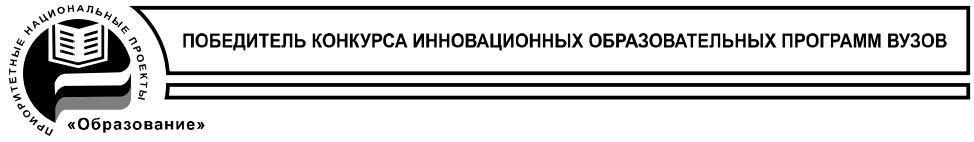
МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ



С.А. Чивилихин

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ В ТЕХНОЛОГИЯХ ПРОГРАММИРОВАНИЯ

ЭЛЕМЕНТЫ ТЕОРИИ И ПРАКТИКУМ

Санкт-Петербург

2008

С.А.Чивилихин. Вычислительные методы в технологиях программирования. Элементы теории и практикум, – СПб: СПбГУИТМО, 2008. –108с.

В пособии рассматриваются методы численного решения нелинейных уравнений, а также систем линейных и нелинейных уравнений, методы интерполяции и аппроксимации функций, методы интегрирования.

Учебное пособие предназначено для студентов СПбГУ ИТМО специальностей NN 2006006802, 010500. Рекомендовано к печати Ученым Советом факультета фотоники и оптоинформатики, протокол N6 от 24 июня 2008 г.

В 2007 году СПбГУ ИТМО стал победителем конкурса инновационных образовательных программ вузов России на 2007–2008 годы. Реализация инновационной образовательной программы «Инновационная система подготовки специалистов нового поколения в области информационных и оптических технологий» позволит выйти на качественно новый уровень подготовки выпускников и удовлетворить возрастающий спрос на специалистов в информационной, оптической и других высокотехнологичных отраслях экономики.

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Важной частью технологии программирования является выбор численного метода, адекватного решаемой задаче. Большинство используемых численных методов являются приближенными. Поэтому, выбрав численный метод, необходимо исследовать его точность, сходимость и устойчивость применительно к решаемой задаче. При численной реализации итерационных алгоритмов, критерий окончания расчетов выбирается на основании апостериорных оценок. Настоящая книга посвящена рассмотрению указанных вопросов. Кроме того, достаточно высокий уровень используемых математических методов способствует повышению общей математической культуры читателей.

Глава 1. Решение уравнений с одной переменной

Наиболее исследованными являются алгебраические уравнения. Известно, что уравнение порядка n имеет n корней с учетом их кратности. Однако общее аналитическое решение алгебраических уравнений существует лишь для уравнений не выше четвертого порядка. Решение уравнений первой и второй степени изучаются в школьном курсе математики. Общее решение уравнения третьей степени довольно громоздко. Существует также общее решение уравнения четвертой степеней, однако оно сложно и неудобно для практического применения. Теория Галуа утверждает, что для уравнений выше четвертой степени не существует общего решения в радикалах. Существуют аналитические решения некоторых типов неалгебраических уравнений, например, тригонометрических уравнений.

Потребность же в решении уравнений весьма велика. В этих условиях большое значение приобретают универсальные вычислительные алгоритмы. В этой главе будут рассмотрены четыре таких алгоритма.

1.1 Метод половинного деления

Рассмотрим уравнение

$$f(x) = 0. \tag{1}$$

В курсе анализа доказывается теорема о существовании корня уравнения (1) для непрерывной функции: Если функция $f(x)$ непрерывна на отрезке $[a,b]$ и принимает на его концах значения разных знаков, то на этом отрезке существует по крайней мере один корень уравнения (1).

Предположим для определенности, что функция $f(x)$ принимает на левом конце отрезка $[a,b]$ отрицательное значение, на правом - положительное:

$$f(a) < 0, \quad f(b) > 0. \tag{2}$$

Тогда, согласно приведенной теореме, на отрезке $[a,b]$ существует корень с функции $f(x)$. Возьмем на отрезке $[a,b]$ среднюю точку $\xi = (b+a)/2$ и вычислим в ней значение функции $f(\xi)$. Если $f(\xi) = 0$, то ξ является искомым корнем. При $f(\xi) > 0$ в качестве нового отрезка $[a_1,b_1]$ выберем отрезок $[a,\xi]$, а при $f(\xi) < 0$ в качестве нового отрезка $[a_1,b_1]$ выберем отрезок $[\xi,b]$. Новый отрезок $[a_1,b_1]$ также содержит корень c, но имеет вдвое меньшую длину. Повторяя эту процедуру n раз, мы получаем отрезок

$[a_n, b_n]$, содержащий корень c и имеющий длину $h_n = \frac{h_0}{2^n}$, где h_0 - длина исходного отрезка $[a,b]$. Длина отрезка h_n представляет собой точность ε, с которой мы знаем положение корня на n-ом шаге процесса половинного деления. Тогда количество шагов, необходимых для определения корня с точностью ε, можно рассчитать как

$$n > log_2\left(\frac{h_0}{\varepsilon}\right). \tag{3}$$

Метод половинного деления требует предварительного отделения корней. Прежде, чем применять этот метод, нам нужно найти интервал $[a,b]$, содержащий один корень уравнения (1) для функции $f(x)$. Однако, если функция $f(x)$ не только непрерывна, но и дифференцируема, то дополнительное ее исследование с помощью производной может во многих случаях решить и этот вопрос. Например, при знакоопределенной производной, функция $f(x)$ является монотонной на отрезке $[a,b]$, поэтому корень у нее может быть только один.

1.2 Метод простых итераций

Приведем уравнение (1) к эквивалентному виду

$$x = \varphi(x). \tag{4}$$

Возьмем произвольную точку x_0 из области определения функции $\varphi(x)$ и будем строить последовательность чисел $\{x_n\}$, определенных с помощью рекуррентной формулы

$$x_{n+1} = \varphi(x_n). \tag{5}$$

Последовательность $\{x_n\}$ называется итерационной. На рис.1 (a - d) представлены примеры применения этого алгоритма. Из рассмотренных примеров видно, что, если абсолютное значение производной $|\varphi'| < 1$, то итерационный процесс сходится к корню c уравнения (4). При $|\varphi'| > 1$ итерационный процесс расходится. Сходимость является монотонной при $\varphi' > 0$ и немонотонной при $\varphi' < 0$.

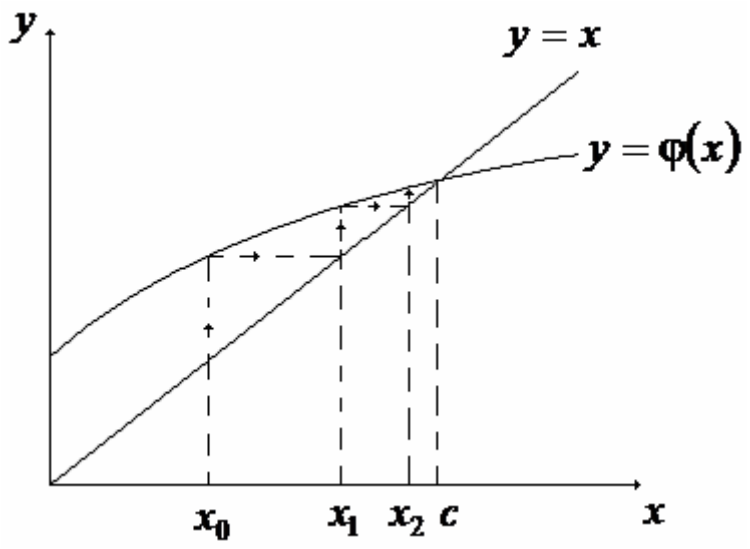


Рис. 1a. Сходящийся итерационный процесс при $0 < \varphi' < 1$

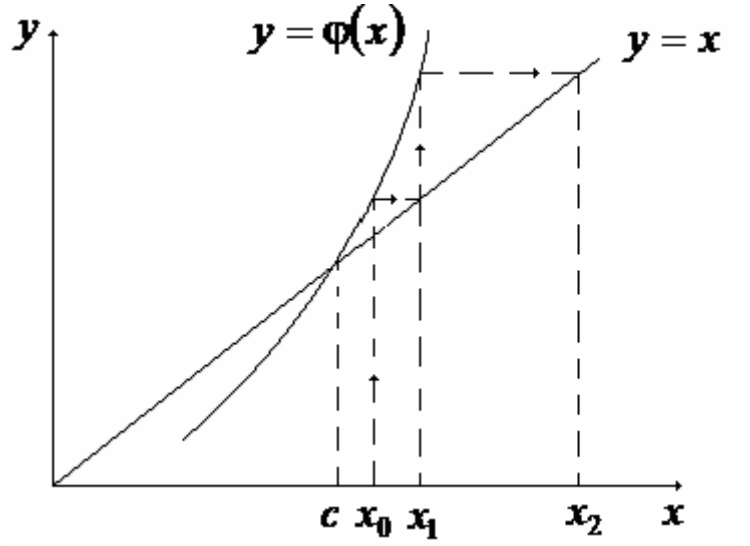


Рис. 1b. Расходящийся итерационный процесс при $\varphi' > 1$

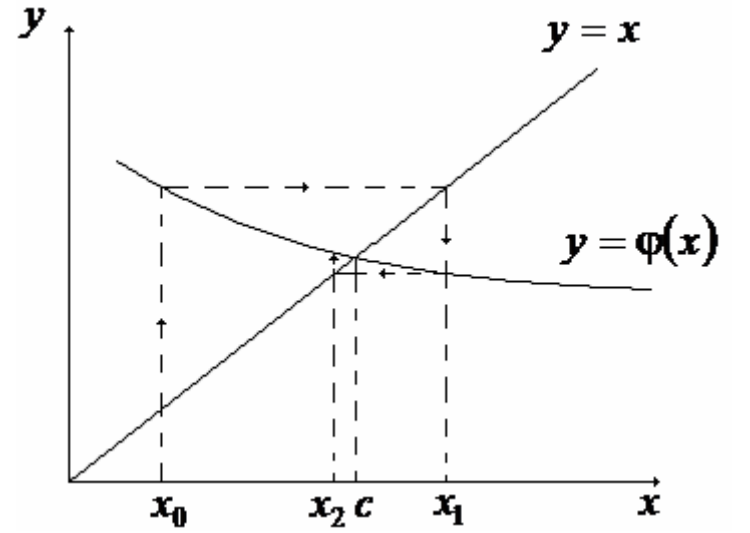


Рис. 1c. Сходящийся итерационный процесс при $-1 < \varphi' < 0$

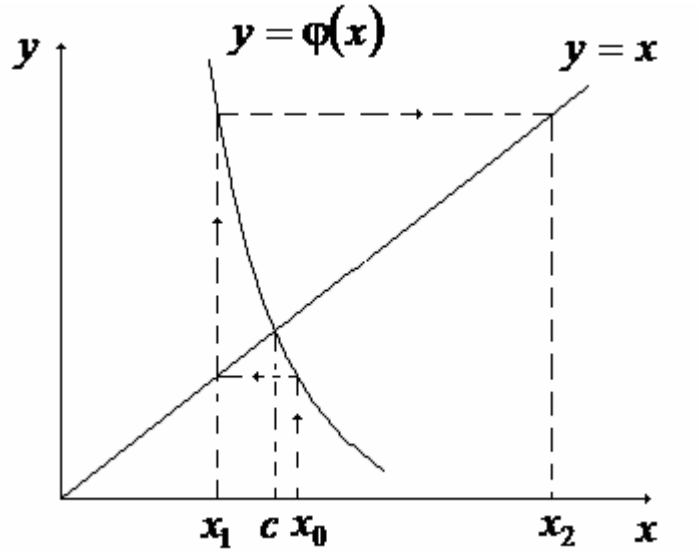


Рис. 1d. Расходящийся итерационный процесс при $\varphi' < -1$

В наших примерах мы рассмотрели случай, когда функция φ имеет производную постоянного знака. Условие сходимости итерационного процесса в общем случае дает следующая теорема.

Теорема о сходимости метода простых итераций. Пусть c - корень уравнения (4) и пусть функция $\varphi(x)$ удовлетворяет на отрезке $[c-\delta, c+\delta]$ условию Липшица с константой $L<1$, т.е.:

$$|y_2 - y_1| = |\varphi(x_2) - \varphi(x_1)| \le L|x_2 - x_1| \quad \forall x_1, x_2 \in [c-\delta, c+\delta]. \qquad (6)$$

Тогда при любом выборе x_0 на отрезке $[c-\delta, c+\delta]$ существует бесконечная итерационная последовательность $\{x_n\}$ (5), сходящаяся к корню уравнения (4) $x = c$, причем этот корень является единственным на отрезке $[c-\delta, c+\delta]$.

Заметим, что условие Липшица (6) будет заведомо выполнено, если функция $\varphi(x)$ имеет на отрезке $[c-\delta, c+\delta]$ непрерывную производную $\varphi'(x)$, модуль которой меньше единицы: $|\varphi'(x)| \le m < 1$. В этом случае согласно формуле конечных приращений Лагранжа будем иметь

$$|y_2 - y_1| = |\varphi'(\xi)(x_2 - x_1)| \le m|x_2 - x_1|. \qquad (7)$$

Мы получили неравенство (6) с константой Липшица $L = m$.

Перейдем теперь к доказательству теоремы. Число c является корнем уравнения (4), так что $c = \varphi(c)$. Возьмем произвольную точку x_0 на отрезке

$[c-\delta, c+\delta]$. Она отстоит от точки c не больше чем на δ: $|x_0 - c| \le \delta$.

Вычислим $x_1 = \varphi(x_0)$. При этом будем иметь

$$|x_1 - c| = |\varphi(x_0) - \varphi(c)| \le L|x_0 - c| \le L\delta. \qquad (8)$$

Неравенство (8) показывает, что точка x_1 принадлежит отрезку $[c-\delta, c+\delta]$ и расположена ближе к корню c, чем x_0.

Продолжим построение итерационной последовательности. Вычислим $x_2 = \varphi(x_1)$. При этом

$$|x_2 - c| = |\varphi(x_1) - \varphi(c)| \le L|x_1 - c| \le L^2|x_0 - c| \le L^2\delta.$$

Точка x_2 тоже принадлежит отрезку $[c-\delta, c+\delta]$ и расположена ближе к точке c, чем x_1. На второй итерации мы опять приблизились к c.

По индукции легко доказать, что все последующие итерации удовлетворяют неравенствам

$$|x_n - c| \le L^n|x_0 - c| \le L^n\delta. \qquad (9)$$

Отсюда следует, что

$$\lim_{n\to\infty}(x_n - c) = 0, \text{ т.е. } \lim_{n\to\infty} x_n = c. \qquad (10)$$

Нам остается доказать, что корень $x = c$ является единственным решением уравнения (4) на отрезке $[c-\delta, c+\delta]$. Предположим, что существует еще один корень $x = c_1$:

$$c_1 = \varphi(c_1), \quad c - \delta \le c_1 \le c + \delta. \qquad (11)$$

Примем c_1 за нулевое приближение и будем строить итерационную последовательность. С учетом (11) получим $x_n = c_1, \quad n = 0,1,2,\ldots$. С другой стороны, по доказанному $\lim_{n\to\infty} x_n = c$, т.е. $c_1 = c$. Таким образом, никаких других решений, кроме $x = c$, уравнение (4) на рассматриваемом отрезке не имеет.

Центральная идея метода простых итераций - сжимающие ото-

бражения - является весьма общей. Для многих сложных нелинейных задач принцип сжимающих отображений оказывается основным методом исследования.

1.3 Метод касательных (метод Ньютона)

Метод касательных является одним из наиболее эффективных численных методов решения уравнения (1). Идея метода состоит в следующем. Предположим, что функция $y = f(x)$, имеющая корень c на отрезке $[a,b]$, дифференцируема на этом отрезке и ее производная $f'(x)$ не обращается на нем в нуль. В качестве нулевого приближения для корня возьмем произвольную точку $x_0 \in [a,b]$ и запишем уравнение касательной к графику функции $f(x)$ в этой точке:

$$y = f(x_0) + f'(x_0)(x - x_0). \qquad (12)$$

В качестве первого приближения для корня выберем точку x_1 пересечения касательной с осью абсцисс (рис. 2). Для определения точки x_1 имеем уравнение

$$f(x_0) + f'(x_0)(x_1 - x_0) = 0,$$

согласно которому

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

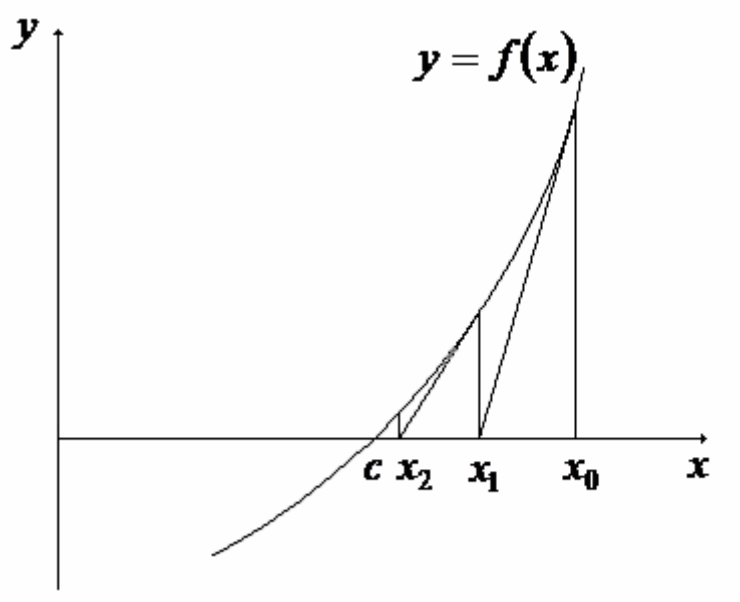


Рис.2. Построение последовательности $\{x_n\}$ по методу касательных

Продолжая этот процесс, получаем последовательность $\{x_n\}$, определенную с помощью рекуррентной формулы

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{13}$$

Условие сходимости итерационного процесса дает следующая теорема.

Теорема о сходимости метода касательных. Пусть c - корень уравнения (1) - является внутренней точкой отрезка $[a,b]$, а функция $f(x)$ дважды непрерывно дифференцируема на данном отрезке, причем ее производные удовлетворяют неравенствам

$$|f'(x)| \geq m > 0, \quad |f''(x)| \leq M, \quad x \in [a,b]. \tag{14}$$

Тогда найдется такое $\delta > 0$, что при любом выборе начального приближения x_0 на отрезке $[c-\delta, c+\delta]$ существует бесконечная итерационная последовательность (13) и эта последовательность сходится к корню c.

В силу предположения о дифференцируемости функции $f(x)$ и неравенстве нулю ее производной уравнение (1) эквивалентно на отрезке $[a,b]$ уравнению

$$x = \varphi(x), \tag{15}$$

где $\varphi(x) = x - \frac{f(x)}{f'(x)}$, так что корень $x = c$ исходного уравнения является одновременно корнем уравнения (15). Исследуем возможность отыскания этого корня с помощью метода простых итераций.

Вычислим и оценим производную функции $\varphi(x)$:

$$\varphi'(x) = \frac{f(x)f''(x)}{(f'(x))^2}, \tag{16}$$

$$|\varphi'(x)| \le \frac{M}{m^2}|f(x)|. \qquad (17)$$

Теперь воспользуемся непрерывностью функции $f(x)$ и ее равенством нулю в точке c. Выберем $\varepsilon = \frac{m^2}{2M}$. Для данного ε можно указать такое δ: $0 < \delta \le min(c-a, b-c)$, что для всех $x \in [c-\delta, c+\delta]$ будет выполняться неравенство

$$|f(x) - f(c)| = |f(x)| \le \varepsilon = \frac{m^2}{2M}. \qquad (18)$$

Учитывая это, получаем окончательную оценку производной:

$$|\varphi'(x)| \le \frac{1}{2}, \quad c-\delta \le x \le c+\delta. \qquad (19)$$

В соответствии с результатами предыдущего параграфа неравенство (19) означает, что уравнение (15) можно решать методом итераций: при любом выборе нулевого приближения на отрезке $[c-\delta, c+\delta]$ существует бесконечная последовательность, сходящаяся к корню $x = c$. Такой итерационной последовательностью для уравнения (15) является последовательность (13) метода касательных.

1.4 Метод секущих

Некоторым компромиссом между методом половинного деления и методом касательных является метод секущих. Как и в случае метода половинного деления решается уравнение (1). Предположим, для определенности, что функция $f(x)$ принимает на левом конце отрезка $[a,b]$ отрицательное значение, на правом - положительное:

$$f(a) < 0, \quad f(b) > 0. \qquad (20)$$

Тогда, согласно теореме из п.1.1, на отрезке $[a,b]$ существует корень c функции $f(x)$. Проведем прямую линию через точки с координатами $(a, f(a))$ и $(b, f(b))$:

$$y = f(a) + \frac{f(b) - f(a)}{b - a}(x - a). \qquad (21)$$

Точка пересечения этой линии с осью абсцисс лежит между точками a и b. Координата этой точки ξ может быть рассчитана как

$$\xi = a - \frac{f(a)}{f(b) - f(a)}(b - a). \qquad (22)$$

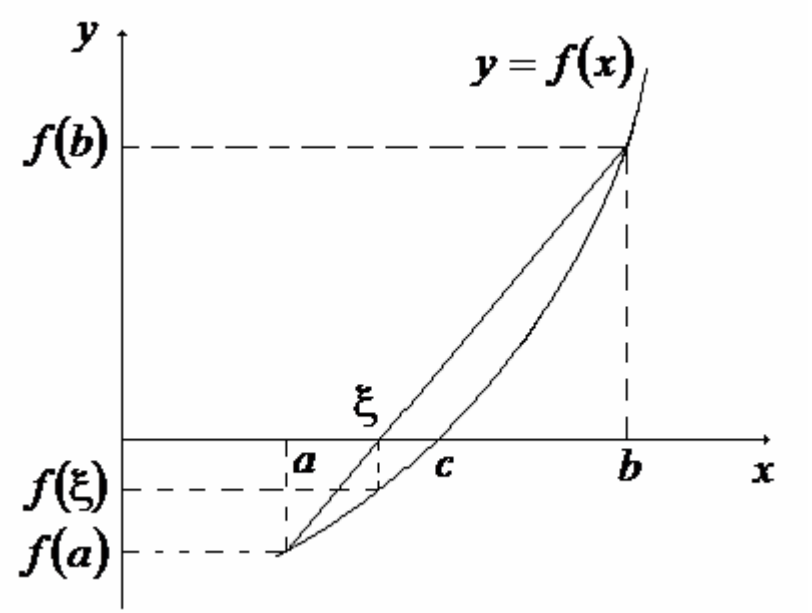


Рис.3. Нахождение промежуточной точки ξ интервала $[a,b]$ по методу секущих

Если $f(\xi) = 0$, то ξ является искомым корнем. При $f(\xi) > 0$, в качестве нового отрезка $[a_1, b_1]$ выберем отрезок $[a, \xi]$, а при $f(\xi) < 0$ в качестве нового отрезка $[a_1, b_1]$ выберем отрезок $[\xi, b]$. Новый отрезок $[a_1, b_1]$ также содержит корень c, но имеет меньшую длину. Повторяя эту процедуру n раз, мы получаем последовательность отрезков убывающей длины и содержащих корень c.

Глава 2. Прямые методы решения систем линейных алгебраических уравнений

В настоящей главе рассматриваются методы решения системы n линейных алгебраических уравнений с n неизвестными

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ &a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{aligned}. \qquad (1)$$

Иногда нам будет удобно представлять систему (1) в виде

$$\sum_{j=1}^{n} a_{ij}x_j = b_i, \quad i = 1,2,\ldots,n$$

или сокращенно

$$A\boldsymbol{X} = \boldsymbol{B},$$

где A – матрица системы, $\mathbf{X}$ и $\mathbf{B}$ – вектор неизвестных и вектор правых частей соответственно:

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_n \end{bmatrix}. \qquad (2)$$

В течение всего изложения мы будем считать, что система невырожденна, т.е. определитель системы отличен от нуля

$$\Delta = det\, A \neq 0.$$

2.1 Формулы Крамера

В курсе линейной алгебры доказывается, что решение системы (1) может быть представлено в виде

$$x_j = \frac{\Delta_j}{\Delta}, \quad j = 1,2,...,n, \tag{3}$$

где Δ - определитель системы (1), а Δ_j - определитель матрицы A_j, которая получается из матрицы A заменой ее j-го столбца столбцом правых частей системы (1):

$$A = \begin{bmatrix} a_{11} & \dots & a_{1,j-1} & f_1 & a_{1,j+1} & \dots & a_{1n} \\ a_{21} & \dots & a_{2,j-1} & f_2 & a_{2,j+1} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{n,j-1} & f_n & a_{n,j+1} & \dots & a_{nn} \end{bmatrix}. \tag{4}$$

Эти соотношения называются формулами Крамера.

При всей своей привлекательности явное решение (3), (4) пригодно лишь для аналитических исследований или для численного решения линейных уравнений низких размерностей. Это связано огромным объемом вычислений, необходимых для реализации рассматриваемого алгоритма. В самом деле, расчет определителя n-го порядка требует $n!$ умножений. Нам же необходимо рассчитать $n+1$ определитель, что требует $N_n = (n+1)!$ умножений. Для оценки факториала при $n >> 1$ воспользуемся формулой Стирлинга

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \quad n >> 1,$$

где e – основание натуральных логарифмов.

Тогда имеем

$$N_n \approx \sqrt{2\pi(n+1)}\left(\frac{n+1}{e}\right)^{n+1}, \quad n >> 1. \tag{5}$$

Согласно (5), для решения системы из 10 уравнений по формулам Крамера требуется примерно $4 \cdot 10^7$ умножений. Полагая для оценки время одной операции равной 10^{-9} сек, получаем время решения порядка 0.04 сек. Однако решение системы из 100 уравнений требует примерно 10^{160} операций, что занимает время порядка 10^{159} сек $\approx 3 \cdot 10^{152}$ лет. Для сравнения, возраст Вселенной составляет примерно $2 \cdot 10^{10}$ лет. Заметим, что современные задачи требуют решения систем линейных уравнений существенно больших размерностей.

