МЕХАНИКО-
МАТЕМАТИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in
ЛЕКЦИИ УЧЕНЫХ МГУ

ЧИСЛЕННЫЕ МЕТОДЫ. ЧАСТЬ 1

КОБЕЛЬКОВ
ГЕОРГИЙ МИХАЙЛОВИЧ

МЕХМАТ МГУ

КОНСПЕКТ ПОДГОТОВЛЕН СТУДЕНТАМИ, НЕ ПРОХОДИЛ ПРОФ. РЕДАКТУРУ И МОЖЕТ СОДЕРЖАТЬ ОШИБКИ. СЛЕДИТЕ ЗА ОБНОВЛЕНИЯМИ НА VK.COM/TEACHINMSU.

ЕСЛИ ВЫ ОБНАРУЖИЛИ ОШИБКИ ИЛИ ОПЕЧАТКИ, ТО СООБЩИТЕ ОБ ЭТОМ, НАПИСАВ СООБЩЕСТВУ VK.COM/TEACHINMSU.

БЛАГОДАРИМ ЗА ПОДГОТОВКУ КОНСПЕКТА
СТУДЕНТКУ ФАКУЛЬТЕТА ВМК МГУ
НЕДОЛИВКО ЮЛИЮ НИКОЛАЕВНУ

Содержание

Лекция 1

Введение

Скажем вначале несколько слов о предмете. Современные численные методы нацелены на решение сложных задач математической физики, статистики, банковского дела, экологии и других предметных областей.

Конечно, численные методы существовали и до появления компьютеров. Зачем же нужна большая мощность? Рассмотрим, например, задачу *прогноза погоды*.

В упрощенном варианте эта задача представляет собой решение системы уравнений динамики океана и уравнений динамики атмосферы. Уравнения динамики океана это 6 нелинейных уравнений в частных производных: 3 уравнения движения, уравнение неразрывности, уравнение для солености и уравнение температуры. Это уравнения от трех пространственных переменных и временной переменной.

Уравнения динамики океана, вообще говоря, надо решать на всем земном шаре. Некоторое время назад эту задачу решали следующим образом. Была построена неравномерная сетка, покрывавшая всю поверхность океана, минимальный шаг от максимального отличался в $2^5 = 32$ раза. Получилось так, что вся поверхность мирового океана была покрыта сеткой с 500000 узлами. Кроме этого, нужно было взять 50-70 уровней по глубине, чтобы считать вертикальную координату. Таким образом, в каждом из 35 миллионов узлов надо аппроксимировать 6 уравнений. Итак, на каждом шаге по времени мы должны решать систему из 200 *миллионов* уравнений.

Понятие погрешности и относительной погрешности

Пусть даны два числа a и $\hat{a}$.

Определение 1.1. Число $\hat{a} - a$ называется *погрешностью*.

Определение 1.2. Число

$$r = \frac{\hat{a} - a}{\hat{a}}$$

называется *относительной погрешностью*.

Интерполяция

Предположим, что в некоторые моменты времени

$$x_1, \ldots, x_n$$

наблюдается функция $f(x)$, то есть имеются наблюдения

$$f(x_1), \ldots, f(x_n).$$

С этой функцией хотелось бы каким-то образом работать – например, интегрировать, дифференцировать и так далее. Для дискретного набора значений такие произвольные операции может быть невозможно или затруднительно.

Поэтому задача сводится к тому, чтобы по исходному набору наблюдений построить функцию, непрерывно дифференциируемую некоторое число раз, так, чтобы в заданных точках x_i она принимала заданные значения $f(x_i)$.

Например, можно поступить следующим образом. Пусть есть некоторые линейно независимые функции

$$\phi_1(x), \ldots, \phi_n(x), \ldots$$

Тогда задача интерполяции ставится следующим образом. Необходимо построить функцию

$$F(x) = \sum_{j=1}^{n} c_j \phi_j(x), \quad F(x_i) = f(x_i),$$

то есть задача сводится к задаче поиска коэффицентов c_j.

В частном случае, когда речь идет об интерполяции многочленами, функции $\phi_j(x)$ имеют вид

$$1, x, \ldots, x^{n-1}, \ldots$$

и искомая функция будет иметь вид

$$L_n(x) = \sum_{j=0}^{n-1} c_j x^j. \tag{1}$$

Данный многочлен называется *интерполяционным многочленом Лагранжа.*

Попробуем поставленную задачу решить. Для значений в заданных точках запишем

$$\begin{cases} c_0 + c_1 x_1 + c_2 x_1^2 + \ldots + c_{n-1} x_1^{n-1} = f(x_1) \\ \ldots\ldots\ldots \\ c_0 + c_1 x_n + c_2 x_n^2 + \ldots + c_{n-1} x_n^{n-1} = f(x_n) \end{cases} \tag{2}$$

Получили систему линейных алгебраических уравнений, причем определитель матрицы этой системы является определителем Вандерморда, поэтому отличен от нуля. Коэффиценты c_j можно найти при любой правой части.

Можем считать, что задача решена.

Заметим, что приведенный выше подход является не очень хорошим с точки зрения вычислений. Для решения системы (2) методом Гаусса потребуется

$$\frac{2}{3}n^3 + O(n^2)$$

арифметических операций.

Решим теперь следующую задачу. Необходимо найти многочлен $\Phi_j(x)$ степени $n-1$ такой, что

$$\Phi_j(x_i) = \sigma_j^i, \tag{3}$$

где σ_j^i – символ Кронекера. Условие (3) означает, что у данного многочлена $n-1$ корень, то есть он имеет вид

$$\Phi_j(x) = \frac{(x-x_1)\ldots(x-x_{j-1})(x-x_{j+1})\ldots(x-x_n)}{(x_j-x_1)\ldots(x_j-x_{j-1})(x_j-x_{j+1})\ldots(x_j-x_n)}. \tag{4}$$

Многочлен (1) имеет вид[1]

$$L_n(x) = \sum_{j=1}^{n} f(x_j)\Phi_j(x). \tag{5}$$

Это очевидно, так как, во-первых, это многочлен степени $n-1$, а во-вторых, с помощью подстановки можно убедиться, что выполняется условие

$$L_n(x_i) = f(x_i).$$

Итого у нас

$$2[2(n-1)+1]n + 2n - 1 = 4n^2 + O(n)$$

арифметических операций.[2] Подчеркнем, что, какая бы форма записи для $L_n(x)$ ни использовалась, речь идет об одном и том же интерполяционном многочлене Лагранжа, то есть из этих форм записи коэффициенты c_j находится однозначно.

Заметим, что, если и в числитель, и в значенатель (4) добавим $(x - x_j)$, для каждой точки x числители всех $\Phi_j(x)$ будут одинаковы. Это поможет сократить количество операций вдвое.

Разделенная разность

Прежде, чем рассмотрим еще одну форму записи для (1), введем понятие *разделенной разности*. По определению, разделенной разностью 0-го порядка называются значения функции

$$f(x_1), \ldots, f(x_n).$$

Разделенная разность 1-го порядка имеет вид

$$f(x_1; x_2) \equiv \frac{f(x_1) - f(x_2)}{x_1 - x_2}.$$

Далее по индукции получим разделенную разность порядка n:

$$f(x_1; x_2; \ldots; x_n) \equiv \frac{f(x_1; \ldots; x_{n-1}) - f(x_2; \ldots; x_n)}{x_1 - x_n}.$$

Каждая разделенная разность следующего порядка по определению зависит от двух предыдущих (рис. 1.1).

Для разделенной разности справедлива следующая формула (другое представление):

$$f(x_1; \ldots; x_n) = \sum_{j=1}^{n} \frac{f(x_j)}{\prod_{i \neq j}(x_j - x_i)}. \tag{6}$$

[1] При этом нам необязательно знать, какой вид имеют c_j, о которых говорилось ранее. Достаточного того, что мы можем вычислить значение многочлена в любой точке x.

[2] Главный член асимптотики (в данном случае это $2n^2$) называется *арифметической сложностью алгоритма.*

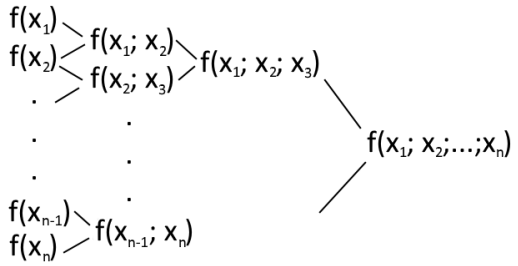

Рис. 1.1. Связь между разделенными разностями

Докажем формулу (6) по индукции. При $n = 2$ имеем:

$$f(x_1; x_2) = \frac{f(x_1)}{x_1 - x_2} + \frac{f(x_2)}{x_2 - x_1} = \frac{f(x_1) - f(x_2)}{x_1 - x_2}.$$

База индукции доказана. Перейдем к шагу индукции. Предположим, что для $n-1$ формула (6) справедлива. Воспользуемся определением и запишем

$$f(x_1; x_2; \ldots; x_n) = \frac{f(x_1; \ldots; x_{n-1}) - f(x_2; \ldots; x_n)}{x_1 - x_n} =$$

$$= \frac{1}{x_1 - x_n} \left[\sum_{j=1}^{n-1} \frac{f(x_j)}{\prod\limits_{\substack{i \neq j \\ i=1}}^{n-1} (x_j - x_i)} - \sum_{j=2}^{n} \frac{f(x_j)}{\prod\limits_{\substack{i \neq j \\ i=2}}^{n} (x_j - x_i)} \right].$$

Вынесем из первой суммы первое слагаемое ($j = 1$), а из второй – последнее ($j = n$). Получим

$$f(x_1; x_2; \ldots; x_n) = \frac{1}{x_1 - x_n} \frac{f(x_1)}{\prod\limits_{i=2}^{n-1} (x_j - x_i)} + \frac{1}{x_n - x_1} \frac{f(x_n)}{\prod\limits_{i=2}^{n-1} (x_j - x_i)} +$$

$$+ \frac{1}{x_1 - x_n} \sum_{j=2}^{n-1} f(x_j) \left[\frac{1}{\prod\limits_{\substack{i \neq j \\ i=1}}^{n-1} (x_j - x_i)} - \frac{1}{\prod\limits_{\substack{i \neq j \\ i=2}}^{n} (x_j - x_i)} \right].$$

Перейдем к общему знаменателю. Первому слагаемому в скобках до него не хватает $(x_j - x_n)$, а второму – $(x_j - x_1)$. Получим

$$f(x_1; x_2; \ldots; x_n) = \frac{f(x_1)}{\prod\limits_{i=2}^{n} (x_j - x_i)} + \frac{f(x_n)}{\prod\limits_{i=2}^{n} (x_j - x_i)} + \sum_{j=2}^{n-1} \frac{f(x_j)}{\prod\limits_{\substack{i \neq j \\ i=1}}^{n} (x_j - x_i)}.$$

Формула (6) доказана.

Из (6) получаем следующие **свойства** разделенной разности:

1. Разделенная разность является линейным функционалом от функции, то есть

$$\alpha f(x_1; x_2; \dots; x_n) + \alpha g(x_1; x_2; \dots; x_n) = \alpha \left(f(x_1; x_2; \dots; x_n) + g(x_1; x_2; \dots; x_n) \right).$$

2. Порядок аргументов не имеет значения, то есть

$$f(x_1; x_2; \dots; x_n) = f(x_{j_1}; x_{j_2}; \dots; x_{j_n}).$$

Оценка погрешности

Будем считать, что

$$x_1, \dots, x_n \in [a, b],$$

а функция f обладает достаточной для доказательства наших утверждений гладкостью.

По значениям

$$f(x_1), \dots, f(x_n)$$

можем построить многочлен (1) $L_n(x)$. Оценим в точке x величину

$$f(x) - L_n(x).$$

Поступим следующим образом. Возьмем функцию

$$\phi(y) = f(y) - L_n(y) - \kappa\omega_n(y), \tag{7}$$

где

$$\omega_n(x) = (x - x_1) \dots (x - x_n).$$

Коэффицент κ выбирается таким образом, чтобы $\phi(x) = 0$ в выбранной для оценки погрешности точке x, то есть

$$\kappa = \frac{f(x) - L_n(x)}{\omega_n(x)}.$$

Заметим, что для всех выбранных точек

$$x_1, \dots, x_n$$

выполнено $\phi(x_i) = 0$.

Так, $\phi(y)$ имеет $n + 1$ корень. Тогда по теоремме Ролля $\phi'(y)$ имеет n корней. Продолжая рассуждения аналогичным образом[3], получим, что $\phi^{(n)}$ имеет 1 корень. Обозначим этот корень через ξ. Рассмотрим

$$\phi^{(n)}(\xi) = 0 = f^{(n)}(\xi) - 0 - \kappa n! = f^{(n)}(\xi) - \frac{f(x) - L_n(x)}{\omega_n(x)} n!$$

[3]Здесь предполагаем, что f имеет достаточно производных.

Отсюда получаем, что[4]

$$f(x) - L_n(x) = \frac{f^{(n)}(\xi)}{n!}\omega_n(x).$$

Отсюда следует оценка

$$\|f(x) - L_n(x)\|_{C[a,b]} \leq \frac{\|f^{(n)}\|_{C[a,b]}}{n!}\|\omega_n\|_{C[a,b]}.$$

Заметим, что если $f(x) = P_n(x)$ – некий многочлен, оценка имеет вид

$$P_n(x) - L_n(x) = \frac{P_n^{(n)}(\xi)}{n!}\omega_n(x),$$

так как $P_n^{(n)}$ – константа.

[4]В правой части равенства стоит некоторая неизвестная точка ξ, значение которой зависит от выбранной для оценки x.

Лекция 2

Краткое повторение материала предыдущей лекции

Вспомним, как ставилась задача.

Даны n точек

$$x_1, \ldots, x_n$$

и значения функции f в этих точках

$$f(x_1), \ldots, f(x_n).$$

Требовалось построить интерполяционный многочлен Лагранжа $L_n(x)$, то есть многочлен степени $n-1$ такой, что

$$L_n(x_j) = f(x_j), \quad j = 1, \ldots, n.$$

Показали, что такой многочлен существует и единственен. Была также выписана формула для этого многочлена[5][6]

$$L_n(x) = \sum_{j=1}^{n} \frac{f(x_j)\omega_n(x)}{\prod\limits_{i \neq j}(x_j - x_i)(x - x_j)}, \tag{8}$$

где

$$\omega(x) = \prod_{i=1}^{n}(x - x_i).$$

В данной лекции речь пойдет о другой форме представления $L_n(x)$. Прежде, чем перейти к обсуждению, вспомним также понятие разделенной разности, выражаемой формулой

$$f(x_1; \ldots; x_n) = \sum_{j=1}^{n} \frac{f(x_j)}{\prod\limits_{i \neq j}(x_j - x_i)}. \tag{9}$$

И, наконец, вспомним, что для погрешности приближения справедлива формула

$$f(x) - L_n(x) = \frac{f^{(n)}(\xi)}{n!}\omega_n(x). \tag{10}$$

Преобразуем формулу (10), разделив обе части на $\omega_n(x)$. Получим

$$\frac{f(x)}{\prod\limits_{i=1}^{n}(x - x_i)} + \sum_{j=1}^{n} \frac{f(x_j)}{(x_j - x)\prod\limits_{i \neq j}(x_j - x_i)} = f(x_1; \ldots; x_n) = \frac{f^{(n)}(\xi)}{n!}. \tag{11}$$

[5] Это представление следует из (5).

[6] Заметим также, что эта формула не очень удобна для вычисления, так как при добавлении новой точки x_{n+1} все придется пересчитывать заново.

Другая запись $L_n(x)$

Будем считать, что $L_k(x)$ – интерполяционный многочлен Лагранжа по k узлам

$$x_1, \ldots, x_k,$$

а $L_{k+1}(x)$ – по $k+1$ узлам

$$x_1, \ldots, x_{k+1}.$$

Тогда справедливо, что

$$L_{k+1}(x) - L_k(x) = A_k \omega_k(x),$$

так как корни этих многочленов в точках $x_1, \ldots, x_k$ совпадают. С учетом формулы (11) можем записать

$$L_{k+1}(x_{k+1}) - L_k(x_{k+1}) = f(x_{k+1}) - L_k(x_{k+1}) = f(x_1; x_2; \ldots; x_{k+1})\omega_k(x_{k+1}).$$

Таким образом,

$$A_k = f(x_1; x_2; \ldots; x_{k+1}).$$

Теперь, можем записать

$$L_n = L_1 + (L_2 - L_1) + \ldots + (L_n - L_{n-1}) =$$

$$f(x_1) + f(x_1; x_2)(x - x_1) + \ldots + f(x_1; \ldots; x_n)\omega_{n-1}(x). \tag{12}$$

Для представления (12) необходимо выполнить (вспомним схему рис. 1.1)

$$3(n-1) + 3(n-2) + \ldots + 3 = \frac{3}{2}n(n-1) \asymp \frac{3}{2}n^2 + O(n)$$

арифметических операций.

Итак, порядок сложности остается прежним, но при добавлении в схему еще одного узла x_{n+1} все предыдущие вычисления остаются актуальными. В этом случае добавляется $O(n)$ операций.

Обобщение задачи

Пусть для каждого из узлов заданы значения функции и некоторого количества ее производных:

$$x_1 : \ f(x_1), f'(x_1), \ldots, f^{(m_1-1)}(x_1)$$

$$\ldots\ldots$$

$$x_n : \ f(x_n), f'(x_n), \ldots, f^{(m_n-1)}(x_n)$$

Обозначим

$$\sum_{i=1}^{n} m_n = N.$$

Задача заключается в том, чтобы построить многочлен[7] $L_N(x)$ степени $N-1$ такой, что

$$L_N(x_1) = f(x_1), \quad L'_N(x_1) = f'(x_1), \dots$$

Сведем данную задачу к предыдущей. Для определенности будем считать, что

$$x_1 < x_2 < \ldots < x_n.$$

Введем новый параметр

$$\varepsilon = \frac{\max_i(x_{i+1} - x_i)}{2N}.$$

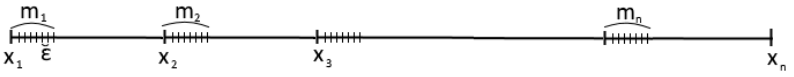

Рис. 2.1. Сетка с дополнительными узлами

Введем дополнительные узлы так, как показано на рис. 2.1.

Для этих узлов можно построить интерполяционный многочлен Лагранжа. По формуле (12), он будет иметь вид

$$L_N^{\varepsilon}(x) = f(x_1) + f(x_1; x_1+\varepsilon)(x-x_1) + f(x_1; x_1+\varepsilon; x_1+2\varepsilon)(x-x_1)(x-(x_1-\varepsilon)) + \dots$$

Найдем теперь, куда перейдут разделенные разности при $\varepsilon \to 0$. Для этого воспользуемся формулой (11).

Возьмем случай $n = 2$. Тогда

$$f(x_1; x_1+\varepsilon) = \frac{f(x_1+\varepsilon) - f(x_1)}{\varepsilon} \to f'(x_1), \quad \varepsilon \to 0.$$

При $n = 3$

$$f(x_1; x_1+\varepsilon; x_1+2\varepsilon) = \frac{f''(\xi)}{2!} \to \frac{f''(x_1)}{2!},$$

так как $\xi \in [x_1, x_1+2\varepsilon]$.

Продолжая рассуждения, получим, что

$$f(\underbrace{x_1; \dots; x_1}_{m_1}) = \frac{f^{(m_1-1)}(x_1)}{(m_1-1)!}.$$

Применим аналогичный предыдущей задаче алгоритм для построения $L_N(x)$.

$$L_N(x) = f(x_1) + f(x_1; x_1)(x-x_1) + \ldots + f(\underbrace{x_1; \dots; x_1}_{m_1})(x-x_1)^{m_1-1} +$$
$$+ f(x_1; \dots; x_1; x_2)(x-x_1)^{m_1} + \ldots +$$
$$+ f(x_1; \dots; x_1; \dots; x_n; \dots; x_n)(x-x_1)^{m_1} \dots (x-x_{n-1})^{m_{n-1}}(x-x_n)^{m_n}. \qquad (13)$$

Многочлен (13) называется *интерполяционным многочленом Лагранжа с кратными узлами.*

[7]Внимательно: используется аналогичное обозначение, но данный многочлен не является интерполяционным многочленом Лагранжа.

