau\Lambda u_j^0 + \frac{\tau^2}{2}\Lambda^2 u_j^0 = (E + \tau\Lambda + \frac{\tau^2}{2}\Lambda^2)\varphi(x_j).$$

2. Зная начальные данные (т.е. решение на «нулевом слое»), находим решение на первом временном слое по какой-либо двухслойной схеме (явной, неявной). Далее используем трехслойную схему.

9.2. Двумерное уравнение теплопроводности

Рассмотрим теперь конечно-разностные схемы для двумерной задачи. Пусть $G = [0, L_x] \times [0, L_y]$ – прямоугольная область на плоскости (x, y), ∂G – граница области G, $u(x, y, t)$ – функция, определенная в области $G \times [0, T]$. Рассмотрим задачу нахождения решения $u(x, y, t)$, удовлетворяющего уравнению

$$\frac{\partial u}{\partial t} = A\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + F(x, y, t), \tag{3.17}$$

дополненному начальными данными

$$u(x, y, 0) = u_0(x, y)$$

и краевыми условиями первого рода

$$u(x, y, t)\,\big|_{\partial G} = \mu(t).$$

Введем в области $G \times [0, T]$ конечно-разностную сетку с шагами $h_x = L_x/N_x$, $h_y = L_y/N_y$ и $\tau = T/M$: $t^n = n \cdot \tau$, $x_i = ih_x$, $y_j = jh_y$, где N_x, N_y – количество разбиений области по осям x и y. Построим семейство двухслойных конечно-разностных схем:

$$\frac{u_{ij}^{n+1} - u_{ij}^{n}}{\tau} = A\sigma\left(\frac{u_{i+1j}^{n+1} - 2u_{ij}^{n+1} + u_{i-1j}^{n+1}}{h_x^2} + \frac{u_{ij+1}^{n+1} - 2u_{ij}^{n+1} + u_{ij-1}^{n+1}}{h_y^2}\right) +$$

$$+ A(1-\sigma)\left(\frac{u_{i+1j}^{n} - 2u_{ij}^{n} + u_{i-1j}^{n}}{h_x^2} + \frac{u_{ij+1}^{n} - 2u_{ij}^{n} + u_{ij-1}^{n}}{h_y^2}\right) + \tag{3.18}$$

$$+ F(x_i, y_j, t^n).$$

Шаблон схемы (3.18), представленный на рис. 3.4, включает 9 точек на неизвестном, $(n + 1)$-м временном слое, и 9 точек на известном n-м слое.

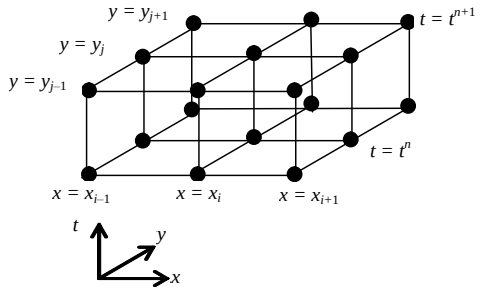


Рис. 3.4. Шаблон схемы (3.18) для уравнения (3.17)

При $\sigma = 0$ схема является явной, и ее решение можно найти по формулам:

$$u_{ij}^{n+1} = u_{ij}^n + \tau A\left(\frac{u_{i+1j}^n - 2u_{ij}^n + u_{i-1j}^n}{h_x^2} + \frac{u_{ij+1}^n - 2u_{ij}^n + u_{ij-1}^n}{h_y^2}\right) +$$

$$+ F(x_i, y_j, t^n), i = 1, 2, ..., N_x; j = 1, 2, ..., N_y; n = 0, 1, 2, ..., M - 1.$$

Явная схема имеет порядок аппроксимации $O(\tau + h_x^2 + h_y^2)$. Как и в случае одной пространственной переменной, схема является условно устойчивой. Для того, чтобы получить устойчивое приближенное решение, шаги разностной сетки должны удовлетворять условию Куранта:

$$A\left(\frac{\tau}{h_x^2} + \frac{\tau}{h_y^2}\right) \le \frac{1}{2}.$$

Свойством безусловной устойчивости схема будет обладать при

$$\sigma > \frac{1}{2} - \frac{h^2}{4\tau A}, \quad h = \max(h_x, h_y).$$

При $\sigma \neq 0$ шаблон схемы (3.18) будет включать 9 точек на верхнем временном слое. Для нахождения решения на неизвестном $(n + 1)$-м слое необходимо решить СЛАУ с заполненной матрицей, для которой экономичный метод прогонки неприменим, что делает процесс решения весьма трудоемким. В этом случае используются так называемые методы дробных шагов, в

которых процесс нахождения решения на новом, $(n + 1)$-м временном слое разбивается на несколько промежуточных (дробных) шагов. Таким образом, на каждом шаге по одному из пространственных направлений схема является явной, а по другому – неявной. Неявность схемы по выбранному направлению делает ее безусловно устойчивой. В то же время для нахождения решения на новом временном слое не требуется решать СЛАУ с заполненной матрицей, а можно найти решение с помощью нескольких прогонок. Эта методика широко используется при решении многомерных уравнений.

Существует много различных схем в дробных шагах для уравнения теплопроводности. Наиболее распространенной является **схема продольно-поперечной прогонки**:

$$\begin{aligned}
&\frac{u_{ij}^{n+\frac{1}{2}} - u_{ij}^{n}}{\tau/2} = A\left(\frac{u_{i+1j}^{n+\frac{1}{2}} - 2u_{ij}^{n+\frac{1}{2}} + u_{i-1j}^{n+\frac{1}{2}}}{h_x^2} + \frac{u_{ij+1}^{n} - 2u_{ij}^{n} + u_{ij-1}^{n}}{h_y^2}\right) + F(x_i, y_j, t^n);\\
&\frac{u_{ij}^{n+1} - u_{ij}^{n+\frac{1}{2}}}{\tau/2} = A\left(\frac{u_{i+1j}^{n+\frac{1}{2}} - 2u_{ij}^{n+\frac{1}{2}} + u_{i-1j}^{n+\frac{1}{2}}}{h_x^2} + \frac{u_{ij+1}^{n+1} - 2u_{ij}^{n+1} + u_{ij-1}^{n+1}}{h_y^2}\right) + F(x_i, y_j, t^{n+\frac{1}{2}}).
\end{aligned} \tag{3.19}$$

Запишем схему в операторном виде:

$$\left(E - \frac{\tau}{2}\Lambda_1\right)u_{ij}^{n+1/2} = \left(E + \frac{\tau}{2}\Lambda_2\right)u_{ij}^{n} + \frac{\tau}{2}F(x_i, y_j, t^n),$$

$$\left(E - \frac{\tau}{2}\Lambda_2\right)u_{ij}^{n+1} = \left(E + \frac{\tau}{2}\Lambda_1\right)u_{ij}^{n+1/2} + \frac{\tau}{2}F(x_i, y_j, t^{n+1}),$$

где $\Lambda_1 u_{ij}^n = A\dfrac{u_{i+1,j}^n - 2u_{i,j}^n + u_{i+1,j}^n}{h_x^2}$, $\Lambda_2 u_{ij}^n = A\dfrac{u_{i,j-1}^n - 2u_{i,j}^n + u_{i,j+1}^n}{h_{yx}^2}$,

E – тождественный оператор. Чтобы исследовать аппроксима-

цию схемы, нужно исключить из нее дробный шаг $u^{n+1/2}$. Применим операторы $\left(E-\frac{\tau}{2}\Lambda_1\right)$, $\left(E+\frac{\tau}{2}\Lambda_1\right)$ к первому и второму уравнениям (3.19), соответственно выразим из уравнений член $\left(E-\frac{\tau}{2}\Lambda_1\right)\left(E+\frac{\tau}{2}\Lambda_1\right)u_{ij}^{n+1/2}$ и приравняем правые части. Таким образом, получим схему в «целых шагах»:

$$\frac{u_{ij}^{n+1}-u_{ij}^{n}}{\tau}=\left(\Lambda_1+\Lambda_1\right)\left(\frac{u_{ij}^{n}+u_{ij}^{n+1}}{2}\right)+\frac{F_{ij}^{n}+F_{ij}^{n+1}}{2}-$$

$$-\frac{\tau^2}{4}\Lambda_1\Lambda_2\left(\frac{u_{ij}^{n}-u_{ij}^{n+1}}{2}\right),$$

откуда следует, что схема (3.19) аппроксимирует уравнение (3.17) с порядком $O(\tau^2 + h^2)$. Схема безусловно устойчива и экономична.

Реализация схемы проводится следующим образом. Пусть решение на n-м временном слое известно. Из первого разностного уравнения с помощью метода прогонки по направлению x находится решение на $(n + ½)$-м шаге. Затем из второго уравнения также с помощью прогонки по направлению y определяется решение на $(n + 1)$-м временном слое. Недостатком схемы является то, что она не обобщается на трехмерный случай.

В случае ненулевых краевых условий или правых частей, граничные условия на дробном шаге необходимо задавать специальным образом. Для этого в обоих уравнениях (3.19) перенесем слагаемые, содержащие $u_{ij}^{n+1/2}$, в левую часть:

$$\left(E-\frac{\tau}{2}\Lambda_1\right)u_{ij}^{n+1/2}=\left(E+\frac{\tau}{2}\Lambda_2\right)u_{ij}^{n}+\frac{\tau}{2}F(x_i,y_j,t^n),$$

$$\left(E+\frac{\tau}{2}\Lambda_1\right)u_{ij}^{n+1/2}=\left(E-\frac{\tau}{2}\Lambda_2\right)u_{ij}^{n+1}-\frac{\tau}{2}F(x_i,y_j,t^{n+1}).$$

Сложив эти уравнения, получим

$$u_{ij}^{n+1/2} = \frac{u_{ij}^{n} + u_{ij}^{n+1}}{2} - \frac{\tau}{4}\Lambda_2\left(\frac{u_{ij}^{n} - u_{ij}^{n+1}}{2}\right) + \frac{\tau}{2}(F_{ij}^{n} - F_{ij}^{n+1}).$$

Последнее равенство позволяет определить решение на дробном шаге на вертикальных границах: $i = 0$ и $i = N_x$, $j = 1, 2, \ldots, N_y$.

Для решения (3.17) используют также **схему расщепления**:

$$\frac{u_{ij}^{n+\frac{1}{2}} - u_{ij}^{n}}{\tau} = A\left(\frac{u_{i+1j}^{n+\frac{1}{2}} - 2u_{ij}^{n+\frac{1}{2}} + u_{i-1j}^{n+\frac{1}{2}}}{h_x^2}\right) + F(x_i, y_j, t^{n+1}),$$

$$\frac{u_{ij}^{n+1} - u_{ij}^{n+\frac{1}{2}}}{\tau} = A\left(\frac{u_{ij+1}^{n+1} - 2u_{ij}^{n+1} + u_{ij-1}^{n+1}}{h_y^2}\right); \qquad (3.20)$$

и **схему предиктор – корректор**:

$$\frac{u_{ij}^{n+\frac{1}{4}} - u_{ij}^{n}}{\tau/2} = \Lambda_1 u_{ij}^{n+\frac{1}{4}} + F(x_i, y_j, t^{n+\frac{1}{2}}),$$

$$\frac{u_{ij}^{n+\frac{1}{2}} - u_{ij}^{n+\frac{1}{4}}}{\tau/2} = \Lambda_2 u_{ij}^{n+\frac{1}{2}}, \qquad (3.21)$$

$$\frac{u_{ij}^{n+1} - u_{ij}^{n}}{\tau} = (\Lambda_1 + \Lambda_2) u_{ij}^{n+\frac{1}{2}} + F(x_i, y_j, t^{n+\frac{1}{2}}).$$

Исследование аппроксимации для этих схем проводится так же, как и для (3.19). Преимуществом схем (3.20), (3.21) по сравнению с (3.19) является то, что они могут быть легко обобщены на пространственный случай.

Тема 10. Уравнение Пуассона

В качестве классического представителя уравнений эллиптического типа рассмотрим двумерное уравнение Пуассона:

$$\Delta u \equiv \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = F(x, y), \qquad (3.22)$$

$F(x, y)$ – функция источников. Уравнение Пуассона описывает, например, распределение электростатического поля или стационарное распределение температуры. Частным случаем этого уравнения является уравнение Лапласа $\Delta u = 0$.

Пусть требуется определить решение в некоторой области G на плоскости (x, y). Корректная постановка задачи требует задания граничных условий на границе этой области ∂G.