2.2 Метод Гаусса с выделением главного элемента

Одним из самых распространенных методов решения систем линейных алгебраических уравнений является метод Гаусса. На первом этапе (прямой ход) система приводится к верхнетреугольному виду. На втором этапе (обратный ход) находятся искомые значения неизвестных.

Прямой ход. Рассмотрим первое уравнение системы (1):

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1. \qquad (6)$$

Поскольку определитель системы не равен нулю, в левой части уравнения (6) есть по крайней мере один не равный нулю коэффициент. Найдем в этом уравнении максимальный по абсолютной величине коэффициент. Перенумеруем неизвестные x_i так, чтобы слагаемое с этим коэффициентом стояло на первом месте в уравнении (6). Это эквивалентно перестановке столбцов матрицы (4) системы (1). Разделим уравнение (6) на этот коэффициент. Тогда (6) приобретает вид

$$x_1 + c_{12}x_2 + \ldots + c_{1n}x_n = f_1, \qquad (7)$$

где

$$c_{12} = \frac{a_{12}}{a_{11}}, \quad c_{13} = \frac{a_{13}}{a_{11}}, \quad \ldots, \quad c_{1n} = \frac{a_{1n}}{a_{11}}, \quad f_1 = \frac{b_1}{a_{11}}.$$

Второе уравнение системы (1) имеет вид

$$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2. \qquad (8)$$

Умножим (7) на первый коэффициент a_{21} уравнения (8) и вычтем из (8). Тогда коэффициент при x_1 в уравнении (8) обращается в нуль, и это уравнение приобретает вид

$$\tilde{a}_{22}x_2 + \tilde{a}_{23}x_3 + ... + \tilde{a}_{2n}x_n = \tilde{b}_2,$$

где

$$\begin{aligned}
&\tilde{a}_{22} = a_{22} - a_{21}c_{12},\\
&\tilde{a}_{23} = a_{23} - a_{21}c_{13},\\
&\dots\dots\dots\dots \qquad .\\
&\tilde{a}_{2n} = a_{2n} - a_{21}c_{1n},\\
&\tilde{b}_2 = b_2 - a_{21}f_1
\end{aligned}$$

Применяя эту процедуру ко всем следующим уравнениям системы, получаем систему вида

$$\begin{aligned}
x_1 + c_{12}x_2 + c_{13}x_3 + ... + c_{1n}x_n &= f_1\\
\tilde{a}_{22}x_2 + \tilde{a}_{23}x_3 + ... + \tilde{a}_{2n}x_n &= \tilde{b}_2\\
\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots&\\
\tilde{a}_{n2}x_2 + \tilde{a}_{n3}x_3 + ... + \tilde{a}_{nn}x_n &= \tilde{b}_n
\end{aligned} \quad . \qquad (9)$$

Значения коэффициентов и правых частей системы (9) вычисляются по формулам

$$\tilde{a}_{ij} = a_{ij} - a_{i1}c_{1j}, \quad \tilde{b}_i = b_i - a_{i1}f_1. \qquad (10)$$

Выделим из (9) систему, содержащую $n-1$ уравнение:

$$
\begin{aligned}
a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n &= \tilde{b}_2 \\
a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n &= \tilde{b}_3 \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\
a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n &= \tilde{b}_n
\end{aligned}
.
$$

Применим к этой системе описанную выше процедуру. И так далее. Продолжая процесс исключения, после $n-1$ -го шага приведем исходную систему к виду

$$
\begin{aligned}
x_1 + c_{12}x_2 + c_{13}x_3 + \ldots + c_{1n}x_n &= f_1 \\
x_2 + c_{23}x_3 + \ldots + c_{2n}x_n &= f_2 \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\
x_{n-1} + c_{n-1,n}x_n &= f_{n-1} \\
x_n &= f_n
\end{aligned}
\qquad (11)
$$

или в матричной форме

$$
C\boldsymbol{x} = \boldsymbol{f} .
$$

Здесь C является верхнетреугольной матрицей с единицами на главной диагонали:

$$
C = \begin{bmatrix}
1 & c_{12} & c_{13} & \ldots & c_{1n} \\
0 & 1 & c_{23} & \ldots & c_{2n} \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & \ldots & 1
\end{bmatrix}. \qquad (12)
$$

Обратный ход состоит в последовательном определении неизвестных из системы (11):

$$
\begin{aligned}
&x_n = f_n \\
&x_{n-1} = f_{n-1} - c_{n-1,n}x_n \\
&x_{n-2} = f_{n-2} - c_{n-2,n-1}x_{n-1} - c_{n-2,n}x_n \qquad (13)\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&x_1 = f_1 - c_{12}x_2 - c_{13}x_3 - \ldots - c_{1n}x_n
\end{aligned}
$$

Оценка числа операций N_n, необходимых для решения системы из n линейных алгебраических уравнений методом Гаусса, дает $N_n = \frac{1}{3}n^3$. Для решения системы 100 линейных уравнений требуется примерно $\frac{1}{3}\cdot 10^6 \approx 3\cdot 10^5$ операций. Полагая, для оценки, время одной операции равной 10^{-9} сек, получаем время решения порядка $3\cdot 10^{-4}$ сек. Напомним, что решение такой системы по методу Крамера требует примерно 10^{160} операций. Для решения системы из 10^4 уравнений по методу Гаусса требуется $3\cdot 10^{11}$ операций или 300 сек $= 5$ мин.

В заключение отметим, что первый этап метода Гаусса может быть использован для вычисления определителя матрицы A. Прямой ход метода Гаусса основан на многократном выполнении операции перестановки столбцов матрицы A и сложения одной из строк матрицы с другой строкой, взятой с некоторым множителем. Обе операции не меняют значения определителя системы. Однако, при делении ведущей строки на ее диагональный элемент, определитель также делится на этот элемент. Поскольку определитель приведенной треугольной системы всегда равен единице, то определитель исходной системы

$$\Delta = det\, A = (-1)^k a_{11}a_{22}^{(1)}\ldots a_{nn}^{(n-1)},$$

где k - число перестановок столбцов в процессе редукции матрицы A к верхнетреугольной матрице C.

2.3 Обусловленность систем линейных алгебраических уравнений

Оценим устойчивость решения системы линейных алгебраических уравнений по отношению к изменению правой части системы. Рассмотрим сначала модельный пример одного линейного уравнения

$$ax = b. \tag{14}$$

Решение этого уравнения имеет вид

$$x = \frac{b}{a}. \tag{15}$$

Малая вариация δb правой части уравнения (14) приводит к вариации решения

$$\delta x = \frac{\delta b}{a}. \tag{16}$$

Используя (15), (16), находим, что относительная вариация решения уравнения (14) равна относительной вариации его правой части

$$\frac{\delta x}{x} = \frac{\delta b}{b}. \tag{17}$$

Вернемся теперь к рассмотрению системы линейных алгебраических уравнений

$$A\boldsymbol{x} = \boldsymbol{b}. \tag{18}$$

Если определитель матрицы A не равен нулю, то существует обратная матрица A^{-1}. Тогда решение системы (18) может быть представлено в виде

$$\boldsymbol{x} = A^{-1}\boldsymbol{b}. \tag{19}$$

Малая вариация $\delta\boldsymbol{b}$ правой части системы (18) приводит к вариации решения

$$\delta \boldsymbol{x} = A^{-1}\delta \boldsymbol{b}. \tag{20}$$

Отсюда получаем связь нормы вариации решения с нормой вариации правой части (смотри Приложение 1)

$$\left\|\delta \boldsymbol{x}\right\| \le \left\|A^{-1}\right\| \cdot \left\|\delta \boldsymbol{b}\right\|. \tag{21}$$

Исходное уравнение (18) позволяет написать неравенство

$$\left\|\boldsymbol{b}\right\| \le \left\|A\right\| \cdot \left\|\boldsymbol{x}\right\|, \tag{22}$$

Перемножив его с неравенством того же знака (21), получим

$$\left\|\boldsymbol{b}\right\| \cdot \left\|\delta \boldsymbol{x}\right\| \le \left\|A\right\| \cdot \left\|A^{-1}\right\| \cdot \left\|\boldsymbol{x}\right\| \left\|\delta \boldsymbol{b}\right\|, \tag{23}$$

Пусть $\boldsymbol{b} \neq 0$, тогда, согласно (19), $\mathbf{x} \neq 0$, и неравенство (23) можно переписать в виде

$$\frac{\left\|\delta \boldsymbol{x}\right\|}{\left\|\boldsymbol{x}\right\|} \le M_A \frac{\left\|\delta \boldsymbol{b}\right\|}{\left\|\boldsymbol{b}\right\|}, \quad M_A = \left\|A\right\| \cdot \left\|A^{-1}\right\|. \tag{24}$$

Число M_A называется стандартным числом обусловленности матрицы A. Покажем, что $M_A \ge 1$. В самом деле, по определению обратной матрицы, $A \cdot A^{-1} = I$, где I – единичная матрица. Тогда имеем

$$1 = \left\|I\right\| = \left\|A \cdot A^{-1}\right\| \le \left\|A\right\| \cdot \left\|A^{-1}\right\| = M_A,$$

что и требовалось доказать.

Согласно (17), число обусловленности системы, состоящей из одного уравнения, равно единице. Если число обусловленности много больше единицы, то малые относительные вариации правой части системы приводят к большим относительным вариациям решения. Такие системы называются плохо обусловленными.

Пример 1

Рассмотрим систему двух уравнений

$$\begin{aligned} &\mathrm{x}_1 + 0 \cdot \mathrm{x}_2 = \mathrm{b}_1, \\ &\mathrm{x}_1 + \varepsilon \cdot \mathrm{x}_2 = \mathrm{b}_2, \end{aligned} \qquad A = \begin{bmatrix} 1 & 0 \\ 1 & \varepsilon \end{bmatrix}, \tag{25}$$

где ε - малый параметр.

Решение системы имеет вид

$$\begin{aligned} x_1 &= b_1 \\ x_2 &= \frac{b_2}{\varepsilon} - \frac{b_1}{\varepsilon} \end{aligned} \quad . \tag{26}$$

Тогда обратная матрица имеет вид

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ -\varepsilon^{-1} & \varepsilon^{-1} \end{bmatrix}. \tag{27}$$

Найдем нормы прямой и обратной матриц и рассчитаем, согласно (24), число обусловленности системы. По определению

$$\|A\| = \max_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\|. \tag{28}$$

Введем в двумерном евклидовом пространстве вектор $\boldsymbol{x}$, норма которого рана единице

$$\boldsymbol{x} = \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix}.$$

Тогда

$$\|A\| = \max_{\varphi} \left\| \begin{bmatrix} 1 & 0 \\ 1 & \varepsilon \end{bmatrix} \begin{pmatrix} \cos\varphi \\ \sin\varphi \end{pmatrix} \right\| = \max_{\varphi} \left\| \begin{pmatrix} \cos\varphi \\ \cos\varphi + \varepsilon \sin\varphi \end{pmatrix} \right\|,$$

$$\|A\| = \max_{\varphi} \sqrt{2\cos^2\varphi + 2\varepsilon \sin\varphi\cos\varphi + \varepsilon^2 \sin^2\varphi} \approx \sqrt{2}.$$

В свою очередь, норма обратной матрицы может быть рассчитана как

$$\left\|A^{-1}\right\| = \max_{\varphi}\left\|\begin{bmatrix} 1 & 0 \\ -\varepsilon^{-1} & \varepsilon^{-1} \end{bmatrix}\begin{pmatrix} cos\,\varphi \\ sin\,\varphi \end{pmatrix}\right\| = \max_{\varphi}\left\|\begin{pmatrix} cos\,\varphi \\ -\varepsilon^{-1}\,cos\,\varphi + \varepsilon^{-1}\,sin\,\varphi \end{pmatrix}\right\|,$$

$$\left\|A^{-1}\right\| = \max_{\varphi}\sqrt{cos^2\,\varphi + \varepsilon^{-2}\left(cos\,\varphi - sin\,\varphi\right)^2}\,,$$

$$\left\|A^{-1}\right\| = \max_{\varphi}\sqrt{cos^2\,\varphi + \varepsilon^{-2}\left(2 - sin\,2\varphi\right)} \approx \varepsilon^{-1}\sqrt{3}\,.$$

Тогда число обусловленности системы

$$M_A \approx \varepsilon^{-1}\sqrt{6}\,. \tag{29}$$

Из (29) видим, что чем меньше ε, тем больше число обусловленности системы (25), т.е. тем хуже она обусловлена. Кроме того, из приведенного примера видно, насколько громоздким является расчет норм прямой и обратной матрицы, исходя из определения (28), даже в простейшем случае системы двух уравнений. В следующем параграфе мы получим гораздо более эффективный способ оценки числа обусловленности системы.

2.4. Оценка числа обусловленности

Пусть λ_{max}- максимальное по модулю собственное число матрицы A, **x** - **с**оответствующий собственный вектор:

$$A\boldsymbol{x} = \lambda_{max}\boldsymbol{x},$$

тогда

$$\left|\lambda_{max}\right|\left\|\boldsymbol{x}\right\| = \left\|A\boldsymbol{x}\right\| \le \left\|A\right\|\left\|\boldsymbol{x}\right\|.$$

Следовательно, поскольку $\left\|\boldsymbol{x}\right\| \neq 0$,

$$\left|\lambda_{max}\right| \le \left\|A\right\|. \tag{30}$$

Аналогично, для минимального собственного числа λ_{min} матрицы A и

соответствующего собственного вектора $\mathbf{y}$, имеем

$$A\boldsymbol{y} = \lambda_{min}\boldsymbol{y}$$

или

$$A^{-1}\boldsymbol{y} = \frac{1}{\lambda_{min}}\boldsymbol{y}.$$

Отсюда следует оценка

$$\frac{1}{|\lambda_{min}|} \le \left\|A^{-1}\right\|. \tag{31}$$

Перемножая неравенства (30) и (31), получаем оценку числа обусловленности матрицы A:

$$M_A = \|A\| \cdot \left\|A^{-1}\right\| \ge |\lambda_{max}| / |\lambda_{min}|. \tag{32}$$

Если матрица симметричная, то все ее собственные числа вещественны, причем

$$\|A\| = |\lambda_{max}| \text{ и } \left\|A^{-1}\right\| = \frac{1}{|\lambda_{min}|}.$$

поэтому для таких матриц

$$M_A = |\lambda_{max}| / |\lambda_{min}|, \tag{33}$$

Из (33) видим, что число обусловленности тем больше, чем больше разброс собственных чисел матрицы. С увеличением размера матрицы число ее обусловленности имеет тенденцию к увеличению.

Возвращаясь к Примеру 1, находим $\lambda_{max} = 1, \quad \lambda_{min} = \varepsilon$. Тогда, согласно (33) ,

$$M_A \ge \frac{\lambda_{max}}{\lambda_{min}} = \varepsilon^{-1},$$

что согласуется с оценкой (29).

2.5 Метод прогонки

Метод прогонки применим при решении систем линейных алгебраических уравнений вида

$$x_0 = p_0 x_1 + q_0, \tag{34}$$

$$A_i x_{i-1} + C_i x_i + B_i x_{i+1} = F_i, \quad i = 1,...,n-1, \tag{35}$$

$$x_n = p_n x_{n-1} + q_n. \tag{36}$$

Такого рода системы возникают, например, при решении краевых задач математической физики. При этом уравнения (35) представляют собой численную аппроксимацию дифференциального уравнения, а уравнения (34) и (36) – аппроксимацию граничных условий. Для простоты изложения, рассмотрим частный случай уравнений (34) и (36):

$$x_0 = q_0, \quad x_n = q_n. \tag{37}$$

Используя (37), приведем систему (35) к виду

$$\begin{gathered} C_1 x_1 + B_1 x_2 = F_1 - A_1 q_0, \\ A_2 x_1 + C_2 x_2 + B_1 x_3 = F_2, \\ \vdots \\ A_{n-1} x_{n-2} + C_{n-1} x_{n-1} = F_{n-1} - B_{n-1} q_n, \end{gathered} \tag{38}$$

Матрица этой системы имеет трехдиагональную структуру:

$$\begin{bmatrix} C_1 & B_1 & 0 & 0 & \dots & 0 & 0 \\ A_2 & C_2 & B_2 & 0 & \dots & 0 & 0 \\ 0 & A_3 & C_3 & B_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & A_{n-1} & C_{n-1} \end{bmatrix}. \tag{39}$$

Это существенно упрощает решение системы (38) благодаря специальному методу, получившему название метода прогонки. Этот метод основан на

предположении, что искомые неизвестные x_i и x_{i+1} связаны рекуррентным соотношением

$$x_i = \alpha_{i+1} x_{i+1} + \beta_{i+1}, \quad 0 \le i \le n-1. \tag{40}$$

Здесь величины $\alpha_{i+1}, \beta_{i+1}$, получившие название прогоночных коэффициентов, подлежат определению, исходя из условий задачи (37), (38). Такая процедура означает замену прямого определения неизвестных x_i задачей определения прогоночных коэффициентов с последующим расчетом по ним величин x_i.

Для реализации описанной программы выразим, с помощью соотношения (40), x_{i-1} через x_{i+1}:

$$x_{i-1} = \alpha_i x_i + \beta_i = \alpha_i \alpha_{i+1} x_{i+1} + \alpha_i \beta_{i+1} + \beta_i$$

и подставим x_{i-1} и x_i, выраженные через x_{i+1}, в исходные уравнения (35). В результате получим

$$\left(A_i \alpha_i \alpha_{i+1} + C_i \alpha_{i+1} + B_i\right) x_{i+1} + A_i \alpha_i \beta_{i+1} + A_i \beta_i + C_i \beta_{i+1} - F_i = 0,$$
$$i = 1,2,\ldots,n-1.$$

Последние соотношения будут заведомо выполняться, и притом независимо от решения, если потребовать, чтобы при $i = 1,2,\ldots,n-1$ имели место равенства

$$A_i \alpha_i \alpha_{i+1} + C_i \alpha_{i+1} + B_i = 0,$$
$$A_i \alpha_{i+1} \beta_{i+1} + A_i \beta_i + C_i \beta_{i+1} - F_i = 0.$$

Отсюда следуют рекуррентные соотношения для прогоночных коэффициентов:

$$\alpha_{i+1} = -\frac{B_i}{A_i \alpha_i + C_i}, \quad \beta_{i+1} = \frac{F_i - A_i \beta_i}{A_i \alpha_i + C_i}, \quad i = 1,2,\ldots,n-1. \tag{41}$$

Граничное условие на левом конце интервала $x_0 = q_0$ и соотношение

$x_0 = \alpha_1 x_1 + \beta_1$ непротиворечивы, если положить

$$\alpha_1 = 0, \quad \beta_1 = q_0 . . \tag{42}$$

Остальные значения коэффициентов прогонки $\alpha_{2,},...,\alpha_n$ и $\beta_{2,},...,\beta_n$ находим из (41), чем и завершаем этап вычисления прогоночных коэффициентов. Далее, согласно граничному условию на правом конце интервала,

$$x_n = q_n . \tag{43}$$

Отсюда можно найти остальные неизвестные $x_{n-1},...,x_1$ в процессе обратной прогонки с помощью рекуррентной формулы (40).
Число операций, которое требуется для решения системы методом прогонки, растет пропорционально размерности системы n. Напомним, что для реализации метода Гаусса, это число растет, как n^3.

Во многих прикладных задачах, которые приводят к системам линейных алгебраических уравнений с трехдиагональной матрицей, ее коэффициенты удовлетворяют неравенствам

$$|C_i| > |A_i| + |B_i| .$$

(т.н. условие диагонального преобладания). Можно показать, что в этом случае прогоночные коэффициенты удовлетворяют неравенствам

$$|\alpha_i| \le 1, \tag{44}$$

что делает прогонку устойчивой. Действительно, предположим, что компонента решения x_i в результате процедуры округления рассчитана с некоторой ошибкой. Тогда при вычислении следующей компоненты x_{i-1} по рекуррентной формуле (40) эта ошибка благодаря неравенствам (44) не будет нарастать.

Глава 3. Итерационные методы решения систем линейных алгебраических уравнений

В главе 2 отмечалось, что если матрица A системы линейных алгебраических уравнений

$$A\boldsymbol{x} = \boldsymbol{B} \tag{1}$$

плохо обусловлена, т.е. число обусловленности $\mathrm{M_A} >> 1$, то при решении (1) методом Гаусса погрешности округления приводят к большим погрешностям решения. Этого недостатка лишены итерационные методы решения систем линейных алгебраических уравнений.

3.1 Итерационные последовательности

Ограничимся рассмотрением линейных одношаговых итерационных алгоритмов

$$C\frac{\boldsymbol{x}^{k+1}-\boldsymbol{x}^{k}}{\tau}+A\boldsymbol{x}^{k}=\boldsymbol{B},\quad det\,C\neq 0,\quad \tau>0, \tag{2}$$

где $\boldsymbol{x}^{k}$ - приближенное значение искомого вектора решения на k-й итерации, τ - итерационный параметр, C - вспомогательная матрица. Разрешая (2) относительно $\boldsymbol{x}^{k+1}$, получаем связь между значениями искомого вектора $\mathbf{x}$ на k+1 и на k-м шаге итерации

$$\boldsymbol{x}^{k+1}=\boldsymbol{x}^{k}+\tau C^{-1}\left(\boldsymbol{B}-A\boldsymbol{x}^{k}\right). \tag{3}$$

Будем говорить, что итерационная последовательность $\boldsymbol{x}^{0},\boldsymbol{x}^{1},\boldsymbol{x}^{2}\ldots$ сходится к вектору $\boldsymbol{x}$ по евклидовой норме (см. Приложение 1), если

$$\lim_{k\to\infty}\left\|\boldsymbol{x}^{k}-\boldsymbol{x}\right\|=\lim_{k\to\infty}\sqrt{\left(x_1^{k}-x_1\right)^2+\left(x_2^{k}-x_2\right)^2+\ldots+\left(x_n^{k}-x_n\right)^2}=0. \tag{4}$$

Для исследования сходимости итерационного процесса введем два понятия – погрешность решения

$$\boldsymbol{z}^k = \boldsymbol{x}^k - \boldsymbol{x} \tag{5}$$

и невязка решения

$$\boldsymbol{\psi}^k = A\boldsymbol{x}^k - \boldsymbol{B}. \tag{6}$$

Сходимость итерационного процесса означает, что $\lim\limits_{k\to\infty}\left\|\boldsymbol{z}^k\right\| = 0$. Норма погрешности показывает, насколько приближенное решение, полученное на k-м шаге итерации, близко к точному решению. К сожалению, эту величину нельзя определить в ходе итераций, поскольку искомое решение неизвестно. Невязка решения показывает, насколько хорошо $\boldsymbol{x}^k$ удовлетворяет системе (1). Эта величина легко рассчитывается на каждом итерационном шаге.

Установим связь между $\boldsymbol{z}^k$ и $\boldsymbol{\psi}^k$:

$$\boldsymbol{\psi}^k = A\boldsymbol{x}^k - \boldsymbol{B} = A(\boldsymbol{z}^k + \boldsymbol{x}) - \boldsymbol{B} = A\boldsymbol{z}^k. \tag{7}$$

Используя обратную матрицу A^{-1}, получаем

$$\boldsymbol{z}^k = A^{-1}\boldsymbol{\psi}^k. \tag{8}$$

Из формул (7) и (8) вытекают неравенства:

$$\left\|\boldsymbol{\psi}^k\right\| \le \|A\|\left\|\boldsymbol{z}^k\right\|, \quad \left\|\boldsymbol{z}^k\right\| \le \left\|A^{-1}\right\|\left\|\boldsymbol{\psi}^k\right\|. \tag{9}$$

Следовательно, погрешность решения стремится к нулю тогда и только тогда, когда стремится к нулю невязка. При исследовании сходимости итерационных методов большую роль играют самосопряженность и знакоопределенность матриц A и C - см. Приложение 2.

3.2 Достаточные условия сходимости итерационного процесса

Теорема Самарского. Пусть A - самосопряженная положительно определенная матрица, $C - \frac{\tau}{2}A$ - положительно определенная матрица, τ - положительное число. Тогда при любом выборе нулевого приближения $\boldsymbol{x}^0$ итерационный процесс (2) сходится к решению системы (1).