Постановка задачи для неравномерной сетки узлов

Итак, есть узлы

$$x_1, \ldots, x_n \in [a, b].$$

Оценка погрешности имеет вид

$$\|f - L_n\|_{C[a,b]} \leq \frac{\|f^{(n)}\|_{C[a,b]}}{n!} \|\omega_n\|_{C[a,b]}.$$

Найдем, при каком выборе узлов правая часть оценки принимает наименьшее значение. От узлов не зависит значение $\|f^{(n)}\|_{C[a,b]}$. Значит, надо выбрать узлы

$$(x_1, \ldots, x_n) = \arg \min_{x_1,\ldots,x_n} \|\omega_n\|_{C[a,b]} = \arg \min_{(x_1,\ldots,x_n)\in[a,b]} \max_{x\in[a,b]} |(x - x_1)\ldots(x - x_n)|.$$

Итак, задача[8] сформулирована.

Многочлены Чебышёва

Напомним, что многочлены Чебышёва выглядят следующим образом:

$$\begin{cases} T_0 \equiv 1, \\ T_1(x) = x, \\ T_{n+1} = 2xT_n(x) - T_{n-1}(x), \quad n = 2, 3, \ldots \end{cases} \tag{14}$$

Обсудим, чем примечательны многочлены (14). Исследуем их **свойства**.

1. T_n – многочлен степени n,

$$T_n(x) = 2^{n-1}x^n + \ldots$$

2. T_{2n} – четная функция, а T_{2n+1} – нечетная.
3. На отрезке $[-1, 1]$ имеет место представление

$$T_n(x) = \cos(n \arccos x).$$

Убедимся в справедливости этого представления. При $n = 0$, $n = 1$ утверждение очевидно. Далее действуем по индукции. Так как

$$T_{n+1}(x) + T_{n-1}(x) = 2xT_n(x),$$

для левой части получим

$$\cos((n + 1) \arccos x) + \cos((n - 1) \arccos x) = 2x \cos(n \arccos x).$$

Здесь воспользовались формулой из школьного курса тригонометрии.

4. Найдем корни.

$$\cos(n \arccos x) = 0,$$

[8] Такие задачи называются *минимаксными*.

тогда

$$n \arccos x = \frac{\pi}{2} + k\pi,$$

$$\arccos x = \frac{\pi}{2n} + \frac{k}{n}\pi,$$

$$x_k = \cos\left(\frac{\pi}{2n} + \frac{k}{n}\pi\right), \quad k = 0, \ldots, n-1.$$

5. Точки экстремума.

$$|\cos(n \arccos x)| = 1,$$

$$x_k = \cos\frac{k\pi}{n}, \quad k = 0, \ldots, n.$$

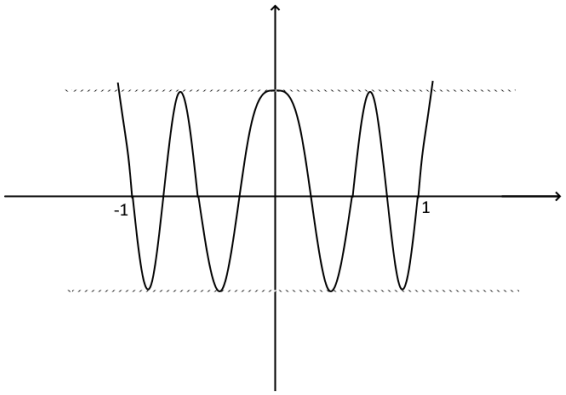

Рис. 2.2. Поведение T_n на $[-1, 1]$

Если построить график T_n на $[-1, 1]$, окажется, что это функция, которая колеблется между своим минимумом и максимумом (рис. 2.2, схематично).

6.Обсудим теперь, что происходит за пределами отрезка $[-1, 1]$. Решим уравнение[9]

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

Решение будем искать в виде

$$T_n(x) = \mu^n,$$

где μ – неизвестное число. Подставив в уравнение, получим

$$\mu^{n+1} - 2x\mu^n + \mu^{n-1} = 0,$$

$$\mu^2 - 2\mu x + 1 = 0,$$

откуда

$$\mu_{1,2} = x \pm \sqrt{x^2 - 1}.$$

[9]Данное уравнение называется *уравнением в конечных разностях.*

Таким образом,

$$T_n(x) = C_1(x + \sqrt{x^2-1})^n + C_2(x - \sqrt{x^2-1})^n.$$

Найдем коэффициенты. При $n = 0$ получим

$$C_1 + C_2 = 1,$$

при $n = 1$

$$(C_1 + C_2)x + (C_1 - C_2)\sqrt{x^2-1} = x.$$

Таким образом, $C_1 = C_2 = 1/2$, и

$$T_n(x) = \frac{1}{2}\left[(x + \sqrt{x^2-1})^n + (x - \sqrt{x^2-1})^n\right].$$

7.

$$T_{3n}(x) = 4T_n(x)\left(T_n(x) - \frac{\sqrt{3}}{2}\right)\left(T_n(x) + \frac{\sqrt{3}}{2}\right).$$

Экстремальные свойства многочленов Чебышёва на $[-1, 1]$

Введем множество многочленов степени n

$$Q = \{P_n(x) = x^n + \ldots\}.$$

Нужно найти $Q_n \in Q$ такой, что

$$Q_n = \arg\min_{P_n \in Q} \|P_n\|_{C[-1,1]} = \arg\min_{P_n \in Q} \max_{x \in [-1,1]} |P_n(x)|.$$

Обозначим

$$\overline{T}_n \equiv 2^{1-n} T_n.$$

Теорема 2.1.

$$\overline{T}_n(x) = \arg\min_{P_n \in Q} \|P_n\|_{C[-1,1]}.$$

Доказательство (От противного) Предположим, что это не так. Тогда $\exists\, Q_n \in Q$ такой, что

$$\|Q_n\| < \|\overline{T}\|.$$

Рассмотрим разность

$$R = \overline{T}_n - Q_n,$$

это многочлен степени $\leq n - 1$. Пусть

$$\xi_0, \ldots, \xi_n$$

– точки экстремума $\overline{T}_n$. Тогда

$$\operatorname{sign} R(\xi_k) = \operatorname{sign} \overline{T}_n(\xi_k).$$

При переходе от одной точки экстремума к другой знак $\overline{T}_n(x)$ меняется, а значит, разность R меняет знак $n+1$ раз, а значит, имеет n корней. Так как R – многочлен степени $\leq n - 1$, получили противоречие по основной теореме алгебры.

Теорема доказана.

Лекция 3

Случай отрезка $[a, b]$

В прошлый раз доказали теорему 2.1. Перейдем теперь от отрезка $[-1, 1]$ к произвольному отрезку $[a, b]$.

Очевидно, в этом случае нужно свести задачу к предыдущему случаю. Запишем

$$T_n\left(\frac{2x-(a+b)}{b-a}\right) = 2^{n-1}\left(\frac{2x-(a+b)}{b-a}\right)^n + \ldots = 2^{n-1}2^n(b-a)^{-n}x^n + \ldots$$

Тогда на отрезке $[a, b]$ среди всех многочленов степени n многочленом с минимальной нормой будет многочлен

$$2^{1-2n}(b-a)^n T_n\left(\frac{2x-(a+b)}{b-a}\right).$$

Оказывается, что корни этого многочлена будут являться оптимальными узлами интерполяции. Многочлен представим в виде

$$2^{1-2n}(b-a)^n T_n\left(\frac{2x-(a+b)}{b-a}\right) = (x-x_1)\ldots(x-x_n) = \omega_n(x).$$

Так как формула для корней многочлена Чебышёва известна, можем выписать их в явном виде. Получим[10]

$$\frac{2x_j-(a+b)}{b-a} = \cos\frac{2j-1}{2n}\pi,$$

$$x_j = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{2j-1}{2n}\pi. \tag{15}$$

Из рис. 3.1 видно, что в середине отрезка корни располагаются почти равномерно, но ближе к краю расстояние между ними уменьшается.

Вернемся к задаче, которую пытались решить. Оценка погрешности для интерполяционного многочлена Лагранжа имеет вид

$$\|f - L_n\|_{C[a,b]} \le \frac{\|f^{(n)}\|_{C[a,b]}}{n!}\|\omega_n\|_{C[a,b]}.$$

Для узлов вида (15) норма $\|\omega_n\|_{C[a,b]}$ будет минимальной. При данном выборе узлов оценка принимает вид

$$\|f - L_n\|_{C[a,b]} \le \frac{\|f^{(n)}\|_{C[a,b]}}{n!}2^{1-2n}(b-a)^n.$$

[10]Будем назвать такие точки *Чебышевским распределением узлов интерполяции.*

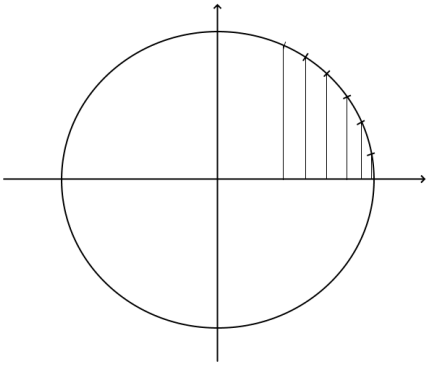

Рис. 3.1. Расположение корней на отрезке

Оценка погрешности для интерполяции с кратными узлами

Напомним, что решали следующую задачу. Для каждой точки x_j были известны значения для функции и нескольких ее производных:

$$\begin{cases} x_1: \ f(x_1), f'(x_1), \ldots, f^{(m_1-1)}(x_1) \\ \ldots\ldots \\ x_n: \ f(x_n), f'(x_n), \ldots, f^{(m_n-1)}(x_n) \end{cases}$$

Был построен многочлен

$$L_N(x): \ L_N^{(k)}(x_j) = f^{(k)}(x_j), \quad j = 1, \ldots, n, \quad k = 0, \ldots, m_j - 1.$$

Для него оказывается справедлива оценка

$$\|f - L_N\|_{C[a,b]} \leq \frac{\|f^{(N)}\|_{C[a,b]}}{N!} \|(x - x_1)^{m_1} \ldots (x - x_n)^{m_n}\|_{C[a,b]}.$$

Постановка задачи наилучшего приближения

Сформулируем постановку задачи в общем случае. Пусть есть некоторое *банахово пространство* B, элемент $f \in B$ и линейно независимые элементы $e_1, \ldots, e_n \in B$.

Задача наилучшего приближения f по $e_1, \ldots, e_n$ формулируется следующим образом:

$$\|f - \sum_{j=1}^{n} c_j e_j\|_B \to \min_{c_1,\ldots,c_n}$$

Если такой минимум существует, элемент, который преставляет минимум нормы разности, называют *элементом наилучшего приближения.*

Теорема 3.1. *Элемент наилучшего приближения существует.*

Доказательство Рассмотрим

$$\Phi(\overline{c}) \equiv \Phi(c_1, \ldots, c_n) = \|f - \sum c_j e_j\|.$$

1. $\Phi(\overline{c})$ – непрерывная функция. Для доказательства воспользуемся неравенством треугольника:

$$\left|\Phi(\overline{c'}) - \Phi(\overline{c})\right| = \left|\|f - \sum c'_j e_j\| - \|f - \sum c_j e_j\|\right| \leq$$

$$\leq \|\sum (c_j - c'_j) e_j\| \leq \sum_{j=1}^{n} |c_j - c'_j| \|e_j\|,$$

откуда и следует непрерывность $\Phi(\overline{c})$.

2. Обозначим

$$|\overline{c}| = \left(\sum_{j=1}^{n} c_j^2\right)^{1/2}.$$

Пусть для некоторого R

$$|\overline{c}| \geq R.$$

Покажем, что за пределами шара радиуса R минимума $\Phi(\overline{c})$ быть не может.

При $|\overline{c}| = 0$ имеем

$$\|\Phi(0, \ldots, 0)\| = \|f\|.$$

Теперь, для $|\overline{c}| \geq R$. имеем[11]

$$\|f - \sum_{c_j} e_j\| = \|f - |\overline{c}| \sum_{j=1}^{n} \frac{c_j}{|\overline{c}|} e_j\| \geq$$

$$\geq |\overline{c}| \|\sum \frac{c_j}{|\overline{c}|} e_j\| - \|f\| \geq R \|\sum \frac{c_j}{|\overline{c}|} e_j\| - \|f\|.$$

Заметим, что коэффициенты $c_j/|\overline{c}|$ пробегают единичную сферу (компакт), а e_j – линейно независимые элементы. Поэтому для

$$\|\sum \frac{c_j}{|\overline{c}|} e_j\|$$

существует некоторый $\min = \mu$, отличный от 0. Тогда, выбрав

$$R \geq 3\|f\|/\mu,$$

получим

$$\|f - \sum_{c_j} e_j\| \geq R\mu - \|f\| \geq 2\|f\|.$$

Таким образом, получими, что за пределами шара значения $\Phi(\overline{c})$ больше, чем в шаре (при $|\overline{c}| = 0$), и значит, минимум достигается внутри шара радиуса R.

Теорема существования минимума доказана.[12]

[11] Здесь снова пользуемся неравенством треугольника.

[12] Заметим, что вопрос единственности минимума требует исследования задачи в каждом конкретном случае

Примеры

Рассмотрим несколько примеров.

- Случай гильбертова пространства $\mathbb{H}$. Рассмотрим

$$\|f - \sum_{j=1}^{n} c_j e_j\|^2 = \|f\|^2 - 2\sum_{j=1}^{n} c_j (f, e_j) + \|\sum c_j e_j\|^2 \to \min$$

Чтобы найти минимум, приравняем к нулю частные производные. Получим

$$\frac{\partial}{\partial c_j}\|f - \sum c_j e_j\|^2 = -2(f, e_j) + 2\sum_{j=1}^{n} c_i (e_i, e_j) = 0.$$

Получаем систему линейных алгебраических уравнений с матрицей

$$a_{i,j} = (e_i, e_j).$$

Матрица такого вида носит название *матрицы Грама* G, а ее определитель отличен от 0. Для $|\overline{c}| \neq 0$

$$\| \underbrace{\sum_{j=1}^{n} c_j e_j}_{\overline{a}} \|^2 = \sum_{i,j} c_i c_j (e_i, e_j) = (G\overline{a}, \overline{a}) > 0.$$

Обратим внимание, что, если система $\{e_j\}$ является ортогональной, матрица G имеет диагональный вид.

- Случай пространства непрерывных функций $C[a, b]$. Задача формулируется следующим образом. Есть элементы

$$f \in C[a, b], \quad g_1(x), \ldots, g_n(x) \in C[a, b].$$

Ищем минимум

$$\|f - \sum_{j=1}^{n} c_j e_j\|_{C[a,b]}.$$

Из общей теоремы следует, что минимум существует. Будем рассматривать частный случай приближения многочленами, то есть задачу

$$\|f - P_n\|_{\to} \min_{\overline{c}}.$$

Здесь

$$g_1 = 0, \quad g_2 = x, \quad \ldots, \quad gn + 1 = x^n.$$

Такой многочлен существует. Обозначим его $Q_n^0(x)$ – многочлен наилучшего равномерного[13] приближения.

[13]Так как используем равномерную метрику.

Теорема об альтернансе

Остановимся подробнее на свойствах многочлена $Q_n^0(x)$.

Теорема 3.2. *(П.Л. Чебышёв) (об альтернансе)* $Q_n^0(x)$ *– многочлен наилучшего равномерного приближения* $\Leftrightarrow \exists\ (n+2)$ *точки* $x_0, \ldots, x_{n+2} \in [a,b]$ *такие, что*

$$f(x_i) - Q_n^0(x_i) = \alpha(-1)^i \|f - Q_n^0\|_{C[a,b]},$$

где $\alpha = 1$ *или* -1.

Обозначим

$$E_n(f) = \|f - Q_n^0\|_{C[a,b]}.$$

Прежде, чем перейти к доказательству теоремы 3.2, докажем следующую лемму.

Лемма 3.1. *(Валле-Пуссен) Пусть* $\exists\ P_n(x)$, $x_0 < x_1 < \ldots < x_{n+1} \in [a,b]$ *такие, что*

$$\operatorname{sign}(-1)^i \left(f(x_i) - P_n(x_i)\right) = \text{const},$$

то есть при переходе от точки к точке данная разность меняет знак. Тогда имеет место оценка

$$E_n(f) \geq \mu = \min_i |f(x_i) - P_n(x_i)|.$$

Доказательство (От противного) Предположим, что это не так, то есть

$$E_n(f) = \|f - Q_n^0\| < \mu \neq 0.$$

Рассмотрим разность

$$Q_n^0(x) - P_n(x)$$

– многочлен степени n. В точках x_i

$$Q_n^0(x_i) - P_n(x_i) = \underbrace{f(x_i) - P_n(x_i)}_{|\cdot| \geq \mu} - \underbrace{\left(f(x_i) - Q_n^0(x_i)\right)}_{|\cdot| < \mu}.$$

Значит, знак разности $Q_n^0(x_i) - P_n(x_i)$ будет определяться знаком первого слагаемого. Так как оно меняет знак на отрезке $[a,b]$ $n+2$ раза по условию, у многочлена $Q_n^0(x) - P_n(x)$ существует $n+1$ корень на этом отрезке. Пришли к противоречию.

Лемма доказана.

Лекция 4

Доказательство теоремы об альтернансе

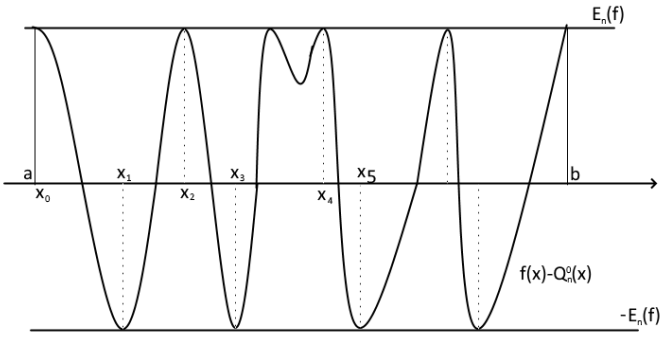

Рис. 4.1. Иллюстрация к т. об альтернансе

Перейдем к доказательству теоремы 3.2. Приведем прежде иллюстрацию к утверждению теоремы (рис. 4.1). Заметим, что в данном примере точка x_4 может быть выбрана двумя способами, так как на промежутке между нулями разности находятся две точки максимума. Обе точки максимума подряд выбирать нельзя, так как между точками альтернанса должна произойти смена знаков разности.

Доказательство

1. Среди всех точек на отрезке $[a, b]$, для которых выполняется

$$|f(x) - Q_n^0(x)| = E_n(f),$$

выделим крайнюю левую точку x_0, то есть

$$|f(x_0) - Q_n^0(x_0)| = E_n(f).$$

2. На отрезке $[x_0, b]$ ищем крайнюю левую точку x_1 из точек, для которых выполнено

$$|f(x_1) - Q_n^0(x_1)| = -E_n(f).$$

3. На $[x_1, b]$ ищем крайнюю левую точку x_2 из всех точек, для которых

$$|f(x_2) - Q_n^0(x_2)| = E_n(f).$$

Продолжая аналогичным образом, строим последовательность точек альтернанса

$$x_0 < x_1 < \ldots < x_k.$$

Предположим, что утверждение теоремы *неверно*, то есть таких точек $k < n + 1$.

Возьмем теперь

$$\xi_1 \in [x_0, x_1]$$

такую, что правее не может быть точек максимума. Затем выбираем

$$\xi_2 \in [x_1, x_2]$$

такую, что правее нет точек минимума (рис. 4.1). Получим последовательность точек

$$\xi_1, \ldots, \xi_k.$$

Отрезок $[a, b]$ оказывается разбит на непересекающиеся отрезки

$$[a, \xi_1] \cup [\xi_1, \xi_2] \cup \ldots \cup [\xi_k, b]. \tag{16}$$

Введем произведение одночленов

$$v(x) = (x - \xi_1)(x - \xi_2) \ldots (x - \xi_k).$$

По нашему предположению (что $k < n + 1$) это многочлен степени $\leq n$. Предположим, что на $[a, \xi_1]$ $v(x)$ принимает положительные значения. Значит, на $[\xi_1, \xi_2]$ этот многочлен будет принимать отрицательные и так далее, на каждом из отрезков (16) $v(x)$ будет менять знак.