В одномерном случае уравнение Пуассона не что иное, как краевая задача первого рода для ОДУ второго порядка, решение которой было рассмотрено в Теме 7.

10.1. Метод установления

Сравнивая уравнение Пуассона и рассмотренное выше двумерное уравнение теплопроводности, можно понять, что уравнение Пуассона является стационарным, т.е. не зависящим от времени вариантом уравнения теплопроводности. Поэтому для решения уравнения Пуассона часто используют так называемый метод установления. Для этого в правую часть уравнения (3.22) добавляют слагаемое $\frac{\partial u}{\partial t}$ и решают полученное уравнение теплопроводности с помощью описанных в Теме 9 методов до тех пор, пока решение не выйдет на стационар, т.е. не перестанет изменяться в зависимости от времени. Время в этой задаче является фиктивным, и в разностных схемах надо использовать максимально возможный шаг. Решение нестационарной задачи стремится к решению стационарной независимо от выбора начальных данных. Процесс установления решения может занять продолжительное время, особенно если используются явные схемы, имеющие жесткое ограничение на временной шаг. В этом случае применение схем дробных шагов помогает существенно сократить время решения.

10.2. Итерационные методы

Для решения уравнения Пуассона используются и другие методы, не связанные со сведением его к уравнению теплопроводности. Как правило, все эти методы приводят к решению

СЛАУ с заполненной матрицей, которая решается одним из итерационных методов.

Пусть областью $G = \{a \le x \le b, c \le y \le d\}$ будет прямоугольник, на границах которого задано гладкое решение $u(x, y)\big|_{\partial G} = \gamma$. Построим в области G прямоугольную расчетную сетку с шагами $h_x = \dfrac{b-a}{N_x}$, $h_y = \dfrac{d-c}{N_y}$. Аппроксимируя вторые производные центральными разностями, имеем:

$$\frac{u_{i-1}^j - 2u_i^j + u_{i+1}^j}{h_x^2} + \frac{u_i^{j-1} - 2u_i^j + u_i^{j+1}}{h_y^2} = F_i^j,$$

$$i = 1, \ldots, N_x - 1,\ j = 1, \ldots, N_y - 1,$$

$$u_0^j = \gamma_1^j,\ u_{N_x}^j = \gamma_2^j,\ u_i^0 = \gamma_3^i,\ u_i^{N_y} = \gamma_4^i.$$

Для прямоугольной области полученную систему линейных алгебраических уравнений можно записать в векторно-матричной форме:

$$A_i \vec{u}_{i-1} - C_i \vec{u}_i + B_i \vec{u}_{i+1} = \vec{F}_i, \qquad (3.23)$$

$$A_i = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & h_y^2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 \end{bmatrix},\ C_i = \begin{bmatrix} 1 & 0 & 0 & \ldots 0 \\ -h_x^2 & 2\left(h_x^2 + h_y^2\right) & -h_x^2 & \ldots 0 \\ \ldots & \ldots & \ldots & \ldots 0 \\ 0 & 0 & 0 & \ldots 1 \end{bmatrix},$$

$$B_i = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & h_x^2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 \end{bmatrix},\ \vec{u}_i = \begin{bmatrix} u_i^0 \\ u_i^1 \\ \ldots \\ u_i^{N_y} \end{bmatrix},\ \vec{F}_i = h_x^2 h_y^2 \begin{bmatrix} \gamma_3^i \\ f_i^1 \\ \ldots \\ \gamma_4^i \end{bmatrix},$$

$$\vec{u}_0 = \gamma_1^j,\ \vec{u}_{N_x} = \gamma_2^j,\ j = 1, \ldots, N_y.$$

Система (3.23) имеет блочно-трехдиагональную матрицу и может быть решена с помощью метода матричной прогонки. Однако этот метод требует больших затрат машинного времени и практически не применяется. Для решения (3.23) обычно используют итерационные методы. Достаточно эффективным

здесь оказывается метод Зейделя в различных его модификациях. Рассмотрим поточечный (классический) и блочный метод Зейделя.

Запишем поточечный метод в виде:

$$2\left(h_x^2+h_y^2\right)u_i^{j\,(n+1)} = h_y^2 u_{i+1}^{j\,(n)} + h_y^2 u_{i-1}^{j\,(n)} + h_x^2 u_i^{j+1(n)} + \\ + h_x^2 u_i^{j-1(n)} + h_x^2 h_x^2 F_i^{\,j},$$

$$i=1,\ldots,N_x-1,\ j=1,\ldots,N_y-1,$$

где n – номер итерации. Начальные значения $u_i^{j(0)}$ могут задаваться произвольно. Можно увидеть, что влияние граничных условий в этом методе распространяется на каждом итерационном шаге на один узел сетки по соответствующей координате, поэтому сходимость оказывается достаточно медленной. Более высокой скоростью сходимости обладает блочный метод Зейделя, который записывается в виде:

$$A_i\vec{u}_{i-1}^{\,(n+1)} = C_i\vec{u}_i^{\,(n)} - B_i\vec{u}_{i+1}^{\,(n)} + \vec{F}_i,\ i=1,\ldots,N_x-1.$$

Если записать его по точкам сетки, то получим следующую формулу:

$$2\left(h_x^2+h_y^2\right)u_i^{j\,(n+1)} = h_y^2 u_{i+1}^{j\,(n)} + h_y^2 u_{i-1}^{j\,(n+1)} + h_x^2 u_i^{j+1(n+1)} + \\ + h_x^2 u_i^{j-1(n+1)} + h_x^2 h_x^2 F_i^{\,j},$$

$$i=1,\ldots,N_x-1,\ j=1,\ldots,N_y-1.$$

Для нахождения $u^{(n+1)}$ на j-й строке нужно решить трехдиагональную СЛАУ методом прогонки и полученное решение разместить на месте $u^{(n)}$ в j-й строке. Видно, что влияние граничных условий по j-й строке распространяется значительно быстрее, чем в поточечном варианте метода, и скорость сходимости оказывается выше.

Если область определения решения является криволинейной, численное решение значительно усложняется. В этом случае решение в приграничных узлах необходимо определять с помощью различных методов интерполяции.

Тема 11. Уравнения гиперболического типа

11.1. Характеристики

Основное свойство уравнений гиперболического типа – наличие полного набора вещественных характеристик, т.е. направлений, вдоль которых уравнение (систему уравнений) можно проинтегрировать (найти первый интеграл). Примеры уравнений гиперболического типа:

– линейное уравнение переноса (адвекции)

$$\frac{\partial u}{\partial t} + C\frac{\partial u}{\partial x} = 0, \quad u = u(x,t), \quad C = \text{const}; \tag{3.24}$$

– квазилинейное уравнение вида закона сохранения

$$\frac{\partial u}{\partial t} + \frac{\partial \varphi(u)}{\partial x} = 0; \tag{3.25}$$

– волновое уравнение

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} = 0, \tag{3.26}$$

а также уравнения теории упругости, мелкой воды, система квазилинейных уравнений типа законов сохранения и еще много других уравнений из различных областей механики и физики.

11.2. Линейное уравнение переноса

Познакомимся с методами приближенного решения гиперболических уравнений на примере простейшего из них – **линейного уравнения переноса**. Рассмотрим уравнение (3.24) при $t \geq 0$. Характеристиками уравнения являются прямые

$$\frac{dx}{dt} = C$$

(рис. 3.5). Поскольку на рис. 3.5 ось t направлена вверх, характеристики имеют угол наклона $1/C$.

Запишем производную от решения по характеристическому направлению:

$$\left.\frac{du(x,t)}{dt}\right|_{\frac{dx}{dt}=C} = \frac{\partial u}{\partial x}\frac{dx}{dt} + \frac{\partial u}{\partial t} = \frac{\partial u}{\partial x}C + \frac{\partial u}{\partial t} = 0$$

в силу уравнения. Таким образом, вдоль данного направления решение уравнения не изменяется. Поэтому говорят, что решение уравнения (3.24) «переносится вдоль характеристик». Допустим, что известны начальные данные $u(x,0)=\varphi(x)$, и необходимо найти решение в точке (x^*, t^*). Построим поле характеристик и определим точку на прямой $t=0$, в которую приходит «выпущенная» из точки (x^*, t^*) характеристика: $x = x_0^*$ (см. рис. 3.5). Поскольку решение сохраняется вдоль характеристик, то $u(x^*,\ t^*)=\varphi(x_0^*)$.

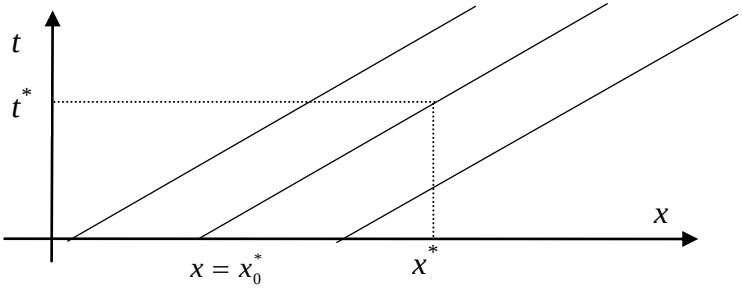


Рис. 3.5. Решение линейного уравнения переноса ($C > 0$)

Если начальные данные заданы на всей прямой $t = 0$, то характеристика, выпущенная из любой точки полуплоскости (x, t), $t > 0$, пересечет линию $t = 0$. Поэтому решение в любой момент времени определяется как $u(x,t)=\varphi(x-Ct)$.

Допустим, что начальные данные заданы в ограниченной области: $0 \le x \le L$, и требуется определить решение при $t \le T$. Выпустим из каждой точки отрезка $[0, L]$ характеристики. В зависимости от знака C они закроют правую или левую часть прямоугольника $[0 < t \le T] \times [0 \le x \le L]$ (рис. 3.6).

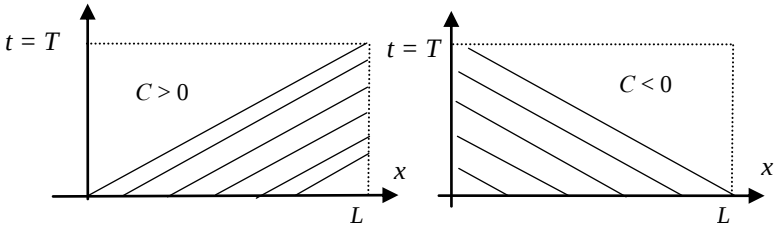


Рис. 3.6. Решение линейного уравнения переноса
в ограниченной области

Для того, чтобы определить решение во всем прямоугольнике, необходимо задать дополнительные (краевые) условия, причем, если $C > 0$, краевые условия задаются на линии $x = 0$, а если $C < 0$ – на линии $x = L$.

Этот простой пример демонстрирует правило корректной постановки начальных и краевых условий для гиперболических уравнений. На каждой границе области решения задачи необходимо задать столько условий, сколько семейств характеристик уходит с этой границы.

Как будет показано ниже, от выбора формулы зависят многие важные свойства приближенного решения: аппроксимация, устойчивость, монотонность.

Построим для уравнения (3.24) несколько простых схем и изучим их свойства. Для аппроксимации производной по пространству можно использовать различные формулы разностного дифференцирования (вперед, назад, центральная разность). При этом необходимо учитывать направление потока – знак константы C. Как следует из рис. 3.6, при $C > 0$ поток направлен слева направо, а при $C < 0$ – справа налево.

Пусть $C > 0$. Построим явную схему, использующую формулу «разность назад», т.е. **против потока**:

$$\frac{u_j^{n+1} - u_j^n}{\tau} + C\frac{u_j^n - u_{j-1}^n}{h} = 0. \qquad (3.27)$$

Выразим искомую величину:

$$u_j^{n+1} = u_j^n - \frac{C\tau}{h}\left(u_j^n - u_{j-1}^n\right) = (1-r)u_j^n + ru_{j-1}^n,$$

где $r = C\tau/h$ – число Куранта.