Обсудим сначала условие $C - \frac{\tau}{2} A > 0$. Оно эквивалентно неравенству

$$(C\boldsymbol{x}, \boldsymbol{x}) > \frac{\tau}{2}(A\boldsymbol{x}, \boldsymbol{x}), \quad \boldsymbol{x} \neq 0. \tag{10}$$

Поскольку A - положительно определенная матрица, из (9) вытекает, что C также является положительно определенной матрицей. Кроме того, (9) определяет интервал, в котором может изменяться параметр τ:

$$0 < \tau < \tau_0 = \inf_{x \neq 0} \frac{2(C\boldsymbol{x}, \boldsymbol{x})}{(A\boldsymbol{x}, \boldsymbol{x})}. \tag{11}$$

Перейдем теперь к доказательству теоремы. Согласно (5)

$$\boldsymbol{x}^k = \boldsymbol{z}^k + \boldsymbol{x}. \tag{12}$$

Используя (12), представим итерационное соотношение (2) в виде

$$C \frac{\boldsymbol{z}^{k+1} - \boldsymbol{z}^k}{\tau} + A\boldsymbol{z}^k = 0. \tag{13}$$

Мы показали выше, что матрица C - положительно определенная. Следовательно, она невырожденная и имеет обратную. Тогда рекуррентное соотношение (13) можно разрешить относительно z_{k+1}:

$$\boldsymbol{z}^{k+1} = \boldsymbol{z}^k - \tau C^{-1} A\boldsymbol{z}^k,$$

$$\boldsymbol{z}^{k+1} = \boldsymbol{z}^k - \tau\boldsymbol{\omega}^k, \tag{14}$$

где

$$\boldsymbol{\omega}^k = C^{-1} A\boldsymbol{z}^k, \text{ так что } A\boldsymbol{z}^k = C\boldsymbol{\omega}^k. \tag{15}$$

Умножая обе части равенства (14) слева на матрицу A, получаем еще одно рекуррентное соотношение:

$$A\boldsymbol{z}^{k+1} = A\boldsymbol{z}^k - \tau A\boldsymbol{\omega}^k . \tag{16}$$

Рассмотрим последовательность положительных функционалов:

$$J_k = \left(A\boldsymbol{z}^k, \boldsymbol{z}^k\right).$$

Составим аналогичное выражение для J_{k+1} и преобразуем его с помощью рекуррентных формул (14) и (16):

$$\begin{aligned} J_{k+1} &= \left(A\boldsymbol{z}^{k+1}, \boldsymbol{z}^{k+1}\right) = \left(A\boldsymbol{z}^k - \tau A\boldsymbol{\omega}^k, \boldsymbol{z}^k - \tau\boldsymbol{\omega}^k\right) = \\ &= \left(A\boldsymbol{z}^k, \boldsymbol{z}^k\right) - \tau\left(A\boldsymbol{\omega}^k, \boldsymbol{z}^k\right) - \tau\left(A\boldsymbol{z}^k, \boldsymbol{\omega}^k\right) + \tau^2\left(A\boldsymbol{\omega}^k, \boldsymbol{\omega}^k\right). \end{aligned} \tag{17}$$

Из самосопряженности матрицы A и формулы (15) следует, что

$$\left(A\boldsymbol{\omega}^k, \boldsymbol{z}^k\right) = \left(A\boldsymbol{z}^k, \boldsymbol{\omega}^k\right) = \left(C\boldsymbol{\omega}^k, \boldsymbol{\omega}^k\right).$$

В результате соотношение (17) принимает вид

$$\begin{aligned} J_{k+1} &= J_k - 2\tau\left(C\boldsymbol{\omega}^k, \boldsymbol{\omega}^k\right) + \tau^2\left(A\boldsymbol{\omega}^k, \boldsymbol{\omega}^k\right) = \\ &= J_k - 2\tau\left(\left(C - \frac{\tau}{2}A\right)\boldsymbol{\omega}^k, \boldsymbol{\omega}^k\right). \end{aligned} \tag{18}$$

Таким образом, последовательность функционалов J_k с учетом условия $C - \frac{\tau}{2}A > 0$ образует монотонно невозрастающую последовательность, ограниченную снизу нулем:

$$J_k \geq J_{k+1} \geq \ldots \geq 0, \tag{19}$$

поэтому она сходится. Далее, по лемме 3 из Приложения 2,

$$\left(\left(C-\frac{\tau}{2}A\right)\boldsymbol{\omega}^k,\boldsymbol{\omega}^k\right)\geq\delta\left\|\boldsymbol{\omega}^k\right\|^2,\quad \delta>0.$$

Тогда, согласно (18),

$$J_k-J_{k+1}=2\tau\left(\left(C-\frac{\tau}{2}A\right)\boldsymbol{\omega}^k,\boldsymbol{\omega}^k\right)\geq 2\tau\delta\left\|\boldsymbol{\omega}^k\right\|^2. \qquad (20)$$

Выше было показано, что последовательности функционалов J_k сходится. Следовательно, $J_k-J_{k+1}\to 0$ при $k\to\infty$. Но тогда и $\left\|\boldsymbol{\omega}^k\right\|\to 0$ при $k\to\infty$. Согласно (14), $\boldsymbol{z}^k=A^{-1}C\boldsymbol{\omega}^k$, так что

$$\left\|\boldsymbol{z}^k\right\|\leq\left\|A^{-1}\right\|\cdot\left\|B\right\|\left\|\boldsymbol{\omega}^k\right\|\to 0,\quad k\to\infty.$$

Теорема доказана.

3.3 Метод простой итерации

Рассмотрим итерационный алгоритм (2). В самом простом случае $C=I$ (I - единичная матрица), при этом получаем

$$\frac{\boldsymbol{x}^{k+1}-\boldsymbol{x}^k}{\tau}+A\boldsymbol{x}^k=\boldsymbol{B},\qquad \boldsymbol{x}^{k+1}=\boldsymbol{x}^k+\tau\left(\boldsymbol{B}-A\boldsymbol{x}^k\right). \qquad (21)$$

Будем считать, что матрица A удовлетворяет условию теоремы Самарского, т.е. $A=A^+>0$, тогда формула (11), определяющая границу интервала сходимости по итерационному параметру τ, принимает вид

$$0<\tau<\tau_0=\inf_{x\neq 0}\frac{2(\boldsymbol{x},\boldsymbol{x})}{(A\boldsymbol{x},\boldsymbol{x})}=\frac{2}{\sup\limits_{x\neq 0}\dfrac{(A\boldsymbol{x},\boldsymbol{x})}{(\boldsymbol{x},\boldsymbol{x})}}. \qquad (22)$$

Пусть $\boldsymbol{e}_1,\boldsymbol{e}_2,\ldots,\boldsymbol{e}_n$ - ортонормированный базис собственных векторов

оператора, соответствующего матрице A:

$$A\boldsymbol{e}_k = \lambda_k \boldsymbol{e}_k, \quad (\boldsymbol{e}_k, \boldsymbol{e}_m) = \delta_{km}.$$

В силу положительной определенности матрицы A все его собственные значения положительны:

$$0 < (A\boldsymbol{e}_k, \boldsymbol{e}_k) = \lambda_k (\boldsymbol{e}_k, \boldsymbol{e}_k) = \lambda_k .$$

Будем считать их занумерованными в порядке убывания:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0. \tag{23}$$

Разложим вектор $\boldsymbol{x} \neq 0$ по базису собственных векторов:

$$\boldsymbol{x} = \xi_1 \boldsymbol{e}_1 + \xi_2 \boldsymbol{e}_2 + \ldots + \xi_n \boldsymbol{e}_n,$$

тогда

$$(\boldsymbol{x}, \boldsymbol{x}) = \xi_1^2 + \xi_2^2 + \ldots + \xi_n^2, \ (A\boldsymbol{x}, \boldsymbol{x}) = \lambda_1 \xi_1^2 + \lambda_2 \xi_2^2 + \ldots + \lambda_n \xi_n^2$$

и

$$\sup_{\boldsymbol{x} \neq 0} \frac{(A\boldsymbol{x}, \boldsymbol{x})}{(\boldsymbol{x}, \boldsymbol{x})} = \sup_{\boldsymbol{x} \neq 0} \frac{\lambda_1 \xi_1^2 + \lambda_2 \xi_2^2 + \ldots + \lambda_n \xi_n^2}{\xi_1^2 + \xi_2^2 + \ldots + \xi_n^2} = \lambda_1.$$

В результате из (22) следует, что метод простой итерации сходится при любом τ, принадлежащем интервалу

$$0 < \tau < \tau_0 = \frac{2}{\lambda_1}. \tag{24}$$

Дальнейшее исследование основано на анализе рекуррентного соотношения (21). Введем матрицу перехода

$$S = I - \tau A, \quad S^+ = S \tag{25}$$

и перепишем (21) в виде

$$\boldsymbol{x}^{k+1} = S\boldsymbol{x}^k + \tau B. \tag{26}$$

Погрешность $\boldsymbol{z}^k = \boldsymbol{x}^k - \boldsymbol{x}$ будет удовлетворять аналогичному рекуррентному соотношению:

$$\boldsymbol{z}^{k+1} = S\boldsymbol{z}^k. \tag{27}$$

Согласно (27), скорость сходимости итерационного процесса определяется значением нормы оператора перехода:

$$\left\|\boldsymbol{z}^{k+1}\right\| \le \|S\| \cdot \left\|\boldsymbol{z}^k\right\|, \text{ тогда } \left\|\boldsymbol{z}^k\right\| \le \|S\|^k \cdot \left\|\boldsymbol{z}^0\right\|. \tag{28}$$

Таким образом, чем меньше $\|S\|$, тем быстрее сходится итерационный процесс. Рассчитаем величину этой нормы. Пусть $\boldsymbol{e}_i$ и λ_i - собственный вектор и соответствующее собственное число матрицы A. Тогда

$$S\boldsymbol{e}_i = (I - \tau A)\boldsymbol{e}_i = (1 - \tau\lambda_i)\boldsymbol{e}_i.$$

Таким образом, $\boldsymbol{e}_i$ является одновременно и собственным вектором матрицы S, а соответствующее собственное число матрицы S имеет вид

$$\mu_i(\tau) = 1 - \tau\lambda_i. \tag{29}$$

При самосопряженной матрице A матрица S также является самосопряженной. Следовательно, ее норма определяется наибольшим по модулю собственным значением $\mu_i(\tau)$:

$$\|S\| = \max_{1\le i\le n}|\mu_i(\tau)| = \max_{1\le i\le n}|1 - \tau\lambda_i|. \tag{30}$$

Найдем значение итерационного параметра τ, при котором $\|S\|$ имеет минимальное значение. Это значение и обеспечивает максимальную скорость сходимости итерационного процесса.

На Рис. 1 представлены графики зависимости от параметра τ функций $|1 - \tau\lambda_i|$ и нормы матрицы перехода $\|S(\tau)\|$. В соответствии с (30), график $\|S(\tau)\|$ является верхней огибающей графиков $|1 - \tau\lambda_i|$. Оптимальное

значение итерационного параметра $\tau = \tau_*$ соответствует наименьшему значению функции $\|S(\tau)\|$. В соответствии с Рис.1, это значение определяется из уравнения $1 - \tau\lambda_n = \tau\lambda_1 - 1$. Следовательно

$$\tau_* = \frac{2}{\lambda_1 + \lambda_2}, \qquad \|S(\tau_*)\| = \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}. \tag{31}$$

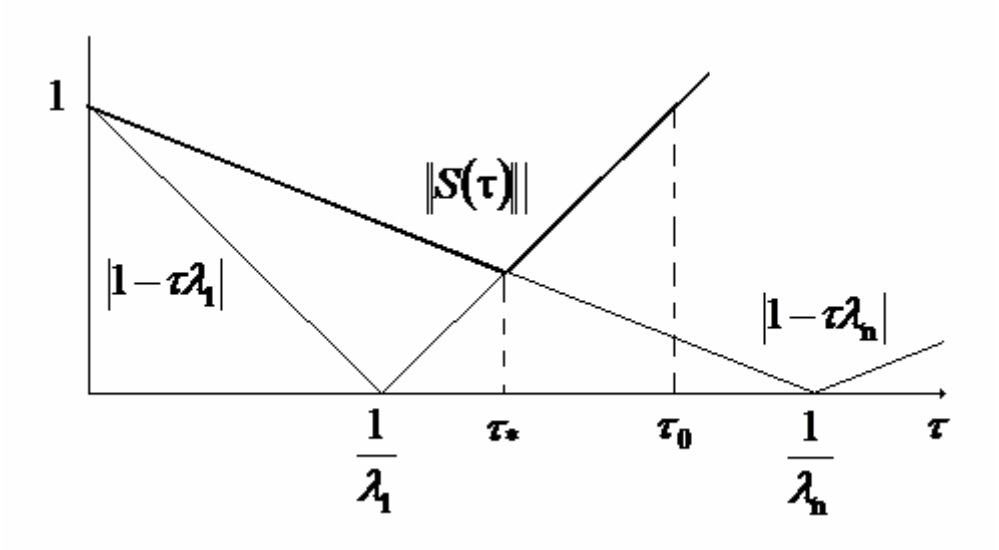


Рис. 1. Определение оптимального значения итерационного параметра τ_*

Используя (2.33), выразим наименьшее значение нормы матрицы перехода через число обусловленности M_A:

$$\|S(\tau_*)\| = \frac{M_A - 1}{M_A + 1}. \tag{32}$$

Таким образом, для плохо обусловленной матрицы $(M_A >> 1)$ даже при оптимальном выборе итерационного параметра $\tau = \tau_*$ норма матрицы S близка к единице, так что сходимость метода простой итерации в этом случае оказывается медленной.

В заключение заметим, что формула (31) для оптимального значения итерационного параметра τ_* представляет прежде всего теоретический интерес. Обычно при решении систем линейных алгебраических уравнений наибольшее и наименьшее собственные числа матрицы A неизвестны, поэтому подсчитать τ_* заранее невозможно. В результате итерационный па-

раметр τ нередко приходится подбирать в процессе вычислений методом проб и ошибок.

Задача 1. *Рассмотреть систему двух уравнений с двумя неизвестными:*

$$\begin{cases} x_1 + x_2 = 0, \\ x_1 + 2x_2 = 1 \end{cases} \tag{33}$$

и построить для нее приближенное решение с помощью метода простой итерации.

Выпишем сразу решение системы (33):

$$x_1 = -1, \quad x_2 = 1, \tag{34}$$

чтобы потом иметь возможность сравнить его с членами итерационной последовательности. Перейдем к решению системы методом простой итерации. Матрица системы имеет вид

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}.$$

Она самосопряженная и положительно определенная, поскольку

$$A\boldsymbol{x}, \boldsymbol{x} = (x_1 + x_2)x_1 + (x_1 + 2x_2)x_2 = (x_1 + x_2)^2 + x_2^2 > 0.$$

Составим характеристическое уравнение для матрицы A и найдем его корни:

$$\begin{vmatrix} 1-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = \lambda^2 - 3\lambda + 1 = 0,$$

$$\lambda_1 = \frac{3+\sqrt{5}}{2} \approx 2.618, \quad \lambda_2 = \frac{3-\sqrt{5}}{2} \approx 0.382.$$

С их помощью можно определить границу интервала сходимости τ_0 и оптимальное значение итерационного параметра τ_*:

$$\tau_0 = \frac{2}{\lambda_1} \approx 0.764, \quad \tau_* = \frac{2}{\lambda_1 + \lambda_2} = \frac{2}{3}.$$

Для построения итерационной последовательности выберем какое-нибудь значение итерационного параметра на интервале сходимости, например $\tau = \frac{1}{2}$. В этом случае рекуррентная формула для членов итерационной последовательности принимает вид

$$\boldsymbol{x}^{k+1} = S\boldsymbol{x}^k + \frac{1}{2}\boldsymbol{B}, \text{ где } S = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 0 \end{bmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Возьмем простейшее начальное приближение $x_0 = 0$ и выпишем несколько первых членов итерационной последовательности x_k, подсчитывая для каждого из них невязку $\boldsymbol{\psi}^k$ (6). В результате получим:

$$x_1 = \left\{0, \frac{1}{2}\right\}, \quad \Psi_1 = \left\{\frac{1}{2}, 0\right\}, \quad \|\Psi_1\| = \frac{1}{2},$$

$$x_2 = \left\{-\frac{1}{4}, \frac{1}{2}\right\}, \quad \Psi_2 = \left\{\frac{1}{4}, -\frac{1}{4}\right\}, \quad \|\Psi_2\| = \frac{1}{2\sqrt{2}},$$

$$x_3 = \left\{-\frac{3}{8}, \frac{5}{8}\right\}, \quad \Psi_3 = \left\{\frac{1}{4}, -\frac{1}{8}\right\}, \quad \|\Psi_3\| = \frac{\sqrt{5}}{2},$$

$$x_4 = \left\{-\frac{1}{2}, \frac{11}{16}\right\}, \quad \Psi_4 = \left\{\frac{3}{16}, -\frac{1}{8}\right\}, \quad \|\Psi_4\| = \frac{\sqrt{13}}{16}.$$

Норма невязок хотя и медленно, но убывает, что говорит о сходимости процесса. Это же видно из сравнения членов итерационной последовательности x_k с решением системы. Медленная сходимость связана с плохой обусловленностью матрицы A:

$$M_A = \frac{\lambda_1}{\lambda_2} \approx 6.854.$$

3.4 Метод Зейделя

Разложим матрицу A системы (1) на сумму трех матриц

$$A = D + T_H + T_B, \tag{35}$$

где D — диагональная часть матрицы A, которая содержит элементы a_{ii}, стоящие на главной диагонали:

$$D_{ij} = a_{ii}\delta_{ij} = \begin{cases} a_{ii}, & i = j \\ 0, & i \neq j \end{cases},$$

T_H - нижнетреугольная матрица:

$$\left(T_H\right)_{ij} = \begin{cases} 0, & i \leq j \\ a_{ij}, & i > j \end{cases},$$

T_B - верхнетреугольная матрица:

$$\left(T_B\right)_{ij} = \begin{cases} a_{ij}, & i < j \\ 0, & i \geq j \end{cases}.$$

Запишем итерационный алгоритм (2), полагая

$$C = D + T_H, \quad \tau = 1. \tag{36}$$

Тогда получаем

$$\left(D + T_H\right)\left(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\right) + A\boldsymbol{x}^k = \boldsymbol{B},$$

или

$$\left(D + T_H\right)\boldsymbol{x}^{k+1} + T_B\boldsymbol{x}^k = \boldsymbol{B}. \qquad (37)$$

Рекуррентная формула (37) задает алгоритм Зейделя.

Перейдем от векторной формы записи (37) к покомпонентной:

$$\begin{aligned}
&a_{11}x_1^{k+1} + a_{12}x_2^k + a_{13}x_3^k + \ldots + a_{1n}x_n^k = b_1\\
&a_{21}x_1^{k+1} + a_{22}x_2^{k+1} + a_{23}x_3^k + \ldots + a_{2n}x_n^k = b_2\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{n1}x_1^{k+1} + a_{n2}x_2^{k+1} + a_{n3}x_3^{k+1}\ldots + a_{nn}x_n^{k+1} = b_n
\end{aligned} \qquad (38)$$

Уравнения (38) позволяют последовательно рассчитать компоненты вектора x_i^{k+1}:

$$x_i^{k+1} = \frac{1}{a_{ii}}\left[f_i - \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k \right], \quad i = 1,\ldots,n. \qquad (39)$$

Формула предполагает, что $a_{ii} \neq 0, \quad 1 \le i \le n$.

Алгоритм в методе Зейделя прост и удобен для вычислений. Он не требует никаких действий с матрицей A. Ранее вычисленные на текущей итерации компоненты $x_j^{k+1} \quad \left(j < i\right)$ сразу же участвуют в расчетах наряду с компонентами $x_j^k \quad \left(j > i\right)$ и, таким образом, не требуют дополнительного резерва памяти, что существенно при решении больших систем.

Сходимость метода Зейделя в случае, когда матрица A удовлетворяет условию теоремы Самарского, т.е. является самосопряженной и положительно определенной, будет доказана в следующем разделе.

Задача 2. *Рассмотреть систему (33) (задача 1) и построить для нее приближенное решение с помощью метода Зейделя.*

В рассматриваемом случае рекуррентные формулы (38)

$$
\begin{aligned}
x_1^{k+1} &= -x_2^k, \\
x_2^{k+1} &= \frac{1}{2}\left(1 - x_1^{k+1}\right).
\end{aligned}
\tag{40}
$$

Принимая, как и при решении задачи 2, за начальное приближение нулевой вектор, подсчитаем по формулам (40) несколько первых итераций, сопровождая этот процесс подсчетом невязки:

$$
\boldsymbol{x}^1 = \begin{pmatrix} 0 \\ 1/2 \end{pmatrix}, \quad \boldsymbol{\psi}^1 = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}, \quad \left\|\boldsymbol{\psi}^1\right\| = 1/2,
$$

$$
\boldsymbol{x}^2 = \begin{pmatrix} -1/4 \\ 3/4 \end{pmatrix}, \quad \boldsymbol{\psi}^2 = \begin{pmatrix} 1/4 \\ 0 \end{pmatrix}, \quad \left\|\boldsymbol{\psi}^2\right\| = 1/4,
$$

$$
\boldsymbol{x}^3 = \begin{pmatrix} -3/4 \\ 7/8 \end{pmatrix}, \quad \boldsymbol{\psi}^3 = \begin{pmatrix} 1/8 \\ 0 \end{pmatrix}, \quad \left\|\boldsymbol{\psi}^3\right\| = 1/8,
$$

Первая норма невязки убывает по закону геометрической прогрессии со знаменателем 1/2, т.е. гораздо быстрее, чем в методе простой итерации.

3.5 Модифицированный метод Зейделя

Запишем итерационный алгоритм (2), полагая

$$
C = D + \tau T_H, \quad \tau > 0. \tag{41}
$$

Тогда получаем

$$
\left(\frac{1}{\tau} D + T_H\right)\left(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\right) + A\boldsymbol{x}^k = \boldsymbol{B},
$$

или

$$D\left[\frac{1}{\tau}\boldsymbol{x}^{k+1}+\left(1-\frac{1}{\tau}\right)\boldsymbol{x}^{k}\right]+T_H\boldsymbol{x}^{k+1}+T_B\boldsymbol{x}^{k}=\boldsymbol{B}. \qquad (42)$$

Перейдем от векторной формы записи (42) к покомпонентной:

$$\begin{aligned}
&a_{11}\left[\frac{1}{\tau}x_1^{k+1}+\left(1-\frac{1}{\tau}\right)x_1^{k}\right]+a_{12}x_2^{k}+a_{13}x_3^{k}+\ldots+a_{1n}x_n^{k}=b_1\\
&a_{21}x_1^{k+1}+a_{22}\left[\frac{1}{\tau}x_2^{k+1}+\left(1-\frac{1}{\tau}\right)x_2^{k}\right]+a_{23}x_3^{k}+\ldots+a_{2n}x_n^{k}=b_2\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&a_{n1}x_1^{k+1}+a_{n2}x_2^{k+1}+a_{n3}x_3^{k+1}\ldots+a_{nn}\left[\frac{1}{\tau}x_n^{k+1}+\left(1-\frac{1}{\tau}\right)x_n^{k}\right]=b_n
\end{aligned} \qquad (43)$$

Уравнения (43) позволяют последовательно рассчитать компоненты вектора x_i^{k+1}:

$$x_i^{k+1}=(1-\tau)x_i^{k}+\frac{\tau}{a_{ii}}\left[b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{k+1}-\sum_{j=i+1}^{n}a_{ij}x_j^{k}\right],\quad i=1,\ldots,n. \qquad (44)$$

Исследуем условия сходимости модифицированного метода Зейделя при дополнительном предположении, что матрица A удовлетворяет условиям теоремы Самарского. Самосопряженность матрицы A означает, что

$$T_H^{+}=T_B,\ T_B^{+}=T_H. \qquad (45)$$

В самом деле, учитывая, что $D^{+}=D$, получаем

$$\begin{aligned}
&(A\boldsymbol{x},\boldsymbol{x})=((T_H+D+T_B)\boldsymbol{x},\boldsymbol{x})=\left(\boldsymbol{x},\left(T_H^{+}+D^{+}+T_B^{+}\right)\boldsymbol{x}\right)=\\
&=(\boldsymbol{x},(T_B+D+T_H)\boldsymbol{x})=(\boldsymbol{x},A\boldsymbol{x})
\end{aligned}.$$

Составим для рассматриваемого случая матрицу $C-\frac{\tau}{2}A$. Согласно (41),

$$C-\frac{\tau}{2}A=\left(1-\frac{\tau}{2}\right)D+\frac{\tau}{2}\left(T_н-T_B\right). \qquad (46)$$

Запишем условие ее положительной определенности:

$$\left(\left(C-\frac{\tau}{2}A\right)x,x\right)=\left(1-\frac{\tau}{2}\right)\left(Dx,x\right)+\frac{\tau}{2}\left(\left(T_H-T_B\right)\boldsymbol{x},\boldsymbol{x}\right)>0. \qquad (47)$$

Второе слагаемое в выражении (47) не дает вклада в квадратичную форму в силу соотношения (45).

Матрица A является по предположению положительно определенной. Следовательно, все ее диагональные элементы строго положительны: $a_{ii}>0,\quad 1\le i\le n$. Это означает положительную определенность матрицы D: $\left(D\boldsymbol{x},\boldsymbol{x}\right)>0$. В результате условие положительной определенности выражения (47) - достаточное условие для сходимости итерационной последовательности - принимает вид

$$0<\tau<2. \qquad (48)$$

Метод Зейделя, соответствующий случаю $\omega=1$, удовлетворяет этому условию. Можно поставить вопрос об оптимальном выборе параметра $\tau=\tau_*$, при котором метод сходится быстрее всего. Теоретическое исследование, на котором мы не будем останавливаться, показывает, что такое значение существует. Однако на практике его приходится подбирать экспериментально.

Задача 3. *Построить приближенное решение системы (33) (задача 1) с помощью модифицированного метода Зейделя, полагая* $\tau=4/3$.

Рекуррентное соотношение (44), записанное покомпонентно, принимает вид

$$x_1^{k+1}=-\frac{1}{3}x_1^k-\frac{4}{3}x_2^k,$$

$$x_2^{k+1}=\frac{2}{3}-\frac{2}{3}x_1^{k+1}-\frac{1}{3}x_2^k.$$

Примем, как и в предыдущих случаях, за начальное приближение нулевой вектор и сделаем три итерации. При этом для каждой из них подсчитаем невязку, позволяющую следить за сходимостью процесса:

$$\boldsymbol{x}^1 = \begin{pmatrix} 0 \\ 2/3 \end{pmatrix}, \quad \boldsymbol{\psi}^1 = \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix}, \quad \left\|\boldsymbol{\psi}^1\right\| = \frac{\sqrt{5}}{3} = 0.745,$$

$$\boldsymbol{x}^2 = \begin{pmatrix} -8/9 \\ 28/27 \end{pmatrix}, \quad \boldsymbol{\psi}^2 = \begin{pmatrix} 4/27 \\ 5/27 \end{pmatrix}, \quad \left\|\boldsymbol{\psi}^2\right\| = \frac{\sqrt{41}}{27} = 0.237,$$

$$\boldsymbol{x}^3 = \begin{pmatrix} -88/81 \\ 256/243 \end{pmatrix}, \quad \boldsymbol{\psi}^3 = \begin{pmatrix} -8/243 \\ 5/243 \end{pmatrix}, \quad \left\|\boldsymbol{\psi}^3\right\| = \frac{\sqrt{89}}{243} = 0.039.$$

Поведение невязок показывает сходимость процесса более быструю, чем в классическом методе Зейделя.