На (a, ξ_1) рассмотрим разность

$$f(x) - Q_n^0(x) - \gamma v(x),$$

где для некоторого $\gamma_1 > 0$ $\forall\ \gamma \in (0, \gamma_1)$ выполняется

$$-E_n(f) < f(x) - Q_n^0(x) - \gamma v(x) < E_n(f).$$

Это верно, так как ξ_1 выбрана таким образом, чтобы на (a, ξ_1) у разности $f(x) - Q_n^0(x)$ не было точек минимума, то есть возможно сдвинуть разность так, чтобы она не достигала максимума, но была больше минимального значения. Для отрезка (ξ_1, ξ_2) выберем $0 < \gamma_2 \leq \gamma_1$ такое, что $\forall\ \gamma \in (0, \gamma_1)$ выполняется[14]

$$-E_n(f) < f(x) - Q_n^0(x) - \gamma v(x) < E_n(f).$$

Продолжаем рассуждения для каждого из отрезков (16). Получаем, что для некоторого $\gamma_k > 0$ $\forall\ \gamma \in (0, \gamma_k)$ выполнено

$$-E_n(f) < f(x) - \underbrace{Q_n^0(x) - \gamma v(x)}_{\text{мн-н ст. } n} < E_n(f).$$

Получаем многочлен, который приближает f лучше, чем $Q_n^0(x)$, то есть противоречие.

Теорема доказана.

[14]Заметим, что, так как на $[\xi_1, \xi_2]$ $v(x)$ по нашему предположению отрицательно, на самом деле к разности $f(x) - Q_n^0(x)$ прибавляется нечто положительное.

Единственность многочлена наилучшего приближения

Теорема 4.1. *Многочлен наилучшего равномерного приближения (НРП) единственен.*

Доказательство (От противного) Предположим, что есть два многочлена НРП: $P_n(x)$ и $Q_n(x)$. Тогда их полусумма тоже будет многочленом НРП, так как

$$\|f - \frac{P_n + Q_n}{2}\| \leq \frac{1}{2}\|f - P_n\| + \frac{1}{2}\|f - Q_n\| = E_n(f).$$

Тогда существуют точки альтернанса

$$x_0 < x_1 < \ldots < x_{n+1}$$

и для них выполняется

$$\left| f(x_i) - \frac{P_n(x_i) + Q_n(x_i)}{2} \right| = \frac{1}{2}| \underbrace{f(x_i) - P(x_i)}_{|\cdot| \leq E_n} + \underbrace{f(x_i) - Q_n(x_i)}_{|\cdot| \leq E_n} | = E_n(f).$$

Такое возможно только тогда, когда обе эти разности по модулю $= E_n(f)$ и их знаки совпадают. Отсюда следует, что

$$P_n(x_i) = Q_n(x_i)$$

для $n + 2$ точек. Значит, многочлен степени $\leq n$

$$P_n(x) - Q_n(x)$$

имеет $n + 2$ нуля. Пришли к противоречию.

Теорема доказана.

Свойства многочлена НРП

1. Верно

$$Q_n^0(x) = L_{n+1}(x).$$

Действительно, так как

$$\text{sign}(-1)^i \left(f(x_i) - Q_n^0(x_i)\right) = \text{const}, \quad i = 0, \ldots, n+1,$$

то существует $n + 1$ корней $f(x) - Q_n^0(x)$:

$$\xi_1, \ldots, \xi_{n+1}.$$

Это значит, что

$$Q_n^0(\xi_i) = f(\xi_i), \quad i = 1, \ldots, n+1,$$

то есть $Q_n^0(x)$ является интерполяционным многочленом Лагранжа f по узлам ξ_i.

2. Пользуясь пунктом 1, можем получить оценку сверху для $Q_n^0(x)$:

$$\|f - Q_n^0\|_{C[a,b]} \leq \frac{\|f^{(n+1)}}{(n+1)!} 2^{-2n-1}(b-a)^{n+1}.$$

3. Пусть $f^{(n+1)}(x)$ сохраняет знак на $[a,b]$. Тогда можно получить оценку снизу. В равенстве

$$f(x) - Q_n^0(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_n(x)$$

возьмем модуль от обеих частей и минимум по $n+1$ производной. Получим, что

$$|f(x) - Q_n^0(x)| \geq \frac{\min_{[a,b]} |f^{(n+1)}(x)|}{(n+1)!}|\omega_{n+1}(x)|.$$

Это неравенство верно для всех x. Возьмем

$$\overline{x} = \arg\max |\omega_{n+1}(x)|.$$

Тогда

$$|f(\overline{x}) - Q_n^0(\overline{x})| \geq \frac{\min |f^{(n+1)}|}{(n+1)!}\|\omega_{n+1}\|.$$

Увеличим теперь левую часть неравенства, взяв максимум по всему $[a,b]$, а правую уменьшим, взяв среди всех ω_{n+1} многочлен с минимальной нормой. Получим

$$\|f - Q_n^0\| \geq \frac{\min |f^{(n+1)}|}{(n+1)!} 2^{-2n-1}(b-a)^{n+1}.$$

4. Рассмотрим для определенности отрезок $[-1,1]$, а $f(x)$ – четная. Тогда $Q_n^0(x)$ – тоже четная функция.

Доказательство элементарно. Так как

$$f(-x) = f(x),$$

то многочлен НРП для этой функции

$$Q_n^0(-x) = Q_n^0(x).$$

Аналогично для нечетной функции.

Рассмотрим следующий пример. На отрезке $[-1,1]$ возьмем $f(x) = \cos x$. Тогда оценка в лоб

$$\|\cos x - Q_2^0(x)\|_{C[-1,1]} \leq \frac{\sin 1}{6} 2^{-5} 2^3 = \frac{\sin 1}{24}$$

получится достаточно грубой. На самом деле, в силу четности,

$$\|\cos x - Q_2^0(x)\|_{C[-1,1]} = \|\cos x - Q_3^0(x)\|_{C[-1,1]} = \frac{\sin 1}{24} 2^{-7} 2^4 = \frac{\sin 1}{24 \cdot 8}.$$

Примеры

- Пусть даны $[a,b]$, $f = P_n(x)$. Требуется построить многочлен $Q^0_{n-1}(x)$. Пусть заданный $P_n(x)$ выглядит следующим образом:

  $$P_n(x) = a_n x^n + \ldots + a_0.$$

  Тогда норма разности будет иметь вид

  $$\|P_n - Q^0_{n-1}\|_{C[a,b]} = \|a_n x^n + \ldots\|,$$

  а значит,

  $$P_n(x) - Q^0_{n-1}(x) = a_n T_n\left(\frac{2x-(a+b)}{b-a}\right) 2^{1-n}2^{-n}(b-a)^n.$$

  Отсюда получаем

  $$Q^0_{n-1}(x) = P_n(x) - a_n 2^{1-2n}(b-a)^n T_n\left(\frac{2x-(a+b)}{b-a}\right).$$

- Даны отрезок $[-1,1]$, $f = x^7$. Построим $Q^0_5(x)$.

  В силу нечетности функции,

  $$Q^0_5(x) = Q^0_6(x),$$

  и задача сводится к предыдущей.

- Даны $[-1,1]$, $f = x^7 + P_5(x)$. Требуется построить $Q^0_5(x)$.

  Строим сначала многочлен НРП для x^7, то есть

  $$Q^0_5(x) = x^7 - 2^{-6}T_7(x).$$

  Тогда для f

  $$Q^0_5(x) = x^7 + P_5(x) - 2^{-6}T_7(x),$$

  так как $f - Q^0_5(x)$ будет многочленом Чебышева, а значит, наименьшим уклоняющимся от 0.[15]

- Даны $[0,1]$, $f = \cos x$, построить $Q^0_1(x)$.

  Задача состоит в том, чтобы построить прямую, аппроксимирующую $\cos x$ так, чтобы было 3 точки альтернанса (4.2). Из рисунка видно, что решением будет прямая, расположенная на равном расстоянии от прямой, соединяющей крайние точки функции, и параллельной ей касательной.

  Запишем решение в аналитическом виде.

  $$Q^0_1(x) = ax + b, \quad a = \cos 1 - 1.$$

[15] Заметим, что в данной задаче существенна степень многочлена $P_5(x)$.

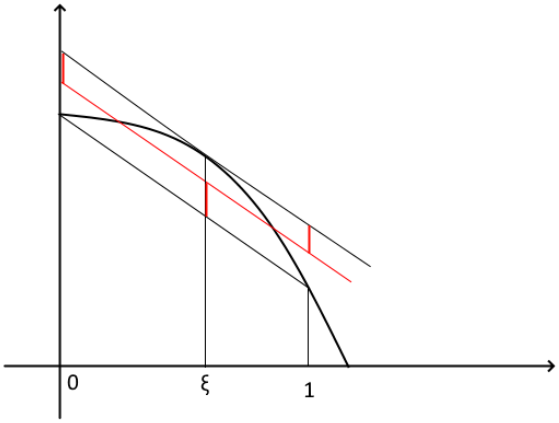

Рис. 4.2. Иллюстрация к решению задачи

Чтобы найти b, запишем условие в точках альтернанса. В точке 0

$$1 - b = a\xi + b - \cos\xi.$$

Так как ξ – внутренняя точка отрезка, а разность в этой точке должна достигать своего экстремального значения, производная разности в этой точке будет 0. Тогда

$$-\sin\xi - a,$$

то есть

$$\xi = \arcsin(\cos 1 - 1).$$

Тогда

$$b = \frac{1}{2}\left(1 - a\xi + \cos\xi\right).$$

Лекция 5

Многочлен наилучшего равномерного приближения

Главный вывод, сделанный на прошлой лекции о многочленах наилучшего равномерного приближения – что он является интерполяционным многочленом по некоторой системе узлов (рис. 5.1). Такое приближение довольно удобно, но, вообще говоря, колебания значений между узлами – не очень хорошее явление.

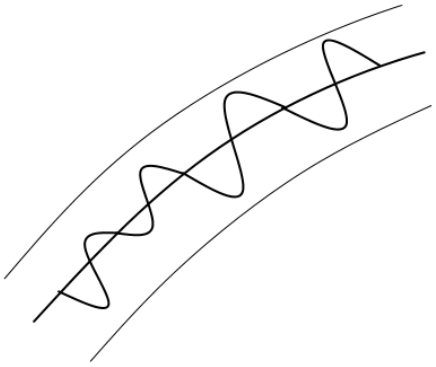

Рис. 5.1. Осцилляции между узлами

Напомним, что погрешность приближения выглядит следующим образом:

$$\|f - Q_n^0\|_{C[a,b]} \leq \frac{\|f^{(n+1)}\|}{(n+1)!} 2^{1+2n} (b-a)^{n+1},$$

то есть зависит от длины интервала $[a, b]$. Было бы логично разбить отрезок с шагом h и на каждом из отрезков строить свой интерполяционный полином. Остается проблемой то, что при таком подходе на стыках отрезков приближение выглядит не очень хорошо.

Сплайны

Пусть есть разбиение отрезка $[0, 1]$ на N отрезков. Длину отрезка $[x_i, x_{i+1}]$ будем обозначать h_{i+1}.

Определение 5.1. *Сплайном m порядка* называется функция $S_m(x)$ такая, что, во-первых, $\forall\ [x_i, x_{i+1}]$ $S_m(x)$ – многочлен степени m, во-вторых,

$$S_m(x_i) = f(x_i),$$

и, в третьих, $S^{(k)}(x)$ непрерывна в x_i, $i = 1, \ldots, N-1$, $k = 1, \ldots, m-1$.

Проверим, совпадает ли число уравнений для сплайнов с числом неизвестных. Отрезков разбиения N штук, на каждом из отрезков $m+1$ неизвестных коэффициента. Всего $(m+1)N$ неизвестных.

Перейдем к числу уравнений. В крайних точках (0 и 1) пишется по 1 уравнению. В каждой внутренней точке разбиения пишем по два уравнения (слева и справа). Кроме того, для каждой из $N-1$ точек надо выписать $m-1$ уравнение непрерывности производной. Всего получаем

$$2 + 2(N-1) + (m-1)(N-1) = 2N + mN - m + 1 = N(m+1) - m + 1.$$

Если $m > 1$, то для определения сплайна нужны дополнительные условия.[16]

Свойства сплайнов

Рассмотрим функционал

$$I_1(s) = \int_0^1 (s')^2 dt$$

и класс

$$K_1 = \{s(t) : \ s(x_i) = f(x_i), \quad s \text{ – непрерывно дифференциируема на } \forall\ (x_i, x_{i+1}\}\,.$$

Будем искать

$$s_1(t) = \arg\min_{s \in K_1} I_1(s),$$

то есть

$$I_1(s_1 + \gamma\delta) \geq I_1(s_1),$$

где γ – любое действительное число, а $\delta = \delta(t)$ – допустимая вариация[17].

Будем считать, что $\delta_i = 0$ вне отрезка $[x_i, x_{i+1}]$. Тогда

$$I_1(s + \gamma\delta) = I_1(s) + 2\gamma \int_{x_i}^{x_{i+1}} s'\delta' dt + \gamma^2 \int_{x_i}^{x_{i+1}} (\delta')^2 dt.$$

Если в $I_1(s)$ достигается минимум, то при $\gamma \to 0$ второе слагаемое является главным членом асимптотики. Так как по условию γ может быть как положительным, так и отрицательным, необходимо, чтобы[18]

$$\int_{x_i}^{x_{i+1}} s'\delta' dt = 0.$$

В предположении, что s имеет вторую производную, можем проинтегрировать по частям:

$$\int_{x_i}^{x_{i+1}} s''(t)\delta(t) dt = 0, \quad \forall\ \delta(t)$$

[16] Заметим, что при $m = 1$ сплайн представляет собой ломанную.

[17] То есть функция, добавление которой оставляет внутри класса. В нашем случае это гладкая функция такая, что $\delta(x_i) = 0$.

[18] Данный интеграл называется *первой вариацией*.

и получить

$$s''(t) = 0,$$

откуда следует, что s – многочлен первой степени.

Повторим те же рассуждения еще раз. Пусть

$$I_2(s) = \int_0^1 (s'')^2 dt$$

и класс

$$K_2 = \{s(t) : \ s(x_i) = f(x_i), \quad s' \text{ – непрерывна, } s'' \text{ – непрерывна } \forall\ [x_i, x_{i+1}]\}.$$

Аналогично предыдущим рассуждениям, выберем $\delta(x)$ – бесконечно дифференцируемую, $\delta(x) = 0$ вне $[x_i, x_{i+1}]$ и

$$\delta(x_i) = \delta(x_{i+1}) = \delta'(x_i) = \delta'(x_{i+1}) = 0.$$

Тогда

$$I_2(s + \gamma\delta) = I_2(s) + 2\gamma \int_{x_i}^{x_{i+1}} s''\delta''dt + \gamma^2 \int_{x_i}^{x_{i+1}} (\delta'')^2 dt \geq I_2(s),$$

если s – точка минимума. Значит,

$$\int_{x_i}^{x_{i+1}} s''\delta''dt = 0.$$

В предположении, что на отрезке $[x_i, x_{i+1}]$ функция s имеет непрерывную четвертую производную, проинтегрируем два раза по частям. Получим, что

$$\int_{x_i}^{x_{i+1}} s^{(4)}(t)\delta(t)dt = 0, \quad \forall\ \delta(t).$$

По основной теореме вариационного счисления,

$$s^{(4)} = 0,$$

то есть s – полином третьей степени.

Рассмотрим теперь $\delta(x)$ следующего вида: $\delta(x) = 0$ вне $[x_{i-1}, x_{i+1}]$, $\delta(x) < 0$ на $[x_{i-1}, x_i]$, $\delta(x) > 0$ на $[x_i, x_{i+1}]$ и

$$\delta(x_i) = \delta(x_{i\pm1}) = \delta'(x_{i-1}) = \delta'(x_{i+1}) = 0.$$

Тогда получим, что

$$I_2(s + \gamma\delta) = I_2(s) + 2\gamma \int_{x_{i-1}}^{x_{i+1}} s''\delta''dt + \gamma^2 \int_{x_{i-1}}^{x_{i+1}} (\delta'')^2 dt \geq I_2(s),$$

если s – точка минимума. Отсюда следует, что

$$\int_{x_{i-1}}^{x_{i+1}} s''\delta''dt = 0.$$

Разобъем данный интеграл на части:

$$\int_{x_{i-1}}^{x_i} s''\delta''dt + \int_{x_i}^{x_{i+1}} s''\delta''dt = s''(x_i-)\delta'(x_i) - s''(x_i+)\delta'(x_i) + \int_{x_{i-1}}^{x_{i+1}} s^{(4)}\delta dt = 0.$$

Отсюда следует, что s'' непрерывна. Получаем, что это *кубический* сплайн.

Рассмотрим еще один случай. Пусть на отрезке $[x_0, x_1]$ $\delta(x) \to 0$ при $x \to 1$. Тогда, повторив рассуждения из предыдущего пункта, получим

$$s''(x_0) = 0, \quad s''(x_N) = 0,$$

то есть получили недостающее условие на сплайн.

Так как s – сплайн третьей степени, его вторая производная будет линейной функцией. Запишем ее в виде

$$s''(x) = M_{i+1}\frac{x - x_i}{h_{i+1}} + M_i\frac{x_{i+1} - x}{h_{i+1}}, \quad x \in [x_i, x_{i+1}], \quad h_{i+1} = x_{i+1} - x_i.$$

Тогда s имеет вид

$$s(x) = M_{i+1}\frac{(x - x_i)^3}{6h_{i+1}} + M_i\frac{(x_{i+1} - x)^3}{6h_{i+1}} + A_i\frac{x - x_i}{h_{i+1}} + B_i\frac{x_{i+1} - x}{h_i}.$$

При этом

$$s(x_i) = f(x_i) = M_i\frac{h_{i+1}^2}{6} + B_i,$$

откуда получаем

$$B_i = f(x_i) - M_i\frac{h_{i+1}^2}{6}.$$

Аналогичным образом получаем, что

$$A_i = f(x_i) - M_{i+1}\frac{h_{i+1}^2}{6}.$$

Найдем значения M_i, воспользовавшись непрерывностью s'.

$$s'(x) = M_{i+1}\frac{(x - x_i)^2}{2h_{i+1}} - M_i\frac{(x_{i+1} - x)^2}{2h_{i+1}} +$$

$$+ \left(f(x_i) - M_{i+1}\frac{h_{i+1}^2}{6}\right)\frac{1}{h_{i+1}} - \left(f(x_i) - M_i\frac{h_{i+1}^2}{6}\right)i\frac{1}{h_i}.$$

Рассмотрим два отрезка $[x_{i-1}, x_i]$ и $[x_i, x_{i+1}]$.

$$s'(x_i-) = M_{i+1}\frac{h_{i+1}}{2} + \left(f(x_i) - M_{i+1}\frac{h_{i+1}^2}{6}\right)\frac{1}{h_{i+1}} - \left(f(x_i) - M_i\frac{h_{i+1}^2}{6}\right)i\frac{1}{h_i},$$

или, по-другому,

$$s'(x_i-) = M_i\frac{h_{i+1}}{3} + f_i\frac{1}{h_i} - \left(f_{i-1} - M_{i-1}\frac{h_i^2}{6}\right)i\frac{1}{h_i}.$$

Аналогично получим, что

$$s'(x_i+) = -M_i\frac{h_{i+1}}{3} + \left(f_{i+1} - M_{i+1}\frac{h_{i+1}^2}{6}\right) i\frac{1}{h_{i+1}} - f_i\frac{1}{h_{i+1}}.$$

Так как s' непрерывна, приравняем $s'(x_i-)$ и $s'(x_i+)$ и сгруппируем все слагаемые, связанные с M_i:[19]

$$\begin{cases} M_{i-1}\frac{h_i}{6} + \frac{h_i+h_{i+1}}{3}M_i + M_{i+1}\frac{h_{i+1}}{6} = \varphi_i, \\ M_0 = M_N = 0, \quad i = 1, \dots, N-1. \end{cases}$$

Получаем линейную систему алгебраических уравнений, решение которой находится методом прогонки за $O(N)$ операций.

Остановимся подробнее на несколько искусственном условии $M_0 = M_N = 0$. Вообще говоря, нужно было взять первые четыре точки (x_0, x_1, x_2, x_3) и построить по ним интерполяционный многочлен $L_3(x)$. Затем полагается

$$M_0 = L_3''(x_0).$$

Аналогично для M_N и правого конца отрезка.

При таком выборе M_0, M_N справедлива оценка

$$\|f^{(k)}(x) - S_2^{(k)}(x)\|_{C[0,1]} \le ch^{1-k}, \quad h = max_i h_i, \quad k = 0, \dots, 3,$$

где c – некоторая константа.