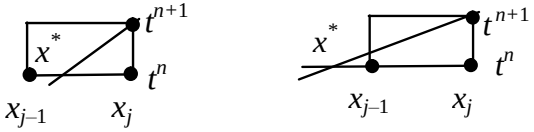


Рис. 3.7. Шаблон схемы (3.27)

Как показывает рис. 3.7, шаблон этой схемы состоит из трех точек, две из которых находятся на явном n-м слое и одна – на неявном $(n + 1)$-м слое. Из точки (t^{n+1}, x_j) выпущена характеристика, которая пересекает линию $t = t^n$ в точке x^*. При $r < 1$ $x^* \in (x_j, x_{j-1})$, при $r = 1$ x^* совпадает с x_{j-1}, а при $r > 1$ находится вне отрезка (x_j, x_{j-1}). В силу рассмотренных выше свойств уравнения переноса $u(t^{n+1}, x_j) = u(x^*)$, следовательно, при $r = 1$ разностная схема (3.27) будет давать точное решение уравнения (3.24). При $r < 1$ для нахождения $u(x^*)$ необходимо использовать интерполяционные формулы, от точности которых зависит порядок аппроксимации разностной схемы. При $r > 1$ значение $u(x^*)$ определяется с помощью экстраполяции, и следует ожидать неустойчивое поведение приближенного решения.

Исследуем устойчивость схемы методом Фурье. Подставим выражение для гармоники $u_j^n = \rho^n e^{ij\varphi}$ в разностную схему:

$$\rho = (1-r) + re^{-i\varphi} = (1-r) + r(\cos\varphi - i\sin\varphi)\,. \qquad (3.28)$$

Геометрическое место точек, описываемых уравнением (3.28), – окружность радиуса r с центром в точке $(1 – r, 0)$. Возможны три геометрические ситуации, представленные на рис. 3.8.

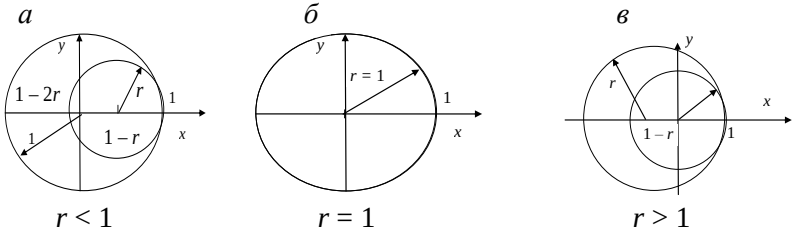


Рис. 3.8. Геометрическая иллюстрация уравнения (3.28)

Рисунки 3.8*а,б* показывают, что при $r \le 1$ окружность (3.28) лежит внутри или совпадает с окружностью $|\rho| = 1$, что означает устойчивость разностной схемы. При $r > 1$ (рис. 3.8*в*) окружность (3.28) является внешней по отношению к $|\rho| = 1$, что свидетельствует о неустойчивости схемы (3.27). К такому же выводу можно прийти аналитически, используя (3.28) и тригонометрические формулы половинного угла:

$$\rho = (1-r) + re^{-i\varphi} = (1-r) + r(\cos\varphi - i\sin\varphi) =$$

$$= 1 - r(1-\cos\varphi) - ir\sin\varphi = 1 - 2r\sin^2\frac{\varphi}{2} - ir\sin\varphi.$$

Оценим модуль комплексного числа ρ:

$$|\rho|^2 = \rho\overline{\rho} = 1 - 4r\sin^2\frac{\varphi}{2} + 4r^2\sin^4\frac{\varphi}{2} + r^2\sin^2\varphi =$$

$$= 1 - 4r\sin^2\frac{\varphi}{2} + 4r^2\sin^4\frac{\varphi}{2} + 4r^2\sin^2\frac{\varphi}{2}\cos^2\frac{\varphi}{2} =$$

$$= 1 - 4r\sin^2\frac{\varphi}{2} + 4r^2\sin^2\frac{\varphi}{2} \le 1, \quad -4r\sin^2\frac{\varphi}{2} + 4r^2\sin^2\frac{\varphi}{2} \le 0,$$

$$-4r\sin^2\frac{\varphi}{2}(1-r) \le 0,$$

что справедливо при $r \le 1$ (**условие Куранта**).

Следующий пример – явная схема **по потоку**:

$$\frac{u_j^{n+1} - u_j^n}{\tau} + C\frac{u_{j+1}^n - u_j^n}{h} = 0, \qquad (3.29)$$

которую можно также записать в виде:

$$u_j^{n+1} = u_j^n - \frac{C\tau}{h}\left(u_{j+1}^n - u_j^n\right) = (1+r)u_j^n - ru_{j+1}^n.$$

Как показывает рис. 3.9*а*, на котором изображен шаблон схемы, характеристика уравнения, выпущенная из точки (t^{n+1}, x_j) ни при каких значениях числа Куранта не пересечет отрезок (x_j, x_{j+1}), следовательно, можно ожидать, что схема будет неустойчивой. Действительно, исследуем устойчивость схемы методом Фурье. Подставив гармонику $u_j^n = \rho^n e^{ij\varphi}$ в разностную схему, получим уравнение окружности радиуса r с центром в точке (1 + r, 0):

$$\rho = (1+r) - re^{i\varphi} = (1+r) - r(\cos\varphi + i\sin\varphi). \qquad (3.30)$$

Рисунок 3.9*б* показывает, что $|\rho| \geq 1$ при любом r, и, следовательно, схема является абсолютно неустойчивой.

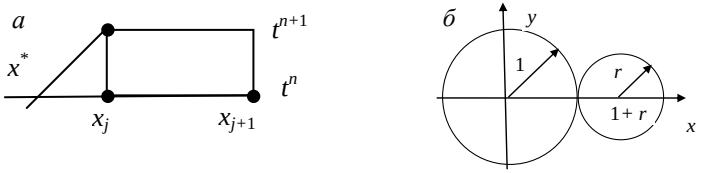


Рис. 3.9. Шаблон схемы (3.29) (*а*)
и иллюстрация уравнения (3.30) (*б*)

Третья явная схема $\frac{u_j^{n+1} - u_j^n}{\tau} + C\frac{u_{j+1}^n - u_{j-1}^n}{2h} = 0$ имеет второй порядок аппроксимации по пространству, однако, как и предыдущая схема по потоку, является абсолютно неустойчивой.

Кроме явных, для решения уравнения (3.24) можно предложить **неявные схемы** с аппроксимацией против потока:

$$\frac{u_j^{n+1} - u_j^n}{\tau} + C\frac{u_j^{n+1} - u_{j-1}^{n+1}}{h} = 0 \qquad (3.31)$$

или по потоку:

$$\frac{u_j^{n+1}-u_j^n}{\tau}+C\frac{u_{j+1}^{n+1}-u_j^{n+1}}{h}=0. \qquad (3.32)$$

Обе схемы имеют первый порядок аппроксимации по времени и пространству. С помощью метода Фурье можно показать, что при $C>0$ схема (3.31) абсолютно устойчива, а схема (3.32) условно устойчива при $r \geq 1$. Если знак константы изменится, необходимо изменить направление аппроксимации производной по пространству.

Рассмотрим теперь линейное уравнение переноса с переменным коэффициентом:

$$\frac{\partial u}{\partial t}+C(x,t)\frac{\partial u}{\partial x}=0,\ 0\le x\le L,\ \ t\ge 0. \qquad (3.33)$$

Для ранее рассмотренных явных и неявных схем устойчивая аппроксимация зависит от знака $C(x, t)$. Если в какой-то части области решения знак функции $C(x, t)$ положителен, а в какой-то – отрицателен, можно использовать **гибридную схему** Куранта – Изаксона – Риса (КИР):

$$\frac{u_j^{n+1}-u_j^n}{\tilde{\tau}}+\frac{C_j+|C_j|}{2}\frac{u_j^n-u_{j-1}^n}{h}+\frac{C_j-|C_j|}{2}\frac{u_{j+1}^n-u_{j-1}^n}{h}=0.$$

Схема КИР устойчива при выполнении условия Куранта $r \leq 1$ при любом знаке коэффициента $C(x, t)$.

Приведем еще несколько явных разностных схем, которые используют для решения уравнения (3.33):

– **симметричная схема**

$$\frac{u_j^{n+1}+u_{j-1}^{n+1}-u_j^n-u_{j-1}^n}{2\tau}+C\frac{u_j^{n+1}+u_j^n-(u_{j-1}^{n+1}+u_{j-1}^n)}{2h}=0; \qquad (3.34)$$

– **схема Лакса**

$$\frac{u_j^{n+1}-u_j^n}{\tau}+C_j\frac{u_{j+1}^n-u_{j-1}^n}{2h}-\frac{h^2}{2\tau}\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{h^2}=0;$$

– **схема Лакса – Вендрофа**

$$\frac{u_{j-1/2}^{n+1/2}-u_{j-1/2}^n}{\tau}+C_j\frac{u_j^n-u_{j-1}^n}{h}=0 \quad \text{(предиктор)},$$

$$\frac{u_j^{n+1}-u_j^n}{\tau}+C_j\frac{u_{j+1/2}^{n+1/2}-u_{j-1/2}^{n+1/2}}{h}=0 \quad \text{(корректор)},$$

где $u_{j-1/2}^n=\dfrac{u_j^n+u_{j-1}^n}{2}$;

– **схема Мак-Кормака**

$$\frac{\tilde{u}_j-u_j^n}{\tau}+C_j\frac{u_{j+1}^n-u_j^n}{h}=0 \qquad \text{(предиктор)},$$

$$\frac{u_j^{n+1}-u_j^n}{\tau}+\frac{C_j}{2}\left(\frac{u_{j+1}^n-u_j^n}{h}+\frac{\tilde{u}_j-\tilde{u}_{j-1}}{h}\right)=0 \quad \text{(корректор)}.$$

11.3. Разрывные решения

При разработке методов приближенного решения необходимо, чтобы приближенные решения сохраняли все качественные свойства исходного уравнения. Уравнение (3.24) допускает разрывные решения. Разрыв может формироваться в начальный момент из-за несогласованности начальных данных и краевых условий на прилегающей границе (рис. 3.10*а*). Если начальные данные содержат разрыв, то он будет «переноситься» по характеристике. В этом случае решение будет иметь вид, представленный на рис. 3.10*б*.

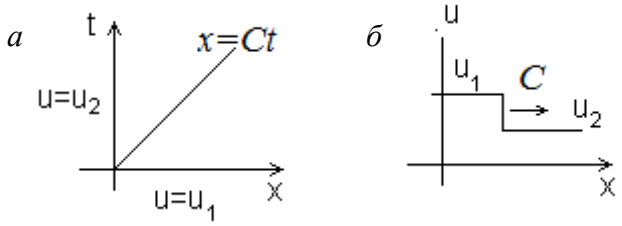


Рис. 3.10. Формирование разрыва из-за несогласованности начальных данных и краевых условий (*а*) и разрывное решение уравнения адвекции (*б*)

Следует заметить, что разрывное решение не может быть решением исходного дифференциального уравнения (3.24), а более общего интегрального уравнения типа закона сохранения, о котором будет рассказано ниже.

11.4. Первое дифференциальное приближение

Способность разностной схемы воспроизводить разрывные решения называется **К-свойством**. Наличие К-свойства очень важно для расчета задач, в которых могут возникать разрывы решений, например, ударные волны в газовой динамике. Далеко не все рассмотренные выше схемы обладают этим свойством. Для исследования данного свойства схем используется **метод первого дифференциального приближения (ПДП)**.

Построим первое дифференциальное приближение схемы (3.27). Для этого разложим решение в ряд Тейлора в окрестности точки (t^n, x_j):

$$u_j^{n+1} = u_j^n + \tau u_t + \frac{\tau^2}{2} u_{tt};\ u_{j-1}^n = u_j^n + h u_x + \frac{h^2}{2} u_{xx}$$

и подставим разложения в разностную схему. В результате получим так называемую **Г-форму ПДП**, в которую кроме членов исходного уравнения входят некоторые добавки порядка аппроксимации схемы:

$$u_t + \frac{\tau}{2} u_{tt} + C\left(u_x - \frac{h}{2} u_{xx} \right) = 0.$$

Преобразуем выражение, используя исходное уравнение:

$$u_t = -C u_x;\ u_{tt} = -C u_{xt} = -C(u_t)_x = C^2 u_{xx},$$

в результате чего получим **П-форму ПДП**:

$$u_t + \frac{\tau}{2} C^2 u_{xx} + C u_x - \frac{Ch}{2} u_{xx} = 0.$$

С учетом обозначения $\alpha = \frac{Ch}{2}\left(\frac{C\tau}{h} - 1 \right)$ ПДП примет вид

$$u_t + C u_x = \alpha u_{xx}. \qquad (3.35)$$

Полученное уравнение отличается от исходного (3.24) наличием члена со второй производной в правой части, имеющего физический смысл вязкости (диффузии, теплопроводности). В этом случае говорят, что схема обладает **схемной вязкостью**, или диффузией. Если $r = \frac{C\tau}{h} < 1$, то $\alpha > 0$, т.е. схема добавляет в исходное уравнение вязкость, которая «размазывает» разрывы и большие градиенты решения. Если $r = \frac{C\tau}{h} = 1$, то $\alpha = 0$ (вязкости нет), и схема (3.27) способна точно воспроизводить разрывные решения. Если $r = \frac{C\tau}{h} > 1$, то $\alpha < 0$, т.е. вязкость отрицательна, что говорит о некорректности уравнения (3.33). Любое численное решение задачи Коши с начальными данными при $t = 0$ будет разрушаться за несколько временных шагов. Это согласуется с тем, что явная схема (3.27) неустойчива при числах Куранта, превышающих единицу.