Глава 4. Решение систем нелинейных уравнений

В настоящей главе рассматриваются методы решения системы n нелинейных уравнений с n неизвестными

$$\begin{aligned} &F_1(x_1, x_2, ..., x_n) = 0 \\ &F_2(x_1, x_2, ..., x_n) = 0 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &F_n(x_1, x_2, ..., x_n) = 0 \end{aligned}. \tag{1}$$

Иногда нам будет удобно представлять систему (1) в виде

$$F_i(x_1, x_2, ..., x_n), \quad i = 1,2,...,n, \tag{2}$$

или сокращенно

$$\boldsymbol{F}(\boldsymbol{x}) = 0, \tag{3}$$

где $\boldsymbol{x}$ - вектор неизвестных. Геометрически каждое из уравнений системы (1) изображается поверхностью в n-мерном пространстве. Решением системы является точка пересечения этих поверхностей.

4.1 Метод простой итерации

Представим систему (3) в виде

$$\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{x}). \tag{4}$$

Пусть $\boldsymbol{x}^0, \boldsymbol{x}^1, \boldsymbol{x}^2, \ldots$ - итерационная последовательность векторов. Зададим начальное приближение $\boldsymbol{x}^0$. Метод простой итерации описывается алгоритмом

$$\boldsymbol{x}^{k+1} = \boldsymbol{g}(\boldsymbol{x}^k). \tag{5}$$

При решении систем нелинейных уравнений, для описания расстояния между векторами $\boldsymbol{x}$ и $\boldsymbol{y}$, удобно использовать метрику $\rho(\boldsymbol{x}, \boldsymbol{y})$.

Пространство, наделенное метрикой, называется метрическим. Пусть B-полное метрическое пространство. Оператор $\boldsymbol{y = g(x)}$, отображает B в себя. Отображение $\boldsymbol{y = g(x)}$ называется сжимающим, если для любых $\boldsymbol{x}^1, \boldsymbol{x}^2$

$$\rho\left(\boldsymbol{g}\left(\boldsymbol{x}^1\right), \boldsymbol{g}\left(\boldsymbol{x}^2\right)\right) \le q\rho\left(\boldsymbol{x}^1, \boldsymbol{x}^2\right), \quad q < 1. \tag{6}$$

Теорема о сходимости метода простой итерации. Если отображение $\boldsymbol{y} = \boldsymbol{g}\left(\boldsymbol{x}\right)$ - сжимающее, то уравнение $\boldsymbol{x} = \boldsymbol{g}\left(\boldsymbol{x}\right)$ имеет решение $\boldsymbol{x}$ и $\rho\left(\boldsymbol{x}^k, \boldsymbol{x}\right) \le \dfrac{q^k}{1-q}\rho\left(\boldsymbol{x}^1, \boldsymbol{x}^0\right)$, где итерационная последовательность $\boldsymbol{x}^k$ задается соотношением (5).

Доказательство.
Рассмотрим два последовательных вектора итерационной последовательности $\boldsymbol{x}^k, \boldsymbol{x}^{k+1}$. Используя (5), (6), получаем

$$\rho\left(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k\right) = \rho\left(g\left(\boldsymbol{x}^k\right), g\left(\boldsymbol{x}^{k-1}\right)\right) \le q\rho\left(\boldsymbol{x}^k, \boldsymbol{x}^{k-1}\right) \le \ldots \le q^k\rho\left(\boldsymbol{x}^1, \boldsymbol{x}^0\right). \tag{7}$$

Рассмотрим теперь два вектора итерационной последовательности $\boldsymbol{x}^m, \boldsymbol{x}^k$, такие, что $m > k$. Используя неравенство многоугольника и соотношение (7), получаем

$$\begin{aligned}&\rho\left(\boldsymbol{x}^m, \boldsymbol{x}^k\right) \le \rho\left(\boldsymbol{x}^m, \boldsymbol{x}^{m-1}\right) + \rho\left(\boldsymbol{x}^{m-1}, \boldsymbol{x}^{m-2}\right) + \ldots + \rho\left(\boldsymbol{x}^{k+1}, \boldsymbol{x}^k\right) \le \\ &\le \left(q^{m-1} + \ldots + q^k\right)\rho\left(\boldsymbol{x}^1, \boldsymbol{x}^0\right) \le q^k\left(1 + q + \ldots\right)\rho\left(\boldsymbol{x}^1, \boldsymbol{x}^0\right)\end{aligned}.$$

Окончательно

$$\rho\left(\boldsymbol{x}^m, \boldsymbol{x}^k\right) \le \frac{q^k}{1-q}\rho\left(\boldsymbol{x}^1, \boldsymbol{x}^0\right). \tag{8}$$

Из полученного неравенства видно, что $\rho\left(\boldsymbol{x}^k, \boldsymbol{x}^m\right) \to 0$ при $k \to \infty$. Согласно критерию Коши, это означает, что существует предел $\boldsymbol{x}$ итерационной последовательности $\boldsymbol{x}^k$. Покажем, что $\boldsymbol{x} = \boldsymbol{g}\left(\boldsymbol{x}\right)$, т.е. предел итерационной последовательности является корнем уравнения (4). Переходя

в (8) к пределу при $m \to \infty$, получаем

$$\rho\left(\boldsymbol{x}, \boldsymbol{x}^k\right) \le \frac{q^k}{1-q}\rho\left(\boldsymbol{x}^1, \boldsymbol{x}^0\right). \tag{9}$$

Найдем расстояние между $\boldsymbol{x}$ и $\boldsymbol{g}(\boldsymbol{x})$. Если $\boldsymbol{x}$ действительно является решением уравнения (4), то это расстояние должно быть равно нулю.
Используя неравенство треугольника и рекуррентное соотношение (5), получаем

$$\rho\left(\boldsymbol{x}, \boldsymbol{g}(\boldsymbol{x})\right) \le \rho\left(\boldsymbol{x}, \boldsymbol{x}^{k+1}\right) + \rho\left(\boldsymbol{x}^{k+1}, \boldsymbol{g}(\boldsymbol{x})\right) = \rho\left(\boldsymbol{x}, \boldsymbol{x}^{k+1}\right) + \rho\left(\boldsymbol{g}\left(\boldsymbol{x}^k\right), \boldsymbol{g}(\boldsymbol{x})\right).$$

Воспользуемся теперь неравенствами (6) и (8). Тогда имеем

$$\rho\left(\boldsymbol{x}, \boldsymbol{g}(\boldsymbol{x})\right) \le \rho\left(\boldsymbol{x}, \boldsymbol{x}^{k+1}\right) + q\rho\left(\boldsymbol{x}^k, \boldsymbol{x}\right) \le \frac{2q^{k+1}}{1-q}\rho\left(\boldsymbol{x}^1, \boldsymbol{x}^0\right). \tag{10}$$

Неравенство (10) справедливо при любых значениях k. Переходя к пределу $k \to \infty$, получаем $\rho(\boldsymbol{x}, \boldsymbol{g}(\boldsymbol{x})) = 0$. Следовательно, $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{x})$, что и требовалось доказать.

Замечание.
Исследуем, насколько удаляются друг от друга точки итерационной последовательности. Первая точка $\boldsymbol{x}^1$ удалена от начальной точки $\boldsymbol{x}^0$ на расстояние $\rho(\boldsymbol{x}^1, \boldsymbol{x}^0)$. Положим в неравенстве (8) $k = 0$. Тогда получаем

$$\rho\left(\boldsymbol{x}^m, \boldsymbol{x}^0\right) \le \frac{1}{1-q}\rho\left(\boldsymbol{x}^1, \boldsymbol{x}^0\right). \tag{11}$$

Таким образом, все элементы итерационной последовательности лежат в шаре с центром $\boldsymbol{x}^0$. Радиус шара в $1/(1-q)$ раз превышает расстояние между первой и нулевой точками последовательности.

4.2 Метод Ньютона

Рассмотрим исходную систему нелинейных уравнений в форме (2). Линеаризуем эту систему, разложив каждое из уравнений в ряд Тейлора около точки $\boldsymbol{x}^k$, удержав члены нулевого и первого порядка малости:

$$F_i\left(\boldsymbol{x}^k\right)+\sum_{j=1}^{n}\frac{\partial F_i\left(\boldsymbol{x}^k\right)}{\partial x_j}\left(x_j-x_j^k\right)=0,\quad i=1,2,...,n. \tag{12}$$

Геометрически это означает замену поверхностей, изображающих уравнения системы (2), на соответствующие касательные плоскости. Систему линейных уравнений (12) запишем в векторном виде

$$\boldsymbol{F}\left(\boldsymbol{x}^k\right)+F'\left(\boldsymbol{x}^k\right)\left(\boldsymbol{x}-\boldsymbol{x}^k\right)=0, \tag{13}$$

где $F'\left(\boldsymbol{x}^k\right)$ - матрица с коэффициентами $\frac{\partial F_i\left(\boldsymbol{x}^k\right)}{\partial x_j}$, $i,j=1,2,...,n$.

Считая эту матрицу невырожденной и вводя соответствующую обратную матрицу $\left(\mathrm{F}'(\mathbf{x}^{\mathrm{k}})\right)^{-1}$, запишем решение системы (13) в виде

$$\boldsymbol{x}=\boldsymbol{x}^k-\left(F'\left(\boldsymbol{x}^k\right)\right)^{-1}\boldsymbol{F}\left(\boldsymbol{x}^k\right). \tag{14}$$

Рассматривая решение (14), как $k+1$ член итерационной последовательности, получаем рекуррентное соотношение метода Ньютона

$$\boldsymbol{x}^{k+1}=\boldsymbol{x}^k-\left(F'\left(\boldsymbol{x}^k\right)\right)^{-1}\boldsymbol{F}\left(\boldsymbol{x}^k\right). \tag{15}$$

Исследуем сходимость метода Ньютона.

Пусть $\Omega_a=\left\{\boldsymbol{x}:\left\|\boldsymbol{x}-\boldsymbol{X}\right\|<a\right\}$ - открытый шар радиуса a с центром в точке $\boldsymbol{X}$.

Пусть при некоторых $a>0,\ a_1\geq 0,\ a_2<\infty$ выполнены условия

1) $\left\|\left(F'\left(\boldsymbol{x}\right)\right)^{-1}\right\|\leq a_1$ при $\boldsymbol{x}\in\Omega_a$,

2) $\left\|\boldsymbol{F}\left(\boldsymbol{u}_1\right)-\boldsymbol{F}\left(\boldsymbol{u}_2\right)-F'\left(\boldsymbol{u}_2\right)\left(\boldsymbol{u}_1-\boldsymbol{u}_2\right)\right\|\leq a_2\left\|\boldsymbol{u}_1-\boldsymbol{u}_2\right\|^2$ при $\mathbf{u}_1,\mathbf{u}_2\in\Omega_{\mathrm{a}}$,

Обозначим $c=a_1a_2,\ b=min\left(a,c^{-1}\right)$.

Теорема о сходимости метода Ньютона.

При условиях 1, 2 и $\boldsymbol{x}^0\in\Omega_b$ итерационный процесс Ньютона сходится к

решению $\boldsymbol{x}$ с оценкой погрешности $\left\|\boldsymbol{x}^k-\boldsymbol{x}\right\| \le c^{-1}\left(c\left\|\boldsymbol{x}^0-\boldsymbol{x}\right\|\right)^{2^k}$.

Доказательство.

1. Пусть $\boldsymbol{x}^0 \in \Omega_b$. Индукцией по k докажем, что все $\boldsymbol{x}^k \in \Omega_b$.

Пусть это утверждение верно при некотором k. Докажем, что оно верно для $\boldsymbol{x}^{k+1}$, т.е. докажем, что из $\boldsymbol{x}^k \in \Omega_b$ следует $\boldsymbol{x}^{k+1} \in \Omega_b$. Так как $b = min\left(a, c^{-1}\right) \le a$, то из $\boldsymbol{x}^k \in \Omega_b$ вытекает $\boldsymbol{x}^k \in \Omega_a$, т.е. $\Omega_b \subset \Omega_a$. Используя условие 2 настоящей теоремы, запишем

$$\left\|\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{F}\left(\boldsymbol{x}^k\right)-F'\left(\boldsymbol{x}^k\right)\left(\boldsymbol{x}-\boldsymbol{x}^k\right)\right\| \le a_2\left\|\boldsymbol{x}-\boldsymbol{x}^k\right\|^2. \tag{16}$$

Поскольку $\boldsymbol{F}\left(\boldsymbol{x}^k\right) = -F'\left(\boldsymbol{x}^k\right)\left(\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\right) = 0$, $\boldsymbol{F}(\boldsymbol{x}) = 0$, неравенство (16) можно переписать в виде

$$\left\|F'\left(\boldsymbol{x}^k\right)\left(\boldsymbol{x}^{k+1}-\boldsymbol{x}\right)\right\| \le a_2\left\|\boldsymbol{x}-\boldsymbol{x}^k\right\|^2. \tag{17}$$

Введем вектор $\boldsymbol{y} = F'\left(\boldsymbol{x}^k\right)\left(\boldsymbol{x}^{k+1}-\boldsymbol{x}\right)$. Тогда $\boldsymbol{x}^{k+1}-\boldsymbol{x} = \left(F'\left(\boldsymbol{x}^k\right)\right)^{-1}\boldsymbol{y}$. Следовательно

$$\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}\right\| \le \left\|\left(F'\left(\boldsymbol{x}^k\right)\right)^{-1}\right\| \cdot \|\boldsymbol{y}\|.$$

Учтем условие 1 настоящей теоремы. Тогда получаем

$$\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}\right\| \le a_1\|\boldsymbol{y}\|.$$

Используя (17), имеем

$$\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}\right\| \le \|\boldsymbol{y}\| \le a_1 a_2\left\|\boldsymbol{x}^k-\boldsymbol{x}\right\|^2. \tag{18}$$

Учитывая, что $a_1 a_2 = c$, $\left\|\boldsymbol{x}^k-\boldsymbol{x}\right\| < b$, представим (18) в виде

$$\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}\right\| \le \|\boldsymbol{y}\| \le cb^2. \tag{19}$$

Поскольку $b = min\left(a, c^{-1}\right) \le c^{-1}$,

$$cb \le 1, \tag{20}$$

видим, что $cb^2 = cb \cdot b \le cb \cdot c^{-1} = b$. Тогда, (19) принимает вид

$$\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}\right\| \le b,$$

следовательно $\boldsymbol{x}^{k+1} \in \Omega_b$. Что и требовалось доказать. Таким образом, при $\boldsymbol{x}^0 \in \Omega_b$, все $\boldsymbol{x}^k \in \Omega_b$.

2. Полагая в (18) $a_1 a_2 = c$, получаем

$$\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}\right\| \le c\left\|\boldsymbol{x}^{k}-\boldsymbol{x}\right\|^2. \tag{21}$$

Пусть

$$q_k = c\left\|\boldsymbol{x}^{k}-\boldsymbol{x}\right\|. \tag{22}$$

Тогда (21) принимает вид

$$q_{k+1} \le q_k^2. \tag{23}$$

Индукцией по k докажем справедливость неравенства

$$q_k \le q_0^{2^k}. \tag{24}$$

При $k = 0$ это неравенство очевидно. Пусть оно верно при некотором k. Докажем, что оно верно для q_{k+1}:

$$q_{k+1} \le q_k^2 \le \left(q_0^{2^k}\right)^2 = q_0^{2^{k+1}}.$$

Таким образом, (24) верно при любом k. Это означает, что

$$\left\|\boldsymbol{x}^k - \boldsymbol{x}\right\| \le c^{-1}\left(c\left\|\boldsymbol{x}^0 - \boldsymbol{x}\right\|\right)^{2^k}, \tag{25}$$

что и требовалось доказать.

3. Согласно (24), для того, чтобы итерационный процесс сходился, требуется, чтобы

$$c\left\|\boldsymbol{x}^0 - \boldsymbol{x}\right\| < 1. \tag{26}$$

Учитывая, что

$$\left\|\boldsymbol{x}^0 - \boldsymbol{x}\right\| < b,$$

умножая последнее неравенство на c и используя (20), получаем (26).

4.3 Модифицированный метод Ньютона

Классический метод Ньютона обеспечивает большую скорость сходимости итерационного процесса. Однако при этом на каждом итерационном шаге необходимо решать систему линейных уравнений.

В отличие от классического, модифицированный метод Ньютона использует фиксированную матрицу $F' = \left\{\dfrac{\partial F_i\left(\boldsymbol{x}^0\right)}{\partial x_j},\quad i,j = 1,2,\ldots,n\right\}$, которая рассчитывается по начальному приближению $\boldsymbol{x}^0$. Эта матрица обращается один раз, и полученная обратная матрица используется на всех шагах итерационного процесса

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \left(F'\left(\boldsymbol{x}^0\right)\right)^{-1}\boldsymbol{F}\left(\boldsymbol{x}^k\right). \tag{27}$$

Для реализации такого алгоритма не требуется решать систему на каждом шаге итерационного процесса. Однако ценой такого упрощения является более низкая скорость сходимости метода.

4.4 Метод Зейделя

Вариант метода Зейделя, применимый к решению системы нелинейных уравнений (1), описывается следующим рекуррентным алгоритмом

$$\begin{aligned} &F_1\left(x_1^{k+1}, x_2^k, \ldots, x_n^k\right) = 0 \\ &F_2\left(x_1^{k+1}, x_2^{k+1}, \ldots, x_n^k\right) = 0 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &F_n\left(x_1^{k+1}, x_2^{k+1}, \ldots, x_n^{k+1}\right) = 0 \end{aligned} . \qquad (28)$$

Таким образом, нахождение нового значения x_i^{k+1} требует решения одного уравнения

$$F_i\left(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i^{k+1}, x_{i+1}^k \ldots, x_n^k\right) = 0. \qquad (29)$$

Уравнение решается одним из методов, развитых для нелинейных уравнений с одним неизвестным.

Глава 5. Минимизация функций

Задачи о нахождение минимума функций одной или многих переменных являются весьма распространенными. Развитые для этой цели методы позволяют также находить решения систем уравнений. В самом деле, система нелинейных уравнений (4.1) позволяет определить функцию

$$\Phi(\boldsymbol{x}) = \sum_{i=1}^{n} \left(F_i(\boldsymbol{x})\right)^2 . \qquad (1)$$

Минимум функции $\Phi(\boldsymbol{x})$ соответствует решению системы (4.1).

Метода нахождения минимума разделяют на методы 0-го, 1-го, 2-го и т.д. порядка. При этом методы 0-го порядка для нахождения минимума функции используют лишь значения этой функции. Методы 1-го порядка кроме вычисления функции требуют расчета ее первой производной, методы 2-го порядка – функции, первой и второй производной и т.д. К сожалению, большинство методов позволяют найти лишь локальный минимум, ничего не говоря о глобальном минимуме.
Методы, развитые для нахождения минимума функций, разумеется, пригодны и для нахождения их максимума.

5.1 Нахождение минимума функции одной переменной

Проиллюстрируем основные идеи минимизации функций ни примере функций одной переменной $f(x)$.

а. Метод перебора. Метод нулевого порядка нахождения минимума функции одной переменной в некотором интервале значений аргумента. Рассчитываются значения функции для некоторого набора значений аргумента (например, с постоянным шагом). Среди найденных значений функции выбирается минимальной. Метод позволяет найти глобальный минимум функции.

б. Симплекс-метод. Метод нулевого порядка, основанный на построении симплекса. В одномерном случае симплекс сводится к отрезку. Пусть начальный отрезок имеет координаты x^0, x^1. Рассчитаем значения минимизируемой функции в этих двух точках. Для построения нового отрезка точку, в которой функция имеет большее из двух значений, отбросим. Вместо нее выберем точку, симметрично отраженную от второй точки. Теперь отрезок состоит из одной новой точки и одной старой.

Рассчитаем значение функции в новой точке. Сравним его с уже рассчитанным значением в старой точке. И так далее. Описанный алгоритм приводит к перемещению отрезка в сторону уменьшения функции. Если отрезок приходит в «колебательный режим», это означает, что мы достигли минимума с точностью, задаваемой длиной нашего отрезка. Метод требует меньшего числа вычислений, чем метод перебора, но позволяет найти лишь локальный минимум.

в. Градиентный метод. Метод первого порядка, основанный на том, что при положительной производной $f'(x)$ функция $f(x)$ возрастает, а при и отрицательной – убывает. Построим итерационную последовательность x^k:

$$x^{k+1} = x^k - \varepsilon f'(x^k), \qquad (2)$$

где $\varepsilon > 0$ - малый параметр. Видно, что при положительной производной $x^{k+1} < x^k$, а при отрицательной - $x^{k+1} > x^k$, т.е. каждый следующий элемент последовательности ближе к минимуму, чем предыдущий. Метод позволяет найти лишь локальный минимум.

г. Метод Ньютона. Метод второго порядка. Для построения итерационной последовательности, разложим функцию $f(x)$ в ряд Тейлора около значения аргумента x^k до членов второго порядка:

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

Мы заменили исследуемую функцию параболой, имеющей в точке x^k то же значение, наклон и кривизну, что и исследуемая функция. Найдем минимум квадратичной функции (3)

$$y' = f'(x^k) + f''(x^k)(x - x^k) = 0, \quad x = x^k - \frac{f'(x^k)}{f''x^k}. \qquad (3)$$

и будем рассматривать его как новый элемент итерационной последовательности:

$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)}. \qquad (4)$$

Построенный итерационный сходится быстрее, чем описанный выше градиентный метод.

5.2 Нахождение минимума функции многих переменных

Рассмотрим основные методы минимизации функций многих переменных

$$y = f(x_1,...,x_n) \equiv f(\boldsymbol{x}).$$

а. Метод перебора. Как и в случае функций одной переменной, метод сводится к расчету набора значений функции в некоторой области и выбору минимального значения. Метод позволяет найти глобальный минимум функции. Для задач с высокой размерностью приводит к недопустимо большому количеству вычислений.

б. Симплекс-метод. Это своеобразный метод нулевого порядка, основанный на построении симплекса – множества равноудаленных точек, в количестве на единицу превышающем размерность пространства. В двумерном случае симплекс – это равносторонний треугольник. В трехмерном случае – правильная треугольная пирамида. На начальном шаге итерационного процесса даются координаты исходного симплекса и в них рассчитываются значения минимизируемой функции. Среди вершин симплекса находится та, в которой функция имеет наибольшее значение. Для построения нового симплекса эта вершина отбрасывается. Вместо нее выбирается новая вершина, симметрично отраженная от плоскости, проведенной через остальные вершины. В новой вершине рассчитывается значение функции. В старых же вершинах, вошедших в новый симплекс, значения функции уже известны. Снова находится вершина, в которой функция имеет наибольшее значение. И так далее. Ключевым моментом является то, что на каждом шаге итерационного процесса требуется расчет функции лишь в одной точке. Для минимизации функций в многомерных пространствах это оказывается очень важным.

в. Градиентный метод. Метод первого порядка, основанный на том факте, что градиент функции направлен в направлении ее наибольшего возрастания. Соответственно, двигаясь в сторону противоположную градиенту, мы приближаемся к минимуму функции. Итерационная последовательность градиентного метода имеет вид

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \varepsilon \cdot grad\left(f\left(\boldsymbol{x}^k\right)\right). \qquad (5)$$

г. Метод Ньютона. Метод второго порядка. Разложим функцию $f(\boldsymbol{x})$ в ряд Тейлора около значения аргумента $\boldsymbol{x}^k$ до членов второго порядка:

$$y = f(\boldsymbol{x}^k) + \sum_{i=1}^{n} \frac{\partial f(\boldsymbol{x}^k)}{\partial x_i^k}\left(x_i - x_i^k\right) + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f\boldsymbol{x}^k}{\partial x_i^k \partial x_j^k}\left(x_i - x_i^k\right)\left(x_j - x_j^k\right). \tag{6}$$

Найдем минимум квадратичной формы (6)

$$\frac{\partial y}{\partial x_m} = \frac{\partial f(\boldsymbol{x}^k)}{\partial x_m^k} + \sum_{i}^{n} \frac{\partial^2 f(\boldsymbol{x}^k)}{\partial x_i^k \partial x_m^k}\left(x_i - x_i^k\right) = 0. \tag{7}$$

Введем матрицу $f''(\boldsymbol{x}^k) = \left\{\frac{\partial^2 f(\boldsymbol{x}^k)}{\partial x_i^k \partial x_m^k}\right\}_{i,m=1}^{n}$ и вектор градиента $\boldsymbol{f}'(\boldsymbol{x}^k) = \left\{\frac{\partial f(\boldsymbol{x}^k)}{\partial x_m^k}\right\}_{m=1}^{n}$. Запишем систему уравнений в виде

$$f'' \cdot \left(\boldsymbol{x} - \boldsymbol{x}^k\right) = -\boldsymbol{f}'. \tag{8}$$

Для того чтобы функция $f(\boldsymbol{x})$ имела минимум в окрестности точки $\mathbf{x}^{\mathrm{k}}$, матрица $f''(\boldsymbol{x}^k)$ должна быть положительно определенной, а, следовательно, невырожденной. Вводя обратную матрицу $\left(f''(\boldsymbol{x}^k)\right)^{-1}$, запишем решение системы (8) в виде

$$\boldsymbol{x} = \boldsymbol{x}^k - \left(f''(\boldsymbol{x}^k)\right)^{-1} \boldsymbol{f}'(\boldsymbol{x}^k). \tag{9}$$

Если бы функция $f(\boldsymbol{x})$ была квадратичной, мы за один шаг получили бы точное положение ее минимума. В общем случае мы будем рассматривать **x** лишь как новый шаг итерационной последовательности

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \left(f''(\boldsymbol{x}^k)\right)^{-1} \boldsymbol{f}'(\boldsymbol{x}^k). \tag{10}$$

Построенный итерационный сходится быстрее, чем описанный выше градиентный метод, однако требует на каждом шаге решения системы линейных алгебраических уравнений.