Таким образом, сплайны приближают не только функцию, но и ее производные.

[19] Оставшиеся слагаемые для краткости обозначим как φ_i.

Лекция 6

В прошлый раз речь шла о построении *интерполяционного сплайна*. Существенным недостатком этого метода является то, что для построения такого сплайна необходимо уже знать все N значений. Возникает вопрос: можно ли построить такое же хорошее приближение, зависимость которого от x носила бы *локальный* характер[20]?

Оказывается, что можно построить хорошую аппроксимацию, если ослабить условие

$$s(x_i) = f(x_i).$$

Такой сплайн будет называться *аппроксимационным*.[21]

Аппроксимационный сплайн

Будем работать с равномерной сеткой на отрезке $[0, 1]$, шаг $h = 1/N$, $x_i = ih$.
Введем функцию $B(x)$ обладающую следующими свойствами:

1. B отлична от 0 на отрезке $[-2, 2]$;
2. B – полином третьей степени на любом $[i, i + 1]$;
3. B, B', B'' – непрерывны, $B''(0)$ не непрерывна.
4. $B(x) = B(-x)$.

Определим $B(x)$ как (рис. 6.1)

$$B(x) = \begin{cases} \frac{2}{3} - x^2 + \frac{1}{2}|x|^3, & |x| \le 1, \\ \frac{1}{6}\left(2 - |x|\right)^3, & 1 < |x| \le 2, \\ 0, & |x| > 2. \end{cases}$$

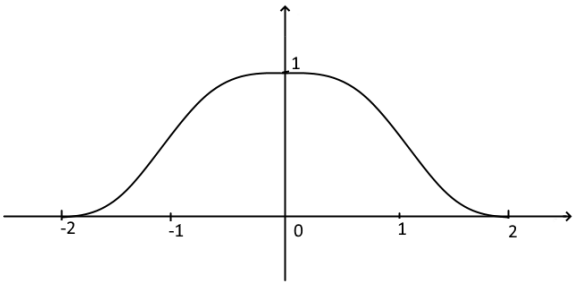

Рис. 6.1. График $B(x)$

[20] То есть зависило бы только от значений функции в нескольких ближайших узлах.
[21] Заметим, что на практике значения функции в точках и так задаются приближенно.

Определение 6.1. *Аппроксимационным сплайном* называется функция[22]

$$B_2^{(k)} \equiv \sum_{j=-2}^{N+2} \alpha_j^{(k)} B\left(\frac{x-jh}{h}\right).$$

При этом полагаем

$$f(x_{N+1}) = L_3(x_{N+1}), \quad f(x_{N+2}) = L_3(x_{N+2}),$$

где $L_3(x)$ – интерполяционный многочлен Лагранжа по узлам x_N, x_{N-1}, x_{N-2}, x_{N-3}. В случае $k = 1$

$$\alpha_j^{(1)} = f(x_j),$$

а в случае $k = 2$

$$\alpha_j^{(2)} = \frac{8f(x_j) - f(x_{j-1}) - f(x_{j+1})}{6}.$$

Для аппроксимационных сплайнов имеют место оценки:

$$\|f^{(k)}(x) - B_2^{(1)}(x)\|_{C[0,1]} \le ch^{2-k}, \quad k = 0, 1;$$

$$\|f^{(k)}(x) - B_2^{(2)}(x)\|_{C[0,1]} \le ch^{4-k}, \quad k = 0, \ldots, 3,$$

где c – некоторая константа.

Быстрое дискретное преобразование Фурье

По сути, быстрое дискретное преобразование Фурье является алгоритмом суммирования ряда Фурье. Прежде, чем перейти непосредственно к нему, вспомним некоторые вещи о рядах Фурье.

Разобьем отрезок $[0, 1]$ с шагом $h = 1/N$. Будем предполагать, что $f(x)$ – непрерывно дифференциируемая функция такая, что

$$\int_0^1 |f| dx < \infty.$$

Этой функции можно поставить в соответствие ряд Фурье:

$$f(x) = \sum_k f_k e^{2\pi i k x}.$$

Из непрерывной дифференциируемости $f(x)$ следует, что

$$\sum |f_k| < \infty.$$

Для произвольного узла сетки $x = nh$ можем записать

$$f(x) = \sum_k f_k e^{2\pi i k (nh)}.$$

[22] В данном случае верхний индекс $B_2^{(k)}$ обозначает индекс 1 или 2, а не производную.

Представим k в виде

$$k = k_1 N + k_2,$$

где $0 \le k_2 \le N-1$.

В силу абсолютной сходимости ряда можем переставлять его члены в произвольном порядке, поэтому можем записать

$$f(x) = \sum_{k_2=0}^{N-1} \left(\sum_{k_1} f_{k_1 N + k_2} e^{2\pi i (k_1 N + k_2)(nh)} \right).$$

Так как $e^{2\pi i k_1 n (Nh)} = 1$,

$$f(x) = \sum_{k_2=0}^{N-1} \underbrace{\left(\sum_{k_1} f_{k_1 N + k_2} \right)}_{F_{k_2}} e^{2\pi i k_2 x}.$$

Таким образом, при равномерной дифференциируемости[23] $f(x)$ на отрезке $[0,1]$ в узлах равномерной сетки справедливо равенство

$$f(x) = \sum_{k=0}^{N-1} f_k e^{2\pi i k_x}, \tag{17}$$

где f_k – переобозначение коэффицентов F_{k_2}.

Рассмотрим теперь ситуацию, когда $f(x)$ задана на сетке. Сведем задачу к предыдущей. Для этого каким-либо гладким образом соединим точки-значения функции[24]. Для такой функции верно сказанное выше, поэтому будет справедливо (17).

Обсудим, как находить коэффициенты f_k из (17). Введем на множестве комплекснозначных функций на сетке $\{0, h, \dots, (N-1)h\}$ скалярное произведение

$$(f, g) \equiv \sum_{j=0}^{N-1} h f(jh) \overline{g}(jh).$$

Утверждение Система функций $\{e^{2\pi i k x}\}$ ортонормирована.

Доказательство Проверяем:

$$(e^{2\pi i k x}, e^{2\pi i m x}) = \sum_{j=0}^{N-1} h e^{2\pi i (k-m) jh} =$$

$$= \begin{cases} 1, & k = m, \\ h(e^{2\pi i (k-m) hN} - 1)/(1 - e^{2\pi i (k-m)h}) = 0, & k \ne m, \end{cases}$$

[23] Вообще говоря, возможно наложить более слабые условия на $f(x)$.
[24] Этот процесс называется *восполнением*.

так как в случае $k \neq m$ запись представляет собой сумму геометрической прогрессии.

Утверждение доказано.

Воспользуемся этим утверждением, чтобы найти f_k:

$$f_k = (f, e^{2\pi ikx}) \equiv \sum_{j=0}^{N-1} hf(jh)e^{-2\pi ik)jh)}. \tag{18}$$

Формула (18) представляет собой *дискретное преобразование Фурье*, а (17) – *обратное дискретное преобразование Фурье*.

Посчитаем теперь количество операций для (17), считая коэффициенты f_k известными. Для каждой из N точек количество операций равно $O(N)$, значит, всего имеет $O(N^2)$ операций.

В 1966 году был предложен алгоритм, названный *быстрым дискретным преобразованием Фурье (FFT, Fast Fourier transform)*, который требует всего $O(N \ln N)$ операций.

Лекция 7

Алгоритм быстрого дискретного преобразования Фурье

В прошлый раз речь шла о том, что в узлах равномерной сетки с шагом h на отрезке $[0,1]$ справедливо представление

$$f(x) = \sum_{n=0}^{N-1} f_n e^{2\pi i n x}, \quad x = jh, \quad j = 0, \ldots, N-1.$$

Алгоритм быстрого преобразования Фурье позволяет вычислить значения функции в узлах сетки за $O(N \ln N)$ операций. Изложим основную идею этого алгоритма.

Пусть N представимо в виде $N = N_1 \cdot N_2$. Разложим

$$n = N_1 n_1 + n_2, \quad 0 \le n_1 < N_2, \quad 0 \le n_2 < N_1;$$

$$j = N_2 j_1 + j_2, \quad 0 \le j_1 < N_1, \quad 0 \le j_2 < N_2.$$

Преобразуем

$$f(x) = \sum_{n_1=0}^{N_2-1} \sum_{n_2=0}^{N_1-1} f_{N_1 n_1 + n_2} \exp\left\{ 2\pi i (N_1 n_1 + n_2) \underbrace{(N_2 j_1 + j_2) h}_{jh = x} \right\}.$$

Заметим, что

$$\exp\{2\pi i N_1 n_1 N_2 j_1 h\} = 1.$$

Тогда

$$f(x) = \sum_{n_2=0}^{N_1-1} \underbrace{\left[\sum_{n_1=0}^{N_2-1} f_{N_1 n_1 + n_2} e^{2\pi i N_1 n_1 j_2 h} \right]}_{A(n_2; j_2)} e^{2\pi i n_2 x}.$$

Алгоритм устроен следующим образом. Вычисляются все $A(n_2; j_2)$, для вычисления каждого из которых нужно $O(N_2)$ операций. Всего таких коэффицентов $N_1 \cdot N_2$. То есть, подсчет всех коэффицентов требует

$$O(N_2^2 N_1) = O(N N_2).$$

Теперь, для каждого узла вычисляется сумма с известными коэффицентами, это $O(N N_1)$ операций. Всего получаем

$$O(N N_2) + O(N N_1) = O(N(N_1 + N_2))$$

операций. Например, если N выбрано таким образом, что

$$N_1 = N_2 = \sqrt{N},$$

то получаем асимптотику $O(N^{3/2})$ операций.

Предположим, что

$$N = 2^l, \quad N_1 = 2, \quad N_2 = 2^{l-1}.$$

Получим

$$f(x) = \sum_{n_2=0}^{1} \underbrace{\left[\sum_{n_1=0}^{2^{l-1}-1} f_{2n_1+n_2} e^{2\pi i 2n_1 j_2 h} \right]}_{A(n_2;j_2)} e^{2\pi i n_2 x}.$$

Для выражения в скобках (то есть для $N_2 = 2^{l-1}$) выполняем аналогичные преобразования. В конечном итоге получим выражение вида

$$f(x) = \underbrace{\sum_0^1 \left(\sum_0^1 \left(\sum_0^1 \dots \right) \right)}_{l}.$$

Итак, получили, что алгоритм требует

$$O(lN) = O(N \log_2 N)$$

операций.

Заметим, что N в таком случае должно принимать одно из значений

$$4, 8, 16, 32, 64, 128, 256, 512, \dots$$

Это не очень удобно, так как если, например, значение 256 не подходит по точности, придется использовать в два раза большее число узлов 512.

Одним из выходов может стать модификация алгоритма, которая рассмотривает число разбиений вида

$$N = 2^l 3^k, \quad k = 0|1.$$

Тогда возможные значения,принимаемые N, выглядят следующим образом:

$$4, 6, 8, 12, 16, 24, 32, 48, 64, \dots$$

Численное дифференцирование

Вообще говоря, малые изменения значений функции могут вести к значительным изменениям значений производной. Например, для функции

$$\widetilde{f}(x) = f(x) + \underbrace{\varepsilon \sin \frac{x}{\varepsilon^2}}_{|\cdot| \le \varepsilon}$$

производная будет иметь вид

$$\widetilde{f}'(x) = f'(x) + \underbrace{\frac{1}{\varepsilon} \cos \frac{x}{\varepsilon^2}}_{O(1/\varepsilon)}.$$

Задача приближенного нахождения производной выглядит следующим образом. Есть

$$x_1, \dots, x_n$$

и значения $f(x)$ в этих точках

$$f(x_1), \dots, f(x_n).$$

Требуется найти значения производной $f'(x)$.

Эта задача сводится к предыдущей. По n точкам строим интерполяционный многочлен Лагранжа и полагаем

$$f'(x) \sim L_n'(x).$$

Пример Пусть даны точки x и $x + h$. Тогда можем положить

$$f'(x) \sim \frac{f(x+h) - f(x)}{h},$$

и если $f(x)$ достаточно гладкая, можем разложить правую часть по параметру h:

$$f'(x) \sim f'(x) + \underbrace{\frac{h}{2} f''(\xi)}_{O(h)}, \quad \xi \in [x, x+h].$$

В этом случае говорят, что $(f(x+h) - f(x))/h$ аппроксимирует $f'(x)$ с первым порядком.

Рассмотрим еще один **пример**. Пусть есть точки 0, h, $2h$, то есть дана равномерная сетка с тремя узлами. Запишем аппроксимацию производной в точке 0. Будем искать приближение в виде

$$\alpha f(0) + \beta f(h) + \gamma f(2h) = f'(0) + r \tag{19}$$

так, чтобы остаток r был наименьшим. Разложим

$$\alpha f(0) + \beta \left(f(0) + h f'(0) + \frac{h^2}{2} f''(0) + O(h^3) \right) +$$

$$+ \gamma \left(f(0) + 2h f'(0) + \frac{(2h)^2}{2} f''(0) + O(h^3) \right) =$$

$$= (\alpha + \beta + \gamma) f(0) + (\beta + 2\gamma) h f'(0) + (\beta/2 + 2\gamma) h^3 f''(0) + (\beta + \gamma) O(h^3).$$

Чтобы выполнялось равенство (19), нужно, чтобы

$$\begin{cases} \alpha + \beta + \gamma = 0, \\ (\beta + 2\gamma) h = 1, \\ \beta/2 + 2\gamma = 0. \end{cases}$$

Решив систему, получим, что

$$f'(0) = \frac{-3f(0) + 4f(h) - f(2h)}{2h} + O(h^2).$$

Перейдем к следующему **примеру**. Пусть дан отрезок $[-h, h]$, разбитый сеткой с шагом h. В этом случае

$$f'(0) = \frac{f(h) - f(-h)}{2h} + O(h^2).$$

Эта формула называется *центральной разностью.*

Заметим, что формула

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

называется *разностью назад*, а

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

– *разностью вперед.*[25]

Оценка погрешности

Рассмотрим следующий простейший случай. Будем работать с функцией вида

$$\widetilde{f}(x) = f(x) + \delta(x), \quad |\delta| < \varepsilon,$$

причем

$$|f''| < M.$$

Запишем

$$r(x) = \frac{\widetilde{f}(x+h) - \widetilde{f}(x)}{h} - f'(x) = \frac{f(x+h) - f(x)}{h} - f'(x) + \frac{\delta(x+h) - \delta(x)}{h}.$$

Найдем минимум ошибки

$$|r(x)| \le \frac{Mh}{2} + \frac{2\varepsilon}{h} \to \min_h.$$

Получим, что

$$\frac{M}{2} - \frac{2\varepsilon}{h^2} = 0,$$

то есть в данном случае оптимальное значение шага

$$h = 2\sqrt{\varepsilon/M},$$

и при этом значении получим оценку погрешности

$$|r(x)| \le 2\sqrt{\varepsilon M}.$$

[25]То, какая из трех формул аппроксимации производных подойдет лучше, зависит от конкретной задачи (несмотря на то, что в формуле центральной разности погрешность имеет вид $O(h^2)$).

Лекция 8

Вычисление формулы для численного дифференцирования

Скажем еще пару слов о численном дифференцировании. Напомним, что в задаче численного дифференцирования по значениям функции f в точках

$$f(x_1), \ldots, f(x_n)$$

требуется получить приближенные значения производной $f'(x)$. Как уже говорилось ранее, в таком случае будет верно

$$f'(x) \sim L_n'(x). \tag{20}$$

При этом

$$L_n'(x) = \sum_{j=1}^{n} f(x_j) \underbrace{\Phi_j'(x)}_{c_j}, \tag{21}$$

то есть производная интерполяционного полинома – линейная комбинация значений функции f в узлах. Заметим, что (20) будет точной в случае, когда $f(x)$ является полиномом степени $n-1$.

Из формулы (21) получим

$$\begin{cases} \sum_{j=1}^{n} c_j = 0, \quad f = 1, \\ \sum c_j x_j = 1, \quad f = x \\ \sum c_j x_j^2 = 2x, \quad f = x^2 \\ \ldots\ldots \\ \sum c_j x_j^{n-1} = (n-1)x^{n-2}, \quad f = x^{n-1}. \end{cases}$$

Решив данную систему, получим значения коэффицентов c_j.

Сжатие информации (одномерный случай)

Обсудим метод, используемый для сжатии информации (например, при передаче изображений).

Рассмотрим сначала одномерный случай. Будем работать с N измерениями на интервалое $[0, 1]$. Обычно в таких случаях передают не сами значения функции, а коэффициенты ее ряда Фурье:

$$f(x) = \sum_{k=0}^{N-1} f_k e^{2\pi i k x}, \quad x = nh, \quad 0 \le n \le N-1.$$

Упорядочив коэффициенты ряда Фурье по убыванию модулей, то есть

$$|f_{k_1}| \ge |f_{k_2}| \ge \ldots$$

можем записать

$$f(x) = \sum_{k=0}^{N-1} f_{k_j} e^{2\pi i k_j x}.$$

Система ортонормирована, имеет место равенство Парсеваля

$$\|f\|^2 \equiv \sum_{n=0}^{N-1} h f^2(nh) = \sum_{j=0}^{N-1} |f_{k_j}|^2.$$

На практике рассматривается не весь ряд, а первые несколько его членов, дающих наиболее существенный вклад, то есть ряд вида

$$\sum_{j=0}^{l} f_{k_j} e^{2\pi i k_j x}, \quad l < N.$$

При этом погрешность

$$r_l = \sum_{j=l+1}^{N-1} f_{k_j} e^{2\pi i k_j x}, \quad \|r_l\|^2 = \sum_{j=l+1}^{N-1} |f_{k_j}|^2.$$

Сжатие информации (двумерный случай)

Рассмотрим следующую задачу. На двумерном квадрате $0 \le x, y \le 1$ определена функция $f(x,y) \in L_2[0,1]^2$. Задача ставится следующим образом:

$$\Phi(\varphi, psi) = \|f(x,y) - \varphi(x)\psi(y)\|^2_{L_2} \to \min_{\varphi,\psi}.$$

Можно переписать задачу в виде

$$\|\Phi - \mu\overline{\varphi}\overline{\psi}\|^2 \to \min_{\substack{\overline{\varphi},\overline{\psi}\in S_1 \\ \mu\in\mathbb{R}}}, \tag{22}$$

где S_1 – единичная сфера,

$$\|\overline{\varphi}\|^2 = \int_0^1 \overline{\varphi}^2(x)dx = 1, \quad \|\overline{\psi}\|^2 = 2.$$

Распишем (22):[26]

$$\Phi = \|f\|^2 - 2\mu \int_0^1 \overline{\varphi}(x) \left(\underbrace{\int_0^1 f(x,y)\overline{\psi}(y)dx}_{K\overline{\psi}} \right) dx + \mu^2 \underbrace{\int_0^1 \overline{\varphi}^2 dx \int_0^1 \overline{\psi} dy}_{=1} =$$

$$= \|f\|^2 + \left(\mu - (\overline{\varphi}, K\overline{\psi})\right)^2 - (\overline{\varphi}, K\overline{\psi})^2 \to \min_{\mu,\overline{\varphi},\overline{\psi}}. \tag{23}$$

[26]Здесь K – интегральный оператор с ядром f.

Задача сводится к следующей:

$$(\overline{\varphi}, K\overline{\psi})^2 \to \max_{\overline{\varphi},\overline{\psi}\in S_1}.$$

Максимум достигается, когда скалярное произведение берется от коллениарных функций, то есть таких, что

$$K\overline{\varphi} = \lambda\overline{\varphi}.$$

Из этого условия следует, что

$$|\lambda| = \|K\overline{\varphi}\|,$$

$$(\overline{\varphi}, K\overline{\psi})^2 = \frac{1}{\lambda^2}\underbrace{(K\overline{\psi}, K\overline{\psi})}_{(K*K\overline{\psi},\overline{\psi})} = \lambda^2 \to \max_{\overline{\psi}}.$$

Максимум достигается на собственной функции ψ_1, соответствующей наибольшему собственному значению оператора $K * K$

$$\lambda_1 = \max \lambda(K * K).$$

Отсюда в (23) получаем

$$\mu_1 = \left(K\overline{\psi_1}, \frac{1}{\lambda_1}K\overline{\psi_1}\right).$$

Теперь повторим все рассуждения для функции

$$f_1(x,y) = f(x,y) - \mu_1\overline{\varphi}_1(x)\overline{\psi}_1(y),$$

затем для функции

$$f_2(x,y) = f(x,y) - \mu_1\overline{\varphi}_1(x)\overline{\psi}_1(y) - \mu_2\overline{\varphi}_2(x)\overline{\psi}_2(y)$$

и так далее. Получим, что f раскладывается в ряд

$$f(x,y) = \sum \mu_n\overline{\varphi}_n(x)\overline{\psi}_n(y),$$

где $\overline{\varphi}_n$ – ортонормированная система собственных функций $K * K$.