Аналогичные выкладки для схемы (3.31) приводят к ПДП

$$u_t + Cu_x = \frac{Ch}{2}\left(1 + \frac{C\tau}{h}\right)u_{xx}, \qquad (3.36)$$

из которого следует, что при любом числе Куранта в уравнении присутствует положительная схемная вязкость. Это, с одной стороны, обеспечивает абсолютную устойчивость неявной схемы, но также делает схему диффузионной.

Найдем ПДП схемы (3.32). Разложим входящие в схему выражения в ряд Тейлора в окрестности точки $(t^{n+1/2}, x_{j-1/2})$, сохраняя члены порядка аппроксимации схемы:

$$\frac{u_j^{n+1} + u_{j-1}^{n+1}}{2} = u_{j-1/2}^{n+1} = u_{j-1/2}^{n+1/2} + \frac{\tau}{2}u_t + \frac{\tau^2}{8}u_{tt} + \frac{\tau^3}{48}u_{ttt},$$

$$\frac{u_j^{n} + u_{j-1}^{n+1}}{2} = u_{j-1/2}^{n} = u_{j-1/2}^{n+1/2} - \frac{\tau}{2}u_t + \frac{\tau^2}{8}u_{tt} - \frac{\tau^3}{48}u_{ttt},$$

$$\frac{u_j^{n+1}+u_j^n}{2}=u_j^{n+1/2}=u_{j-1/2}^{n+1/2}+\frac{h}{2}u_x+\frac{h^2}{8}u_{xx}+\frac{h^3}{48}u_{xxx},$$

$$\frac{u_{j-1}^{n+1}+u_{j-1}^n}{2}=u_{j-1}^{n+1/2}=u_{j-1/2}^{n+1/2}-\frac{h}{2}u_x+\frac{h^2}{8}u_{xx}-\frac{h^3}{48}u_{xxx}$$

и подставим эти выражения в (3.32):

$$\frac{u+\frac{\tau}{2}u_t+\frac{\tau^2}{8}u_{tt}+\frac{\tau^3}{48}u_{ttt}-u+\frac{\tau}{2}u_t-\frac{\tau^2}{8}u_{tt}+\frac{\tau^3}{48}u_{ttt}}{\tau}+$$

$$+C\frac{u+\frac{h}{2}u_x+\frac{h^2}{8}u_{xx}+\frac{h^3}{48}u_{xxx}-u+\frac{h}{2}u_x-\frac{h^2}{8}u_{xx}+\frac{h^3}{48}u_{xxx}}{h}=0.$$

После преобразований получим Г-форму ПДП:

$$u_t+Cu_x-\frac{C^3\tau^2}{24}u_{ttt}+\frac{Ch^2}{24}u_{xxx}=0,$$

а с учетом уравнения и его следствий

$$u_t=-Cu_x;\ u_{tt}=C^2u_{xx};\ u_{ttt}=-C^3u_{xxx}$$

– П-форму ПДП:

$$u_t+Cu_x-\frac{Ch^2}{24}\left(\frac{C^2\tau^2}{r^2}-1\right)u_{xxx}=0. \qquad (3.37)$$

В отличие от (3.35) и (3.36), уравнение (3.37) не содержит члена со второй производной по пространству (схемную вязкость). Однако в уравнении появился член с третьей производной по пространству, имеющий физический смысл **численной дисперсии**. Наличие дисперсионного члена означает зависимость $\omega-$ скорости распространения гармоники $e^{ikx+\omega t}$ от ее длины волны k. Если начальные данные представляют собой суперпозицию нескольких гармоник с различными длинами волн, то при расчете с использованием схемы (3.37) численное решение распадется на несколько отдельных возмущений, двигающихся с различными скоростями. Численная дисперсия приводит к появлению мелких высокочастотных возмущений впереди или позади областей больших градиентов или разрывов решения.

11.5. Монотонность численного решения

Важным свойством точного решения является сохранение монотонности начальных данных. Это означает, что если в начальных данных отсутствовали локальные экстремумы, то решение останется монотонным во все последующие моменты времени. Схемы, сохраняющие это свойство уравнения, называются **монотонными**. С.К. Годунов доказал, что любая явная двухслойная линейная схема, которую можно записать в виде

$$y_i^{n+1} = \sum_l \beta_l y_{i+l}^n , \qquad (3.38)$$

где $l_0 \le l \le l_1$ – номера точек шаблона на явном слое, будет монотонна тогда и только тогда, если все коэффициенты β_l неотрицательны.

Например, явную схему (3.27) можно записать как

$$u_j^{n+1} = ru_{j-1}^n + (1-r)u_j^n .$$

Здесь $\beta_0 = r > 0; \beta_1 = 1-r \ge 0$. Следовательно, если число Куранта меньше 1 и схема устойчива, она монотонна.

Неявную схему (3.31) можно записать в виде

$$u_j^{n+1}(h + C\tau) = hu_j^n + C\tau u_{j-1}^{n+1} .$$

Выражая искомую функцию на неявном слое, получим бесконечный ряд:

$$u_j^{n+1} = \frac{1}{h+C\tau}\left(C\tau u_{j-1}^{n+1} + hu_j^n\right) = \frac{1}{h+C\tau}\left(C\tau\left(\frac{1}{h+C\tau}u_{j-2}^{n+1} + hu_{j-1}^n\right) + h_j^n\right) = \ldots$$

$$\ldots \frac{h}{h+C\tau}\sum_{l=0}^{\infty}\left(\frac{C\tau}{h+C\tau}\right)^l u_{j-l}^n ; \beta_l \ge 0.$$

Известна доказанная С.К. Годуновым теорема («**барьер Годунова**») о том, что монотонная двухслойная линейная схема не может иметь порядок аппроксимации выше первого.

11.6. Схемы высокого порядка

Как показывают численные эксперименты, при использовании схем, имеющих второй и выше порядок аппроксимации

(например, рассмотренная выше схема (3.34)), на монотонном начальном профиле появляются осцилляции, т.е. решение становится немонотонным. Эта особенность численного решения связана с численной дисперсией, т.е. наличием члена с третьей производной, который присутствует в ПДП схем с порядком аппроксимации выше первого.

Необходимость точного воспроизведения в расчетах разрывов и областей решения с большими градиентами при сохранении монотонности начального профиля привела к идее создания так называемых **TVD-схем**, т.е. схем с невозрастающей полной вариацией (*Total Variation Diminishing*). Полная вариация численного решения на *n*-м временном слое определяется как

$$TV\left(u^n\right) = \sum_{l=-\infty}^{\infty} \left| u_j^n - u_{j-1}^n \right| .$$

Схемы, у которых полная вариация решения не возрастает при переходе на следующий временной слой, называются TVD-схемами. TVD-свойство означает, что в расчете не возникают новые локальные максимумы или минимумы, существующие локальные максимумы не увеличиваются, а минимумы – не убывают.

Можно доказать, что линейная схема (3.38) обладает TVD-свойством, если $\beta_l \geq 0$ и $\sum_l \beta_l \leq 1$. Таким образом, все монотонные схемы обладают TVD-свойством, обратное в общем случае неверно.

Чтобы преодолеть «барьер Годунова», т.е. построить TVD-схему, имеющую, например, второй порядок аппроксимации, в исходную разностную схему вводят так называемые **ограничители** (*limiters*).

Любая двухслойная схема для (3.24) может быть записана как

$$u_j^{n+1} = u_j^n - r(u_j^n - u_{j-1}^n) + \Phi_{j-1/2} L(u_j^n, u_{j\pm1}^n, u_{j\pm2}^n, \ldots),$$

где первые два члена правой части представляют собой устойчивую противопоточную аппроксимацию, а перед дополнитель-

ным членом L, обеспечивающим более высокий порядок аппроксимации, введен ограничитель

$$\Phi_j = \Phi_j(S_j), \text{ где } S_j = \frac{u_j^n - u_{j-1}^n}{u_{j+1}^n - u_j^n}.$$

Ограничитель обращается в ноль там, где его аргумент отрицателен. В случае появления осцилляций (как правило, это происходит в областях с большими градиентами или разрывами решений) исходная схема превращается в устойчивую монотонную схему первого порядка аппроксимации, и присущая ей схемная вязкость гасит нежелательные осцилляции.

В литературе можно найти много образцов ограничителя, приведем наиболее распространенные:

– *minmod*: $\Phi(S) = \max[0, \min(1, S)]$;

– *superbee*: $\Phi(S) = \max[0, \min(1, 2S), \min(2, S)]$;

– *van Leer*: $\Phi(S) = \dfrac{|S| + S}{|S| + 1}$.

Следует заметить, что точно доказать TVD-свойство можно лишь в простейшем линейном случае. Однако и в нелинейных задачах используют аналогичный прием. В частности, TVD-схемы нашли широкое распространение при расчете задач газовой динамики, в которых часто встречаются разрывные решения (ударные волны, контактные разрывы).

11.7. Уравнение Хопфа. Градиентная катастрофа

Познакомимся со свойствами нелинейных гиперболических уравнений на примере уравнения (3.25), которое в частном случае $\varphi(u) = \dfrac{u^2}{2}$ можно записать в виде

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0. \qquad (3.39)$$

Характеристики уравнения (3.39) находятся из уравнения $\frac{dx}{dt} = u(x,t)$, следовательно, если $u(x, t)$ отлично от константы, характеристиками будут кривые линии.

Будем решать уравнение (3.39) в области $\mathbf{\Omega} = [0, L] \times [0, T]$. Пусть заданы начальные данные и краевые условия

$$u(x,0) = \varphi(x), \quad \varphi(x) > 0, \quad \varphi'(x) > 0, \tag{3.40}$$

$$u(0,t) = \psi(t), \quad \psi(t) > 0, \quad \psi'(t) < 0,$$

причем начальные данные и краевые условия согласованы: $\varphi(0) = \psi(0)$. Поле характеристик для этой задачи представляет собой расходящийся веер кривых (рис. 3.11*а*). Из каждой внутренней точки области $\mathbf{\Omega}$ можно опустить характеристику, которая придет на границу $x = 0$ или $t = 0$. Следовательно, существует гладкое решение задачи, которое может быть найдено с помощью приближенных методов, описанных выше.

Ситуация изменится, если начальные данные или краевые условия будут разрывными или для них не будет выполнено условие согласования, например:

$$u(x, 0) = a = \text{const}, \quad u(0, t) = b = \text{const}. \tag{3.41}$$

Пусть $b < a$. Характеристиками являются параллельные прямые линии с наклонами $1/b$ и $1/a$ (рис. 3.11*б*), однако в области, ограниченной прямыми ($t = x/a$, $t = x/b$), характеристики отсутствуют. В этом случае пустую область можно заполнить расходящимся веером характеристик, которые определяют решение в виде *волны разрежения*

$$u(x,t) = \begin{cases} a, & x < at \\ \dfrac{x}{t}, & at \le x \le bt \\ b & x \ge bt \end{cases}. \tag{3.42}$$

Решение непрерывно, но на линиях сопряжения рвется производная (слабый разрыв).