д. Метод покоординатного спуска. Напоминает метод Зейделя решения нелинейных систем уравнений. На первом итерационном шаге фиксируются все переменные кроме первой. По выбранной переменной находится минимум, используя методы минимизации функций одной переменной. На втором итерационном шаге находится минимум функции по второй переменной, и т.д. После n шагов итерации мы вновь возвращаемся к первой переменной.

Глава 6. Интерполяция функций

Пусть функция $y = f(x)$ задана на интервале $[a,b]$, но ее значения известны лишь в конечном наборе точек. Требуется восстановить функцию по этой информации. Точнее, требуется найти такую функцию $F(x)$, которая на заданном наборе точек $x_0, x_1,\ldots,x_n$ принимала бы те же значения, что и исходная функция, а в остальных точках интервала была бы близка к ней. Такая функция $F(x)$ называется интерполирующей, а точки $x_i, i = 0,1,\ldots,n$ - узлами интерполяции.

Подобные задачи часто возникают на практике, например при обработке экспериментальных данных, когда значение переменной y, зависящей от x, измеряется в конечном числе точек или при работе с табличными функциями, если требуется вычислить $y = f(x)$ при значениях аргумента, не совпадающего ни с одним из табличных значений.

Ограничимся рассмотрением линейной интерполяции. В этом случае интерполирующая функция ищется в виде линейной комбинации некоторого набора функций $\varphi_0(x),\varphi_1(x),\ldots,\varphi_n(x)$:

$$F(x) = \sum_{i=0}^{n} c_i \varphi_i(x), \tag{1}$$

Коэффициенты c_i подбираются так, чтобы интерполирующая функция совпадала с исходной в узлах интерполяции:

$$\sum_{i=0}^{n} c_i \varphi_i(x_j) = f(x_j), \quad j = 0,1,\ldots,n. \tag{2}$$

Отсюда видно, что количество коэффициентов c_i должно равняться количеству узлов интерполяции. Для того, чтобы система линейных уравнений (2) была разрешимой относительно коэффициентов c_i, ее определитель должен быть отличен от нуля:

$$\begin{vmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \ldots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_n(x_1) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \cdots & \varphi_n(x_n) \end{vmatrix} \neq 0. \tag{3}$$

Необходимым (но не достаточным) условием для этого является линейная независимость набора функций $\varphi_0(x), \varphi_1(x), \ldots, \varphi_n(x)$. В частности, широкое распространение получило интерполирование с помощью степенных функций:

$$\varphi_0(x) = 1, \quad \varphi_1(x) = x, \quad \varphi_2(x) = x^2, \ldots, \varphi_n(x) = x^n .$$

В этом случае интерполирующая функция представляет собой полином степени n:

$$F(x) = P_n(x) = \sum_{i=0}^{n} c_i x^i \tag{4}$$

с неизвестными коэффициентами $c_i, \quad i = 0,1,\ldots,n$. Тогда уравнение (2) для нахождения коэффициентов c_i приобретает вид

$$\sum_{i=0}^{n} c_i x_j^i = f(x_j), \quad j = 0,1,\ldots,n . \tag{5}$$

Определителем этой системы является определитель Ван-дер-Монда:

$$\begin{vmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{vmatrix} .$$

В нашем случае этот определитель отличен от нуля, поскольку все узлы интерполирования различны между собой.

6.1. Интерполяционный полином Лагранжа

Задача линейной интерполяции полиномами требует решения системы линейных уравнений (5) для определения коэффициентов полиномов. Лагранж предложил способ построения интерполирующих полиномов, не требующий решения системы уравнений.

Рассмотрим сначала простейший пример. Пусть значения функции $f(x)$ известны в двух точках x_0, x_1. Построить линейную функцию, принимающую в этих точках значения $f(x_0), f(x_1)$. Построим сначала две вспомогательные линейные функции

$$Q_{1,0}(x) = \frac{x - x_1}{x_0 - x_1}, \quad Q_{1,1}(x) = \frac{x - x_0}{x_1 - x_0}. \tag{6}$$

Первая из функций $Q_{1,0}(x)$ равна единице в точке x_0 и нулю в точке x_1. Вторая же функция $Q_{1,1}(x)$ равна единице в точке x_1 и нулю в точке x_0. Тогда искомый полином запишем в виде:

$$P_1(x) = f(x_0)Q_{1,0}(x) + f(x_1)Q_{1,1}(x). \tag{7}$$

Видно, что построенный полином имеет первый порядок и принимает в точках x_0 и x_1 требуемые значения.

Для того чтобы построить интерполяционный полином второго порядка методом Лагранжа, введем три вспомогательные квадратичные функции

$$\begin{aligned} Q_{2,0}(x) &= \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \; Q_{2,1}(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}, \\ Q_{2,2}(x) &= \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} \end{aligned}. \tag{8}$$

Каждая из этих функций равна единице в одном из узлов, а в двух других равна нулю. Искомый полином может быть представлен в виде:

$$P_2(x) = f(x_0)Q_{2,0}(x) + f(x_1)Q_{2,1}(x) + f(x_3)Q_{2,3}(x). \tag{9}$$

Рассмотрим, наконец, построение полинома Лагранжа n- го порядка. Для этого введем набор полиномов n- го порядка:

$$Q_{n,0}(x)=\frac{(x-x_1)(x-x_2)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_n)},$$

$$Q_{n,i}(x)=\frac{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}, \qquad (10)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$Q_{n,n}(x)=\frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})}.$$

или

$$Q_{n,i}(x)=\prod_{\substack{j=0\\ j\neq i}}^{j=n}\frac{(x-x_j)}{(x_i-x_j)}. \qquad (11)$$

Видно, что $Q_{n,i}(x)$ - полиномы степени n, такие, что

$$Q_{n,i}(x)=\begin{cases}1, & x=x_i\\ 0, & x=x_j,\quad j\neq i\end{cases}, \qquad (12)$$

Иногда нам будет удобно записывать $Q_{n,i}(x)$ в виде

$$Q_{n,i}(x)=\frac{(x-x_0)\ldots[i]\ldots(x-x_n)}{(x_i-x_0)\ldots[i]\ldots(x_i-x_n)}. \qquad (13)$$

Искомый полином $P_n(x)$ представим в виде

$$P_n(x)=\sum_{i=0}^{n}f(x_i)Q_{n,i}(x), \qquad (14)$$

Из (10) и (14) видно, что построенный полином $P_n(x)$ действительно является интерполяционным полиномом для функции $f(x)$на сетке с

узлами $x_0, x_1, \ldots, x_n$. Его принято называть интерполяционным полиномом в форме Лагранжа. Возможны и другие эквивалентные представления интерполяционного полинома $P_n(x)$. С одним из них мы познакомимся в следующем разделе.

6.2. Интерполяционный полином Ньютона

Интерполяционный полином в форме Лагранжа неудобен для вычислений тем, что при увеличении числа узлов интерполяции приходится перестраивать весь полином заново. Перепишем интерполяционный полином Лагранжа в эквивалентной форме:

$$P_n(x) = P_0(x) + \sum_{i=1}^{n} \left(P_i(x) - P_{i-1}(x)\right), \quad n \geq 1, \tag{15}$$

где $P_i(x)$- полиномы Лагранжа степени $i \leq n$, соответствующие узлам интерполирования $x_0, x_1, \ldots, x_i$. Здесь

$$P_0(x) \equiv f(x_0) \tag{16}$$

- полином нулевой степени. Полином

$$Q_i(x) = P_i(x) - P_{i-1}(x) \tag{17}$$

имеет степень i и по построению обращается в нуль при $x = x_0, x = x_1, \ldots, x = x_{i-1}$, поэтому его можно представить в виде

$$Q_i(x) = A_i(x - x_0)(x - x_1)\ldots(x - x_{i-1}), \tag{18}$$

где A_i- числовой коэффициент при x^i. Поскольку $P_{i-1}(x)$не содержит степени x^i, то A_i просто совпадает с коэффициентом при x^i в полиноме $P_i(x)$. Согласно (10) и (14), его можно записать в виде

$$A_i = \sum_{k=0}^{i} \frac{f(x_k)}{\omega_{k,i}}, \tag{19}$$

где

$$\omega_{k,i} = (x_k - x_0)...(x_k - x_{k-1})(x_k - x_{k+1})...(x_k - x_i). \qquad (20)$$

При этом

$$A_0 = f(x_0). \qquad (21)$$

Формулы (17) и (18) позволяют написать рекуррентное соотношение для полинома $P_n(x)$:

$$P_n(x) = P_{n-1}(x) + A_n(x - x_0)...(x - x_{n-1}). \qquad (22)$$

Используя (16), (22), получаем окончательную формулу для полинома $P_n(x)$:

$$\begin{aligned} P_n(x) &= A_0 + A_1(x - x_0) + A_2(x - x_0)(x - x_1) + ... \\ &+ A_i(x - x_0)...(x - x_{i-1}) + ...A_n(x - x_0)...(x - x_{n-1}) \end{aligned}. \qquad (23)$$

Представление (23) удобно для вычислителя, поскольку увеличение n на единицу требует только добавления к «старому» многочлену одного дополнительного слагаемого. Такое представление интерполяционного полинома $P_n(x)$ называют интерполяционным полиномом в форме Ньютона.

6.3. Погрешность интерполяции

Оценим отклонение интерполяционного полинома $P_n(x)$ от интерполируемой функции $f(x)$ на отрезке $[a,b]$. Для этого введем погрешность (остаточный член)

$$R_n(x) = f(x) - P_n(x), \quad x \in [a,b]. \qquad (24)$$

По определению интерполяционного полинома

$$R_n(x_i) = 0, \quad i = 0,1,...,n, \qquad (25)$$

поэтому речь идет об оценке $R_n(x)$ при значениях $x \neq x_i$.

Для того чтобы это сделать, введем дополнительно предположение о гладкости функции $f(x)$. Пусть эта функция имеет непрерывную производную порядка $n+1$ на отрезке $[a,b]$.

В силу (25) $R_n(x)$ можно представить в виде

$$R_n(x) = \omega_{n+1}(x) r_n(x), \tag{26}$$

где $\omega_{n+1}(x)$ - полином $n+1$-й степени:

$$\omega_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n). \tag{27}$$

Зафиксируем произвольное значение $x \in [a,b]$ и рассмотрим вспомогательную переменную t и функцию от нее:

$$g(t) = f(t) - P_n(t) - \omega_{n+1}(t) r_n(x),$$

заданную на отрезке $[a,b]$ и содержащую переменную x в качестве параметра. В силу своего определения функция $g(t)$ должна обращаться в нуль в узлах интерполирования $t = x_i$. Кроме того, согласно (24), (26), $g(t)$ должна обращаться в нуль при $t = x$. Таким образом, $g(t)$ имеет по крайней мере $n+2$ нуля:

$$g(x_i) = 0, \quad i = 0,1,\ldots,n, \quad g(x) = 0. \tag{28}$$

Если $x \in [x_0, x_n]$, то все эти нули также лежат на отрезке $[x_0, x_n]$. Если $x < x_0$, то эти нули, вообще говоря, принадлежат отрезку $[x, x_n]$, а если $x > x_n$, то они находятся на отрезке $[x_0, x]$. Объединяя эти три случая, скажем, что указанные нули функции $g(t)$ принадлежат отрезку $[\alpha,\beta]$, где $\alpha = min(x_0, x) \geq a, \quad \beta = max(x_n, x) \leq b$.

Согласно теореме Ролля, можно утверждать, что производная $g'(t)$ имеет по крайней мере $n+1$ нуль на отрезке $[\alpha,\beta]$ (эти нули перемежаются с нулями самой функции $g(t)$. Повторяя это рассуждение, видим, что $g''(t)$ имеет по крайней мере n нулей на отрезке $[\alpha,\beta]$ и, наконец, $g^{(n+1)}(t)$ обращается хотя бы один раз в нуль в некоторой точке $t = \xi \in [\alpha,\beta]$, т.е.

$$g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - P_n^{(n+1)}(\xi) - (n+1)! \cdot r_n(x) = 0.$$

Учитывая, что $(n+1)$-я производная полинома степени n тождественно равна нулю, получаем, что

$$r_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad \xi \in [\alpha, \beta]. \tag{29}$$

и, согласно (26),

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x). \tag{30}$$

Формула (30) не позволяет вычислить погрешность, поскольку точное значение аргумента ξ нам неизвестно. Однако с ее помощью погрешность можно оценить:

$$|R_n(x)| \le \frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)|. \tag{31}$$

где

$$M_{n+1} = \max_{x \in [\alpha,\beta]} |f^{(n+1)}(x)| \le \max_{x \in [a,b]} |f^{[n+1]}(x)|. \tag{32}$$

Обсудим роль полинома $\omega_{n+1}(x)$ (27) в оценке (31). На отрезке $[x_0, x_n]$ он имеет $(n+1)$ нуль, а его значения между этими нулями сравнительно невелики, но когда точка x выходит за пределы отрезка $[x_0, x_n]$ и удаляется от точки x_0 влево или от точки x_n вправо, оценка (31) ухудшается из-за быстрого роста функции $|\omega_{n+1}(x)|$. Итак, если $x \in [x_0, x_n]$, то множитель $|\omega_{n+1}(x)|$ не ухудшает оценку (31). Такой случай называют собственно интерполяцией функции $f(x)$. Противоположный случай, когда точка x лежит вне отрезка, называют экстраполяцией функции $f(x)$. Отмеченная выше особенность поведения полинома $\omega_{n+1}(x)$ резко ухудшает оценку (32) при экстраполяции.

6.4 Сходимость интерполяционного процесса

Поставим вопрос: будут ли сходиться интерполяционные полиномы $P_n(x)$ к интерполируемой функции $f(x)$ на отрезке $[a,b]$ при неограниченном возрастании числа узлов n?

Упорядоченное множество точек $x_i, \quad i = 0,1,...,n$, назовем сеткой на отрезке $[a,b]$ и обозначим Ω_n. Рассмотрим последовательность сеток с возрастающим числом узлов

$$\Omega_0 = \left\{x_0^{(0)}\right\}, \quad \Omega_1 = \left\{x_0^{(1)}, x_1^{(1)}\right\},..., \quad \Omega_n = \left\{x_0^{(n)}, x_1^{(n)},..., x_n^{(n)}\right\},...$$

и отвечающую ей последовательность интерполяционных полиномов $P_n(x)$, построенных для фиксированной непрерывной на отрезке $[a,b]$ функции $f(x)$.

Интерполяционный процесс для функции сходится в точке $x_* \in [a,b]$, если существует предел $\lim\limits_{n\to\infty} P_n(x_*) = f(x_*)$. Наряду с обычной (поточечной) сходимостью, часто рассматривается сходимость в различных нормах. Так, равномерная сходимость на отрезке $[a,b]$ означает, что $\max\limits_{x\in[a,b]} |f(x) - P_n(x)| \to 0$ при $n \to \infty$.

Сходимость или расходимость интерполяционного процесса зависят как от выбора последовательности сеток, так и от гладкости функции $f(x)$. Если $f(x)$- целая аналитическая функция, то при произвольном расположении узлов на отрезке $[a,b]$, интерполяционный многочлен $P_n(x)$ равномерно сходится к $f(x)$ при $n \to \infty$.

Положение резко меняется, если производные функции разрывны или не существуют в отдельных точках. Например, для функции $f(x) = |x|$ на отрезке $[-1, 1]$, покрытом равномерной сеткой узлов, значения $P_n(x)$ между узлами интерполяции неограниченно возрастают при $n \to \infty$. Вместе с тем для заданной непрерывной функции $f(x)$ за счет выбора сеток можно добиться сходимости, притом равномерной на $[a,b]$. Однако построить такие сетки довольно сложно, и, главное, такие сетки «индивидуальны» для каждой конкретной функции.

Объем вычислений при построении интерполяционного полинома быстро нарастает с ростом n. Поэтому на практике избегают пользоваться интерполяционными полиномами высокой степени. Более распространенной является интерполяция сплайнами, которую мы обсудим ниже.

6.5 Интерполяционный полином Эрмита

Расширим постановку задачи об интерполяции. Ранее полагалось, что в узлах интерполяции заданы только значения функции $f(x)$. Пусть теперь в узлах $x_k \in [a,b], \quad k = 0,1,...,m$, среди которых нет совпадающих, заданы значения функции $f(x_k)$ и ее производных $f^{(i)}(x_k), \quad i = 1,2,..., N_k - 1$ до $N_k - 1$-го порядка включительно. Числа N_k при этом называют кратностью узла x_k. В каждой точке x_k, таким образом, задано N_k величин:

$$f(x_k), f'(x_k),..., f^{(N_k-1)}x_k .$$

В общей сложности на всей совокупности узлов $x_0, x_1, ..., x_m$ известно $N_0 + N_1 + ... + N_m$ величин, что дает возможность ставить вопрос о построении полинома $H_n(x)$степени

$$n = N_0 + ... + N_m - 1, \tag{33}$$

удовлетворяющего требованиям

$$H_n^{(i)}(x_k) = f^{(i)}(x_k), \quad k = 0,1,...,m; \quad i = 0,1,..., N_k - 1. \tag{34}$$

Такой полином называется интерполяционным полиномом Эрмита для функции $f(x)$. Рассмотренный ранее вариант построения интерполяционного полинома $P_n(x)$по известным значениям функции $f(x)$в узлах интерполяции является частным случаем построения полинома Эрмита при условии, что все узлы простые: $N_k = 1, \quad k = 0,1,...,m.$

Докажем, что интерполяционный полином Эрмита существует и является единственным. Представим его в стандартном виде:

$$H_n(x) = a_0 + a_1 x + ... + a_n x^n .$$

Наше утверждение будет справедливо, если показать, что коэффициенты $a_0, a_1, ..., a_n$ определяются из условий (34), и притом единственным образом. Условия представляют собой систему линейных алгебраических уравнений относительно этих коэффициентов, причем число уравнений равно числу неизвестных, а именно: $N_0 + N_1 + ... + N_m = n + 1$. Для того,

чтобы эта система была однозначно разрешима, соответствующая ей однородная система должна иметь только тривиальное (нулевое) решение. Рассмотрим соответствующую однородную систему

$$\overline{H}_n^{(i)}(x_k) = 0, \quad k = 0,1,\ldots,m, \quad i = 0,1,\ldots,N_{k-1}. \tag{35}$$

Уравнения (35) указывают на то, что числа x_k являются корнями полинома $H_n(x)$ кратности N_k. Мы видим, таким образом, что полином $H_n(x)$ имеет, с учетом кратности, не менее $N_0 + N_1 + \ldots + N_m = n+1$. Поскольку его степень равна n, то он должен тождественно равняться нулю. Это означает, что $a_0 = a_1 = \ldots = a_n = 0$, т.е. однородная система уравнений (35) имеет только тривиальное решение. Отсюда следует, что неоднородная система (34) при любой правой части разрешима, и притом единственным образом.

Исследование погрешности интерполирования полинома Эрмита

$R_n(x) = f(x) - H_n(x)$ почти дословно повторяет проведенное ранее исследование для полинома с простыми узлами x_k, в которых заданы только $f(x_k)$. Достаточно представить $R_n(x)$ в виде

$$R_n(x) = r_n(x)\omega_{n+1}(x), \tag{36}$$

где

$$\begin{aligned} &\omega_{n+1}(x) = (x - x_0)^{N_0}(x - x_1)^{N_1}\ldots(x - x_m)^{N_m}, \\ &n+1 = N_0 + \ldots + N_m, \end{aligned} \tag{37}$$

и рассмотреть функцию

$$g(t) = f(t) - H_n(t) - r_n(t)\omega_{n+1}(t).$$

Применяя теорему Ролля к функции $g(t)$ и ее производным, с учетом кратности корней в узлах $t = x_k$ и условия $g(x) = 0$, придем к формуле

$$f(x) - H_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x). \tag{38}$$

С ее помощью можно написать оценку типа (31):

$$|R_n(x)| \le \frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)|. \qquad (39)$$

где M_{n+1}- максимальное значение модуля функции $f^{(n+1)}(x)$. Здесь полином $\omega_{n+1}(x)$ (37) является обобщением полинома (27) на случай кратных корней.

Построение полинома Эрмита в общем случае при произвольном числе узлов и их произвольной кратности приводит к довольно громоздким выражениям и редко используется.

6.6 Интерполирование сплайнами

Увеличение степени интерполяционного полинома может оказаться невыгодным из-за быстрого роста объема вычислений. К тому же далеко не всегда оно приводит к повышению точности. Во второй половине XX века с появлением компьютеров и развитием современной вычислительной математики при обработке больших таблиц получила развитие новая идея - строить приближение функций с помощью кусочно-полиномиальной интерполяции с использованием полиномов сравнительно невысоких степеней. Наиболее удобными оказались полиномы третьей степени. Такие конструкции получили название кубических сплайнов.

Пусть на отрезке $[a,b]$задана функция $y = f(x)$. Рассмотрим сетку узлов

$$a = x_0 < x_1 < x_2 < \ldots x_n = b \qquad (40)$$

и обозначим через h_i расстояние между смежными узлами:

$$h_i = x_i - x_{i-1}, \quad i = 1,\ldots,n. \qquad (41)$$

Определение. Назовем кубическим сплайном функции $y = f(x), \quad x \in [a,b]$ на сетке (41) функцию $S(x)$, удовлетворяющую условиям:

1. На каждом отрезке $[x_{i-1}, x_i]$ функция $S(x)$ является полиномом третьей степени.

2. Функция $S(x)$, ее первая $S'(x)$и вторая $S''(x)$производные непрерывны

на сегменте $[a,b]$.

3. $S(x_i) = f(x_i) = f_i, \quad i = 0,1,...,n$.

4. На концах сегмента $[a,b]$ функция $S''(x)$удовлетворяет дополнительным условиям $S''(a) = S''(b) = 0$.

Замечание. На концах сегмента $[a,b]$ могут быть заданы в принципе и другие условия, например: $S''(a) = A, \ S''(b) = B$.

Построение сплайна. Сведем задачу построения сплайна к отысканию коэффициентов упомянутых полиномов третьей степени на каждом из отрезков $[x_{i-1}, x_i]$. Для этого сопоставим отрезку $[x_{i-1}, x_i]$ полином $S_i(x)$, для удобства записанный в виде

$$S_i(x) = a_i + b_i(x - x_i) + \frac{c_i}{2}(x - x_i)^2 + \frac{d_i}{6}x - x_i^3, \quad x \in [x_{i-1}, x_i], \quad i = 1,...,n. \tag{42}$$

При этом очевидно, что

$$S_i'(x) = b_i + c_i(x - x_i) + \frac{d_i}{2}(x - x_i)^2, \tag{43}$$

$$S_i''(x) = b_i + d_i(x - x_i), \tag{44}$$

так что

$$S_i(x_i) = a_i, \quad S_i'(x_i) = b_i, \quad S_i''(x_i) = c_i. \tag{45}$$

Для выполнения требования 3 в узлах интерполяции с номерами $i = 1,...,n$ следует положить

$$a_i = f(x_i), \quad i = 1,...,n.. \tag{46}$$

Требуя непрерывности сплайна в узлах $x_i, \quad i = 1,...,n-1$, и выполнения условия 3 при $i = 0$, получаем

$$S_i(x_{i-1}) = f_{i-1}, \quad i = 1,...,n, \tag{47}$$

или

$$f_i + b_i(x_{i-1} - x_i) + \frac{c_i}{2}|x_{i-1} - x_i|^2 + \frac{d_i}{6}(x_{i-1} - x_i)^3 = f_{i-1}, \quad i = 1,...,n,$$

Это равенство можно переписать следующим образом:

$$b_i h_i - \frac{c_i}{2} h_i^2 + \frac{d_i}{6} h_i^3 = f_i - f_{i-1}, \quad i = 1,...,n. \tag{48}$$

Условие 2 непрерывности первой производной $S'(x)$в узлах $x_i, \quad i = 1,...,n-1$, принимает вид

$$S_i'(x_{i-1}) = S_{i-1}'(x_{i-1}) = b_{i-1}, \quad i = 2,...,n, \tag{49}$$

и приводит к соотношениям

$$b_i - c_i h_i + \frac{d_i}{2} h_i^2 = b_{i-1}, \quad i = 2,...,n,$$

или

$$c_i h_i - \frac{d_i}{2} h_i^2 = b_i - b_{i-1}, \quad i = 2,...,n, \tag{50}$$

Аналогичным образом условия непрерывности второй производной $S''(x)$в тех же узлах

$$S_i''(x_{i-1}) = S_{i-1}''(x_{i-1}) = c_{i-1}, \quad i = 2,...,n, \tag{51}$$

означают, что

$$d_i h_i = c_i - c_{i-1}, \quad i = 2,...,n. \tag{52}$$

Наконец, дополнительные граничные условия 4 дают еще два уравнения:

$$\begin{aligned} S_1''(x_0) &= S_1''(a) = c_1 - d_1 h_1 = 0, \\ S_n''(x_n) &= S_n''(b) = c_n = 0. \end{aligned} \tag{53}$$

В итоге мы получили замкнутую систему (48), (50, (52) ,(53), содержащую в сумме $3n$ линейных уравнений для отыскания $3n$ неизвестных: $b_i, c_i, d_i, \quad i = 1,2,...,n$.

Приведение системы уравнений для коэффициентов сплайна к трехдиагональному виду. Удобно формально ввести еще одно неизвестное c_0, положив при этом, что оно равно нулю, и первое уравнение в (53) переписать в виде $d_1 h_1 = c_1 - c_0$, т.е. в форме, аналогичной (52).

Теперь уравнения (52) и (53) можно представить в единообразном виде:

$$d_i h_i = c_i - c_{i-1}, \quad i = 1,...,n. \tag{54}$$

$$c_0 = 0, \quad c_n = 0. \tag{55}$$

Обратим внимание на то, что из системы (54) можно выразить все коэффициенты d_i через разности $c_i - c_{i-1}$, а затем из системы (48) выразить через c_i и c_{i-1} коэффициенты b_i.