Перейдем к матрицам. Имеем функцию $f(x,y)$, где $x = 1,\dots,n$, $y = 1,\dots,n$, которую хотим аппроксимировать произведением функций одного переменного

$$\overline{\varphi} = (\varphi_1,\dots,\varphi_n)^T, \quad \overline{\psi} = (\psi_1,\dots,\phi_m).$$

Их произведение $\overline{\varphi}\overline{\psi}$ является матрицей ранга 1.

$K\overline{\psi}$ представляет собой матрицу f

$$f = \begin{pmatrix} f_{11} & \dots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{m1} & \dots & f_{mn} \end{pmatrix},$$

ранг f равен q, примененную к вектору $\overline{\psi}$. Так как результатом применения будет вектор, скалярное произведение в данном случае – скалярное произведение векторов.

Получаем разложение

$$f(x,y) = \sum_{j=1}^{q} \mu_j \underbrace{\overline{\varphi}^{(j)}\overline{\psi}^{(j)}}_{\text{матрица ранга 1}}. \tag{24}$$

Рассмотрим также частичную сумму

$$s_k = \sum_{j=1}^{k} \mu_j \overline{\varphi}^{(j)}\overline{\psi}^{(j)}, \quad k \le q.$$

Так как разложение (24) ортогонально, справедливо равенство Парсеваля

$$\|f\|^2 = \sum_{j=1}^{q} \mu_j^2.$$

Тогда

$$\|f - s_k\|^2 = \sum_{j=k+1}^{q} \mu_j^2.$$

Лекция 9

Численное интегрирование

Задача состоит в приближенном вычислении интеграла

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

где a, b для начала полагаем конечными. Фиксированная функция $p(x)$ называется *весовой функцией*.

Рассмотрим следующий алгортм:

1. Фиксируем $x_1, \dots, x_n \in [a, b]$ – *узлы квадратурной формулы*
2. Строим $L_n(x)$.
3. Полагаем[27]

$$I(f) \sim \int_a^b p(x)L_n(x)dx = K_n(f).$$

Запишем в явном виде

$$K_n(f) = \int_a^b p(x)L_n(x)dx = \sum_{j=1}^n f(x_j) \underbrace{\int_a^b p(x)\Phi_j(x)dx}_{c_j},$$

где

$$\Phi_j(x) = \frac{\omega_n(x)}{(x - x_j)\prod\limits_{i\neq j}(x_j - x_i)}.$$

Таким образом,

$$K_n(f) = \sum_{j=1}^n f(x_j)c_j. \tag{25}$$

Формула (25) называется *квадратурной формулой*. Заметим, что из построения $L_n(x)$ следует

$$I(P_{n-1}) = K_n(P_{n-1}).$$

Поэтому, подставляя последовательно вместо f функции

$$1, x, x^2, \dots, x^{n-1},$$

получим систему уравнений[28]

$$\begin{cases} \int_a^b p(x)dx = \sum_{j=1}^n c_j, \\ \int_a^b p(x)xdx = \sum_{j=1}^n c_j x_j, \\ \dots\dots \\ \int_a^b p(x)x^{n-1}dx = \sum_{j=1}^n c_j x_j^{n-1}, \end{cases}$$

для коэффицентов c_j.

[27] Здесь неявно полагаем, что интеграл $K_n(f)$ можно вычислить точно, То есть накладываем некоторые ограничения на $p(x)$.

[28] Легко заметить, что определитель данной системы является определителем Вандерморда.

Примеры

- Пусть $n = 1$, $x_1 = \frac{a+b}{2}$, а $p(x) \equiv 1$.

$$K_1(f) = (b-a) f\left(\frac{a+b}{2}\right) \tag{26}$$

Данная формула называется *формулой прямоугольников* (рис. 9.1).

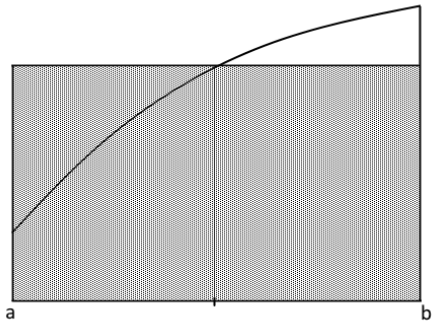

Рис. 9.1. Формула прямоугольников

Формула (26), очевидно, будет точна для полиномов 0-й степени. На самом деле, (26) будет точна и для полиномов степени 1, поскольку такой полином имеет вид[29]

$$P_1(x) = \alpha\left(x - \frac{a+b}{2}\right) + \beta.$$

- Пусть $n = 2$, $x_1 = a$, $x_2 = b$, $p \equiv 1$. Тогда

$$K_2(f) = \frac{b-a}{2}\left[f(a) + f(b)\right]. \tag{27}$$

Формула (27) называется формулой трапеций (рис. 9.2). Заметим, что (27) будет точна для полиномов степени 1.

- Пусть $n = 3$, $x_1 = a$, $x_2 = \frac{a+b}{2}$, $x_3 = b$, а весовая функция $p(x) \equiv 1$. Тогда

$$K_3(f) = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]. \tag{28}$$

Формула (28) называется *формулой Симпсона*. Эта формула точна для многочленов степени 3.

[29] Заметим, что задача вообще говоря состоит в нахождении коэффицентов c_j таких, чтобы точность приближения была как можно выше. В случае полиномов степени $n-1$ это выполняется гарантировано, но, вполне возможно, будет выполняться и для других функций. В этом примере формула получилось точной для полинома степени $n = 1$.

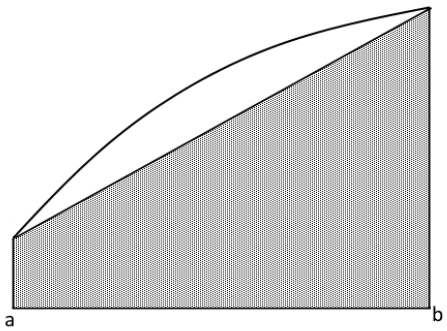

Рис. 9.2. Формула трапеций

Оценка погрешности

Пусть $K_n(f)$ точна для многочленов степени m. Тогда[30]

$$I(f - P_m) = I(f) - I(P_m) = I(f) - K_n(P_m).$$

Тогда погрешность

$$r(f) = I(f) - K_n(f) = I(f - P_m) - K_n(f - P_m).$$

Возьмем в качестве P_m многочлен наилучшего равномерного приближения. Получим

$$r(f) = I(f - Q_m^0) - K_n(f - Q_m^0).$$

Оценим

$$|I(f - Q_m^0)| = |\int_a^b p(x)\left(f(x) - Q_m^0(x)\right)dx| \le \int_a^b |p(x)|dx \cdot E_m(f),$$

где $E_m(f)$ – максимальное значение разности. Напомним, что

$$E_m(f) = \|f - Q_m^0\| \le \frac{\|f^{(m+1)}}{(m+1)!}(b-a)^{m+1}2^{-(2m+1)}.$$

Кроме того, верна оценка

$$|K_n(f - Q_m^0)| = |\sum c_j\left(f(x_j) - Q_m^0(x_j)\right)| \le \sum_{j=1}^{n} |c_j|E_m(f).$$

Наконец, получим окончательную оценку

$$|I(f) - K_n(f)| \le \left(\int_a^b |p|dx + \sum |c_j|\right) E_m(f).$$

[30]Как убедились в примерах выше, m не обязательно равна $n-1$.

Заметим, что если $p(x) \geq 0$ в конечном числе точек и $c_j > 0$, то

$$|I(f) - K_n(f)| \leq 2 \int_a^b p(x)dx E_m(f).$$

Заметим, что заменой переменных можно свести

$$I(f) = \int_a^b p(x)f(x)dx = \int_{-1}^1 \dots,$$

поэтому можно рассматривать квадратурную формулу только на отрезке $[-1, 1]$.

Ортогональный многочлен

Далее нам понадобится понятие ортогонального многочлена.

Пусть $p(x) \geq 0$, $p(x) = 0$ в конечном числе точек. Скалярное произведение двух функций вводится как

$$(f, g) \equiv \int_a^b p(x)f(x)g(x)dx.$$

Вспомним процесс *ортогонализации Грамма-Шмидта*. Рассмотрим линейно независимые функции

$$1, x, x^2, \dots, x^n, \dots$$

Тогда положим

$$P_0 = 1,$$
$$P_1 = x + \gamma_1^{(1)} P_0$$

так, чтобы $P_1 \perp P_0 \Leftrightarrow$

$$(P_1, P_0) = 0 = \underbrace{(x, 1)}_{\int_a^b p(x)xdx} + \gamma_1^{(1)} \|P_0\|^2,$$

откуда находим коэффицент $\gamma_1^{(1)}$. Далее положим

$$P_2 = x^2 + \gamma_2^{(2)} P_1 + \gamma_2^{(2)} P_0$$

и так далее.

Пример Рассмотрим $[-1, 1]$,

$$p(x) = (1 - x)^\alpha (1 + x)^\beta, \quad \alpha, \beta > -1. \tag{29}$$

В этом случае получим ортогональную систему полиномов $H_n(x)$, называемых *многочленами Якоби*.

Пример Частный случай (29) при $\alpha = \beta = -1/2$, то есть

$$p(x) = \frac{1}{\sqrt{1 - x^2}}.$$

В этом случае отрогональную систему образуют *многочлены Чебышёва I рода*

$$T_n(x) = \cos\left(n \arccos x\right).$$

Пример Частный случай (29):

$$p(x) = \sqrt{1-x^2}, \quad \alpha = \beta = 1/2.$$

Ортогональную систему образуют *многочлены Чебышёва II рода.*

Пример Пусть $[0, \infty)$,

$$p(x) = x^\alpha e^{-x}, \quad \alpha > -1.$$

Тогда ортогональной системой многочленов будут *многочлены Лаггера.*

Пример Для частного случая (29) при $\alpha = \beta = 0$ получим *многочлены Лежандра.*

Свойства ортогональных многочленов

Теорема 9.1. *Если $P_n(x)$ – ортогональный многочлен, $\exists$ n различных корней $\in (a, b)$.*

Доказательство (От противного). Предположим, что утверждение неверно. Тогда $x_1, \ldots, x_m$ – корни нечетной кратности, $m < n$. По теореме Безу

$$P_n(x) = (x - x_1)^{k_1} \ldots (x - x_m)^{k_m} r(x),$$

где остаток $r(x)$ сохраняет знак.

Так как P_n – ортогональный многочлен,

$$(P_n, \omega_m) = 0 = \int_a^b \underbrace{p(x)}_{=0 \text{ в кон. ч. т.}} \underbrace{(x - x_1)^{k_1+1} \ldots (x - x_m)^{k_m+1}}_{\geq 0} r(x)dx > 0.$$

Получаем противоречие. Теорема доказана.

Теорема 9.2. *Имеет место формула для ортогональных многочленов*

$$P_{n+1}(x) = (x + \alpha_n)P_n(x) + \beta_n P_{n-1}(x).$$

Доказательство Распишем

$$(P_{n+1}, P_n) - 0 = (xP_n, P_n) + \alpha_n \|P_n\|^2,$$

откуда возможно найти α_n. Аналогично, из

$$(P_{n+1}, P_{n-1}) - 0 = (xP_n, P_{n-1}) + \beta_n \|P_{n-1}\|^2$$

однозначно находится β_n. Проверим теперь, что найденный таким образом P_{n+1} будет ортогонален всем многочленам степени $j < n - 1$. Вычислим

$$(P_{n+1}, P_j) = (xP_n, P_j) + \alpha_n \underbrace{(P_n, P_j)}_{=0} + \beta_n \underbrace{(P_{n-1}, P_n)}_{=0} = (P_n, xP_j).$$

Так как xP_j – многочлен степени $< n$, получим, что

$$(P_{n+1}, P_j) = 0.$$

Теорема доказана.

Лекция 10

Постановка задачи

Вспомним постановку задачи. Есть определенный интеграл

$$I(f) = \int_a^b p(x)f(x)dx, \quad p(x) \geq 0,$$

где $p(x) = 0$ в конечном числе точек. Хотим построить квадратурную формулу

$$K_n(f) = \sum_{j=1}^{n} c_j f(x_j),$$

то есть найти $x_1, \ldots, x_n \in [a, b]$ и $c_1, \ldots, c_n$ (всего $2n$ неизвестных) такие, чтобы K_n была наиболее точной.

Предположим, что $K_n(f)$ точна для многочленов степени $2n - 1$. Произвольный многочлен такой степени можно представить в виде

$$P_{2n-1}(x) = \omega_n(x)Q_{n-1}(x) + r_{n-1}(x).$$

Тогда

$$I(P_{2n-1}) = \int_a^b p(x)\omega_n(x)Q_{n-1}(x)dx + \int_a^b p(x)r_{n-1}(x)dx =$$

$$= \sum_{j=1}^{n} c_j r_{n-1}(x_j), \quad \forall\, Q_{n-1}, \quad r_{n-1}.$$

Это выполнено $\Leftrightarrow$

$$\int_a^b p(x)\omega_n(x)Q_{n-1}(x)dx = 0,$$

то есть $\omega_n(x)$ – ортогональный многочлен степени n.

Квадратурная формула Гаусса

Опираясь на соображения выше, сформулируем алгоритм *нахождения квадратурной формулы Гаусса*:

1. Строим ортогональный многочлен степени n.

2. Находим корни $x_1, \ldots, x_n \in (a, b)$.

3. Так как $x_1, \ldots, x_n$ – узлы $K_n(f)$, находим коэффициенты $c_1, \ldots, c_n$, подставляя в формулу функции

$$1, x, \ldots, x^{n-1}.$$

Полученная таким образом формула называется *квадратурой Гаусса*. С помощью подстановки легко убедиться, что она будет точна для многочлена степени $2n - 1$.

Свойства квадратурной формулы Гаусса

1. Убедимся сначала, что квадратурная формула Гаусса не точна для многочленов степени $2n$. Рассмотрим многочлен

$$P_{2n}(x) = \omega_n^2(x) = (x - x_1)^2 \dots (x - x_n)^2.$$

Получим, что

$$I(P_{2n}) = \int_a^b p(x)\omega_n^2(x)dx > 0.$$

С другой стороны,

$$K_n(f) = \sum_{j=1}^{n} c_j \underbrace{\omega_n^2(x_j)}_{=0} = 0,$$

то есть для данного многочлена квадратурная формула не является точной.

2. Покажем, что $c_j > 0$. Рассмотрим многочлен

$$P_{2n-2}(x) = \frac{\omega_n^2(x)}{(x - x_i)^2},$$

$$I(P_{2n-2}) = \int_a^b p(x)\frac{\omega_n^2}{(x - x_i)^2}dx > 0,$$

$$K_n(P_{2n-2}) = \sum_{j=1}^{n} c_j \frac{\omega_n^2(x_j)}{(x_j - x_i)^2} = c_i \prod_{i \neq j}(x_i - x_j)^2 > 0.$$

Для данного многочлена квадратурная формула будет точна, а значит, все $c_j > 0$.

3. Для квадратурной формулы Гаусса верна оценка сверху погрешности

$$|R_n(f)| \leq 2\int_a^b p(x)dx \cdot E_{2n-1}(f).$$

Обобщение задачи

Рассмотрим в некотором роде обобщение предыдущей задачи. Нужно вычислить интеграл

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

где $f(x)$ вычисляется очень сложно, но известны значения $f(a)$ и $f(b)$. Тогда

$$a, b, x_2, \dots, x_{n-1}$$

– узлы, а $c_1, \dots, c_n$ – коэффициенты. Требуется найти узлы $x_2, \dots, x_{n-1}$ и коэффициенты $c_1, \dots, c_n$ (всего $2n-2$ неизвестных) такие, чтобы K_n была точна для многочлена наиболее высокой степени.

Попробуем для этого сделать рассуждения, аналогичные рассуждениям из предыдущей темы. Предположим, что квадратурная формула $K_n(f)$ точна для многочленов степени $2n-3$. Представим произвольный многочлен такой степени в виде[31]

$$P_{2n-3}(x) = \omega_n(x)Q_{n-3}(x) + r_{n-1}(x),$$

$$\omega_n(x) = (x-a)(b-x)\prod_{j=2}^{n-1}(x-x_j).$$

Теперь,

$$I(P_{2n-3}) = \int_a^b p(x)(x-a)(b-x)\prod_{j=2}^{n-1}(x-x_j)Q_{n-3}(x)dx + \int_a^b p(x)r_{n-1}(x)dx =$$

$$= \sum_{j=1}^{n} c_j r_{n-1}(x_j), \quad \forall\ Q_{n-3}, \quad r_{n-1},$$

причем $x_1 = a$, $x_n = b$.

$$\int_a^b p(x)\omega_n(x)Q_{n-1}(x)dx = 0,$$

то есть $\omega_n(x)$ – ортогональный многочлен степени n.

Для того, чтобы это выполнялось, первый из интегралов должен быть равен нулю. Обозначим

$$G_{n-2}(x) = \prod_{j=2}^{n-1}(x-x_j),$$

$$q(x) = p(x)(x-a)(b-x) \geq 0$$

– новая весовая функция и введем скалярное произведение

$$(f,g) \equiv \int_a^b q(x)f(x)g(x)dx. \tag{30}$$

Тогда можем переписать условие на интеграл в виде

$$(G_{n-2}, Q_{n-3}) = 0, \quad \forall Q_{n-3},$$

откуда следует, что G_{n-2} – ортогональный многочлен в скалярном произведении (30).

Алгоритм решения обобщенной задачи.

1. Берем новую весовую функцию

$$q(x) = p(x)(x-a)(b-x).$$

[31] Здесь в представлении ω_n поменяли знак на противоположный. Для разложения P_{2n-3} это роли не играет.

2. Строим ортогональный многочлен степени $n-2$ $G_{n-2}(x)$.

3. Ищем его корни $x_2, \ldots, x_{n-1}$. Они будут являться узлами.

4. Находим $c_1, \ldots, c_n$, подставив в формулу последовательно функции

$$1, x, \ldots, x^{n-1}$$

и решив полученную систему уравнений.

При помощи подстановки можно убедиться, что полученная формула действительно точна для многочлена степени $2n-3$.

Заметим, что такая квадратура называется *квадратурой Лобатто*.

Точная формула погрешности

Пусть квадратура K_n точна для многочленов степени m. Оценим

$$R_n(f) = I(f) - K_n(f).$$

Предположим, что $b - a \equiv h$ – малый параметр.[32] Распишем

$$R_n(f) = R_n(f - Q_m^0) =$$

$$= \int_a^b (f(x) - Q_m^0(x))p(x)dx - \sum_{j=1}^{n} c_j(f(x_j) - Q_m^0(x_j)).$$

Можем записать

$$f(x) - Q_m^0(x) = \frac{f^{(m+1)}}{(m+1)!}(x-a)^{m+1} + o((x-a)^{m+1}).$$

Тогда оценка будет иметь вид

$$R_n(f) = d_1 h^{m+2} + o(h^{m+2}) + d_2 h^{m+2} = Mh^{m+2} + o(h^{m+2}).$$

Рассмотрим теперь следующий подход к вычислению интегралов. Делим отрезок $[a, b]$ на N частей с шагом $h = (b-a)/N$ и на каждом отрезке применяем какую-то квадратурную формулу.

1. **Составная квадратурная формула прямоугольников**. На каждом подотрезке в качестве единственного узла берем его середину и применяем формулу

$$K_1^{(N)}(f) = \sum_{j=0}^{N-1} hf\left(a + \frac{h}{2} + jh\right).$$

Пусть

$$|f''| \le M, \quad p(x) \equiv 1.$$

[32] Для нас это будет значит, что разрешается использовать разложение по h.