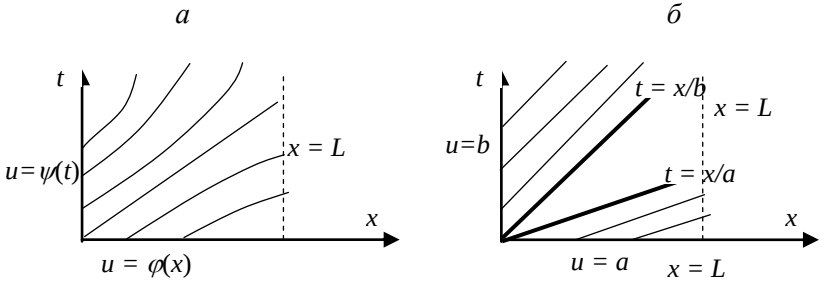


Рис. 3.11. Поле характеристик задачи (3.40) и (3.41) при $b < a$

Если в (3.41) предположить, что $b > a$, то картина характеристик кардинально перестроится. В некоторой подобласти $\mathbf{\Omega}_1$ характеристики, выпущенные с линий $x = 0$ и $t = 0$, пересекутся (рис. 3.12*а*). Это означает, что в данной подобласти существует два различных решения, принесенных с разных границ $\mathbf{\Omega}$. Пересечение характеристик одного семейства называется **градиентной катастрофой**. В последующие за градиентной катастрофой моменты времени непрерывное решение не существует.

Для корректного определения единственного решения разделим область пересечения характеристик на два сектора, в каждом из которых решение определяется единственным образом. С одной стороны от линии разрыва решения $x = Dt$ решение равно b, а с другой – a (рис. 3.12*б*). На плоскости (u, x) решение представляет собой *ударную волну*, т.е. кусочно-постоянную функцию с разрывом, двигающимся слева направо со скоростью D:

$$u(x,t) = \begin{cases} b, & x < Dt \\ a, & x > Dt \end{cases}. \qquad (3.43)$$

Как будет показано ниже, скорость разрыва D не произвольна, а выбирается из некоторых дополнительных соображений.

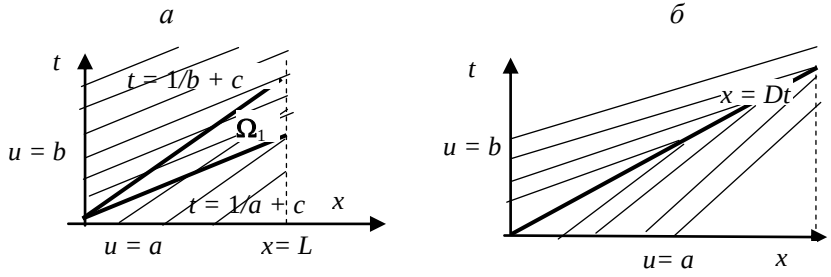


Рис. 3.12. Поле характеристик задачи (3.40) и (3.41) при $b > a$: пересечение характеристик (*а*) и линия разрыва (*б*)

Поскольку разрывные функции нельзя дифференцировать, то построенная функция не может быть решением исходного дифференциального уравнения. В этом случае говорят об **обобщенном решении**, удовлетворяющем более общему интегральному уравнению

$$\iint_S \left[\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2} \right) \right] dxdt = 0,$$

где S – произвольная замкнутая область на плоскости (x, t). Пусть dS – граница области S. По теореме Грина

$$\oint_{dS} P dx + Q dt = \iint_S \left[\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial t} \right] dxdt .$$

Положим $Q = \frac{u^2}{2}$, $P = -u$, тогда

$$\iint_S \left[\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2} \right) \right] dxdt = \oint_{dS} \frac{u^2}{2} dt - u dx = 0. \qquad (3.44)$$

Будем называть обобщенным решением (3.39) функции, удовлетворяющие интегральному уравнению (3.44). Для гладких функций уравнения (3.39) и (3.44) эквивалентны, однако разрывные функции могут быть только обобщенными решениями.

Уравнение (3.44) позволяет определить скорость движения разрыва D для задачи с условиями (3.41).

Выберем в качестве S прямоугольную область со сторонами h и τ, изображенную на рис. 3.13. Линия разрыва является диагональю прямоугольника, значит, $h = D\tau$. Найдем интеграл (3.44) по границе области S с учетом направления обхода от точки начала координат против часовой стрелки:

$$\oint_{dS} \frac{u^2}{2}dt - udx = -ah + \frac{a^2}{2}\tau + bh -$$

$$-\frac{b^2}{2}\tau = (b-a)h + \frac{a^2-b^2}{2}\tau = 0,$$

$$(b-a)\tau\left(D - \frac{a+b}{2}\right) = 0,$$

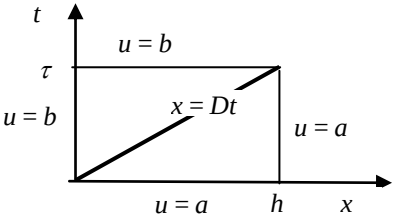


Рис. 3.13. Область интегрирования

поскольку $b \neq a$, то

$$D = \frac{a+b}{2},$$

т.е. скорость разрыва равна полусумме значений решения по разные стороны от разрыва.

Пусть $u > 0$ во все моменты времени. По аналогии с линейным уравнением построим для уравнения (3.34) противопотоковые схемы:

$$\frac{u_j^{n+1} - u_j^n}{\tau} + u_{j-1}^n \frac{u_j^n - u_{j-1}^n}{h} = 0, \qquad (3.45)$$

$$\frac{u_j^{n+1} - u_j^n}{\tau} + u_j^n \frac{u_j^n - u_{j-1}^n}{h} = 0, \qquad (3.46)$$

$$\frac{u_j^{n+1} - u_j^n}{\tau} + \frac{u_j^n + u_{j-1}^n}{2} \cdot \frac{u_j^n - u_{j-1}^n}{h} = 0. \qquad (3.47)$$

Посмотрим, будут ли указанные схемы правильно воспроизводить разрывное решение уравнения (3.39) с условиями (3.41) при $b = 1$, $a = 0$ (**первый пример Лакса**).

Точное решение представляет собой ступенчатую функцию (3.43), скорость движения разрыва $D = 0.5$. Однако в расчетах по

схеме (3.45) разрыв стоит на месте. При использовании (3.46) разрыв движется с большей скоростью: $D = 1$, и только схема (3.47) дает правильную скорость движения разрыва $D = 0.5$.

Причиной успеха схемы является то, что она аппроксимирует исходное уравнение (3.39), записанное в дивергентной, или консервативной форме:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0,$$

из которой следует определение обобщенного решения:

$$\frac{u_j^{n+1} - u_j^n}{\tau} + \frac{u_j^n + u_{j-1}^n}{2}\,\frac{u_j^n - u_{j-1}^n}{h} = \frac{u_j^{n+1} - u_j^n}{\tau} + \frac{\left(u_j^n\right)^2 - \left(u_{j-1}^n\right)^2}{2h} = 0.$$

Такие схемы называются **консервативными**, и только такие схемы гарантируют правильное воспроизведение скорости разрывных решений.

Пусть теперь $a = 1$, $b = 0$ (**второй пример Лакса**). Выше для этих условий было построено непрерывное решение (3.42) – «волна разрежения». Попробуем построить еще одно решение по формуле (3.43). В исходное поле характеристик (рис. 3.14*а*) введем линию разрыва $x = Dt$ и достроим характеристики, выходящие из этой линии (рис. 3.14*б*).

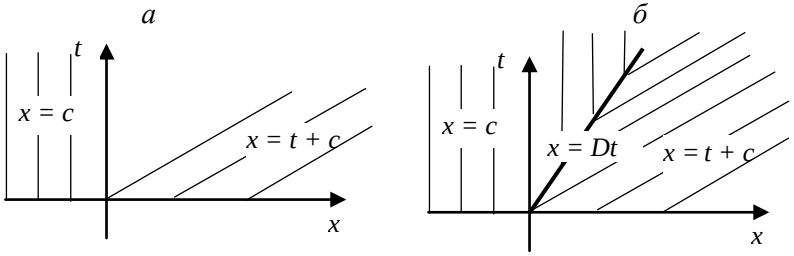


Рис. 3.14. Исходное (*а*) и достроенное (*б*) поле характеристик второго примера Лакса

Несмотря на то, что формально это решение построено по аналогии с ударной волной, оно является нефизичным. В отличие от ударной волны (рис. 3.12*б*), для которой характеристики

«входят» в линию разрыва и приносят на нее информацию о решении, в последнем случае характеристики «выходят» из разрыва и уносят информацию в окружающую среду. В газовой динамике существование таких решений запрещает второй закон термодинамики, предписывающий неубывание энтропии замкнутой системы. В связи с этой аналогией в вычислительной математике используют термин «энтропийное» решение. При построении разностных схем для уравнения (3.34) надо следить за тем, чтобы в процессе решения не появлялись нефизичные, т.е. «неэнтропийные» решения. Доказано, что решения, полученные с помощью монотонных разностных схем, являются энтропийными.

Второй способ построения правильных «энтропийных» решений – это использование «псевдовязкости», т.е. искусственного сглаживания решений уравнения (3.39) путем введения в него дополнительного слагаемого, например

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \varepsilon \frac{\partial^2 u}{\partial x^2}, \tag{3.48}$$

где ε – малый параметр, имеющий физический смысл коэффициента вязкости. При $\varepsilon \to 0$ уравнение (3.48) переходит в (3.39), следовательно, решение уравнения (3.39) может быть получено как предел решений (3.48).

Приведем еще один – **третий пример Лакса** для уравнения (3.39) с начальными данными:

$$\varphi_0 = \begin{cases} 1, & x \le 0 \\ 1 - x, & 0 < x < 1 \\ 0, & x \ge 1 \end{cases} \tag{3.49}$$

Точное решение задачи при $t < 1$ имеет вид:

$$u(x,t) = \begin{cases} 1, & x \le t \\ \dfrac{(1-x)}{1-t}, & t < x < 1, \\ 0, & x \ge 1 \end{cases} \qquad u(x,t) = \begin{cases} 1, & x \le 1 + t/2 \\ 0, & x \ge 1 + t/2 \end{cases}$$

При $0 < t < 1$ решение непрерывно. В момент времени $t = 1$ происходит градиентная катастрофа, т.е. производная решения

обращается в бесконечность, и при $t > 1$ решение является разрывным. Это решение моделирует процесс образования ударных волн и поэтому является хорошим тестом для уравнений газовой динамики.

В заключение приведем TVD-схему для уравнения (3.39), которую можно использовать для расчетов разрывных решений. Пусть $F = u^2/2$. С учетом направления поток раскладывается на две составляющие $F = F^- + F^+$,

где $F^+ = \begin{cases} u^2/2, u > 0 \\ 0, u < 0 \end{cases}$, $F^- = \begin{cases} u^2/2, u < 0 \\ 0, u > 0 \end{cases}$. Аппроксимация повышенного порядка строится по следующим формулам:

$$\frac{\partial F(u)}{\partial x} \approx \frac{F^+_{j+1/2} - F^+_{j-1/2} + F^-_{j+1/2} - F^-_{j-1/2}}{\Delta x},$$

$$F^-_{j+1/2} = F^-_{j+1} - \frac{1}{4}\left\{(1-k)\Delta^+\left(F^-_{j+1}\right) + (1+k)\Delta^-\left(F^-_{j+1}\right)\right\},$$

$$F^+_{j+1/2} = F^+_j + \frac{1}{4}\left\{(1-k)\Delta^-\left(F^+_j\right) + (1+k)\Delta^+\left(F^+_j\right)\right\},$$

$$\Delta^+\left(F_j\right) = F_{j+1} - F_j, \quad \Delta^-\left(F_j\right) = F_j - F_{j-1}.$$

Для сохранения монотонности решения в областях разрывов и локальных экстремумов для схем повышенного порядка аппроксимация снижается путем применения ограничителей. Один из возможных – это замена операторов Δ^+ и Δ^- операторами δ^+ и δ^- соответственно:

$$\delta^+ = \begin{cases} 0, & sign\Delta^+ sign\Delta^- \le 0 \\ \min\left(\left|\Delta^+\right|, \Theta\left|\Delta^-\right|\right) & sign\Delta^+ sign\Delta^- \ge 0 \end{cases},$$

$$\delta^- = \begin{cases} 0, & sign\Delta^+ sign\Delta^- \le 0 \\ \min\left(\left|\Delta^-\right|, \Theta\left|\Delta^+\right|\right) & sign\Delta^+ sign\Delta^- \ge 0 \end{cases},$$

k – параметр схемы, от которого зависит порядок аппроксимации: второй при $k = -1$, третий при $k = 1/3$, Θ лежит в пределах $1 \le \Theta \le \frac{3-k}{1-k}$.

11.8. Волновое уравнение

Типичным представителем уравнений гиперболического типа является так называемое волновое уравнение, описывающее распространение различных волн:

$$u_{tt} = g^2(x,t)u_{xx} + f(x,t). \tag{3.50}$$

Пусть требуется решить уравнение в области G: $x \in (0,1)$, $t \in (0,T)$.