Подставляя полученные выражения в (50), приходим к системе линейных уравнений для c_i:

$$\frac{1}{3} c_{i-2} h_{i-2} + \frac{2}{3} c_{i-1}(h_{i-1} + h_i) + \frac{1}{3} c_i h_i = 2\left(\frac{f_i - f_{i-1}}{h_i} - \frac{f_{i-1} - f_{i-2}}{h_{i-1}} \right), \tag{56}$$

$$i = 2,3,...,n$$

Сдвигая индекс i на единицу, получаем симметричную форму записи уравнений (56):

$$h_i c_{i-1} + 2(h_i + h_{i+1}) c_i + h_{i+1} c_{i+1} = 6\left(\frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i} \right), \tag{57}$$

$$i = 1,2,...,n-1.$$

Кроме того, согласно (55)

$$c_0 = c_n = 0. \tag{58}$$

Система (57) содержит $[n-1]$ уравнение с $(n-1)$-й неизвестной:

$c_1, c_2, \ldots, c_{n-1}$. Величины c_0 и c_n определены дополнительными соотношениями (58). Если сетка (40) равномерная, т.е. $h_i = h = const$, то уравнения (57) принимают особенно простой вид:

$$c_{i-1} + 4c_i + c_{i+1} = 6\frac{f_{i-1} - 2f_i + f_{i+1}}{h^2}. \tag{59}$$

Для уравнений системы (57) выполнено условие диагонального преобладания. Отсюда следует существование и единственность решения задачи (57), (58). Зная величины c_i, можно рассчитать остальные коэффициенты сплайна по формулам

$$d_i = \frac{c_i - c_{i-1}}{h_i}, \quad i = 1,\ldots,n, \tag{60}$$

$$b_i = \frac{1}{2}h_i c_i - \frac{1}{6}h_i{}^2 d_i + \frac{f_i - f_{i-1}}{h_i}, \quad i = 1,\ldots,n, \tag{61}$$

завершив тем самым построение сплайна.

Замечание о решении системы. Уравнения (57) имеют так называемую трехточечную структуру, общий вид таких систем

$$A_i y_{i-1} + C_i y_i + B_i y_{i+1} = F_i, \quad i = 1,2,\ldots,n-1, \tag{62}$$

$$y_0 = 0, \quad y_n = 0 \tag{63}$$

соответствует системе линейных уравнении с трехдиагональной матрицей T для определения вектора неизвестных $\boldsymbol{y} = (y_1, y_2, \ldots, y_{n-1})$:

$$T\boldsymbol{y} = \boldsymbol{F}, \tag{64}$$

При этом легко видеть, что в нашем случае

$$|C_i| > |A_i| + |B_i|, \quad i = 1,\ldots,n-1, \tag{65}$$

поскольку

$$C_i = 2(h_i + h_{i+1}), \quad A_i = h_i, \quad B_i = h_{i+1}. \tag{66}$$

Как было показано в гл. 2, решение подобных систем эффективно осуществляется методом прогонки.

Сходимость и точность интерполирования сплайнами. При обсуждении эффективности численного метода в первую очередь обращают внимание на две характеристики.
1. Сходимость означает, что приближенное решение задачи стремится к точному решению.
2. Точность - это характеристика близости приближенного решения к точному.

Пусть на сегменте $[a,b]$ задана функция $f(x)$ и построена сетка

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b; \quad h_i = x_i - x_{i-1} > 0.$$

Определим максимальный шаг сетки

$$h = \max_{1 \le i \le n} h_i.$$

Приведем без доказательства две теоремы.
Теорема 1. Пусть $f(x)$ непрерывна на сегменте $[a,b]$, тогда для любого $\varepsilon > 0$ можно указать $\delta(\varepsilon) > 0$ такое, что при любой сетке, удовлетворяющей условию $h < \delta$, справедливо неравенство

$$|f(x) - S(x)| < \varepsilon, \quad \forall x \in [a,b],,$$

иными словами, $S(x)$ при $h \to 0$ равномерно сходится к функции $f(x)$.

Теорема 2. Пусть $f(x)$ имеет на сегменте $[a,b]$ четыре непрерывных производных и дополнительно удовлетворяет условию $f''(a) = f''(b) = 0$.
Тогда имеют место неравенства:

$$|f(x) - S(x)| < M_4 h^4, \quad \forall x \in [a,b], \tag{67}$$

$$|f'(x) - S'(x)| < M_4 h^3, \quad \forall x \in [a,b], \tag{68}$$

$$|f''(x) - S''(x)| < M_4 h^2, \quad \forall x \in [a,b], \tag{69}$$

$$M_4 = \max_{x\in[a,b]} \left| f^{(4)} x \right|. \qquad (70)$$

Глава 7. Аппроксимация функций

Пусть имеется семейство функций (например, параметризованное некоторым набором параметров) и функция F, вообще говоря, не принадлежащая этому семейству. Требуется найти функцию из данного семейства , наиболее близкую к функции F. Такая задача является весьма общей. Можно рассматривать функции одной переменной или многих переменных. Функция F может быть известной во всех точках области определения или же в некотором достаточно большом наборе точек. При этом, в отличие от задачи интерполяции, не предполагается, что аппроксимирующая функция будет в этом наборе точек принимать те же значения, что и исходная функция.

7.1 Метод наименьших квадратов

В этом случае задача построения функции непрерывного аргумента по дискретной информации характеризуется двумя особенностями.

1.Число точек, в которых проводятся измерения, обычно бывает достаточно большим.

2. Значения аппроксимируемой функции y_i в точках сетки определяются приближенно в связи с неизбежными ошибками измерения.

С учетом этих обстоятельств, строить аппроксимирующую функцию в виде суммы большого числа слагаемых и добиваться ее точного равенства значениям аппроксимируемой функции в точках сетки, как это делалось при интерполировании, становится нецелесообразным.

В методе наименьших квадратов аппроксимирующая функция $y = F(x)$ ищется в виде суммы, содержащей сравнительно небольшое число слагаемых:

$$F(x) = \sum_{k=0}^{m} a_k \varphi_k(x), \quad m < n, \qquad (1)$$

в частности, возможен вариант $m << n$.

Предположим, что мы каким-то образом выбрали коэффициенты a_k, тогда в каждой точке сетки x_i можно подсчитать погрешность

$$\delta_i = y_i - F(x_i) = y_i - \sum_{k=0}^{m} a_k \varphi_k(x_i), \quad i = 0,1,2,\ldots,n. \tag{2}$$

Сумма квадратов этих величин называется суммарной квадратичной погрешностью:

$$J = \sum_{i=0}^{n} \left(y_i - \sum_{k=0}^{m} a_k \varphi_k(x_i) \right)^2. \tag{3}$$

Она дает количественную оценку того, насколько близки значения функции $F(x)$ в точках сетки к величинам y_i.

Меняя значения коэффициентов a_k, будем менять погрешность J, которая является их функцией. В результате естественно возникает задача найти такой набор коэффициентов a_k, при которых суммарная квадратичная погрешность J оказывается минимальной.

Функцию $F(x)$ (1) с набором коэффициентов, удовлетворяющих этому требованию, называют наилучшим приближением по методу наименьших квадратов.

Построение наилучшего приближения сводится к классической задаче математического анализа об экстремуме функции нескольких переменных. Необходимым условием экстремума является равенство нулю в экстремальной точке всех первых частных производных рассматриваемой функции. В случае (3) это дает

$$\frac{\partial J}{\partial a_j} = -2\sum_{i=0}^{n} \left(y_i - \sum_{k=0}^{m} a_k \varphi_k(x_i) \right) \varphi_j(x_i) = 0, \quad j = 0,1,\ldots,m. \tag{4}$$

Оставим члены, содержащие a_k, слева и поменяем в них порядок суммирования по индексам i и k. Члены, содержащие y_i перенесем направо. В результате уравнения (4) примут вид

$$\sum_{k=0}^{m} \gamma_{jk} a_k = b_j, \quad j = 0,1,\ldots,m, \tag{5}$$

где

$$\gamma_{jk} = \sum_{i=0}^{n} \varphi_j(x_i)\varphi_k(x_i), \tag{6}$$

$$b_j = \sum_{i=0}^{n} \varphi_j(x_i) y_i, \tag{7}$$

Мы получили систему линейных алгебраических уравнений (5), в которой роль неизвестных играют искомые коэффициенты разложения $a_0, a_1, \ldots, a_m$. Числа уравнений и неизвестных в этой системе совпадают и равны $(m+1)$. Матрица коэффициентов системы Γ состоит из элементов γ_{lk}, которые определяются формулой (6). Ее называют матрицей Грама системы функций $\varphi_0(x), \varphi_1(x), \ldots, \varphi_m(x)$на сетке. Отметим, что матрица Грама является симметричной: для ее элементов, согласно (6), справедливо равенство $\gamma_{lk} = \gamma_{kl}$. Числа b_j, стоящие в правой части уравнений (5), вычисляются по формуле (7) через значения y_i аппроксимируемой функции.

Предположим, что функции $\varphi_0(x), \varphi_1(x), \ldots, \varphi_m(x)$выбраны такими, что определитель матрицы Грама отличен от нуля:

$$\Delta = det\,\Gamma \neq 0. \tag{8}$$

В этом случае при любой правой части система имеет единственное решение

$$\bar{a}_0, \bar{a}_1, \ldots, \bar{a}_m, \tag{9}$$

Рассмотрим наряду с набором коэффициентов (9), полученных в результате решения системы (5), любой другой набор коэффициентов $a_0, a_1, \ldots, a_m$. Представим числа a_k в виде

$$\begin{aligned} &a_0 = \bar{a}_0 + \Delta a_0, \quad a_1 = \bar{a}_1 + \Delta a_1, \ldots, a_m = \bar{a}_m + \Delta a_m, \\ &(\Delta a_0)^2 + (\Delta a_1)^2 + \ldots + (\Delta a_m)^2 \geq 0 \end{aligned} \tag{10}$$

и сравним значения суммарной квадратичной погрешности J для функций $F(x)$, построенных с помощью коэффициентов (9) и (10).

Квадрат погрешности в точке $x = x_i$ для функции $F(x)$ (9) с коэффициентами (10) можно записать в виде

$$\begin{aligned}
&\delta_i^2 = \left\{ y_i - \sum_{k=0}^{m} (\bar{a}_k + \Delta\bar{a}_k)\varphi_k(x_i) \right\}^2 = \\
&= \left\{ \left(y_i - \sum_{k=0}^{m} \bar{a}_k \varphi_k(x_i) \right) - \sum_{k=0}^{m} \Delta\bar{a}_k \varphi_k(x_i) \right\}^2 = \\
&= \left(y_i - \sum_{k=0}^{m} \bar{a}_k \varphi_k(x_i) \right)^2 - 2\left(y_i - \sum_{k=0}^{m} \bar{a}_k \varphi_k(x_i) \right) \sum_{l=0}^{m} \Delta\bar{a}_j \varphi_j(x_i) + \\
&\left(\sum_{k=0}^{m} \Delta\bar{a}_k \varphi_k(x_i) \right)^2 .
\end{aligned} \tag{11}$$

Здесь в среднем слагаемом мы заменили в одной из сумм индекс суммирования k на j, чтобы не использовать один и тот же индекс в двух разных суммах и иметь возможность перемножить их почленно.

Для получения суммарной квадратичной погрешности нужно просуммировать выражения (11) для δ_i^2 по индексу i. Первые слагаемые не содержат Δa_k. Их сумма дает погрешность J, вычисленную для функции (1) с коэффициентами $\bar{a}_k$ (9).

Рассмотрим теперь сумму вторых слагаемых, которые зависят от Δa_l линейно:

$$\begin{aligned}
&-2\sum_{i=1}^{n} \left\{ \left(y_i - \sum_{k=0}^{m} \bar{a}_k \varphi_k(x_i) \right) \sum_{l=0}^{m} \Delta\bar{a}_l \varphi_j(x_i) \right\} = \\
&= -2\sum_{l=0}^{m} \Delta a_j \left\{ \sum_{i=1}^{n} y_i \varphi_j(x_i) - \sum_{k=0}^{m} \bar{a}_k \sum_{i=0}^{n} \varphi_k(x_i)\varphi_j(x_i) \right\} = \\
&= -2\sum_{l=0}^{m} \Delta a_j \left[b_j - \sum_{k=0}^{m} \bar{a}_k \gamma_{jk} \right] = 0.
\end{aligned} \tag{12}$$

С учетом (12) будем иметь

$$J\left(\overline{a}_0+\Delta a_0,\overline{a}_1+\Delta a_1,\ldots,\overline{a}_m+\Delta a_m\right)=$$

$$=J\left(\overline{a}_0,\overline{a}_1,\ldots,\overline{a}_m\right)+\sum_{l=0}^{n}\left(\sum_{k=0}^{m}\Delta a_k\varphi_k\left(x_i\right)\right)^2\geq J\left(\overline{a}_0,\overline{a}_1,\ldots,\overline{a}_m\right). \quad (13)$$

Формула (13) показывает, что функция $F\left(x\right)$(1) с коэффициентами $\overline{a}_k$ (9), полученными в результате решения уравнений (5), действительно минимизирует суммарную квадратичную погрешность J. Если взять любой другой набор коэффициентов (10), отличный от (9), то, согласно формуле (13), к погрешности $y(\overline{a}_0,\overline{a}_1,\ldots,\overline{a}_m)$добавится положительное слагаемое и она увеличится.

Итак, чтобы построить наилучшее приближение (1) аппроксимируемой функции по методу наименьших квадратов, нужно взять в качестве коэффициентов разложения a_k решение системы линейных уравнений (5).

Глава 8. Численное интегрирование функций

В курсе математического анализа описывается способ расчета определенных интегралов с помощью формулы Ньютона-Лейбница:

$$I = \int_a^b f(x)dx = F(b) - F(b),$$

а где $F(x)$- первообразная подынтегральной функции $f(x)$. Однако, существует много простых функций, первообразные которых не выражаются через элементарные функции. В качестве примера можно привести такие функции, как e^{-x^2} или $\frac{\sin x}{x}$. Многие из таких интегралов подробно исследованы и названы новыми - специальными функциями. С ними можно работать так же, как и с обычными функциями. В частности, интеграл от e^{-x^2} называется интегралом вероятности, а интеграл от $\frac{\sin x}{x}$ - интегральным синусом. Однако, даже класс специальных функций недостаточно широк, чтобы охватить все практически важные случаи.

Универсальные алгоритмы вычисления определенных интегралов дают формулы численного интегрирования или, как их часто называют, квадратурные формулы. Квадратурные формулы имеют вид

$$I = \int_a^b f(x)dx = \sum_{i=0}^{n} c_i f(x_i) + R_n . \qquad (1)$$

Здесь точки $x_i \in [a,b]$ называют узлами, коэффициенты c_i - весовыми множителями или весами, величину R_n - остаточным членом или погрешностью. Узлы и веса подбираются таким образом, чтобы выполнялось предельное равенство $\lim\limits_{n\to\infty} R_n = 0$, так что

$$\lim_{n\to\infty} \sum_{i=0}^{n} c_i f(x_i) = I . \qquad (2)$$

Условие (2), которое называют условием сходимости, позволяет сделать погрешность в равенстве (1) меньше любого наперед заданного числа за счет выбора достаточно большого n. Таким образом, открывается возможность

вычислить интеграл I с любой наперед заданной точностью. Чем выше требование точности, тем больше слагаемых следует удерживать в сумме. За точность приходится платить увеличением объема вычислений.

Замечание. Подставляя в формулу (1) функцию $f(x) \equiv 1$, получаем

$$b - a = \sum_{i=0}^{n} c_i + R_n .$$

Обычно весовые коэффициенты c_i, подбираются так, чтобы выполнялось равенство

$$b - a = \sum_{i=0}^{n} c_i ,$$

т.е. чтобы при интегрировании константы равенство (1) было не приближенным, а точным.

В следующих параграфах этой главы мы обсудим методы построения квадратурных формул и получим оценки их точности.

8.1 Квадратурные формулы прямоугольников, трапеций и Симпсона

Возьмем произвольное целое число n и разобьем отрезок $[a,b]$, по которому ведется интегрирование, на n равных отрезков длиной $h = \dfrac{b-a}{n}$ точками

$$x_i = a + ih, \quad 0 \le i \le n. \tag{3}$$

Для дальнейшего нам также понадобятся средние точки этих отрезков:

$$\xi_i = a + \left(i - \frac{1}{2}\right)h, \quad \xi_i \in [x_{i-1}, x_i], \quad 1 \le i \le n. \tag{4}$$

Формула прямоугольников. Построим с помощью проведенного разбиения интегральную сумму, в которой значения функции $f(x)$ для каждого отрезка $[x_{i-1}, x_i]$ вычисляются в его средней точке ξ_i :

$$P_n = \frac{b-a}{n}\sum_{i=1}^{n} f(\xi_i). \tag{5}$$

Принимая во внимание то, что интегральная сумма дает приближенное значение интеграла, можно написать

$$I = P_n + \alpha_n, \tag{6}$$

где α_n - остаточный член.

Формулу (5) называют формулой прямоугольников. Причина такого названия имеет простой геометрический смысл. Величина P_n представляет собой сумму площадей прямоугольников с одинаковыми основаниями $h = \frac{b-a}{n}$ и высотами $f(\xi_i)$. Она аппроксимирует с точностью до α_n площадь криволинейной трапеции, соответствующей исходному интегралу (рис.1).

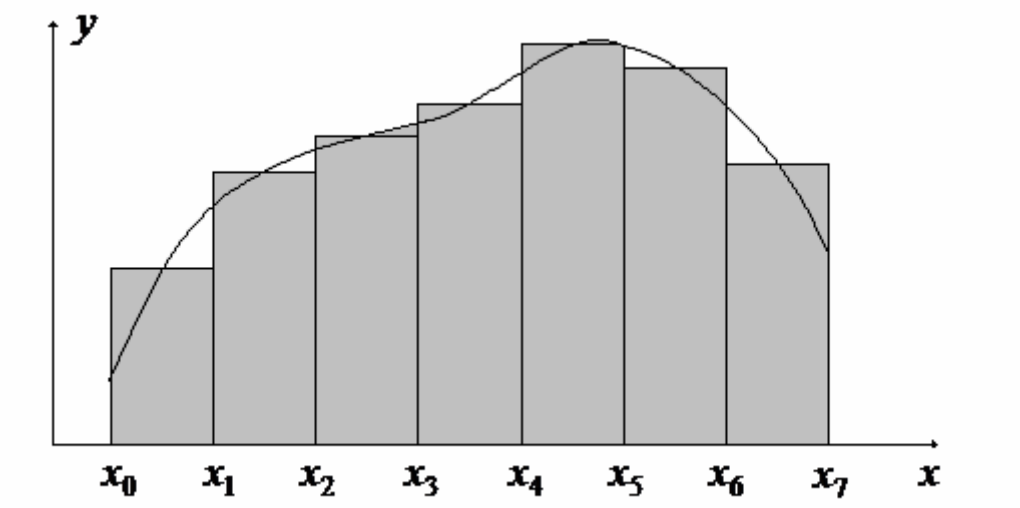


Рис.1.Геометрическая интерпретация формулы прямоугольников

Формула трапеций. В этом случае в качестве аппроксимирующей функции $g_n(x)$ берется кусочно-линейная функция. На каждом из частичных сегментов $[x_{i-1}, x_i]$ она задается формулой

$$g_n(x) = f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{h}(x - x_{i-1}),$$
$$x \in [x_{i-1}, x_i], \quad 1 \le i \le n. \tag{7}$$

В граничных точках отрезка $x = x_{i-1}$ и $x = x_i$ функция $g_n(x)$ принимает те же значения, что и функция $f(x)$:

$$g_n(x_{i-1}) = f(x_{i-1}), \quad g_n(x_i) = f(x_i), \tag{8}$$

она осуществляет кусочно-линейную интерполяцию функции $f(x)$на отрезке $[a,b]$ (рис.2). Вычислим интеграл: y

$$\int_{x_{i-1}}^{x_i} g_n(x)dx = \int_{x_{i-1}}^{x_i} \left\{ f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{h}(x - x_{i-1}) \right\} dx = \\ = \frac{h}{2}(f(x_{i-1}) + f(x_i)). \tag{9}$$

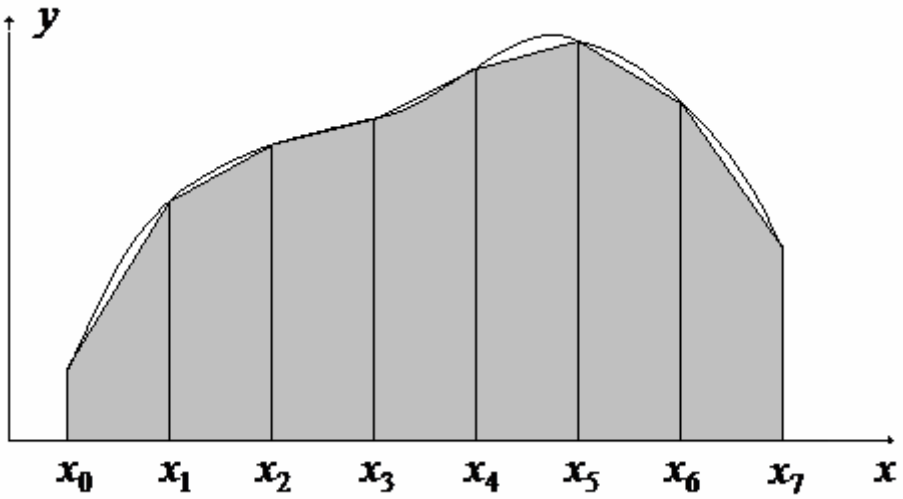


Рис. 2. Геометрическая интерпретация формулы трапеций

Этот результат имеет простой геометрический смысл: фигура, ограниченная снизу отрезком $[x_{i-1}, x_i]$ оси x, сверху отрезком прямой (7), с боков вертикальными прямыми $x = x_{i-1}$ и $x = x_i$, представляет собой трапецию, площадь которой определяется формулой (9).

Интеграл от функции $g_n(x)$по всему отрезку $[a,b]$ является суммой интегралов (9):

$$T_n = \int_a^b g_n(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} g_n(x)dx = \\ = \frac{b-a}{n}\left\{\frac{1}{2}f(a) + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2}f(b)\right\} \quad . \quad (10)$$

Он дает приближенное значение интеграла I :

$$I = \int_a^b f(x)dx = T_n + \beta_n . \quad (11)$$

В квадратурной формуле (11) узлами являются точки x_i (3). Все весовые коэффициенты, кроме двух, одинаковы и равны $h = \frac{b-a}{n}$, а весовые коэффициенты при $i = 0$ и $i = n$ имеют вдвое меньшие значения. Для остаточного члена введено специальное обозначение β_n. Формулу (11) называют квадратурной формулой трапеций. С точностью до β_n она выражает площадь криволинейной трапеции, соответствующую интегралу I через сумму площадей трапеций (10) (см. рис.2). Формула (5) для величины P_n изначально строилась как интегральная сумма. При выводе формулы (10) для величины $\mathrm{T_n}$ понятие интегральной суммы не использовалось. Однако теперь, когда формула уже получена, видно, что величину T_n тоже можно интерпретировать как интегральную сумму. Чтобы убедиться в этом, рассмотрим разбиение отрезка $[a,b]$ на частичные отрезки точками ξ_i (4). Оно дает $n+1$ отрезок. Два крайних $[a,\xi_1]$ и $[\xi_n,b]$ имеют длину $h/2$, а остальные - h. Выберем для образования интегральной суммы на крайних отрезках значения функции $f(x)$ в точках a и b, а на остальных отрезках $[\xi_i,\xi_{i+1}]$ - значения функции $f(x)$ в их средних точках x_i, $1 \le i \le n-1$. Образованная таким образом интегральная сумма соответствует выражению (10) для T_n.

Формула Симпсона. Вывод квадратурной формулы Симпсона развивает описанный подход дальше. Теперь для аппроксимации функции $f(x)$ используется не кусочно-линейное, а кусочно-квадратичное интерполирование.

Будем считать n четным и сгруппируем отрезки $[x_{i-1},x_i]$ парами: первая пара $[a,x_1],[x_1,x_2]$ вторая пара $[x_2,x_3],[x_3,x_4]$ и т.д. Для каждого

двойного отрезка $\left[x_{2j-2}, x_{2j}\right]$ построим интерполяционный полином второй степени в форме Лагранжа, принимающий в узлах $x_{2j-2}, x_{2j-1}, x_{2j}$ значения функции $f(x)$. В результате получим аппроксимирующую функцию $g_n(x)$ на отрезке $[a,b]$ в виде кусочно-квадратичной функции:

$$
\begin{aligned}
g_n(x) &= f\left(x_{2j-2}\right)\frac{\left(x-x_{2j-1}\right)\left(x-x_{2j}\right)}{2h^2} + \\
&+ f\left(x_{2j-1}\right)\frac{\left(x-x_{2j-2}\right)\left(x-x_{2j}\right)}{\left(-h^2\right)} + \\
&+ f\left(x_{2j}\right)\frac{\left(x-x_{2j-2}\right)\left(x-x_{2j-1}\right)}{2h^2}, \\
x &\in \left[x_{2j-2}, x_{2j}\right], \quad 1 \le j \le \frac{n}{2}.
\end{aligned}
\tag{12}
$$

Проинтегрировав полином второй степени (12) по отрезку $\left[x_{2j-2}, x_{2j}\right]$ получим

$$
\int_{x_{2j-2}}^{x_{2j}} g_n(x)dx = \frac{h}{3}\left\{f\left(x_{2j-2}\right) + 4f\left(x_{2j-1}\right) + f\left(x_{2j}\right)\right\}, \quad h = \frac{b-a}{n}. \tag{13}
$$

Интеграл от функции $g_n(x)$ по всему отрезку $[a,b]$ равен сумме интегралов

$$
\begin{aligned}
S_n &= \int_a^b g_n(x)dx = \sum_{j=1}^{n/2}\int_{x_{2j-2}}^{x_{2j}} g_n(x)dx = \\
&= \frac{b-a}{3n}\left\{f(a) + 4f\left(x_1\right) + 2f\left(x_2\right) + \ldots + 2f\left(x_{n-2}\right) + 4f\left(x_{n-1}\right) + f(b)\right\}.
\end{aligned}
\tag{14}
$$

Напомним, что число n должно быть обязательно четным. Величина S_n в (14) дает приближенное значение интеграла I:

$$
I = \int_a^b f(x)dx = S_n + \gamma_n. \tag{15}
$$

Узлами квадратурной формулы (14), как и формулы трапеций (10), являются точки x_i (3). Весовые коэффициенты в узлах с четными и нечетными номерами имеют разные значения. Для остаточного члена введено обозначение γ_n. Формула (14) называется квадратурной формулой Симпсона.