Тогда на $[ih, (i+1)h]$ имеет место оценка

$$|R_1(f)| \leq \frac{M}{24}h^3.$$

Общая погрешность в таком случае имеет вид

$$|R_1^{(N)}(f)| \leq \frac{M}{24}h^3 \sum_{j=0}^{N-1} 1 = \frac{M(b-a)}{24}h^2.$$

2. **Квадратурная формула трапеций**. Аналогично предыдущему случаю, можем записать

$$K_2^{(N)}(f) = \frac{h}{2}(f(a) + f(b)) + sum_{j=1}^{N-1} hf(x_j).$$

В этом случае общая погрешность

$$|R_2^{(N)}(f)| \leq \frac{M(b-a)}{12}h^2.$$

3. **Квадратурная формула Симпсона**.

Опустим точную запись квадратурной формулы и запишем только оценку:

$$|R^{(N)}(f)| \leq \frac{\|f^{(4)}\|}{2880}(b-a)h^4.$$

Лекция 11

Вычисление интегралов в нерегулярных случаях

Рассмотрим несколько специальных случаев.

1. Речь пойдет о вычислении интегралов вида

$$I(f) = \int_0^1 e^{i\omega x} f(x)dx, \quad \omega \gg 1,$$

а $f(x)$ – гладкая функция. Главный прием, использующийся в таких задачах – правильное выделение весовой функции.[33] Положим

$$p(x) = e^{i\omega x},$$

то есть выделим неприятные особенности подынтегральной функции в весовую функцию. Далее можем построить квадратурную формулу с 1, 2 или 3 узлами[34], воспользовавшись стандартным алгоритмом.

Квадратурная формула с 3 узлами называется *формулой Филона*. Как правило, строится составная квадратурная формула Филона при помощи разбиения отрезка $[0, 1]$ с шагом h. В этом случае оценка погрешности имеет вид

$$|R(x)| \leq \frac{\max|f''|}{2} h^3 \left(\underbrace{\int_0^1 |e^{i\omega x}|dx}_{=1} + \sum |c_j| \right).$$

2. Рассмотрим интеграл вида

$$I(f) = \int_0^1 \frac{f(x)}{\sqrt{\sin x}} dx.$$

Брать в качестве $p(x)$ функцию $1/\sqrt{\sin x}$ в данном случае плохо, так как (см. предыдущие лекции) интеграл от функции вида $p(x)P_m(x)$ должен вычисляться точно. Запишем

$$I(f) = \int_0^1 \frac{1}{\sqrt{x}} \left(\sqrt{\frac{x}{\sin x}} f(x) \right) dx.$$

Положим

$$p(x) = \frac{1}{\sqrt{x}}$$

и будем строить составную квадратурную формулу Гаусса от функии

$$g(x) = \sqrt{\frac{x}{\sin x}} f(x),$$

[33] Например, если бы здесь предположили $p(x) \equiv 1$, то полином, который мы должны были бы построить, должен был бы аппроксимировать степенную функцию с очень большим показателем степени ω (например, 10^6). Это очень затратная задача.

[34] Большее число узлов на практике обычно не используется.

разбив отрезок $[0,1]$ на N равных частей. Заметим, что $g(x)$ будет достаточно гладкой, так как производная от первого сомножителя имеет асимптотику

$$\frac{1}{2}\sqrt{\frac{\sin x}{x}}\frac{\sin x - x\cos x}{\sin^2 x} \sim \frac{1}{2}\frac{x + x^3/6 - x + x^3/2 + \dots}{x^2 + \dots} \sim \frac{x}{6} + \dots.$$

Для того, чтобы найти коэффициенты c_i, вычислим

$$\int_{ih}^{(i+1)h} 1\frac{1}{\sqrt{x}}dx = 2\sqrt{x}\Big|_{ih}^{(i+1)h} = 2\left[\sqrt{(i+1)h} - \sqrt{ih}\right] = c_i.$$

Узлы x_i найдем из уравнений

$$\int_{ih}^{(i+1)h} \frac{1}{\sqrt{x}}x dx = \frac{2}{3}\left(((i+1)h)^{3/2} - (ih)^{3/2}\right) = c_i x_i.$$

3. Рассмотрим еще один пример. Вычислим интеграл

$$I = \int_0^1 \frac{\ln x}{1+x^2}dx.$$

Сделаем замену $x = t^k$ и получим

$$I = \int_0^1 \frac{k^2 t^{k-1}\ln t}{1+t^{2k}}dt,$$

сведя тем самым задачу к стандартной при $k > 1$.

4. В случае, если интеграл имеет вид

$$\int_0^\infty \dots dx = \int_0^1 \dots dx + \int_1^\infty \dots dx,$$

заменой переменных (например, $1/x$) можно свести его к стандартному случаю.

Оптимальная квадратурная формула

Введем класс функций

$$K = \{f : \ |f'| < M\}$$

и класс квадратурных формул S. Требуется вычислить интеграл

$$I(f) = \int_0^1 f(x)dx.$$

Пусть $s_n \in S$ – квадратурная формула. Для функции f обозначим погрешность квадратурной формулы

$$|s_n(f) - I(f)| = \varepsilon(f).$$

Определим

$$e(s) = \max_{f\in K}\varepsilon(f),$$

$e(s)$ зависит от квадратуры s. Задача состоит в нахождении

$$\min_{s\in S} l(s) = \min_{s\in S}\max_{f\in K}\varepsilon(f,s).$$

Оказывается, что наилучшей квадратурой для класса функций K оказывается формула прямоугольников.

Обобщение задачи

Предположим, отрезок $[a, b]$ разбит на k отрезков,

$$x_{i+1} - x_i = l_{i+1},$$

и на каждом отрезке $|f''| \leq A_{i+1}$. На каждом $[x_i, x_{i+1}]$ выберем N_i узлов и будем использовать составную формулу трапеций. Общее число узлов фиксировано и равно N, то есть

$$\sum_{j=1}^{k} N_j - N = 0. \tag{31}$$

Погрешность приближения на $[a, b]$ имеет вид

$$\Phi = \frac{A_1}{12}\frac{l_1^3}{N_1^2} + \ldots + \frac{A_k}{12}\frac{l_k^3}{N_k^2}.$$

Задача состоит в выборе N_i таким образом, чтобы погрешность была минимальной, то есть

$$\Phi \to \min$$

при выполнении (31).

Таким образом, это задача на условный экстремум. Составляем функцию Лагранжа

$$\mathcal{L} = \sum_{i=1}^{k} \frac{A_i}{12}\frac{l_i^3}{N_i^2} + \lambda\left(\sum_{j=1}^{k} N_j - N\right).$$

Из системы уравнений

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial N)i} = -\frac{A_i}{6}\frac{l_i^3}{N_i^3} + \lambda, \\ \sum_{j=1}^{k} N_j - N = 0. \end{cases}$$

находим N_i. Заметим, что N_i будут иметь вид

$$N_i = l_i\left(\frac{A_i}{6\lambda}\right)^{1/3},$$

то есть будут пропорциональны длине отрезка и кубическому корню из A_i.

Оценка главного члена погрешности

Пусть дан интеграл

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

который оцениваем с помощью $K_n(f)$ – составной квадратурной формулы трапеций с шагом $h = (b-a)/n$. Тогда ошибка оценки будет иметь асимптотику

$$\varepsilon_h(f) = I(f) - K_n(f) = Mh^2 + o(h^2).$$

Запишем ошибку как

$$\varepsilon_{h/2}(f) = I(f) - K_{2n}(f) = M\left(\frac{h}{2}\right)^2 + o(h^2).$$

Отсюда можем оценить главный член погрешности, исходя из предположения, что все находится в зоне действия асимптотики:

$$Mh^2(1 - 1/4) = K_{2n}(f) - K_n(f),$$

$$Mh^2 \asymp \varepsilon_h(f) \asymp \frac{4}{3}\left(K_{2n}(f) - K_n(f)\right).$$

Автоматический выбор шага

Пусть заданы границы интервала a, b, точность ε и начальный шаг h_0. Алгоритм будет устроен следующим образом. По шагу h вычисляем квадратурную формулу $K_n(f)$, по шагу $h/2$ – квадратурную формулу $K_{2n}(f)$. Найдя[35] Mh^3 и $M(h/2)^3$, можем вычислить $\varepsilon_h[a, a + h]$.

Теперь, на отрезке длины $b - a$ требуется точность ε. Значит, на отрезке длины h требуемая точность будет равна

$$\frac{\varepsilon}{b-a}h.$$

Тогда оценка точности должна удовлетворять

$$0.1\frac{\varepsilon h}{b-a} \le \varepsilon_h[a, a + h] \le \frac{\varepsilon}{b-a}h.$$

В случае, если не выполняется правое неравенство, $\varepsilon_h[a, a + h]$ слишком велика. В этом случае от шага h необходимо перейти к шагу $h/2$ и повторить все вычисления. В случае, когда не выполняется левое неравенство[36], шаг оказывается слишком мелким и его можно увеличить в два раза, то есть выполнить вычисления для $2h$.

Возможен случай, когда подыинтегральная функция имеет проблемы (рис. 11.1). В таком случае сказанное выше не выполняется, так как не выполнено предположение о том, что наши вычисления находятся в зоне действия асимптотики. Такие функции подлежат специальному исследованию.

[35] В асимптотической формуле погрешности главный член имеет вид Mh^2, так как $M \sim (b - a)$, а на каждом трезке длины h речь идет о h^3.

[36] Значение 0.1 взято для примера, значение константы может быть другим.

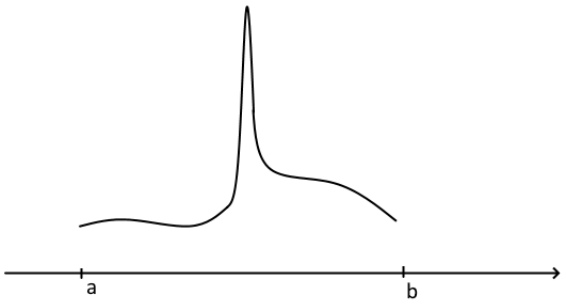

Рис. 11.1. Пример проблемной функции

Лекция 12

Погрешность приближения многочленов

Предположим, есть многочлен, хранящийся в памяти в виде массива коэффицентов

$$P_n(x) = a_0 + a_1 x + \ldots + a_n x^n.$$

Все коэффициенты хранятся в памяти компьютера в приближенном виде, то есть

$$\widetilde{a_i} = (1 + \delta_i)a_i, \quad |\delta_i| \leq \delta,$$

где δ_i – относительная погрешность, которую полагаем не большей некоторого маленького числа δ. Рассмотрим отрезок $[-1, 1]$.

Получим, что

$$\widetilde{P}_n(1) = P_n(1) + \underbrace{\sum_{i=0}^{n} a_i \delta_i}_{\text{погрешность}}.$$

Оценим погрешность сверху. Предположим, что

$$\delta_i = \delta \operatorname{sign} a_i, \quad \forall\ a_i.$$

Получим, что погрешность

$$r = \sum_{i=0}^{n} |a_i| \delta$$

зависит суммы коэффицентов.

Обсудим математическую сторону данной проблемы. Обозначим через G_φ область точек комплексной плоскости $\mathbb{C}$, из которых отрезок $[-1, 1]$ виден под углом, большим φ (рис. **??**).

На отрезке $[-1, 1]$ рассмотрим непрерывную функцию $f(x)$ и последовательность многочленов[37]

$$P_m(x) = a_0^{(m)} + \ldots + a_{n(m)}^{(m)} x^{n(m)}$$

[37]Здесь m – номер, а не степень.

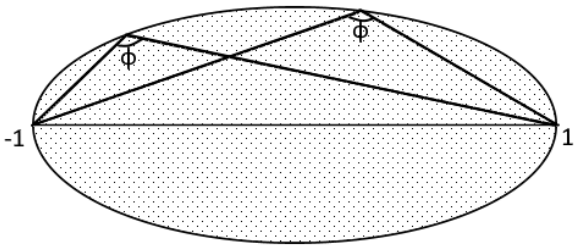

Рис. 12.1. Область G_φ

такую, что, во-первых,

$$\|f - P_m\|_{C[-1,1]} \leq 2^{-m},$$

и, во-вторых, $\exists\ q$ такая, что

$$\sum_{j=0}^{n(m)} |a_j^{(m)}| \leq M \cdot 2^{qm}, \tag{32}$$

где M – некоторая константа. Тогда оказывается, что[38] $f(x)$ можно продолжить аналитически на G_φ, где[39]

$$\varphi = \pi \frac{2q+1}{2q+2}. \tag{33}$$

Пример Рассмотрим

$$f(x) = \frac{1}{1+25x^2}.$$

У ее продолжения на комлпексную плоскость

$$f(z) = \frac{1}{1+25z^2}$$

есть два полюса $z = \pm i/5$. По выше описанной схеме находим, что

$$\operatorname{tg}\varphi = 5,$$

откуда из (33) можем найти q и оценить сумму коэффицентов (32).

Заметим, что при работе с многочленами невысокой степени эту проблему можно считать несущественной.

Разложение многочлена по ортогональному базису

Рассмотрим альтернативный вариант записи многочлена. Представим

$$P_n(x) = \sum_{j=0}^{n} d_j T_j(x),$$

[38] Данный факт опирается на теорему из курса теории функций комплексного переменного.
[39] Заметим, что при больших значениях q область G_φ будет «сплющиваться», так как $\varphi \to \pi$.

где $T_j(x)$ – многочлены Чебышёва. Так как скалярное

$$(T_i, T_j) \equiv \frac{2}{\pi} \int_{-1}^{1} \frac{T_i(x)T_j(x)}{\sqrt{1-x^2}} dx = \begin{cases} 2, & i = j = 0, \\ \delta_i^j, & i^2 + j^2 \neq 0, \end{cases}$$

где δ_i^j – символ Кронекера, многочлены Чебышёва образуют ортогональную систему. Получим оценку[40]

$$\|P_n\|^2 = 2d_0^2 + \sum_{j=1}^{n} d_j^2 \underset{\to}{n} \|f\|^2,$$

то есть сумма квадратов коэффицентов разложения ограничена. Тогда

$$\left(\sum_{j=0}^{n} |d_j| \right)^2 \leq (n+1) \sum_{j=0}^{n} d_j^2,$$

откуда получаем, что

$$\sum_{j=0}^{n} |d_j| \asymp \sqrt{n}.$$

Многомерный случай

Обсудим, какие проблемы возникают в многомерном случае. Для определенности рассмотрим случай $n = 2$. Если область задана квадратом, можно построить сетку и считать ее точки пересечения узлами интерполяции. Гораздо более сложный случай, если область произвольная, возможно, не является односвязной и значения функции даны в произвольных точках.

Общих алгоритмов для таких случаев не существует. В некоторых случаях, например, такая область тоже разбивается на сетку. Проблемой может быть то, что в некоторые ячейки разбиения может попасть несколько точек, а в другие – ни одной. В некоторых случаях область разбивается не прямоугольной сеткой (например, треугольной). В этом случае аппроксимация строится не с использованием многочленов.

Численные методы линейной алгебры

В численных методах линейной алгебры можно выделить следующие две задачи:

1. Решение системы линейных алгебраических уравнений.
2. Нахождение собственных значений матрицы.[41]

Будем заниматься решением системы вида

$$Ax = b, \quad A = (a_i j)$$

[40] Здесь норма берется по скалярному произведению, описанному выше.
[41] В курсе практически не будет разбираться.

, где A – матрица размера $m \times m$. Будем рассматривать A как линейный оператор в m-мерном пространстве. Норма в пространстве $\|x\|$ порождает норму оператора $\|A\|$. Можно рассматривать, например, следующие нормы:

$$\|x\|_\infty = \max_i |x_i|,$$

$$\|x\|_1 = \sum_{i=1}^{m} |x_i|,$$

$$\|x\|_2^2 = \sum_{i=1}^{m} x_i^2.$$

Метод Гаусса

Повторим кратко классический метод Гаусса решения СЛАУ. Пусть дана система

$$\begin{cases} a_{11}a_1 + a_{12}x_2 + \ldots + a_{1m}x_m = b_1, \\ a_{21}a_1 + a_{22}x_2 + \ldots + a_{2m}x_m = b_1, \\ \ldots\ldots \end{cases}$$

Предположим, что $a_{11} \neq 0$. Тогда

1. Делим первую строку на a_{11}.

2. Для каждой i-й строки расширенной матрицы умножаем 1-ю строку на a_{i1} и вычитаем результат из i-й строки.

Рассмотрим, какое количество арифметических операций было совершено. На шаге 1 было выполнено m операций. На втором шаге $m+m$ операций выполняется для каждой из $m-1$ строк. Всего получим порядка

$$m^2 + m^2$$

операций.

На следующем шаге алгоритме все те же операции совершаются для матрицы меньшей на 1 размерности, то есть выполняем порядка

$$(m-1)^2 + (m-1)^2$$

операций.

Всего, чтобы привести методом Гаусса матрицу к треугольному виду, потребуется порядка

$$2m^2 + 2(m-1)^2 + \ldots + 2 = 2\frac{m(m+1)(2m+1)}{6} = \frac{2}{3}m^3 + O(m^2).$$

операций. Заметим, что решение приведенной СЛАУ с треугольной матрицей требует $O(m^2)$ операций. Это означает, что основной вклад в решение СЛАУ дает именно приведение матрицы A к треугольному виду.

Заметим, что оценка снизу будет иметь порядок $O(m^2)$, так как при решении системы используются все элементы матрицы A.[42]

Обычно используют *модифицированный метод Гаусса* с выбором главного элемента по Жордану. Ищется

$$i = \arg\max_i |a_{1i}|,$$

это требует $O(m)$ операций. Затем i-й и 1-й столбец переставляются местами. Далее выполняется алгоритм, описанный выше.

Метод отражений

Предположим, есть

$$w = (w_1, \ldots, w_m)^T, \quad |w| = \left(\sum w_i^2\right)^{1/2} = 1.$$

Рассмотрим матрицу (рис. 12.2)

$$U_w \equiv I - 2ww^T, \tag{34}$$

где I – единичная матрица. Тогда

$$U_w w = w - 2ww^T w = w - 2w\underbrace{(w^T w)}_{1} = -w.$$

Пусть $v \perp w$. Тогда

$$U_w v = v - 2w\underbrace{(w^T v)}_{=0} = v.$$

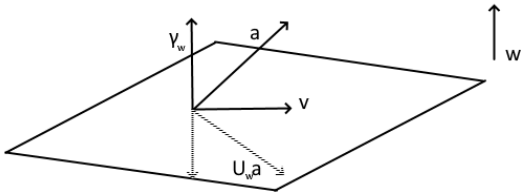

Рис. 12.2. Действие матрицы U_w

Поэтому матрица (34) называется *матрицей отражений*. Заметим, что

$$U_w^2 = I.$$

[42]То есть, например, для решения СЛАУ с 10^8 неизвестными в общем случае даже самый оптимальный метод будет требовать порядка 10^{16} операций.

Рассмотрим следующую задачу. Пусть заданы $\overline{a}$ и $\overline{e}$, как правило $\overline{e}$ – единичной длины. Требуется найти w такое, что

$$U_w\overline{a} = \lambda\overline{e}.$$

Заметим, что, так как матриа ортогональна,

$$\lambda = |\overline{a}|/|\overline{e}|.$$

Предположим, что такой w существует. Тогда

$$\begin{cases} \overline{a} = \gamma\overline{w} \oplus v, \\ U_w\overline{a} = -\gamma\overline{w} + v = \lambda\overline{e}. \end{cases}$$

Вычитая одно уравнение системы из другого, получим

$$\overline{a} - \lambda\overline{e} = 2\gamma\overline{w}.$$

Так как w должен быть единичной длины, получим

$$\overline{w} = (\overline{a} - \lambda\overline{e})/|\overline{a} - \lambda\overline{e}|$$

На его вычисление потребуется $O(m)$ операций.

Перейдем к **методу отражений**.

Полагаем

$$a^{(1)} = (a_{11}, \ldots, a_{m1})^T, \quad e^{(1)} = (1, 0, \ldots, 0)^T$$

и находим $w^{(1)}$. Применим $U_{w^{(1)}}$ к СЛАУ:

$$U_{w^{(1)}}Ax = U_{w^{(1)}}b.$$

При реализации вычисления должны производиться в следующем порядке:

$$(I - 2w^{(1)}w^{(1)T})A = A - 2w^{(1)}(w^{(1)T}A),$$

так как такой порядок требует $O(m^2)$ операций.

В результате обнулили все элементы первого столбца, кроме самого верхнего. Далее повторяем те же действия для матрицы $A^{(2)}$ размерности $m - 1$ и так далее.