Дополним уравнение начальными данными

$$u(x,0) = \sigma_1(x); \quad u_t(x,0) = \sigma_2(x) \tag{3.51}$$

и краевыми условиями

$$\begin{aligned} p_0 u(0,t) + p_1 u_x(0,t) &= A(t), \\ s_0 u(1,t) + s_1 u_x(1,t) &= B(t). \end{aligned} \tag{3.52}$$

Явная конечно-разностная схема

Для приближенного решения будем использовать конечно-разностный метод. Для этого введем в области G разностную сетку, в качестве которой используем совокупность точек пересечения прямых $x = ih$, $t = j\tau$, $i = 0, 1, ..., N$; $j = 0, 1, ..., M$, где h и τ – шаги сетки по пространственной и временной координатам. Если положить, что шаги h и τ связаны соотношением $\tau = rh$, $r = \text{const}$, то сетка будет зависеть только от одного параметра h.

Через u_i^j обозначим значение сеточной функции в точке (x_i, t^j). Аппроксимируем входящие в (3.50) – (3.52) производные конечно-разностными соотношениями второго порядка точности:

$$\frac{\partial^2 u}{\partial t^2}(x_i,t^j) \approx \frac{u(x_i,t^{j+1}) - 2u(x_i,t^j) + u(x_i,t^{j+1})}{\tau^2} = \frac{u_i^{j+1} - 2u_i^{j} + u_i^{j-1}}{\tau^2},$$

$$\frac{\partial^2 u}{\partial x^2}(x_i,t^j) \approx \frac{u(x_{i+1},t^j) - 2u(x_i,t^j) + u(x_{i-1},t^j)}{h^2} = \frac{u_{i+1}^{j} - 2u_i^{j} + u_{i-1}^{j}}{h^2}.$$

Подставив эти выражения в (3.50), получим явную разностную схему:

$$\frac{u_i^{j+1} - 2u_i^j + u_i^{j-1}}{\tau^2} = (g_i^j)^2 \frac{u_{i+1}^j - 2u_i^j + u_{i-1}^j}{h^2} + f_i^j. \qquad (3.53)$$

Схеме (3.53) отвечает шаблон типа «крест», изображенный на рис. 3.15. Он иллюстрирует тот факт, что для вычисления значения искомой функции на временном слое $j + 1$ необходимо знать значения этой функции на двух предыдущих слоях j и $j - 1$. Следовательно, для того, чтобы начать расчет, необходимо знать значения сеточной функции на первых двух временных слоях.

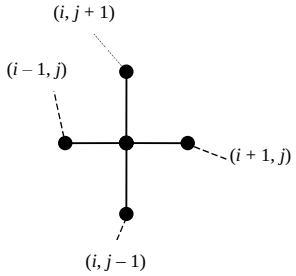


Рис. 3.15. Шаблон явной схемы для волнового уравнения

Решение на временном слое $t = t_0$ определено начальными данными (3.51): $u_i^0 = \sigma_1(x_i)$. Чтобы вычислить решение при $t = t_1$ воспользуемся формулой Тейлора, а также начальными данными (3.51) и уравнением (3.50):

$$u(t^1, x) = u(t^0, x) + \tau u_t(t^0, x) + \frac{\tau^2}{2} u_{tt}(t^0, x) + \frac{\tau^3}{6} u_{ttt}(t^0, x) + \ldots =$$

$$= \sigma_1(x) + \tau\sigma_2(x) + \frac{\tau^2}{2}\left(g^2(t^0, x) \frac{d^2\sigma_1(x)}{dx^2} + f(t^0, x) \right) + C\tau^3.$$

Для нахождения значений сеточной функции u_i^j во внутренних точках $x_i = ih$, $i = 1, \ldots, N - 1$, на временных слоях t_j, $j = 2, 3, \ldots, M$, используем разностную схему (3.53):

$$u_i^{j+1} = 2u_i^j - u_i^{j-1} + r^2 g^2 (u_{i+1}^j - 2u_i^j + u_{i-1}^j) + \tau^2 f_i^j,$$

$$i = 1, 2, \ldots, N - 1. \qquad (3.54)$$

Для определения искомой сеточной функции на линиях $x = x_0$, $x = x_N$ воспользуемся краевыми условиями. В случае первой краевой задачи ($p_1 = 0$, $s_1 = 0$) значения функции в гранич-

ных точках задаются точно: $u_0^j = A(t^j)$, $u_N^j = B(t^j)$. Если $p_1 \neq 0$, $s_1 \neq 0$ (вторая или третья краевая задача), производные в (3.52) необходимо заменить конечно-разностными соотношениями. Используем формулы первого порядка аппроксимации:

$$p_0 u_0^j + p_1 \frac{u_1^j - u_0^j}{h} = A^j, \quad s_0 u_N^j + s_1 \frac{u_N^j - u_{N-1}^j}{h} = B^j, \qquad \text{откуда}$$

$$u_0^j = \frac{p_1 u_1^j - A^j h}{p_1 - p_0 h}, \quad u_N^j = \frac{B^j h + s_1 u_{N-1}^j}{h s_0 + s_1}, \quad j = 2, 3, \ldots, M. \quad (3.55)$$

Можно показать, что схема (3.54) имеет второй порядок аппроксимации относительно h. Однако соотношения (3.55) имеют лишь первый порядок аппроксимации, что, несомненно, снижает общую точность полученного приближенного решения. Чтобы общий порядок аппроксимации задачи не понижался, для аппроксимации производных в граничных условиях (3.52) необходимо использовать соотношения второго порядка аппроксимации:

$$p_0 u_0^j + p_1 \frac{-3u_0^j + 4u_1^j - u_2^j}{2h} = A^j, \quad s_0 u_N^j + s_1 \frac{3u_N^j - 4u_{N-1}^j + u_{N-2}^j}{2h} = B^j,$$

откуда легко получить выражения для u_0^j, u_N^j, имеющие второй порядок аппроксимации.

Исследование устойчивости разностной схемы

Для того, чтобы решение задачи (3.54) сходилось к решению исходной задачи, требуется устойчивость этой схемы. Опишем один из методов исследования устойчивости. Рассмотрим задачу Коши:

$$\frac{\partial^2 u}{\partial t^2} - g^2 \frac{\partial^2 u}{\partial x^2} = 0, \quad -\infty < x < \infty, \quad 0 < t < T, \ g = \text{const},$$
$$u(x,0) = \psi_1(x), \quad \frac{\partial u(x,0)}{\partial t} = \psi_2(x), \quad -\infty < x < \infty, \qquad (3.56)$$

которую аппроксимируем разностной схемой

$$\frac{u_m^{j+1}-2u_m^j+u_m^{j-1}}{\tau^2}-g^2\frac{u_{m+1}^j-2u_m^j-u_{m-1}^j}{h^2}=0,$$

$$j=1,2,\ldots,M-1, \tag{3.57}$$

$$u_m^0=\psi_1(x_m),\quad \frac{u_m^1-u_m^0}{\tau}=\psi_2(x_m),\quad m=0,\pm1,\ldots$$

Для устойчивости разностной схемы относительно возмущения начальных данных необходимо, чтобы решение задачи (3.57) удовлетворяло условию

$$\max_m\left|u_m^j\right|\le C\cdot\max_m\left|u_m^0\right|,\ \ j=0,1,\ldots,M, \tag{3.58}$$

при произвольной ограниченной функции $\psi_1(x_m)$, в частности, для $\psi_1=e^{i\alpha m}$, где α – вещественный параметр. Тогда решение задачи (3.56) можно искать в виде

$$u_m^j=\lambda^j e^{i\alpha m}, \tag{3.59}$$

где $\lambda=\lambda(\alpha)$. Условие (3.59) выполняется, если числа $\lambda(\alpha)$ лежат внутри круга единичного радиуса, т.е.

$$\left|\lambda(\alpha)\right|\le 1. \tag{3.60}$$

Неравенство (3.60) выражает необходимое условие устойчивости Неймана. Подставив (3.59) в (3.57), для определения $\lambda(\alpha)$ получим уравнение

$$\lambda^2-2\left(1-2r^2g^2\sin^2\frac{\alpha}{2}\right)\lambda+1=0\,. \tag{3.61}$$

По теореме Виета произведение корней этого уравнения равно 1, т.е. для выполнения условия (3.60) требуется, чтобы корни $\lambda_{1,2}$ уравнения (3.61) были комплексно сопряженными и лежали на единичной окружности. Для этого, в свою очередь, необходимо, чтобы дискриминант $D(\alpha)$ уравнения (3.51) был отрицателен:

$$D(\alpha)\equiv 4r^2g^2\sin^2\frac{\alpha}{2}\left(r^2g^2\sin^2\frac{\alpha}{2}-1\right)<0.$$

Данное неравенство выполняется при всех α, если $gr\le 1$. Следовательно, условием устойчивости схемы (3.57) будет

$$\tau \le \frac{h}{g}. \qquad (3.62)$$

Пусть теперь $g = g(x, t) \neq \text{const}$. В этом случае применяется принцип «замороженных коэффициентов», в соответствии с которым необходимое условие устойчивости Неймана можно записать в виде

$$\tau \le \frac{h}{g_*}, \quad g_* = \max_{x,t} g(x,t). \qquad (3.63)$$

В заключение отметим, что вопрос влияния граничных условий на устойчивость разностной схемы здесь не рассматривается.

Неявная разностная схема

При построении схемы (3.53) производная u_{xx} была заменена на конечную разность на временном слое $t_j = j\tau$. Если же использовать значения с временного слоя t_{j+1}, то получим схему

$$\frac{u_i^{j+1} - 2u_i^j + u_i^{j-1}}{\tau^2} = g_i^{j+1^2} \frac{u_{i+1}^{j+1} - 2u_i^{j+1} + u_{i-1}^{j+1}}{h^2} + \tau^2 f_i^{j+1}, \quad (3.64)$$

которой соответствует шаблон, изображенный на рис. 3.16.

Из уравнения (3.64) невозможно явно выразить u_i^{j+1} через значения функции u с предыдущих слоев по времени (j и $j-1$), поскольку в (3.61) наряду с u_i^{j+1} входят неизвестные u_{i+1}^{j+1} и u_{i-1}^{j+1}. Поэтому данная схема называется неявной.

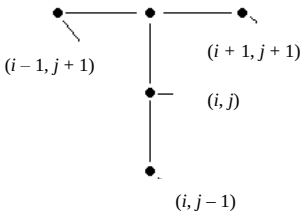


Рис. 3.16. Шаблон неявной схемы для волнового уравнения

Анализ устойчивости показывает, что неявная схема безусловно устойчива, т.е. обеспечивает сходимость разностной задачи к решению соответствующей дифференциальной при любом отношении τ/h. Решение на первых двух временных слоях определяется из начальных данных так же, как это сделано для явной схемы. Обозначив $\gamma = g_i^{j+1^2} r^2$, перепишем (3.64) в виде

$$\gamma u_{i+1}^{j+1} - (1+2\gamma)u_i^{j+1} + \gamma u_{i-1}^{j+1} = -2u_i^j + u_i^{j-1} + \tau^2 f_i^{j+1}, \quad (3.65)$$
$$i = 1, 2, \ldots, N-1.$$

Дополнив (3.65) формулами, аппроксимирующими краевые условия, получим СЛАУ с трехдиагональной матрицей, которая решается с помощью метода прогонки (см. Раздел 1).

Тема 12. Современные пакеты прикладных программ для решения задач строительства*

В настоящее время невозможно представить решение сложных инженерных задач без помощи методов численного моделирования. В каждой из прикладных областей существует иерархия подходов для решения различных проблем, предлагаемых практикой. На их основе разрабатываются как коммерческие, так и свободно распространяемые программы. Так, в строительной области на сегодняшний день завоевывают популярность пакеты ABAQUS, SCAD, LS-DYNA, ЛИРА, COSMOS, ANSYS, NEiNASTRAN и др.

Основными схемами, используемыми для расчета работы конструкций в этих программных системах, или САПР (системы автоматизированного проектирования; англ. CAE – Computer Aided Engineering), являются конечно-элементные методы (МКЭ). Их описание приведено в Разделе 2.

Напомним, что метод конечных элементов предназначен для решения дифференциальных уравнений. Конечно-элементный анализ позволяет моделировать нагрузку на конструкцию и определять ее реакцию на эту нагрузку. Конструкция моделируется набором дискретных блоков, которые называются элементами. Число этих блоков конечно, в то время как физическая постановка содержит бесконечное количество элементов. Поэтому естественно, что в результате мы получаем приближенное решение, и очень важно, насколько точно мы определяем искомую реакцию конструкции на нагрузку.