Представление (14) для S_n можно рассматривать как интегральную сумму. Для ее построения нужно разбить отрезок $[a,b]$ на $n+1$ частичный отрезок с помощью n внутренних точек:

$$\eta_{2j-1} = x_{2j-1} - \frac{2h}{3}, \quad \eta_{2j} = x_{2j-1} + \frac{2h}{3}, \quad 1 \le j \le \frac{n}{2}, \tag{16}$$

и двух граничных точек:

$$\eta_0 = a \quad \text{и} \quad \eta_{n+1} = b. \tag{17}$$

В результате получаются отрезки $[\eta_{i-1}, \eta_i]$, $1 \le j \le n+1$ различной длины. Два крайних отрезка $[a, \eta_1]$ и $[\eta_n, b]$ имеют длину $\frac{h}{3}$; отрезки, в центре которых лежат точки x_i с четными номерами, - длину $\frac{2h}{3}$; отрезки, в центре которых лежат точки x_i с нечетными номерами, - длину $\frac{4h}{3}$.

Для построения интегральной суммы, соответствующей данному разбиению, возьмем для крайних отрезков значения функции $f(x)$ в точках a и b, для остальных отрезков - значение функции $f(x)$ в их средних точках x_i. В результате получим интегральную сумму в виде выражения (15). Разные длины частичных отрезков приводят к своеобразному чередованию коэффициентов в виде двоек, четверок и единиц в крайних точках.

Заканчивая обсуждение формул (10) для T_n и (14) для S_n, установим полезную для дальнейшего связь между этими величинами:

$$S_n = \frac{4}{3} T_n - \frac{1}{3} T_{n/2}. \tag{18}$$

Здесь $T_{n/2}$ - сумма (10) с вдвое меньшим числом слагаемых и,

соответственно, вдвое большим шагом. Благодаря этому, при ее образовании в качестве узлов используются точки x_i (3) только с четными номерами. Поскольку в формуле Симпсона n предполагается обязательно четным, то $\frac{n}{2}$ - целое число, так что выражение $T_{n/2}$ определено.

Проверим (18). Из (10) следует, что

$$\frac{4}{3}T_n = \frac{b-a}{3n}\{2f(a)+4f(x_1)+4f(x_2)+\ldots+4f(x_{n-1})+2f(b)\},$$

$$\frac{1}{3}T_{n/2} = \frac{b-a}{3n}\{f(a)+2f(x_2)+4f(x_2)+\ldots+2f(x_{n-2})+f(b)\}.$$

Вычитая теперь вторую строку из первой, получаем равенство (18).

8.2 Сходимость и точность квадратурных формул прямоугольников, трапеций и Симпсона

Сходимость. После того как мы установили, что величины P_n, T_n, S_n являются интегральными суммами, проблема сходимости рассмотренных методов численного интегрирования решается элементарно. Их сходимость имеет место для любой интегрируемой функции:

$$\lim_{n\to\infty}\alpha_n = 0, \quad \lim_{n\to\infty}P_n = I, \tag{19}$$

$$\lim_{n\to\infty}\beta_n = 0, \quad \lim_{n\to\infty}T_n = I, \tag{20}$$

$$\lim_{n\to\infty}\gamma_n = 0, \quad \lim_{n\to\infty}S_n = I. \tag{21}$$

Этот вывод является прямым следствием определения интегрируемости.

Предельные соотношения (19)—(21) доказывают принципиальную возможность вычисления интеграла от произвольной интегрируемой функции каждым из трех методов с любой точностью ε за счет выбора достаточно большого n и соответственно малого шага $h = \frac{b-a}{n}$.

Точность. После общего вывода о сходимости методов перейдем к обсуждению основного вопроса, связанного с организацией реального вычислительного процесса: каким нужно взять n, чтобы добиться при

вычислении интеграла нужной точности. Ответ на него требует анализа остаточных членов. При этом на функцию $f(x)$ приходится накладывать некоторые дополнительные ограничения.

Начнем с обсуждения остаточных членов в квадратурных формулах прямоугольников и трапеций. Предположим, что функция $f(x)$ дважды непрерывно дифференцируема на отрезке $[a,b]$. В курсе математического анализа при этом предположении устанавливаются формулы

$$\int_{x_{i-1}}^{x_i} f(x)dx = f(\xi_i)h + \frac{h^3}{24} f''(\eta_i^*), \tag{22}$$

$$\int_{x_{i-1}}^{x_i} f(x)dx = \frac{f(x_{i-1}) + f(x_i)}{2} h - \frac{h^3}{12} f''(\eta_i^{**}), \tag{23}$$

где η_i^* и η_i^{**} - некоторые точки отрезка $[x_{i-1}, x_i]$. Существование таких точек гарантировано, но их точное положение неизвестно.
Суммируя равенства (22) и (23) по i, получаем формулы (11) и (15) со следующими выражениями для остаточных членов:

$$\alpha_n = \frac{h^3}{24} \sum_{i=1}^{n} f''(\eta_i^*), \tag{24}$$

$$\beta_n = -\frac{h^3}{12}. \tag{25}$$

Рассмотрим суммы

$$h\sum_{i=1}^{n} f''(\eta_i^*) \quad \text{и} \quad h\sum_{i=1}^{n} f''(\eta_i^{**}). \tag{26}$$

Функция $f''(x)$ по предположению непрерывна и, следовательно, интегрируема на отрезке $[a,b]$. Следовательно, выражения (26) можно рассматривать как интегральные суммы для интеграла $\int_a^b f''(x)dx$.

Отсюда следует вывод:

$$\lim_{n\to\infty} h\sum_{i=1}^{n} f''\left(\eta_i^*\right) = \int_a^b f''(x)dx = f'(b) - f'(a), \tag{27}$$

$$\lim_{n\to\infty} h\sum_{i=1}^{n} f''\left(\eta_i^{**}\right) = \int_a^b f''(x)dx = f'(b) - f'(a). \tag{28}$$

Предельные равенства (27) и (28) позволяют записать остаточные члены квадратурных формул прямоугольников и трапеций в виде

$$\alpha_n = \frac{1}{n^2}(A + \mu_n), \tag{29}$$

$$\beta_n = \frac{1}{n^2}(B + \nu_n), \tag{30}$$

где

$$A = \frac{(b-a)^2}{24}[f'(b) - f'(a)], \tag{31}$$

$$\mu_n = \frac{(b-a)^2}{24}\left[h\sum_{i=1}^{n} f''\left(\eta_i^*\right) - \int_a^b f''(x)dx\right] \to 0, \quad n \to \infty, \tag{32}$$

$$B = -\frac{(b-a)^2}{12}[f'(b) - f'(a)], \tag{33}$$

$$\nu_n = -\frac{(b-a)^2}{12}\left[h\sum_{i=1}^{n} f''\left(\eta_i^{**}\right) - \int_a^b f''(x)dx\right] \to 0, \quad n \to \infty, \tag{34}$$

Формулы (29) и (30) выделяют в остаточных членах главные слагаемые $\frac{A}{n^2}$ и $\frac{B}{n^2}$, которые при возрастании n стремятся к нулю как n^{-2}. Важно

подчеркнуть, что коэффициенты A (31) и B (32) от n не зависят. Дополнительные слагаемые $\frac{\mu_n}{n^2}$ и $\frac{\nu_n}{n^2}$ являются бесконечно малыми более высокого порядка. Если ими пренебречь по сравнению с главными слагаемыми, то получатся простые асимптотические представления остаточных членов:

$$\alpha_n = \frac{A}{n^2}, \quad \beta_n = \frac{B}{n^2}. \tag{35}$$

Их относительная точность возрастает при увеличении n.

Перейдем к обсуждению остаточного члена γ_n в методе Симпсона, которое проведем при предположении о четырехкратной непрерывной дифференцируемости подынтегральной функции $f(x)$. Напомним, что в методе Симпсона число точек выбирается четным, поэтому $\frac{n}{2}$ является целым числом.

Рассмотрим отрезок двойной длины $2h$, расположенный между точками разбиения (3) с четными номерами $\lfloor x_{2j-2}, x_{2j} \rfloor$, $1 \le j \le n/2$. В курсе математического анализа выводится формула

$$\int_{x_{2j-2}}^{x_{2j}} f(x)dx = \frac{h}{3}\left[f(x_{2j-2}) + 4f(x_{2j-1}) + f(x_{2j})\right] - \frac{h^5}{90} f^{(4)}(\eta_j), \tag{36}$$

где $\eta_j \in \lfloor x_{2j-2}, x_{2j} \rfloor$. Существование такой точки гарантировано, но ее точное положение на отрезке неизвестно.

Суммируя равенства (36) по j, получаем квадратурную формулу (15) со следующим выражением для остаточного члена:

$$\gamma_n = -\frac{h^5}{90} \sum_{j=1}^{n/2} f^{(4)}(\eta_j). \tag{37}$$

Из формулы (37), можно вывести различные представления остаточного

члена и изучить его свойства. Рассмотрим сумму

$$2h\sum_{j=1}^{n/2} f^{(4)}(\eta_j). \tag{38}$$

Функция $f^{(4)}(x)$ предполагается непрерывной и, следовательно, интегрируемой на отрезке $[a,b]$. С учетом этого (38) можно рассматривать как интегральную сумму для интеграла $\int_a^b f^{(4)}(x)dx$. Отсюда следует вывод:

$$\lim_{n\to\infty} 2h = \int_a^b f^{(4)}(x)dx = f'''(b) - f'''(a). \tag{39}$$

Предельное равенство (39) позволяет записать остаточный член квадратурной формулы Симпсона (37) в виде

$$\gamma_n = \frac{1}{n^4}(C + \sigma_n), \tag{40}$$

$$C = -\frac{(b-a)^4}{180}\{f'''(b) - f'''(a)\}, \tag{41}$$

$$\sigma_n = -\frac{(b-a)^4}{180}\left\{2h\sum_{j=1}^{n/2} f^{(4)}(\eta_j) - \int_a^b f^{(4)}(x)dx\right\} \to 0, \quad n \to \infty. \tag{42}$$

Эта формула выделяет в остаточном члене γ_n главное слагаемое $\frac{C}{n^4}$, которое стремится к нулю как n^{-4}. Коэффициент C (41) не зависит от n. Дополнительное слагаемое $\frac{\sigma_n}{n^4}$ является бесконечно малой более высокого порядка. Если им пренебречь, то получится асимптотическое представление остаточного члена:

$$\gamma_n = \frac{C}{n^4}. \tag{43}$$

Его относительная точность возрастает с увеличением n.

Метод Симпсона является методом более высокого порядка точности - четвертого. В этом его преимущество перед методами прямоугольников и трапеций. Правда, приведенные выше оценки остаточного члена требуют большей гладкости подынтегральной функции — она должна быть четыре раза непрерывно дифференцируема.

8.3 Апостериорные оценки погрешности численного интегрирования

В латинском языке существуют два термина - антонима: априори (*a priori*) и апостериори (*a posteriori*). Первый означает изначально, независимо от опыта, второй — на основании опыта. Оба они часто используются в вычислительной математике, подразделяя информацию на ту, которая известна до начала вычислений, и ту, которая получается в процессе вычислений.

Оценки погрешности квадратурных формул прямоугольников (29), трапеций (30), Симпсона (40) называют *априорными*. Они справедливы изначально и предсказывают точность вычисления интеграла независимо от того, будем мы фактически проводить вычисления или нет. Эти результаты позволяют понять структуру остаточных членов, определить скорость их убывания при возрастании n.

Начнем обсуждение идеи апостериорных оценок погрешности с методов второго порядка - прямоугольников и трапеций. Предположим, что мы провели расчеты по методу прямоугольников с числом точек $n/2$ (n — четное число), а потом с числом точек n и в результате получили два числа - $P_{n/2}$ и P_n. Согласно формулам (6) и (29), это позволяет написать соотношения

$$I = P_{n/2} + \frac{4}{n^2}\left(A + \mu_{n/2}\right), \quad I = P_n + \frac{1}{n^2}\left(A + \mu_n\right). \tag{44}$$

Вычитая теперь второе равенство из первого, получаем

$$P_{n/2} - P_n + \frac{3}{n^2}A + \frac{1}{n^2}\left(4\mu_{n/2} - \mu_n\right) = 0,$$

или

$$\alpha_n = \frac{1}{n^2}\left(A + \mu_n\right) = \frac{1}{3}\left(P_n - P_{n/2}\right) + \frac{4}{3n^2}\left(\mu_n - \mu_{n/2}\right). \tag{45}$$

Первый член в правой части этого представления остаточного члена нам известен из результатов вычислений. Он является главным. Второй член неизвестен, но он по сравнению с первым представляет собой бесконечно малую более высокого порядка. Если им пренебречь, то для погрешности получится простая асимптотическая формула:

$$\alpha_{\mathrm{n}} \approx \frac{1}{3}\left(\mathrm{P}_{\mathrm{n}} - \mathrm{P}_{\mathrm{n}/2}\right). \tag{46}$$

Ее относительная точность возрастает при увеличении n.
Аналогичные формулы имеют место для погрешности метода трапеций:

$$\beta_n \approx \frac{1}{3}\left(T_n - T_{n/2}\right) + \frac{4}{3n^2}\left(\nu_n - \nu_{n/2}\right) \approx \frac{1}{3}\left(T_n - T_{n/2}\right). \tag{47}$$

Для метода Симпсона, который является методом четвертого порядка, формулы немного изменяются. Теперь соотношения, аналогичные (44), будут иметь вид

$$I = S_{n/2} + \frac{16}{n^4}\left(C + \sigma_{n/2}\right), \quad I = S_n + \frac{1}{n^4}\left(C + \sigma_n\right). \tag{48}$$

Здесь число n предполагается кратным четырем, поэтому $n/2$ — четное число. Проводя в (48) вычитание второй строки из первой, получаем

$$\gamma_n = \frac{1}{n^4}\left(C + \sigma_n\right) = \frac{1}{15}\left(S_n - S_{n/2}\right) + \frac{16}{15n^4}\left(\sigma_n - \sigma_{n/2}\right). \tag{49}$$

Здесь опять первый член в правой части равенства известен из вычислений. Он является главным. Второй член неизвестен, но он представляет собой бесконечно малую более высокого порядка по сравнению с первым. Если им пренебречь, то получим асимптотическую формулу для приближенного вычисления погрешности по результатам двух вычислений:

$$\gamma_n = \frac{1}{15}\left(S_n - S_{n/2}\right). \tag{50}$$

Ее относительная точность возрастает с увеличением n.
Обычно апостериорные оценки погрешности с помощью асимптотических формул (46), (47), (50) включаются в компьютерные программы численного интегрирования. Они служат критерием для завершения вычислений после того, как нужная точность достигнута.

9. Лабораторный практикум

9.1 Нахождение корня уравнения

Метод половинного деления

Рассмотрим уравнение

$$f(x)=0. \tag{1}$$

Если функция $f(x)$ непрерывна на отрезке $[a,b]$ и принимает на его концах значения разных знаков, то на этом отрезке существует, по крайней мере, один корень уравнения (1).

Предположим, для определенности, что функция $f(x)$принимает на левом конце отрезка $[a,b]$ отрицательное значение, на правом - положительное:

$$f(a)<0, \quad f(b)>0. \tag{2}$$

Возьмем на отрезке $[a,b]$ среднюю точку $\xi=(b+a)/2$ и вычислим в ней значение функции $f(\xi)$. Если $f(\xi)=0$, то ξ является искомым корнем. При $f(\xi)>0$, в качестве нового отрезка $[a_1,b_1]$ выберем отрезок $[a,\xi]$, а при $f(\xi)<0$ в качестве нового отрезка $[a_1,b_1]$ выберем отрезок $[\xi,b]$. Новый отрезок $[a_1,b_1]$ также содержит корень c, но имеет вдвое меньшую длину. Повторяя эту процедуру n раз мы получаем отрезок $[a_n,b_n]$, содержащий корень c и имеющий длину $h_n=\dfrac{h_0}{2^n}$, где h_0 - длина исходного отрезка $[a,b]$. Длина отрезка h_n представляет собой точность ε, с которой мы знаем положение корня на n - шаге процесса половинного деления.

Метод касательных (метод Ньютона)

Метод касательных, является одним из наиболее эффективных численных методов решения уравнения (1). Идея метода состоит в следующем. Предположим, что функция $f(x)$, имеющая корень c на отрезке $[a,b]$, дифференцируема на этом отрезке и ее производная $f'(x)$не обращается на нем в нуль. Пусть приближенное значение корня функции,

рассчитанное на n-м итерационном шагу равно x_n. Для получения $n+1$-го приближения, запишем уравнение касательной к графику функции $f(x)$ в этой точке:

$$y = f(x_n) + f'(x_n)(x - x_n). \tag{3}$$

В качестве $n+1$-го приближения для корня выберем точку x_{n+1} пересечения касательной с осью абсцисс. Согласно (3)

$$f(x_n) + f'(x_n)(x_{n+1} - x_n) = 0.$$

Следовательно, рекуррентная формула нахождения корня по методу касательных имеет вид

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{4}$$

Метод секущих

Предположим, для определенности, что функция $f(x)$ принимает на левом конце отрезка $[a,b]$ отрицательное значение, на правом - положительное:

$$f(a) < 0, \quad f(b) > 0. \tag{5}$$

Тогда на отрезке $[a,b]$ существует корень с функции $f(x)$. Проведем прямую линию через точки с координатами $(a, f(a))$ и $(\mathrm{b}, \mathrm{f}(\mathrm{b}))$:

$$y = f(f) + \frac{f(b) - f(a)}{b - a}(x - a). \tag{6}$$

Точка пересечения этой линии с осью абсцисс лежит между точками a и b. Координата этой точки ξ может быть рассчитана как

$$\xi = a - \frac{f(a)}{f(b) - f(a)}(b - a). \tag{7}$$

Если $f(\xi) = 0$, то ξ является искомым корнем. При $f(\xi) > 0$, в качестве нового отрезка $[a_1, b_1]$ выберем отрезок $[a, \xi]$, а при $f(\xi) < 0$ в качестве

нового отрезка выберем отрезок $[\xi, b]$. Новый отрезок $[a_1, b_1]$ также содержит корень c, но имеет меньшую длину. Повторяя эту процедуру n раз мы получаем последовательность отрезков убывающей длины и содержащих корень c.

Численная реализация методов нахождения корня уравнения

Задание. Используя описанные выше методы, составить программу нахождения корня одного скалярного уравнения. Исследовать работу программы и сравнить эффективность используемых методов нахождения корня уравнения.

На рис.1 - 3 представлены результаты работы такой программы.

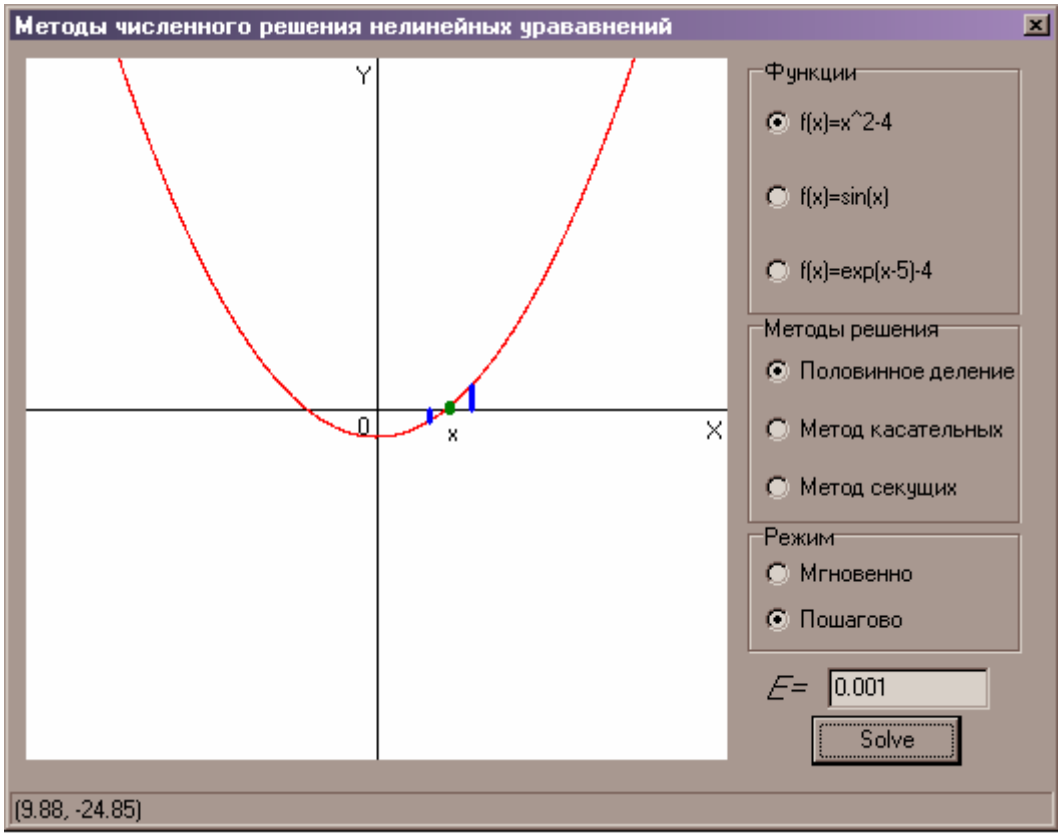


Рис.1. Решение уравнения методом половинного деления

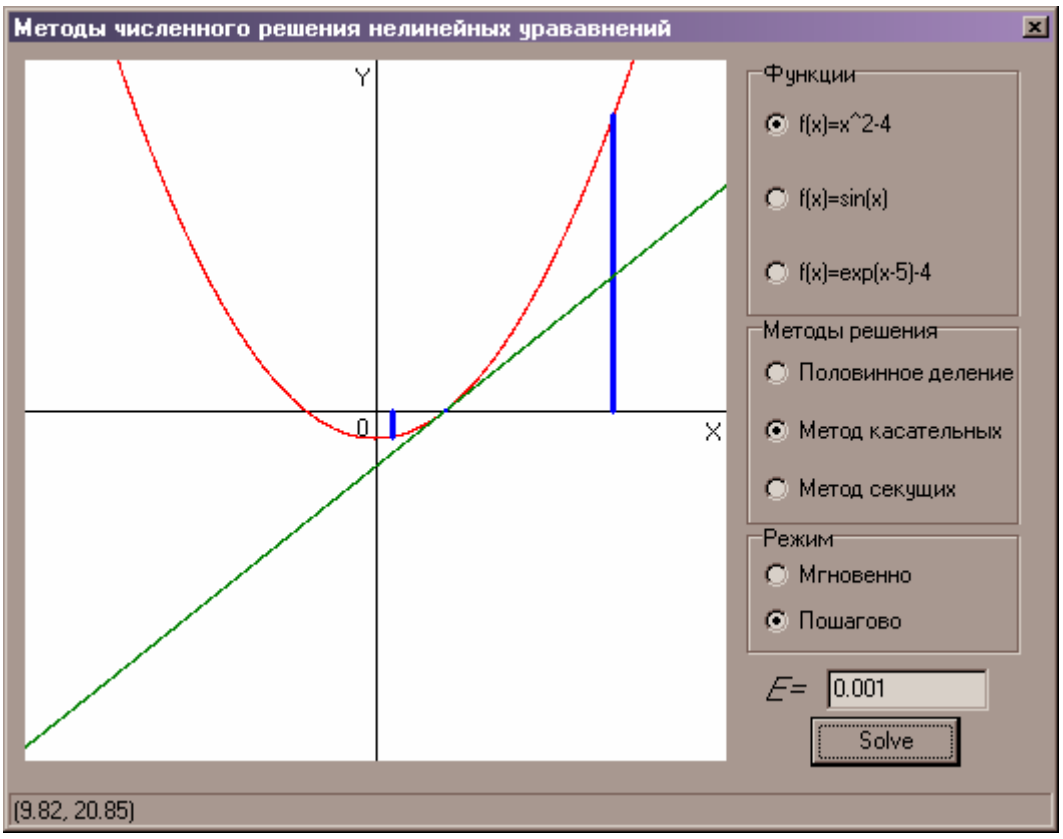


Рис.2. Решение уравнения методом касательных

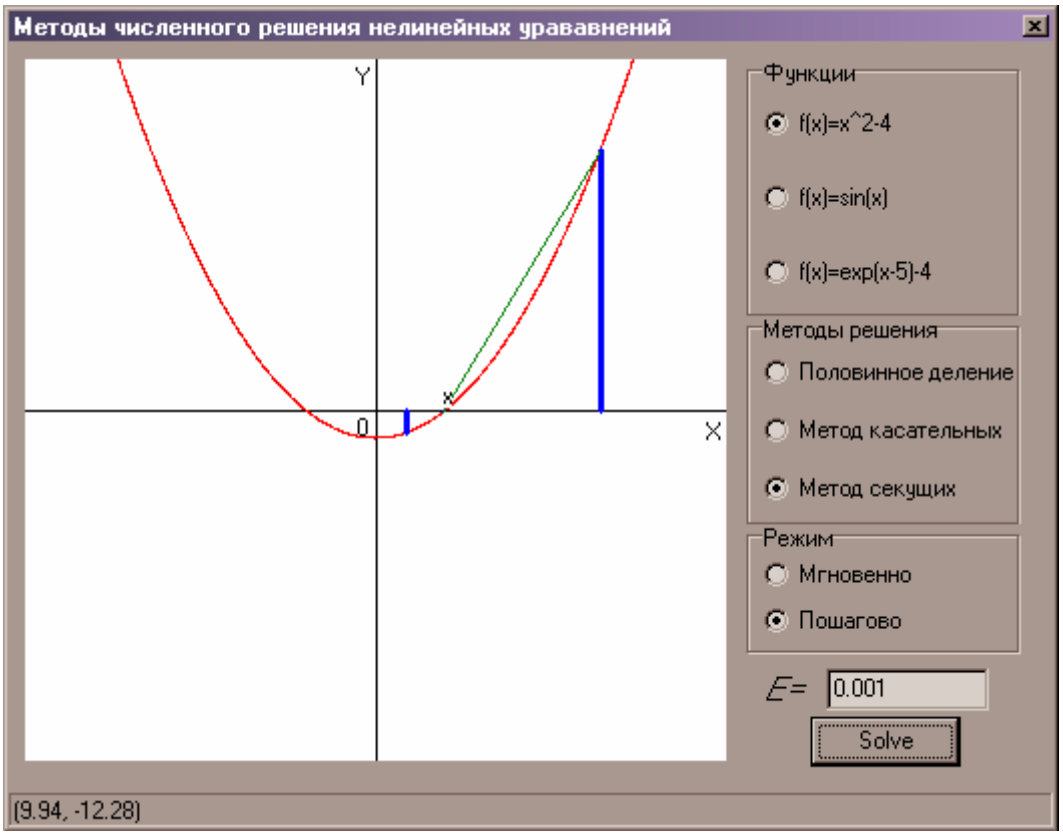


Рис.3. Решение уравнения методом секущих

9.2 Интерполяция функции

Построение интерполяционного полинома в форме Лагранжа

Представим искомый полином $P_n(x)$ в виде

$$P_n(x) = \sum_{i=0}^{n} f(x_i) Q_{n,i}(x), \tag{8}$$

где $Q_{n,i}(x)$ - полиномы степени n:

$$Q_{n,0}(x) = \frac{(x-x_1)(x-x_2)\ldots(x-x_n)}{(x_0-x_1)(x_0-x_2)\ldots(x_0-x_n)},$$

$$Q_{n,i}(x) = \frac{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}, \tag{9}$$

……

$$Q_{n,m}(x) = \frac{(x-x_0)(x-x_1)\ldots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ldots(x_n-x_{n-1})}.$$

Видно, что

$$Q_{n,i}(x) = \begin{cases} 0, & x = x_j \\ 1, & x = x_i \end{cases} \quad \forall j \neq i. \tag{10}$$

Численная реализация метода интерполяции с помощью полинома Лагранжа

Задание. Составить программу построения интерполяционного полинома Лагранжа по заданным значениям функции в узлах интерполяции. Исследовать работу программы. Рассмотреть сходимость выбранного метода интерполяции при увеличении числа узлов.