Лекция 13

На прошлой лекции разобрали два метода решения СЛАУ вида

$$Ax = b, \quad x = (x_1, \dots, x_m)^T.$$

Метод Гаусса требует порядка

$$\frac{2}{3|m^3 + O(m^2)}$$

операций, а метод отражений порядка

$$m^3 + O(m^2)$$

операций, то есть в общем случае метод отражений менее предпочтителен. Оказывается, что когда A близка к вырожденной, метод отражений будет предпочтительнее метода Гаусса.

Вышеупомянутые методы являются *прямыми методами*, то есть за конечное число арифметических операций при отсутствии ошибок округления получим точное решение СЛАУ.

Итерационные методы

Вспомним сначала, каким образом задается норма линейного оператора в конечномерном пространстве. Если

$$\|x\|_\infty = \max_i |x_i|,$$

то

$$\|A\|_\infty = \max_i \sum_{j=1}^{m} |a_i j|.$$

Для

$$\|x\|_1 = \sum_{i=1}^{m} |x_i|$$

норма линейного оператора равна

$$\|A\|_1 = \max_j \sum_{i=1}^{m} |a_i j|.$$

Наконец, для

$$\|x\|_2^2 = \sum_{i=1}^{m} x_i^2$$

норма линейного оператора имеет вид

$$\|A\|_2^2 = \max \lambda(A^T A).$$

Заметим, что при $A = A^T$

$$\|A\|_2 = \max |\lambda(A)|.$$

Метод простой итерации

Предположим, что система

$$Ax = b$$

может быть преобразована к эквивалентному виду

$$x = Bx + c. \tag{35}$$

Методом простой итерации для решения системы уравнений (35) называется метод[43]

$$x^{n+1} = Bx^n + c. \tag{36}$$

Теорема 13.1. *Пусть*[44] $\|B\| = q < 1$. *Тогда (36) сходится со скоростью геометрической прогрессии для любого начального приближения* x^0.

Доказательство 1. Решение (35) существует и единственно. Предположим обратное. Пусть $x \neq 0$ – решение однородной системы

$$x = Bx,$$

тогда

$$\|x\| = \|Bx\| \leq \|B\|\|x\| = q\|x\|,$$
$$(1-q)\|x\| \leq 0,$$

откуда следует, что $\|x\| = 0$.

2. Перейдем к уравнению для ошибки. Пусть $\overline{x}$ – решение (35). Тогда вычтем из (36) уравнение для точного решения

$$\overline{x} = B\overline{x} + c.$$

Получим уравнение для ошибки

$$r^{n+1} = Br^n, \quad r^n = x^n - \overline{x}.$$

Отсюда получем, что

$$\|r^{n+1}\| \leq \|B\|\|r^n\|,$$
$$\|r^n \leq q^n\|r^0\|n \underset{\to}{\to} \infty 0$$

Теорема доказана.

Обсудим теперь следующий вопрос. Хотим узнать, сколько требуется сделать шагов, чтобы найти решение с точностью ε, то есть чтобы

$$\frac{\|r^n\|}{\|r^0\|} \leq \varepsilon.$$

Получим, что

$$n > \ln \varepsilon^{-1} / \ln q^{-1}.$$

[43] Здесь n означает номер итерации, а не показатель степени.
[44] Заметим, что это условие является достаточным.

Теорема 13.2. *Метод (33) сходится* $\Leftrightarrow |\lambda(B)| < 1$.

Доказательство 1. Достаточность. Пусть

$$|\lambda(B)| \leq q < 1.$$

Вспомним теорему о приведении матрицы к жордановой форме. $\exists\ D$ такое, что

$$DBD^{-1} = \left(\begin{array}{c|c|c|c} J_1 & 0 & \dots & 0 \\ \hline 0 & J_2 & \dots & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \dots & J_m \end{array}\right), J_k = \begin{pmatrix} \lambda_k & 1 & \dots & 0 \\ 0 & \lambda_k & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k \end{pmatrix}. \tag{37}$$

Докажем, что в представлении (37) вместе 1 можно поставить любое число. Для этого рассмотрим матрицу

$$\mu B, \mu \in \mathbb{R}.$$

Для нее выполняется теорема о жордановой форме, то есть $\exists\ D$ такое, что в представлении (37) в J_k по диагонали будут стоять числа $\lambda_k \cdot \mu$. Умножив обе части представления на μ^{-1}, получим, что матрицы по диагонали будут иметь вид

$$J'_k = \begin{pmatrix} \lambda_j & \mu^{-1} & \dots & 0 \\ 0 & \lambda_j & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_j \end{pmatrix}.$$

Факт доказан.

Вернемся к доказательству теоремы. Погрешность метода итераций удовлетворяет соотношению

$$r^{n+1} = Br^n.$$

Из написанного выше следует, что $\exists$ матрица D такая, что

$$D^{-1}\Lambda D = B.$$

Тогда

$$r^n = D^{-1}\Lambda D r^{n-1} = D^{-1}\Lambda D D^{-1}\Lambda D r^{n-2} =$$
$$= D^{-1}\Lambda^2 D r^{n-2} = D^{-1}\Lambda^n D r^0.$$

Теперь, по условию

$$|\lambda_i| \leq q < 1.$$

Выберем μ из соотношения

$$q + \mu^{-1} < 1.$$

Тогда

$$\|\Lambda\|_\infty < 1,$$

откуда получим, что

$$\|r^n\|_\infty \leq \|D^{-1}\|\|\Lambda\|^n_\infty \|D\|_\infty \|r^0\|_\infty \to 0.$$

Достаточность доказана.

2. Необходимость. Допустим, что $|\lambda_i| \geq 1$. Пусть e_1 – собственный вектор, соответствующий λ_1. Предположим, что $r^0 = e_1$. Тогда

$$r^n = \lambda_1^n e_1 \not\to 0.$$

Теорема доказана.

В случае, когда $B = B^T$, $\|B\|_2 < 1$,

$$r^0 = \sum c_i e_i,$$

где e_i – собственные вектора B, получим, что

$$\|r^0\|_2^2 = \sum c_i^2,$$

$$r^n = \sum c_i \lambda_i^n e_i,$$

$$\|r^n\|_2^2 = \sum c_i^2 \lambda_i^{2n} \to 0,$$

причем убывает норма погрешности монотонно.

Рассмотрим теперь пример, иллюстрирующий проблему, которая может возникнуть, если B несиметрична. Пусть в

$$x = Bx + c$$

$$B = \begin{pmatrix} \alpha & \beta & \dots & 0 \\ 0 & \alpha & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha \end{pmatrix}$$

– двухдиагональная матрица, $|\alpha| < 1$, а

$$c = (0, \dots, 0, 1)^T.$$

Выпишем уравнение для x_m:

$$x_m = \alpha x_m + 1,$$

$$x_m = \frac{1}{1-\alpha}.$$

Для индексов $k < m$:

$$x_k = \alpha x_k + \beta x_{k+1} = \frac{\beta}{1-\alpha} x_{k+1}.$$

Будем рассматривать случай, когда число итераций в методе итераций не превышает размерности ($n < m$). Точное решение системы имеет вид

$$\overline{x} = \frac{1}{1-\alpha} \left(\frac{\beta}{1-\alpha}^{m-1}, \dots, 1 \right)^T,$$

$$\|\overline{x}\|_1 = \left|\frac{1}{1-\alpha}\right| \frac{1-(\beta/(1-\alpha))^m}{1-(\beta/(1-\alpha))}.$$

При соблюдении условий

$$|\alpha| < 1, \quad \left|\frac{\beta}{1-\alpha}\right| < 1 \tag{38}$$

$$\|x\|_1 \asymp \frac{1}{1-\alpha-\beta}.$$

Поставим третье условие

$$\frac{1}{1-\alpha-\beta} \asymp 1. \tag{39}$$

Предположим, что погрешность на начальном шаге

$$r^0 = (0, \ldots, 0, 1)^T.$$

Легко проверить, что

$$B^n = \begin{pmatrix} \alpha^n & C_n^1\alpha^{n-1}\beta & C_n^2\alpha^{n-2}\beta^2 & \ldots & \beta^n & 0 & \ldots \\ 0 & \alpha^n & \ldots & \ldots & \ldots & 0 & \ldots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \ldots & 0 & \ldots & \ldots \end{pmatrix}.$$

Оценим

$$\|B^n\|_1 = (|\alpha| + |\beta|)^n.$$

Наложим условие

$$\alpha| + |\beta| > 1. \tag{40}$$

Подберем α и β, исходя из условий (38, 39, 40). Пусть

$$\alpha = -0.9, \beta = 1.8.$$

С этими значениями получим, что

$$r^n = B^n r^0 = (0, \ldots, 0, \beta^n, \ldots, C_n^1\alpha^{n-1}\beta, \alpha^n)^T,$$

$$\|r^n\|_1 = (|\alpha| + |\beta|)^n = 2.7^n.$$

Погрешность будет возрастать экспоненциально до шага $n = m$. Это создает проблему. Например, при $m = 100$

$$\|r^100\|_1 = 2.7^{100},$$

а значит, хотя бы одна из компонент вектора погрешности больше, чем $2.7^{100}/100$.

Замечание В случае, когда матрица B несимметрична, метод хорошо сходится, если это матрица с диагональным преобладанием.

Переход к эквивалентной системе

Обсудим, как от системы

$$Ax = b \tag{41}$$

перейти к эквивалетной системе

$$x = Bx + c,$$

чтобы к ней уже применить итерационный метод. Прибавим и вычтем к (41) Dx, где D – невырожденная матрица. Получим

$$Dx = (D - A)x + b,$$

$$x = D^{-1}(D - A)x + D^{-1}b = \underbrace{(I - D^{-1}A)}_{=B}x + \underbrace{D^{-1}b}_{=c}.$$

Если, например, возьмем D такую, что $D^{-1} = \tau I$, получим

$$x = (I - \tau A)x + \tau b.$$

Лекция 14

СЛАУ с симметричной положительно определенной матрицей

Будем рассматривать систему вида

$$Ax = b, \quad A = A^T > 0,$$

собственные числа $\lambda(A) \in [\mu, M]$.

Выполнив простые преобразования (см. предыдущую лекцию), сведем систему к виду

$$x = \underbrace{(I - \tau A)}_{B} x + \underbrace{\tau b}_{c}.$$

Метод простой итерации будет выглядеть следующим образом:

$$x^{n+1} = (I - \tau A)x^n + \tau b. \tag{42}$$

Перепишем (42) в виде

$$\frac{x^{n+1} - x^n}{\tau} + Ax^n = b.$$

Если λ – собственное значение A, то $1-\lambda\tau$ – собственное значение $(I-\tau A)$. Требуется выяснить, при каких $\tau > 0$

$$|1 - \tau\lambda| < 1, \quad \lambda \in [\mu, M].$$

Это выполняется при $\tau < 2/M$. Таким образом, доказали следующий результат.

Теорема 14.1. *Метод (42) сходится, если $\tau \in (0, 2/M)$.*

Обсудим теперь, какое значение τ надо выбрать, чтобы метод (42) сходился наилучшим образом, то есть

$$\tau_0 = \arg \min_{\tau \in (0,2/M)} \|I = \tau A\|_2 = \arg \min_{\tau \in (0,2/M)} \max_{\lambda \in [\mu, M]} |1 - \tau\lambda|. \tag{43}$$

Решим задачу (43) в явном виде. Обратим внимание, как ведет себя $|1 - \tau\lambda|$ на заданном отрезке (рис. 14.1). Отсюда

$$\tau_0 = \arg \min_{\tau \in (0,2/M)} \max \left\{ |1 - \tau\mu|, |1 - \tau M| \right\}.$$

Решив задачу графически (рис. 14.1), найдем

$$\tau_0 = \frac{2}{\mu + M}.$$

При этом норма оператора будет равна

$$1 - \frac{2}{M + \mu}\mu = \frac{M - \mu}{M + \mu},$$

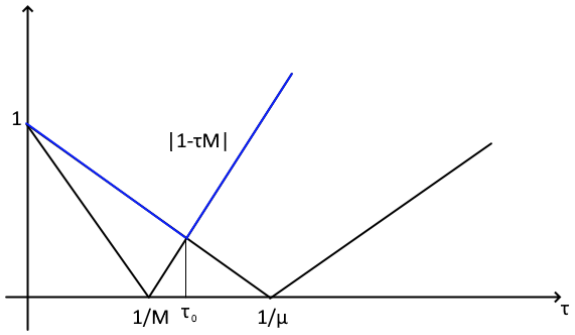

Рис. 14.1. Решение минимаксной задачи

и, следовательно, для погрешности будет справедлива оценка

$$\|r^n\|_2 \le \left(\frac{M-\mu}{M+\mu}\right)^n \|r^0\|_2.$$

Рассмотрим теперь случай, когда разброс спектра велик, то есть

$$M/\mu \gg 1.$$

В этом случае

$$\frac{M-\mu}{M+\mu} = (1-\mu/M)(1+\mu/M)^{-1} =$$
$$= (1-\mu/M)(1-\mu/M+o(\mu/M)) = 1-2\mu/M+o(\mu/M).$$

Посчитаем количество итераций n, необходимых, чтобы

$$\frac{\|r^n\|_2}{\|r^0\|_2} \le (1-2\mu/M)^n \le \varepsilon.$$

Получим

$$n = \ln^{-1}(1-2\mu/M)\ln\varepsilon \asymp \frac{M}{\mu}\ln\varepsilon^{-1}.$$

Таким образом, в случае $M/\mu \gg 1$ для нахождения решения с точностью ε необходимо $n \asymp O(M/\mu \ln\varepsilon^{-1})$ итераций.

Чебышёвское ускорение итерационного процесса

Далее обсудим, как можно уменьшить число итераций. Для каждой итерации будем выбирать свое τ_n.

Зафиксируем N – число итераций. Погрешность имеет вид

$$r^N = x^N - \overline{x} = (I-\tau_N A)(I-\tau_{N-1}A)\dots(I-\tau_1 A)r^0.$$

Оператор перехода будет иметь вид

$$B_N \equiv (I - \tau_N A)(I - \tau_{N-1} A) \dots (I - \tau_1 A).$$

Поставим задачу следующим образом. Требуется найти

$$(\tau_1, \dots, \tau_N) = \arg \min_{\tau_1, \dots, \tau_N} \|B_N\|_2. \tag{44}$$

Заметим, что, раскрыв скобки, получим следующее представление:

$$B_N = P_N(A), \quad P_N(0) = 1.$$

Кроме того, если λ – собственное значение матрицы A, то $P_N(\lambda)$ – собственное значение $P_N(A)$.

С учетом этих фактов задачу (44) можно записать в виде

$$(\tau_1, \dots, \tau_N) = \arg \min_{1, \dots, \tau_N} \max_{\lambda \in [\mu, M]} |P_N(\lambda)|.$$

Ограничения в задаче (44) являются очень сильными. Сформулируем поэтому более общую задачу.

Пусть K – класс многочленов степени N таких, что $P_N(0) = 1$. Требуется найти

$$Q_N(\lambda) = \arg \min_{P_N \in K} \max_{\mu \le \lambda \le M} |P_N(\lambda)|. \tag{45}$$

Теорема 14.2. *Решением задачи (45) является многочлен Чебышёва:*

$$Q_N(\lambda) = T_N \left(\frac{M + \mu - 2\lambda}{M - \mu} \right) / T_N \left(\frac{M + \mu}{M - \mu} \right).$$

Доказательство (От противного) Допустим, утверждение неверно, то есть $\exists$ многочлен $P_N(\lambda) \in K$ такой, что

$$\|P_N\|_{C[\mu, M]} < \|Q_N\|.$$

Пусть $\xi_0 < \xi_1 < \dots < \xi_N$ – точки экстремума $Q_N(\lambda)$ на отрезке $[\mu, M]$.

В этих предположениях

$$\operatorname{sign} (Q_N(\xi_i) - P_N(\xi_i)) =_N (\xi_i).$$

Тогда

$$Q_N(\lambda) - P_N(\lambda) = a_1 \lambda + \dots + a_N \lambda^N. \tag{46}$$

На отрезке $[\mu, M]$ имеем $N + 1$ перемену знака данного многочлена, то есть имеем N корней. Кроме того, из вида (46) следует, что 0 тоже является корнем, причем $0 \not\in [\mu, M]$. Получили противоречие.

Теорема доказана.

Корни многочлена $P_N(\lambda)$ имеют вид

$$\frac{M + \mu - 2\lambda_k}{M - \nu} = \cos \frac{2k - 1}{2N} \pi,$$

откуда получаем

$$\lambda_k = \frac{M+\mu}{2} - \frac{M-\mu}{2}\cos\frac{2k-1}{2N}\pi.$$

Так как многочлен имеет вид

$$P_N(\lambda) = (1-\tau_N\lambda)\dots(1-\tau_1\lambda),$$

можем выразить

$$\tau_k = \frac{2}{M+\mu-(M-\mu)\cos\frac{2k-1}{2N}\pi}. \tag{47}$$

Задача (45) решена в явном виде.

Выясним теперь, насколько лучше приближение при значениях (47). Напомним, что при $x > 1$

$$T_N(x) = \frac{1}{2}\left[(x+\sqrt{x^2-1})^N + (x-\sqrt{x^2-1})^N\right].$$

Вычислим

$$T_N\left(\frac{M+\mu}{M-\mu}\right) = \frac{1}{2}\left[\left(\frac{M+\mu}{M-\mu} + \sqrt{\frac{4M\mu}{(M-\mu)^2}}\right)^N + \left(\frac{M+\mu}{M-\mu} - \frac{2\sqrt{M\mu}}{M-\mu}\right)^N\right] =$$

$$= \frac{1}{2}\left[\left(\underbrace{\frac{\sqrt{M}+\sqrt{\mu}}{\sqrt{M}-\sqrt{\mu}}}_{\lambda_0}\right)^N + \left(\frac{\sqrt{M}-\sqrt{\mu}}{\sqrt{M}+\sqrt{\mu}}\right)^N\right] = \frac{1}{2}\left[\lambda_0^N + \lambda_0^{-N}\right].$$

Отсюда следует, что

$$\|Q_N(A)\|_2 = \frac{2}{\lambda_0^N + \lambda_0^{-N}} < \frac{2}{\lambda_0^N},$$

«средняя скорость сходимости» равна

$$2^{1/N}\frac{\sqrt{M}-\sqrt{\mu}}{\sqrt{M}+\sqrt{\mu}} \asymp 2^{1/N}\left(1-2\sqrt{\mu/M}\right).$$

Таким образом, для нахождения решения с точностью ε нам требуется

$$O(\sqrt{M/\mu}\ln\varepsilon^{-1})$$

итераций.[45]

Такой алгоритм называется *Чебышёвским ускорением итерационного процесса.*

[45] Заметим, что результат существенно отличается от предыдущего.

Перестановки τ_n

Обсудим еще один момент. (47) на каждом шаге могут выбираться в любом порядке. Обсудим перестановку $\{\tau_{j_k}\}$, предложенную В.И. Лебедевым (2008 г.), что выполняется

$$\|B_k\|_2 \leq 1.$$

Положим $N = 2^l$, тогда перестановки определяются как

$$(2,1), \quad l = 1,$$

$$(b_1^{l-1}, b_2^{l-1}, \ldots, b_{2^l-1}^{l-1}), \quad N = l - 1,$$

тогда при $N = 2^l$ нужная перестановка имеет вид

$$(2^l + 1 - b_1^{l-1}, b_1^{l-1}, 2^l + 1 - b_2^{l-1}, b_2^{l-1}, \ldots).$$

Выпишем несколько перестановок. При $l = 2$ перестановка имеет вид

$$(3, 2, 4, 1),$$

при $l = 3$

$$(6, 37, 2, 5, 4, 8, 1).$$

Лекция 15

Линейный оптимальный процесс

На прошлой лекции речь шла о решении системы уравнений вид

$$Ax = b, \quad A = A^T > 0, \quad \lambda(A) \in [\mu, M]$$

при помощи итерационного процесса, называемого *процессом Ричардсона* и имеющего вид

$$\frac{x^{n+1} - x^n}{\tau_{n+1}} + Ax^n = b.$$

При достаточно большой фиксированном N и специальном образом выбранных τ_n можно достичь «средней» скорости сходимости за

$$O(\sqrt{M/\mu}|\ln\varepsilon|) \tag{48}$$

итераций.