* Авторы выражают благодарность за помощь в подготовке Темы 12 аспирантке кафедры прикладной математики НГАСУ (Сибстрин) С.А. Вальгер.

С математической точки зрения суть метода заключается в разбиении области решения на конечное число подобластей (элементов) и построении внутри них аппроксимирующей функции некоторого порядка (базисной функции, функции формы). Значения функции на границах элементов являются решениями исходной задачи. В результате получается система линейных алгебраических уравнений с разреженной матрицей, которая затем решается одним из известных способов.

В данном пособии в качестве основы для получения представлений о современных методах численного моделирования и об использовании программных пакетов расчета задач строительства выбран пакет ANSYS. Данный коммерческий продукт, помимо расчетных возможностей в области строительства и механики твердого тела, включает также линейку программ, предназначенных для решения задач газодинамики, электромагнетизма, теплообмена и микроэлектроники. В каждом из них используется свой набор численных методов. Мы сосредоточимся на задачах строительства и МКЭ. Поскольку интерфейс программы не русифицирован, для удобства читателя названия большинства терминов и объектов будут дублироваться на английском языке.

12.1. Базовые сведения

Современные версии ANSYS предлагают пользователю общую платформу для работы с различными приложениями, имеющимися в пакете. Эта платформа называется ANSYS Workbench; она является средой для разработки инженерных проектов, которые связываются воедино даже в случае междисциплинарных расчетов, включающих разные аспекты одной проблемы (прочность, теплообмен, аэродинамика). Удобный и понятный для инженеров интерфейс платформы позволяет обновлять данные на уровне проекта, проводить параметрические и оптимизационные расчеты, обеспечивает связь с САПР системами. Таким образом, многие приложения, ранее являвшиеся независимыми, были интегрированы в среду Workbench. Мо-

дуль для расчета прочности и устойчивости конструкций носит название ANSYS Mechanical и доступен как в виде самостоятельного приложения, так и в составе Workbench. Выполнение сложных прочностных и многодисциплинарных расчетов доступно пока лишь в традиционном интерфейсе ANSYS и носит название Mechanical APDL. APDL (ANSYS Parametric Design Language) – это язык, созданный для написания программ, скриптов, пользовательских окон данного решателя.

В среде Mechanical возможно проведение следующих типов расчетов конструкций на прочность:

1. Статический (или переходный к динамическому) анализ (Static Analysis) определяет деформации, напряжения и т.п. при статических нагрузках. Сюда включаются как линейные, так и нелинейные постановки. Нелинейность может включать явления пластичности, больших деформаций, упрочнения, ползучести, поверхности контакта и т.п.

2. Динамический анализ (Transient Dynamic Analysis) используется для определения отклика конструкции на произвольные нагрузки, изменяющиеся во времени. Все нелинейности, рассматриваемые в статическом анализе, также могут быть учтены.

3. Модальный анализ (Modal Analysis) рассчитывает естественные частоты и формы колебаний конструкции. Имеет дополнение в виде спектрального анализа (Spectrum Analysis).

4. Гармонический анализ (Harmonic Analysis) призван определять отклик конструкции при воздействии нагрузок, зависящих от времени по гармоническому закону.

5. Расчет потери устойчивости (Buckling Analysis).

6. Оптимизация формы.

7. Расчет контактных задач: скольжение с разделением, трение, уплотнение.

В дополнение к прочностным расчетам имеется возможность провести (отдельно или совместно, т.е. с учетом взаимовлияния) тепловой расчет (Thermal) – стационарный и неста-

ционарный, а также решить задачи проводимости, конвекции, излучения, радиации, учесть фазовые переходы.

Также имеются возможности междисциплинарных исследований (мультифизичные модели), включая:

1. Акустический/прочностной анализ.
2. Тепловой/прочностной анализ.
3. Тепловой/электрический анализ*.
4. Тепловой/электрический/прочностной анализ*.
5. Пьезоэлектрический анализ*.
6. Пьезорезистивный анализ*.
7. Взаимодействие конструкции с текучей средой*.
8. Универсальный модуль расчета связанных полей*.
9. Моделирование магнитных полей конструкции (Magnetostatic).

Примечание. Расчеты, отмеченные звездочкой, требуют приобретения дополнительных продуктов пакета.

Для решения системы уравнений строительной механики используется метод конечных элементов, упомянутый выше. Таким образом, решение задачи включает несколько основных этапов:

1. Физическая постановка задачи.
2. Выбор математической модели (типа решателя).
3. Построение геометрии задачи.
4. Построение сетки (разбиение на конечные элементы).
5. Задание граничных условий.
6. Проведение расчета.
7. Анализ результатов.

Физическая постановка задачи определяется пользователем на основе исходных данных. В зависимости от постановки выбирается один из перечисленных выше типов расчета (анализа). Решается, будет ли объектом исследования некоторая деталь либо это сборка, являются ли элементы оболочками (Surface) или телами (Solid).

Рассмотрим последующие этапы решения задачи в ANSYS более подробно.

12.2. Основной интерфейс и запуск Workbench

Запуск среды Workbench осуществляется через последовательность меню Windows ПУСК → Программы → ANSYS → Workbench. При этом открывается окно Workbench, разделенное на несколько полей: поле инструментов (слева; разделы можно свернуть/развернуть, нажав на значок «–»/«+»), поле для создания проекта (справа вверху) и поле вывода записей об осуществленных операциях (внизу). Последнее поле выключается/включается нажатием нижней правой кнопки Hide Progress/Show Progress.

Раздел «Инструменты» содержит четыре группы вкладок:

1. Analysis Systems: готовые шаблоны основных задач.
2. Component Systems: различные приложения ANSYS.
3. Custom Systems: предопределенные расчетные приложения для сопряженных расчетов (FSI, thermal-stress, др.). Также можно создавать свои комбинации задач.
4. Design Exploration: параметрическая оптимизация решений.

Имеется возможность объединять разные компоненты проекта (рис. 3.17).

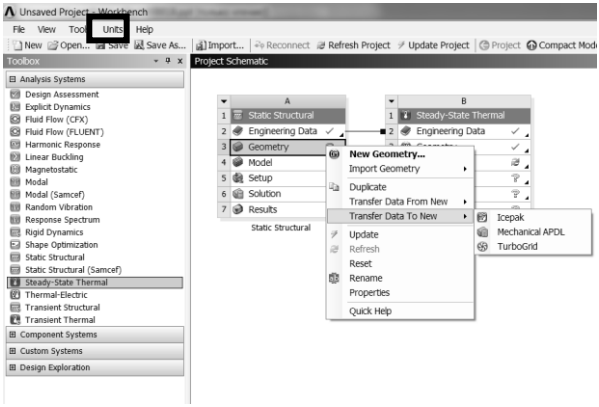


Рис. 3.17. Пример поля проектов

Выбор системы единиц измерения производится в меню Units (рис. 3.17, выделено в жирную рамку). Есть возможность создания собственной системы путем комбинации существующих.

После выбора типа анализа необходимо задать свойства и модели материалов. Для этого можно воспользоваться приложением Engineering Data, которое является составляющей любого проекта независимо от типа выбранного анализа. Это приложение можно также вызвать отдельно. При запуске приложения открывается окно, содержащее несколько полей, включая панель инструментов, свойства материала, отображение редактируемых данных в табличном и графическом виде. В ANSYS имеется библиотека, содержащая основные свойства большинства используемых в строительстве материалов. При необходимости свойства материала можно изменить и сохранить в библиотеке. Для использования материала в текущем расчете (проекте) его необходимо добавить из библиотеки.

12.3. Построение геометрии

В пакете ANSYS предусмотрена возможность импортирования существующей геометрии из другой САПР. Поддерживаются многие популярные форматы, такие как CATIA, Parasolid, Autodesk, SolidWorks и т.д. Также можно создать геометрию при помощи встроенного построителя DesignModeller (DM).

DM работает в двух основных режимах: создания двумерных эскизов (Sketching) и трехмерных моделей (Modelling). В режиме эскизов имеется пять различных панелей инструментов для создания эскизов путем добавления или удаления двумерных граней:

1. Панель рисования линий, прямоугольников, кривых – Draw Toolbox.

2. Панель модификации для выравнивания, обрезания и склеивания объектов – Modify Toolbox.

3. Панель размеров определяет размеры длин, расстояний, диаметров и углов – Dimensions Toolbox.

4. Панель связей накладывает связи в виде касательных линий, симметрии и соосности – Constraints Toolbox.

5. На установочной панели указываются параметры сетки и ее размеры – Settings Toolbox.

На основе двумерных эскизов можно создавать трехмерные модели. Режим создания моделей позволяет путем штампования или вращения профилей из эскизов получать трехмерные объекты (модели). Для создания новой геометрии необходимо начать с рисования эскиза (Sketch).

Все возможности и инструменты DM доступны также через выпадающие меню в панели основного меню. Перечислим основные разделы и некоторые функции данного меню (более подробную информацию можно почерпнуть в разделе помощи):

1. Файловое меню File.
2. Меню конструирования Create.
3. Концептное меню Concept.
4. Меню инструментов Tools.
5. Меню просмотра View.
6. Помощь Help.

После выбора необходимых материалов и создания либо импорта геометрии модели следующим этапом является построение расчетной сетки. При запуске приложения моделирования (Model) в окне проекта открывается новое окно, в левой части которого расположено дерево проекта, которое позволяет отследить все шаги, необходимые для проведения расчета (рис. 3.18): геометрию (Geometry), систему координат (Coordinate System), сетку (Mesh), расчет (например, Static Structural). В геометрии может присутствовать до трех типов тел: твердотельные (Solid), поверхностные или оболочки (Surface Body) и линейные или балочно-стержневые (Line Body). Твердотельные тела могут быть двух- (2D) или трехмерными (3D).

Дискретизация трехмерных тел производится при помощи тетраэдрических и гексагональных элементов. Двумерные тела дискретизируются треугольными и четырехугольными конечными элементами. Каждый узел имеет три степени свободы (перемещения) (Degree Of Freedom – DOF) для прочности. В случае теплового расчета узел имеет одну степень свободы – температуру.

Поверхности (оболочки) дискретизируются линейными оболочечными элементами, а балки и стержни – линейными балочными элементами. В обоих случаях в каждом узле имеется шесть степеней свободы.

При создании геометрии в DesignModeller тело может быть одной деталью, а может являться сборкой деталей, имеющих поверхность контакта. Для каждой детали сборки можно назначить свой материал во вкладке Свойства (Preferences). Каждая деталь может иметь свою сетку, узлы сеток на поверхности контакта не обязаны совпадать.

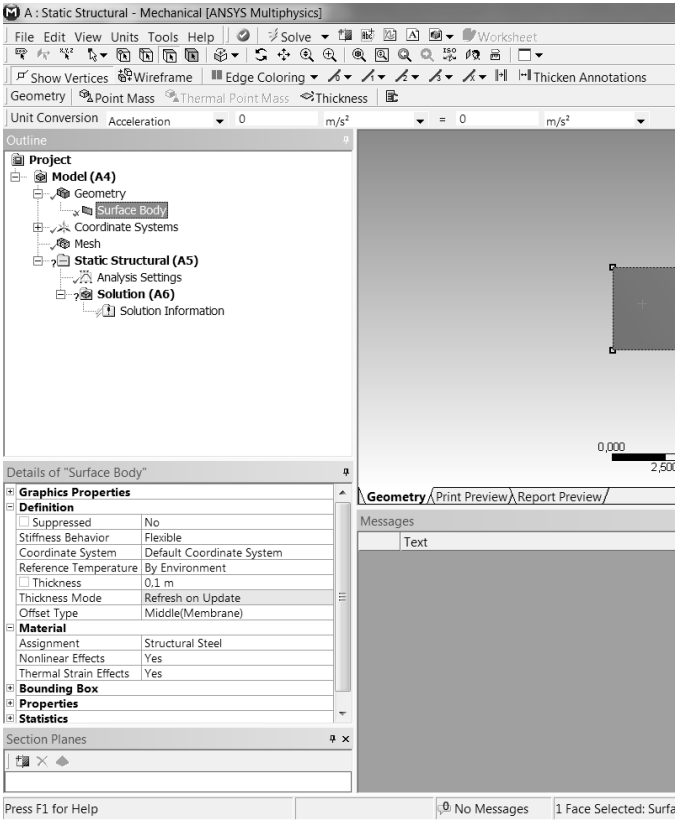


Рис. 3.18. Окно расчета

Генерация сетки происходит автоматически при запуске задачи на расчет. Тем не менее, сгенерировать сетку можно и до расчета при помощи команды Generate Mesh, задав свои пользовательские параметры сетки. Основные параметры доступны во вкладке Mesh Defaults. Здесь задается тип решаемой задачи (в нашем случае – Mechanical), а также параметр Relevance, который определяет степень измельчения сетки. Отрицательные значения этого параметра приводят к созданию более грубой сетки (–100 – самая грубая сетка, с наименьшим возможным количеством элементов), положительные – к созданию более мелкой сетки. В расширенных настройках также можно задать размер элемента и наличие либо отсутствие срединной точки элемента и т.п. Во вкладке Mesh Control на панели инструментов можно выбрать тип используемой дискретизации, задать средний размер ребра элемента (Sizing), измельчить исходную сетку (Refinement), упорядочить разбиение (Mapped Face Meshing), создать дополнительные более мелкие слои около поверхностей (Inflation). В случае возникновения ошибок в геометрии или настройках сетки появится сообщение об ошибке и сетка не будет создана. Ошибки могут возникать в следующих случаях:

- ✓ некорректно задан размер на поверхностях, что вызывает появление вырожденных элементов;
- ✓ сложная CAD-геометрия – мелкие фаски, радиусы, двойные поверхности и др.;
- ✓ жесткие условия проверки качества элементов (включена опция Aggressive).