На рис.4 представлены результаты работы такой программы.

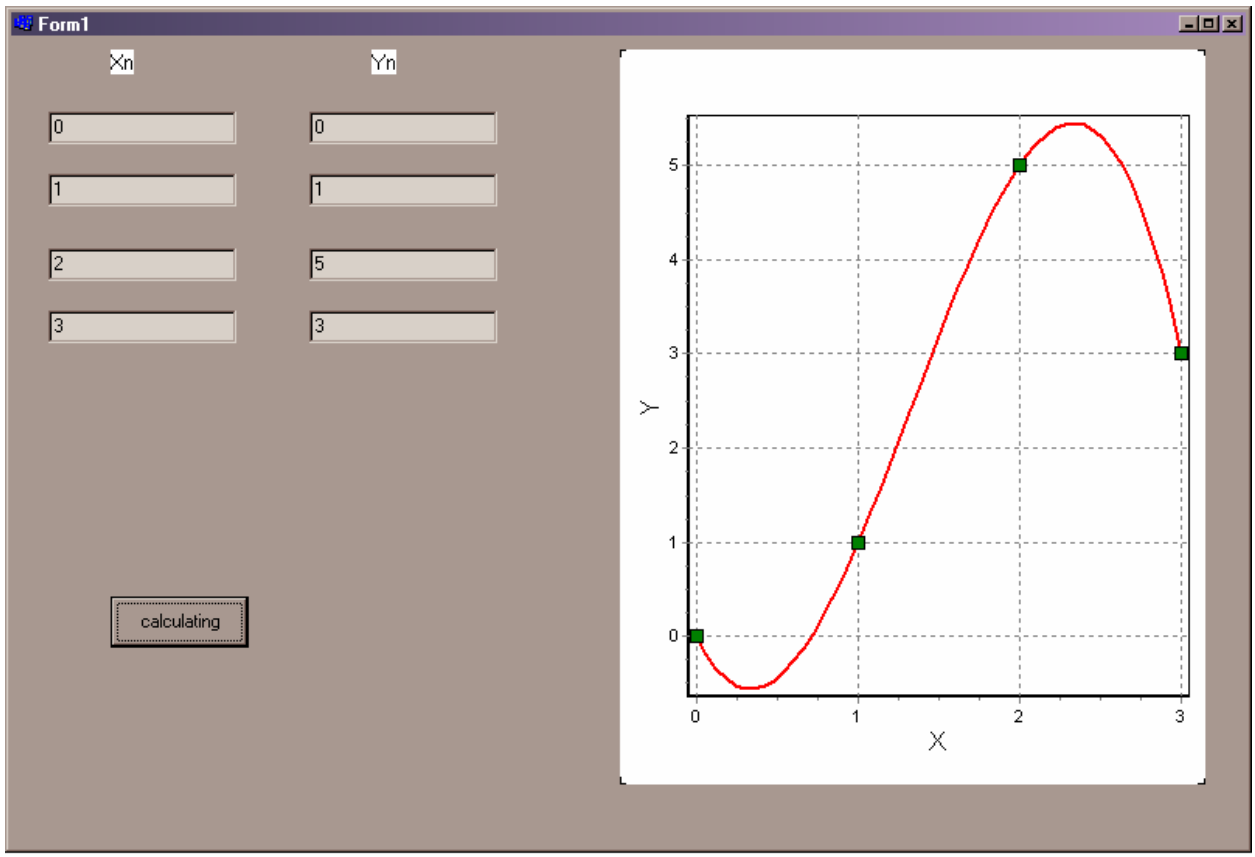


Рис.4. Интерполяция функции с помощью полинома Лагранжа

9.3 Нахождение минимума функции нескольких переменных

Градиентный метод нахождения минимума

Рассмотрим скалярную функцию нескольких переменных,

$$z = f(\boldsymbol{r}), \quad \boldsymbol{r} = (x_1, x_2, ..., x_n), \tag{11}$$

имеющую локальный минимум. Изменение этой функции при малом изменении аргумента можно приближенно записать в виде

$$\delta f = \nabla f \cdot \delta \boldsymbol{r}\,. \tag{12}$$

Таким образом, перемещаясь в направлении, противоположном градиенту функции, мы приближаемся к ее минимуму:

$$\boldsymbol{r}_{n+1} = \boldsymbol{r}_n - \varepsilon \nabla f\,, \tag{13}$$

где ε - малый параметр. В координатном представлении (13) имеет вид

$$x_{i,n+1} = x_{i,n} - \varepsilon \frac{\partial f}{\partial x_i}. \qquad (14)$$

Численная реализация градиентного метода нахождения минимума функции нескольких переменных

Задание. Составить программу нахождения минимума функции двух переменных градиентным методом.

На рис.5 представлены результаты работы такой программы. Исследовать работу программы. Используя составленную программу, проанализировать эффективность градиентного метода нахождения минимума функции при различных значениях ε.

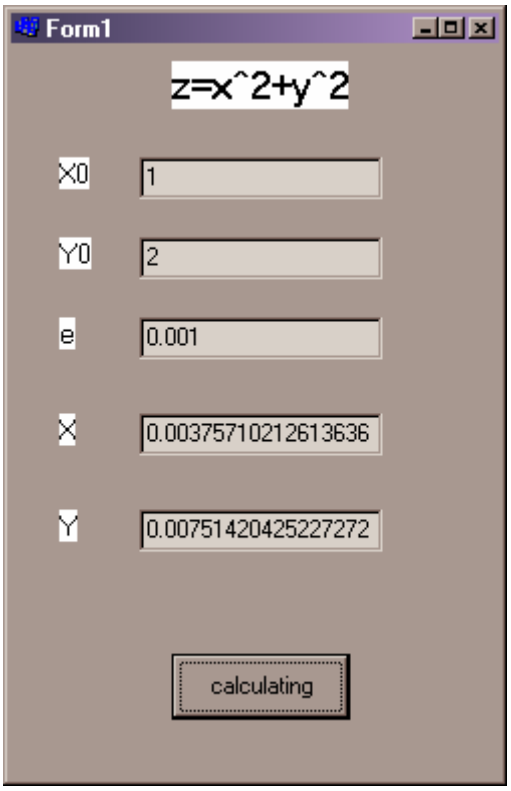


Рис.5. Нахождение минимума функции $z = x^2 + y^2$ градиентным методом

9.4 Численное интегрирование

Возьмем произвольное четное число n и разобьем отрезок $[a,b]$, по

которому ведется интегрирование, на n равных отрезков длиной $h = \dfrac{b-a}{n}$ точками

$$x_i = a + ih, \quad 0 \le i \le n. \tag{15}$$

Идея вывода квадратурных формул заключается в том, чтобы сопоставить подынтегральной функции $f(x)$ близкую ей $g_n(x)$, которую можно проинтегрировать, и приближенно заменить искомый интеграл I интегралом от этой функции. В методе прямоугольников это кусочно-постоянная функция, в методе трапеций - кусочно-линейная функция, в методе Симпсона второго порядка эта функция на каждом отрезке является полиномом второго порядка. Квадратурные формулы прямоугольников P_n, трапеций T_n и Симпсона второго порядка S_n имеют вид:

$$P_n = \frac{b-a}{n}\sum_{i=1}^{n} f(\xi_i), \;\; \xi_i = a + \left(i - \frac{1}{2}\right)h, \tag{16}$$

$$T_n = \frac{b-a}{n}\left[\frac{1}{2}f(a) + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2}f(b)\right], \tag{17}$$

$$S_n = \frac{b-a}{3n}\left[f(a) + 4f(x_1) + 2f(x_2) + \ldots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(b)\right]. \tag{18}$$

Остаточные члены для первых двух формул убывают как n^{-2}, а для третьей формулы, как n^{-4}. Отметим полезную связь между квадратурными формулами:

$$S_n = \frac{4}{3}T_n - \frac{1}{3}T_{n/2}. \tag{19}$$

Численная реализация методов прямоугольников, трапеций и метода Симпсона второго порядка

Задание. Составить программу нахождения интеграла функции одной переменной используя методы прямоугольников, трапеций и Симпсона второго порядка. Исследовать работу программы при разных шагах сетки. Проверить зависимость остаточных членов от количества узлов. Проверить выполнение соотношения (19). На рис.6 представлены результаты работы такой программы.

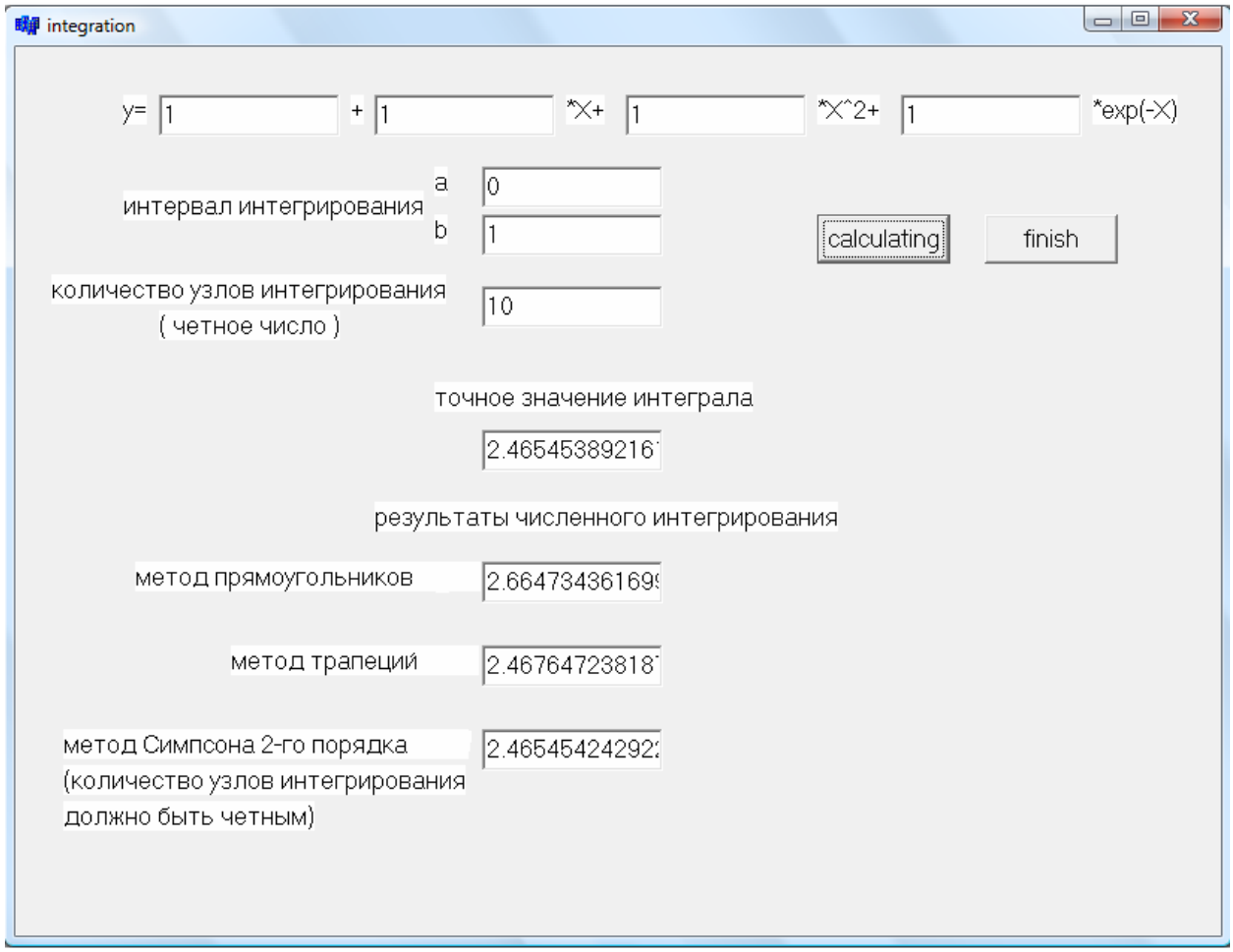


Рис.6. Нахождение определенного интеграла методами прямоугольников, трапеций и Симпсона второго порядка

10. Приложения

Приложение1. Норма матрицы

Рассмотрим линейное вещественное евклидово пространство E_n, элементами которого являются вектора в виде упорядоченной системы n чисел $\boldsymbol{x} = \{x_1,...,x_n\}$. В пространстве E_n определены скалярное произведение

$$(\boldsymbol{x}, \boldsymbol{y}) = x_1 y_1 + ... + x_n y_n \tag{1}$$

и евклидова норма

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})} = \sqrt{x_1^2 + ... + x_n^2}\,, \tag{2}$$

удовлетворяющая трем аксиомам нормы:

1) $\|\boldsymbol{x}\| \geq 0, \quad \|\boldsymbol{x}\| = 0$ тогда и только тогда, когда $\boldsymbol{x} = 0$;

2) $\|\alpha\boldsymbol{x}\| = |\alpha| \cdot \|\boldsymbol{x}\|, \quad \forall \alpha, \boldsymbol{x}$;

3) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ (неравенство треугольника).

Для скалярного произведения справедливо неравенство Коши-Буняковского $|(\boldsymbol{x}, \boldsymbol{y})| \leq \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\||$.

Рассмотрим квадратную матрицу A размером $n \times n$. Она определяет в пространстве E_n линейное преобразование

$$\boldsymbol{y} = A\boldsymbol{x} \tag{3}$$

или

$$y_i = \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1,...,n.$$

Введем величину

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}, \tag{4}$$

которую принято называть *нормой матрицы* A, подчиненной норме вектора $\|x\|$. Записывая ненулевой вектор x в виде

$$x = \|x\| z,$$

где z- вектор единичной длины: $\|z\| = 1$, получаем представление для нормы, эквивалентное (4):

$$\|A\| = \sup_{\|z\|=1} \|Az\|. \tag{5}$$

Отсюда следует, что в конечномерном пространстве норма матрицы ограничена, причем на единичной сфере всегда найдется такой вектор z_0, что $\|A\| = \|Az_0\|$. Наконец, из определения нормы (4) следует, что

$$\|Ax\| \leq \|A\| \cdot \|x\|. \tag{6}$$

Это простое неравенство лежит в основе всех дальнейших оценок.

Приложение 2. Самосопряженность и знакоопределенность матриц

В вещественном евклидовом пространстве E_n для каждого линейного преобразования существует единственное сопряженное ему линейное преобразование, определяемое тождественным равенством скалярных произведений:

$$(Ax, y) = (x, A^+ y), \quad \forall x, y \in E_n. \tag{7}$$

В частности,

$$(Ax, x) = (x, A^+ x), \quad \forall x \in E_n.$$

Преобразование называется *самосопряженным*, если

$$(Ax, y) = (x, Ay), \quad \forall x, y \in E_n. \tag{8}$$

Матрицы сопряженных преобразований в ортонормированном базисе связаны простым транспонированием:

$$a_{ij}^{*} = a_{ji}, \quad \forall i, j = 1,\ldots,n.$$

Как известно, любая матрица представима в виде

$$A = \overline{A} + \widetilde{A}, \tag{9}$$

где

$$\overline{A} = \frac{A + A^{+}}{2} = \overline{A}^{+}, \quad \widetilde{A} = \frac{A - A^{+}}{2} = -\widetilde{A}^{+}. \tag{10}$$

Нетрудно видеть, что

$$\begin{aligned} &(Ax, x) = (A^{+}x, x) = (\overline{A}x, x), \\ &(\widetilde{A}x, x) = 0. \end{aligned} \tag{11}$$

В дальнейшем будем опираться на следующие важные свойства самосопряженных преобразований:
а) все собственные значения самосопряженного линейного преобразования (характеристические числа матрицы A) вещественны;
б) самосопряженное линейное преобразование всегда имеет полный набор линейно независимых собственных векторов, из которых можно образовать ортонормированный базис пространства E_n. В этом базисе матрица линейного преобразования принимает диагональный вид, причем на диагонали стоят все собственные значения этого преобразования с учетом их кратности.

Наконец, матрица линейного преобразования A называется *положительно определенной*, если для любого, отличного от нуля $\boldsymbol{x} \in E_n$ справедливо неравенство $(A\boldsymbol{x}, \boldsymbol{x}) > 0$. В ортонормированном базисе это означает:

$$\sum_{i,j=1}^{n} a_{ij} x_i x_j > 0, \quad \forall x \in E_n, \quad x \neq 0. \tag{12}$$

Для краткости, если это не вызывает недоразумений, будем часто писать $A > 0$.

Необходимым и достаточным условием положительной определенности самосопряженной матрицы A является критерий Сильвестра, из которого, в частности, следует строгая положительность всех диагональных элементов:

$$a_{ii} > 0, \quad 1 \le i \le n. \tag{13}$$

Условимся обозначать собственные векторы линейного преобразования с матрицей A как $\boldsymbol{e}_i$, ее характеристические числа как λ_i, координаты произвольного вектора $\boldsymbol{x}$ в ортонормированием базисе из собственных векторов $\boldsymbol{e}_i$, как ξ_i.

Для дальнейшего рассмотрения будут полезны три леммы.

Лемма 1. *Для того чтобы самосопряженная ($A = A^+$) матрица была положительно определенной, необходимо и достаточно, чтобы все ее характеристические числа были положительны:* $\lambda_i > 0$.

Необходимость. Выберем любой собственный вектор $\boldsymbol{e}_i$ линейного преобразования с матрицей A, тогда

$$(A\boldsymbol{e}_i, \boldsymbol{e}_i) = \lambda_i > 0.$$

Достаточность. Расположим для определенности все характеристические значения матрицы $A = A^+$ в порядке убывания: $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n > 0$. Поскольку по условию леммы $\lambda_i > 0$, то в ортонормированном базисе из собственных векторов преобразования с матрицей A для любого $\boldsymbol{x} \ne 0$ имеем

$$(A\boldsymbol{x}, \boldsymbol{x}) = \sum_{i=1}^{n} \lambda_i \xi_i^2 > 0, \quad \forall \{\xi_i\}, \quad \sum_{i=1}^{n} \xi_i^2 > 0.$$

Поэтому очевидно, что $A > 0$.

Лемма 2. Пусть $A = A^+ > 0$ и $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n > 0$ - упорядоченный набор характеристических чисел этой матрицы, тогда

$$\lambda_n \|\boldsymbol{x}\|^2 \le (A\boldsymbol{x}, \boldsymbol{x}) \le \lambda_1 \|\boldsymbol{x}\|^2. \tag{14}$$

Доказательство предлагается провести самостоятельно.

Лемма 3. Если $A > 0$, то всегда найдется постоянное число $\delta > 0$, такое, что

$$(A\boldsymbol{x}, \boldsymbol{x}) \geq \delta \|\boldsymbol{x}\|^2, \quad \forall \boldsymbol{x} \in E_n. \tag{15}$$

Доказательство. Если $A = A^+$, то достаточно положить $\delta = \lambda_n$. В общем случае напомним, что согласно (11) $(A\boldsymbol{x}, \boldsymbol{x}) = (\overline{A}\boldsymbol{x}, \boldsymbol{x}) > 0$, где $\overline{A} = \overline{A}^*$, поэтому согласно предыдущей лемме

$$(A\boldsymbol{x}, \boldsymbol{x}) = (\overline{A}\boldsymbol{x}, \boldsymbol{x}) \geq \overline{\lambda}_n \|\boldsymbol{x}\|^2,$$

где $\overline{\lambda}_n > 0$ - минимальное характеристическое число матрицы $\overline{A} = \dfrac{A + A^*}{2}$. Полагая $\delta = \overline{\lambda}_n$, приходим к требуемому неравенству (15).

Литература

1. Н.С. Бахвалов, Р.П. Жидков, Г.М. Кобельков. Численные методы. М., Наука, 1987.
2. Д.П. Костомаров, А.П. Фаворский. Вводные лекции по численным методам. М., Логос, 2004.
3. Н.С. Бахвалов, Н.П. Жидков, Г.М. Кобельков. Численные методы. М., Наука, 1987.
4. А.А. Самарский. Введение в численные методы. - СПб., Лань, 2005.
5. Л. Коллатц. Функциональный анализ и вычислительная математика. М., Мир, 1969.
6 С.В. Поршнев, И.В. Беленкова. Численные методы на базе Mathcad. СПб., БХВ-Петербург, 2005.
7. А.Н. Боровский. C++ и Borland C++ Builder. СПб., Питер, 2005.
8. Э.Э. Шноль. Семь лекций по вычислительной математике. М, Едиториал УРСС, 2004.
9. Д. Каханер, К. Моулер, С. Неш. Численные методы и математическое обеспечение. М., Мир, 1998.
10. А.А. Самарский, А.П. Михайлов. Математическое моделирование. М., Физматлит, 2005.

11. Х. Гулд, Я. Тобочник. Компьюторное моделирование в физике. Часть1,2. М., Мир, 1990.

В 2007 году СПбГУ ИТМО стал победителем конкурса инновационных образовательных программ вузов России на 2007–2008 годы. Реализация инновационной образовательной программы «Инновационная система подготовки специалистов нового поколения в области информационных и оптических технологий» позволит выйти на качественно новый уровень подготовки выпускников и удовлетворить возрастающий спрос на специалистов в информационной, оптической и других высокотехнологичных отраслях экономики.

КАФЕДРА ФОТОНИКИ И ОПТОИНФОРМАТИКИ

Кафедра "Фотоника и оптоинформатика" была создана летом 2002 года. Одной из ее важнейших задач является организация учебного процесса и подготовка специалистов по оптоинформатике – стремительно развивающейся новой области науки и техники, в которой разрабатываются оптические технологии сверхбыстрой передачи, обработки и записи информации, создаются быстродействующие оптические компьютеры и системы искусственного интеллекта. Разработка таких оптических информационно-телекоммуникационных технологий, представляющих собой информационные технологии нового поколения, является приоритетным направлением развития российской науки, техники и технологий.

В рамках образовательного направления 200600 «Фотоника и оптоинформатика» студентам читаются лекционные курсы по оптической физике, теории информации и кодирования, архитектуре вычислительных систем, технологии программирования, цифровым оптическим вычислениям, оптическим технологиям искусственного интеллекта, голографическим системам записи информации и другим актуальным проблемам оптоинформатики, а также по квантовой информатике. Эти лекционные курсы поддержаны эксклюзивными учебно-научными экспериментальными практикумами.

Среди научных подразделений кафедры – научно-образовательный центр фемтосекундной оптики и фемтотехнологий, лаборатория компьютерного моделирования и параллельных вычислений, проблемная лаборатория волновых процессов. Среди студентов и аспирантов кафедры – стипендиаты Президента и Правительства Российской Федерации, победители конкурсов научных работ, проводимых Российской Академией наук, крупнейшими мировыми научными обществами.

Сергей Анатольевич Чивилихин

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ В ТЕХНОЛОГИЯХ ПРОГРАММИРОВАНИЯ

ЭЛЕМЕНТЫ ТЕОРИИ И ПРАКТИКУМ

В авторской редакции компьютерный набор, верстка, дизайн С.А.Чивилихин

Редакционно-издательский отдел СПбГУ ИТМО

Зав. РИО Н.Ф.Гусарова

Подписано к печати 23.12.08

Заказ N Тираж 100 экз.

Отпечатано на ризографе

Редакционно-издательский отдел
Санкт-Петербургского государственного
университета информационных
технологий, механики и оптики
197101, Санкт-Петербург, Кронверкский пр., 49