Обсудим теперь, можно ли добиться того, чтобы итерационный метод при любом n давал оценку (48). В таком случае должно выполняться

$$r^{n+1} = T_{n+1}\left(\frac{(M+\mu)I - 2A}{M-\mu}\right) t_{n+1}^{-1} r^0,$$

где

$$t_n = T_n\left(\frac{M+\mu}{M-\mu}\right).$$

По-другому,

$$t_{n+1} r^{n+1} n = T_{n+1}\left(\frac{(M+\mu)I - 2A}{M-\mu}\right) r^0, \tag{49}$$

$$t_n r^n = T_n\left(\frac{(M+\mu)I - 2A}{M-\mu}\right) r^0, \tag{50}$$

$$t_{n-1} r^{n-1} = T_{n-1}\left(\frac{(M+\mu)I - 2A}{M-\mu}\right) r^0. \tag{51}$$

Вспомним, что многочлены Чебышёва удовлетворяют реккурентному соотношению

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0,$$

откуда при $x = (M+\mu)/(M-\mu)$ получим

$$t_{n+1} - 2\left(\frac{M+\mu}{M-\mu}\right) t_n + t_{n-1} = 0. \tag{52}$$

Применим к (50) оператор $-2x\left(\frac{(M+\mu)I-2A}{M-\mu}\right)$ и прибавим к нему операторы (43,51). Получим

$$t_{n+1} r^{n+1} - \frac{2t_n}{M-\nu}\left(\frac{(M+\mu)I - 2A}{M-\mu}\right) r^n + t_{n-1} r^{n-1} = 0.$$

Отсюда получим соотношение

$$t_{n+1}x^{n+1} - \frac{2t_n(M+\mu)}{M-\mu}x^n + \frac{4t_n}{M-\mu}(Ax^n - b) + t_{n-1}x^{n-1} - \underbrace{\left(t_{n+1} - 2t_n\frac{M+\mu}{M-\mu} + t_{n-1}\right)}_{=0}\overline{x} = 0. \tag{53}$$

Итерационная формула для алгоритма построена. Введем обозначение

$$\frac{t_n}{t_{n+1}} = \omega_n.$$

Тогда из (53) получим

$$x^{n+1} - (1 + \omega_{n-1}\omega_n)x^n + \frac{2(1+\omega_{n-1}\omega_n}{M+\mu}(Ax^n - b) + \omega_{n-1}\omega_n x^{n-1} = 0. \tag{54}$$

Формула (54) называется *линейным оптималъным процессом.*

Запишем формулу для вычисления ω_n. При $n = 0$

$$\omega_0 = \frac{M-\mu}{M+\mu}.$$

Запишем (52) в виде

$$1 - 2\frac{M+\mu}{M-\mu}\omega_n + \omega_{n-1}\omega_n = 0,$$

откуда получим

$$\omega_n = \frac{1}{2\frac{M+\mu}{M-\mu} - \omega_{n-1}}.$$

В формуле (54) x^0 – задается, а x^1 вычисляется из соотношения

$$\frac{x^1 - x^0}{\tau_0} + Ax^0 = b, \quad \tau_0 = \frac{2}{M+\mu}.$$

Варционный итерационный процесс

Задача, как и в предыдущих случаях, ставится следующим образом:

$$Ax = b, \quad A = A^T > 0, \quad \lambda(A) \in [\mu, M],$$

причем хотим найти решение, не вычисляя значеняи μ, M.

Введем *квадратичный функцинал*

$$F(x) = (Ax, x) - 2(b, x). \tag{55}$$

Так как $A = A^T > 0$, можем ввести *энергетическую норму, индуцированную матрицей* A:

$$\|x\|_A^2 \equiv (Ax, x).$$

Проверим, что минимум квадратичного функционала $F(x)$ достигается на решении СЛАУ. Оказывается, что $F(x)$ представимо в виде

$$F(x) = \|x - \overline{x}\|_A^2 - \|\overline{x}\|_A^2.$$

Действительно,

$$\|x - \overline{x}\|_A^2 - \|\overline{x}\|_A^2 = (A(x - \overline{x}), x - \overline{x}) - (A\overline{x}, \overline{x}) =$$

$$= (Ax, x) - (A\overline{x}, x) - (Ax, \overline{x}) + (A\overline{x}, \overline{x}) - (A\overline{x}, \overline{x}) = (Ax, x) - 2(b, x).$$

Отсюда видно, что

$$\overline{x} = \arg\min_x F(x).$$

Будем искать минимум в виде

$$x^{n+1} = x^n - \alpha_n \nabla F(x^n), \tag{56}$$

где $\nabla F(x)$ – градиент, а α_n – некоторый параметр. Вычислим значение градиента:

$$\frac{\partial}{\partial x_k}\left(\sum_{i,j=1}^m a_{ij}x_i x_j - 2\sum_{i=1}^m b_i x_i\right) = 2\left(\sum_{j=1}^m a_{kj} - b_k\right),$$

откуда

$$\nabla F(x) = 2(Ax - b).$$

Обозначив $2\alpha_n = \delta_n$, запишем (56) в виде

$$x^{n+1} = x^n - \delta_n \underbrace{(Ax^n - b)}_{\xi^n}.$$

Заметим, что полученное выражение по структуре напоминает итерационную формулу предыдущего метода. Здесь δ_n выбирается так, тобы $F(x)$ был минимален. Распишем

$$F(x^{n+1}) = F(x^n - \delta_n\xi^n) = (A(x^n - \delta_n\xi^n), x^n - \delta_n\xi^n) - 2(b, x^n - \delta_n\xi^n) =$$

$$= (Ax^n, x^n) - \delta_n(A\xi^n, x^n) - \delta_n(Ax^n, \xi^n) + \delta_n^2(A\xi^n, \xi^n) - 2(b, x^n) + 2\delta_n(b, \xi^n).$$

$$\frac{dF(x^{n+1}}{d\delta_n} = 2\delta_n\|\xi^n\|_A^2 - 2(Ax^n, \xi^n) + 2(b, \xi^n) =$$

$$= 2\left[\delta_n\|\xi^n\|_A^2 - \|\xi^n\|^2\right].$$

Из соотношения

$$\frac{dF(x^{n+1}}{d\delta_n} = 0$$

находим

$$\delta_n = \frac{\|\xi^n\|^2}{\|\xi^n\|_A^2}.$$

Метод

$$x^{n+1} = x^n - \delta_n \underbrace{(Ax^n - b)}_{\xi^n}, \quad \delta_n = \frac{\|\xi^n\|^2}{\|\xi^n\|_A^2} \tag{57}$$

называется *методом наискорейшего градиентного спуска.*

Скорость сходимости вариационного итерационного метода

Сделаем один шаг по методу Ричардсона с оптимальным параметром

$$\frac{y - x^n}{\tau_0} + Ax^n = b,\ \tau_0 = \frac{2}{M + \mu}.$$

Обозначим $y - \overline{x} = z$, тогда

$$z = (I - \tau_n A)r^n,$$

$$r^{n+1} = (I - \delta_n A)r^n.$$

$$F(x^{n+1}) = \| \underbrace{x^{n+1} - \overline{x}}_{r^{n+1}} \|_A^2 - \|\overline{x}\|_A^2 = \|r^{n+1}\|_A^2 - \|\overline{x}\|_A^2 \le F(y) = \|z\|_A^2 - \|\overline{x}\|_A^2.$$

Отсюда следует, что погрешность метода наискорейшего градиентного спуска мажорируется сверху погрешностью метода Ричардсона:

$$\|r^{n+1}\|_A \le \|z\|_A.$$

Разложим

$$r^n = \sum c_i e_i,$$

тогда

$$z = \sum (1 - \tau_0 \lambda_i) c_i e_i,$$

$$\|z\|_A^2 = \sum \lambda_i (1 - \tau_0 \lambda_i)^2 c_i^2 \le \left(\frac{M - \mu}{M + \mu}\right)^2 \sum \lambda_i c_i^2 = \left(\frac{M - \mu}{M + \mu}\right)^2 \|r^n\|_A^2.$$

Таким образом, справедлива оценка

$$\|Rr^{n+1}\|_A \le \frac{M - \mu}{M + \mu} \|r^n\|_A.$$

Итак, получили, что скорость сходимости у метода наискорейшего градиентного спуска такая же, но при этом нам не нужно знать границы спектра $[\mu, M]$.

Уменьшение числа операций

Заметим, что в методе наискорейшего градиентного спуска нам потребуется в два раза больше операций, чем в методе Ричардсона. Обсудим, как можно уменьшить количество операций в этом методе. Применим к (57) матрицу A:

$$\underbrace{Ax^{n+1} - b}_{\xi^{n+1}} = \underbrace{Ax^n - b}_{\xi^n} - \delta_n A\xi^n.$$

Вычисляя на каждом шаге

$$\begin{cases} x^{n+1} = x^n - \delta_n \xi^n, \\ \xi^{n+1} = \xi^n - \delta_n A\xi^n, \end{cases}$$

получим такое же количество операций, как и в методе Ричардсона.

Метод Зейделя

Обсудим еще один метод. Будем считать, что диагональные элементы матрицы СЛАУ $a_{ii} \neq 0$. Представим

$$A = B + C,$$

где B – нижнетреугольная матрица, включающая в себя диагональ (и, очевидно, невырожденная), а C – верхнетреугольная, $c_{ii} = 0$.

Итерационный метод вида

$$Bx^{n+1} + Cx^n = b \tag{58}$$

называется *методом Зейделя.*

Теорема 15.1. *Если $A = A^T > 0$, то метод Зейделя (58) сходится.*

Доказательство Рассмотрим квадратичный функционал $F(x)$ (55). Будем искать x_1^{n+1}, для которого функционал будет минимален:

$$F(x) = \sum_{i,j=1}^{m} a_{ij}x_i x_j - 2\sum_{i=1}^{m} b_i x_i,$$

$$a_{11}(x_1^{n+1})^n + 2\sum_{j=2}^{m} a_{1j}x_1^{n+1}x_j^n - 2b_1 x_1^{n+1} - 2\sum_{j=2}^{m} b_i x_i^n \to \min_{x_1^{n+1}}.$$

Взяв производную и приравняв ее к 0, получим

$$a_{11}x_1^{n+1} + \sum_{j=2}^{m} a_{ij}x_j^n = b,$$

то есть первое уравнение метода Зейделя.

Аналогично с остальными переменными. Для $A = A^T > 0$ получается метод покоординатного спуска.

Отсюда получаем, что, если x^n не является решением СЛАУ, верно

$$F(x^{n+1}) < F(x^n).$$

Это выполняется $\Leftrightarrow$ когда

$$\|r^{n+1}\|_A^2 - \|\overline{x}\|_A^2 < \|r^n\|_A^2 - \|\overline{x}\|_A^2,$$

откуда получаем соотношение

$$\frac{\|r^{n+1}\|_A}{\|r^n\|_A} < 1.$$

Погрешность метода Зейделя имеет вид

$$r^{n+1} = -B^{-1}Cr^n,$$

а значит,

$$\frac{\|B^{-1}Cr^n\|_A}{\|r^n\|_A} < 1, \quad \forall\, r^n \neq 0.$$

Отсюда

$$\sup_{\|y\|=1} \frac{\|B^{-1}Cy\|_A}{\|y\|_A} < 1,$$

и значит, метод Зейделя сходится.

Теорема доказана.

Лекция 16

Применение переобуславливателя в методе Ричардсона

Как и ранее, рассматриваем задачу

$$Ax = b, \quad A = A^T > 0.$$

Пусть дан итерационный метод Ричардсона

$$\frac{x^{n+1} - x^n}{\tau} + Ax^n = b,$$

начальное приближение x_0 задано. Нашей целью является получение решения за как можно меньшее число операций.

Ранее, если разброс собственных значений A был большим, вводили дополнительные параметры τ_n. Введем теперь модификацию

$$B\frac{x^{n+1} - x^n}{\tau} + Ax^n = b, \tag{59}$$

где на B наложим следующие условия:

1. B – «легко обратимая», то есть СЛАУ

$$Bx = c$$

легко решается;

2. $B = B^T > 0$.

В таком случае B называется *переобуславливателем (preconditioner)*. Напомним, что, так как $B = B^T > 0$, можно ввести понятие энергетической нормы

$$\|y\|_B^2 \equiv (By, y).$$

Кроме того, $\exists\ \gamma_1,\ \gamma_2 > 0$ такие, что

$$\gamma_1 B \le A \le \gamma_2 B, \tag{60}$$

то есть

$$\gamma_1(By, y) \le (Ay, y) \le \gamma_2(By, y).$$

Теперь, если мы рассматриваем класс итерационных задач и γ_1 и γ_2 не зависят от размерности m, то A и B называются *спектрально эквивалентными*.

Для ошибки

$$r^n = x^n - \overline{x}$$

уравнение имеет вид

$$B\frac{r^{n+1} - r^n}{\tau} + Ar^n = 0.$$

Преобразовав, получим

$$\frac{r^{n+1} - r^n}{\tau} + B^{-1}Ar^n = 0 \tag{61}$$

– уравнение, похожее на уравнение метода простой итерации. Проблема состоит в следущем. Произведение двух симметричных матриц $B^{-1}A$ не обязательно является симмтеричной матрицей.

Введем следующее обозначение:

$$B^{1/2}r^n = v^n,$$

тогда (61) можно записать в виде[46]

$$\frac{v^{n+1} - v^n}{\tau} + B^{-1/2}AB^{-1/2}v^n = 0.$$

Так как $B^{-1/2}AB^{-1/2}$ является симметричной положительно определенной матрицей, свели задачу к методу простой итерации. Проблема решена.

Исследуем теперь спектр оператора $B^{-1/2}AB^{-1/2}$.

$$\max\lambda(B^{-1/2}AB^{-1/2}) = \max_{y\neq 0}\frac{(B^{-1/2}AB^{-1/2}y, y)}{(y,y)}.$$

Обозначим $B^{-1/2}y = x$. Тогда

$$\max\lambda(B^{-1/2}AB^{-1/2}) = \max_{x\neq 0}\frac{(Ax,x)}{(Bx,x)} \leq \gamma_2$$

в силу (60). Аналогично получим, что

$$\min\lambda(B^{-1/2}AB^{-1/2}) = \gamma_1.$$

По теореме 14.1 о сходимости метода Ричардсона метод (59 сходится, если $\tau \in (0, 2/\gamma_2)$.

Если положим $\tau = \tau_0 = \frac{2}{\gamma_1+\gamma_2}$, имеет место оценка

$$\|v^{n+1}\| \leq \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1}\|v^n\|.$$

Отсюда с учетом

$$\|v^n\| = (B^{-1/2}r^n, B^{1/2}r^n) = \|r^n\|_B.$$

получим, что

$$\|r^{n+1}\|_B \leq \frac{\gamma_2 - \gamma_1}{\gamma_2 + \gamma_1}\|r^n\|_B.$$

Так как все нормы в пространстве эквивалентны, оценка для r^n будет сходиться и в любой другой норме.

[46]Заметим, что $B^{1/2}$ нужна нам для доказательства, но при подсчетах данное представление использовать неудобно.

Метод градиентного спуска

е

На прошлой лекции рассматривали одношаговый метод наискорейшего градиентного спуска. Данную задачу можно обобщить для n шагов. Постановка будет выглядеть следующим образом:

$$r^n = P_n(A)r^0, \quad P_n(0) = 1.$$

$$F(x^n) \to \min_{P_n}.$$

К рассмотрению данной задачи вернемся во второй части курса.

Некорректные задачи

Рассмотрим следующие случаи для СЛАУ

$$Ax = b. \tag{62}$$

1. (62) не имеет решения в классическом смысле.
2. Пусть $A = A^T > 0$ и $\lambda_{\max}(A) \asymp 1$, $\lambda_{\min}(A) \ll 1$.

Если

$$b = \sum_{k=1}^{m} b_k e_k,$$

то решение можно записать как

$$\overline{x} = \sum_{k=1}^{m} \frac{b_k}{\lambda_k} e_k.$$

Если b задано с ошибкой $\widetilde{b} = b + \delta b$, тогда приближенное решение будет иметь вид

$$\widetilde{x} = \sum_{k=1}^{m} \frac{b_k}{\lambda_k} e_k + \sum_{k=1}^{m} \frac{\delta b_k}{\lambda_k} e_k.$$

Тогда в условиях пункта 2 малому изменению правой части b соответствует большое[47] изменение решения. Такого рода задачи называются *некорректными*. Для решения таких задач необходимы дополнительные условия.

Метод Тихонова

Будем рассматривать следующую задачу. Пусть собственные значения A (62) имеют вид

$$\lambda_1, \ldots, \lambda_k \asymp 1$$

[47] Что именно считать малым и большим изменениями, необходимо оговаривать для каждой задачи отдельно.

– назовем их *гладкой частью спектра*, и

$$\lambda_{k+1}, \dots, \lambda_m \ll 1$$

– *сильно осциллирующей.*

Наложим дополнительное условие. Будем считать, что решение находится в гладкой части спектра:

$$\overline{x} = \sum_{j=1}^{k} x_j e_j.$$

Предположим, что евклидова норма правой части мала, то есть

$$\|\delta b\| \le \delta.$$

Кроме того, «испортим» еще и матрицу A. Как это сделать, обсудим позднее. Вместо задачи

$$A\widetilde{x} = \widetilde{b}$$

будем решать задачу

$$A_\alpha x_\alpha \equiv (A + \alpha I)x_\alpha = \widetilde{b}.$$

Решение выписывается в явном виде:

$$x_\alpha = \sum_{j=1}^{m} \frac{b_j + \delta b_j}{\lambda_j + \alpha} e_j.$$

Рассмотрим теперь погрешность

$$x_\alpha - \overline{x} = \sum_{j=1}^{k} b_j \underbrace{\left(\frac{1}{\lambda_j + \alpha} - \frac{1}{\lambda_j} \right)}_{=\alpha/(\lambda_j(\lambda_j+\alpha))} e_j + \sum_{j=1}^{k} \frac{\delta b_j}{\lambda_j + \alpha} e_j +$$

$$+ \sum_{j=k+1}^{m} \frac{\delta b_j}{\lambda_j + \alpha} e_j \equiv s_1 + s_2 + s_3,$$

где s_i – каждая из трех сумм. Оценим каждое из слагаемых отдельно. Получим

$$\|s_1\|^2 \asymp \alpha^2 \|b\|^2, \quad \|s_1\| \le c_1 \alpha \|b\|,$$

$$\|s_2\| \le \delta,$$

$$\|s_3\| \le \frac{\delta}{\alpha}.$$

Будем искать минимум для погрешности, то есть задача выглядит следующим образом:

$$\|x^\alpha - \overline{x}\|^2 \le c_1^2 \alpha^2 + \left(1 + \frac{1}{\alpha^2}\right) \delta^2 \to \min_\alpha.$$

Взяв производную, получим

$$c_1^2 - \frac{\delta^2}{\alpha^4} = 0,$$

откуда

$$\alpha = c\sqrt{\delta}.$$

Оценка погрешности

$$\|x^\alpha - \overline{x}\| \leq c\delta.$$

Итак, оказывается, если в исходной некорректной задаче отодвинуть от 0 спектр на величину, пропорциональную $\sqrt{\delta}$, возможно найти приблизительное решение.

Решение некорректных задач в общем случае

Умножим в (62) обе части на A^T слева. Получим

$$A^T Ax = A^T b. \tag{63}$$

Если бы A была симметричной, собственные значения λ_j, и, соответственно, отношение $\lambda_{\max}/\lambda_{\min}$, от которого зависисит скорость сходимости итерационных процессов, возвелись бы в квадрат, что сильно бы ухудшило свойства матрицы системы.

Однако, в случае, когда размерность СЛАУ невелика, преобразование (63) иногда используют. Оказывается, что система (63) всегда имеет решение. Покажем это.

1. $\ker(A^T A) = \ker A$. Если $x \in \ker A$, очевидно, $x \in \ker(A^T A)$. Если $x \in \ker(A^T A)$, то

$$0 = (A^T Ax, x) = \|Ax\|^2,$$

откуда следует, что $x \in \ker A$.

2. Убедимся, что правая часть (63) ортогональна ядру $A^T A$. Пусть

$$x \in \ker(A^T A),$$

тогда $x \in \ker A$. Отсюда получаем, что

$$(A^T b, x) = (b, Ax) = 0.$$

Утверждение доказано.

МЕХАНИКО-
МАТЕМАТИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in

ЛЕКЦИИ УЧЕНЫХ МГУ