Возможные пути устранения ошибок генерации:

- ✓ задание корректных размеров элементов на геометрии;
- ✓ задание возможно меньших размеров элементов, что позволит сеточному генератору получать более качественные элементы;
- ✓ избавление от мелких линий, двойных поверхностей и других второстепенных геометрических объектов в CAD-системе;
- ✓ применение виртуальной топологии для формирования укрупнения поверхностей.

12.4. Нагрузки и закрепления

После создания расчетной сетки можно приступать к приложению нагрузок (Loads) и закреплений (Supports). Закрепления ограничивают перемещение узлов и различаются своим назначением. Нагрузки бывают инерционные (действуют внутри рассчитываемой конструкции), прочностные (усилия и моменты, действующие на конструкцию) и тепловые. Инерционные нагрузки можно задать при помощи:

- ✓ ускорения (Acceleration): определяет прочностные изменения при воздействии постоянного линейного ускорения;
- ✓ стандартной земной гравитации (Standard Earth Gravity);
- ✓ скорости вращения (Rotational Velocity): отвечает за прочностные изменения части конструкции, вращающейся с постоянной скоростью.

Прочностные нагрузки делятся на:

- ✓ давление (Pressure): такой тип нагрузки задает величину постоянного либо переменного давления в одном направлении на поверхность детали;
- ✓ гидростатическое давление (Hydrostatic Pressure): моделирует давление, создаваемое столбом жидкости;
- ✓ усилие (Force): распределяет усилие по поверхности, на грани или в узле;
- ✓ удаленную нагрузку (Remote Force): задается на поверхности или грани аналогично усилию, но с добавлением момента; таким образом, можно прикладывать рычаговые усилия без построения дополнительных деталей;
- ✓ распределенную переменную нагрузку по поверхности цилиндра (Bearing Load);
- ✓ приложение осевой нагрузки предварительного напряжения на цилиндрическую поверхность (Bolt Pretension): доступно только в трехмерном расчете;
- ✓ приложение момента (Moment);
- ✓ распределение давления вдоль линии (Line Pressure);
- ✓ вибрационные нагрузки (PSD Base Excitation, RS Base Excitation);

- ✓ нагрузки соединений (Joint Load);
- ✓ задание температурной нагрузки (Thermal Condition).

Тепловые нагрузки:

- ✓ Температура задается на поверхности, грани или в узле (Temperature).
- ✓ Моделирование конвективного теплообмена на поверхности контакта с жидкостью или газом (Convection).
- ✓ Моделирование радиационного теплообмена (Radiation).
- ✓ Задание теплового потока на поверхности, грани или в узле (Heat Flow).
- ✓ Отмена всех тепловых нагрузок (Perfectly Insulated).
- ✓ Генерация равномерного потока тепла внутри тела (Internal Heat Generation).

Также есть возможность задания различных разновидностей электрических, магнитных, взрывных и импортированных из других типов расчетов нагрузок.

Закрепления, в свою очередь, делятся на:

- ✓ жесткие: ограничиваются все степени свободы (Fixed);
- ✓ заданные перемещения (Displacement);
- ✓ закрепление по нормали к поверхности (Frictionless);
- ✓ закрепление в направлении сжимающего усилия (Compression Only Support);
- ✓ цилиндрические (Cylindrical);
- ✓ ограничение линейных перемещений с разрешенными моментами (Simply Supported);
- ✓ ограничение моментов с разрешенными линейными перемещениями (Fixed Rotation);
- ✓ упругие, позволяющие плоскостям и ребрам упруго деформироваться (по закону упругой пружины) (Elastic).

Нагрузки и закрепления можно разложить по осям активной системы координат, локальной или глобальной, и задать значения каждой из компонент.

После задания нагрузок можно переходить к расчету. Это осуществляется нажатием кнопки Solve. Задание пользовательских параметров расчета проводится в меню Analysis Settings.

Как известно, в линейном статистическом анализе перемещения δ рассчитываются из уравнения

$$[K]\{\delta\} = [F],$$

где $[K]$ – матрица жесткости, константа, $[F]$ – приложенные нагрузки. Предполагается, что материал ведет себя по линейно-упругому закону. В результате дискретизации конструкции конечными элементами мы получаем систему линейных алгебраических уравнений размерности, зависящей от количества элементов. Для решения этой СЛАУ применяется либо точный (прямой) метод (Sparse Solver в ANSYS), либо итерационный решатель разреженных матриц (PCG Solver в ANSYS). Тип решателя можно выбрать в настройках Analysis Settings, подменю Solver type.

Для анализа полученных данных необходимо в ветке Solution выбрать требуемые для вывода результаты. При этом на панели инструментов появляется набор доступных для расчета параметров: деформации, напряжения, энергия. Добавление параметров в ветку Solution позволяет получить эти величины в результате прочностного расчета и отобразить их в удобном виде при помощи возможностей обработки результатов в ANSYS (раздел Postprocessing).

12.5. Просмотр и анализ результатов расчетов

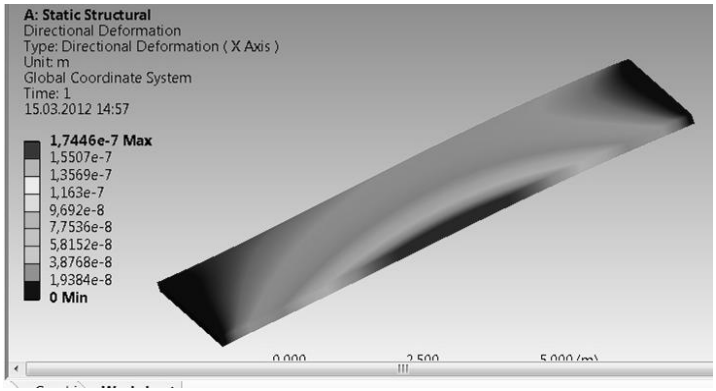
В рамках расчета в пакете ANSYS для просмотра доступны следующие численные результаты: деформации (полные (Total) и покомпонентные (Directional)); компоненты, главные напряжения (Principal) или инварианты напряжений и деформаций; результаты контактного моделирования; реакции в опорах. После расчета все необходимые данные (Evaluate Results) можно визуализировать в соответствующем окне. Отображение результатов можно осуществлять в контурном или векторном виде (рис. 3.19).

Управление параметрами контуров, масштабом отображения деформаций, шагом вывода параметров, отображением сетки или недеформированного тела, указанием точек максимального и минимального нагружения конструкции производится

при помощи панели инструментов. Доступно к выводу изменение параметров не только по поверхности, но и на гранях и в узлах. Окно Animation позволяет наблюдать расчет в динамике.

Дополнительно к стандартным могут быть использованы пользовательские форматы вывода результатов, куда включаются формулы, комбинирующие разные параметры расчета (User Defined Result).

а



б

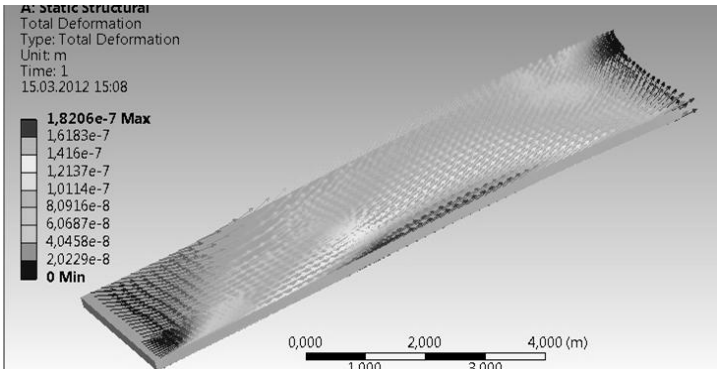


Рис. 3.19. Представление покомпонентных (*а*) и полных (*б*) деформаций в контурном (*а*) и векторном (*б*) виде

Данные расчета можно экспортировать в текстовом виде для обработки или использования в дальнейших расчетах, а также в процессе последующего анализа.

Широкий набор инструментов для анализа результатов позволяет оценить запасы прочности по различным теориям (Stress Tool). Можно также оценить прочностной расчет на ошибку (Error Estimation) и выявить области модели, где ошибка максимальная. В данном случае рекомендуется применить измельчение сетки в проблемных местах с целью достижения нужного результата и снижения погрешности. В этой ситуации также может оказаться полезным инструмент, отслеживающий сходимость результатов (Convergence). В зависимости от заданной степени сходимости будет автоматически выполнено несколько итераций с измельчением расчетной сетки до достижения заданной погрешности.

ЗАКЛЮЧЕНИЕ

В пособии приведены сведения из линейной алгебры, функционального анализа, дифференциальных уравнений, уравнений в частных производных и численных методов.

Авторы надеются, что предложенный материал поможет начинающим исследователям получить представление о математических моделях, возникающих в процессе научной и проектной деятельности; научит подбирать и модифицировать методы прикладной математики для решения поставленной задачи, а также анализировать полученное решение.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Вержбицкий В. М. Основы численных методов : учебник для вузов / В. М. Вержбицкий. – М. : Высшая школа, 2009. – 840 с.
2. Бахвалов Н. С. Численные методы / Н. С. Бахвалов, Н. П. Жидков, Г. М. Кобельков. – М. : Бином. Лаборатория знаний, 2008. – 640 с.
3. Бедарев И. А. Методы вычислений : учеб. пособие / И. А. Бедарев, Ю. В. Кратова, Н. Н. Федорова. – Новосибирск : НГАСУ (Сибстрин), 2009. – 112 с.
4. Деммель Дж. Вычислительная линейная алгебра / Дж. Деммель ; пер. с англ. Х. Д. Икрамова. – М. : Мир, 2001. – 430 с.
5. Стренг Г. Линейная алгебра и ее применения / Г. Стренг. – М. : Мир, 1980. – 454 с.
6. Беклемишев Д. В. Курс аналитической геометрии и линейной алгебры / Д. В. Беклемишев. – М. : ФИЗМАТЛИТ, 2005. – 304 с.
7. Канатников А. Н. Линейная алгебра : учебник для вузов / А. Н. Канатников, А. П. Крищенко. – М. : Изд-во МГТУ им. Н.Э. Баумана, 2002. – 336 с.
8. Математика и САПР : в 2 кн. / П. Жермен-Лакур, П. Л. Жорж, Ф. Пистр, П. Безье. – М. : Мир, 1989. – Кн. 2. – 264 с.
9. Сабоннадьер Ж.-К. Метод конечных элементов и САПР / Ж.-К. Сабоннадьер, Ж.-Л. Кулон. – М. : Мир, 1989. – 190 с.
10. Поршнев С. В. Численные методы на базе Mathcad : учеб. пособие / С. В. Поршнев, И. В. Беленкова. – СПб. : БХВ-Петербург, 2005. – 456 с.
11. Портал компании ANSYS [Электронный ресурс]. – Режим доступа: http://ansys.com/ (дата обращения: 15.06.2012).