ОГЛАВЛЕНИЕ

Пришлось потрудится, чтобы все сделать.

Глава 1
Приближение функций и интерполяция

§1. Равномерное приближение функций. Полиномы Чебышева.

Идеи приближения функций пронизывают всю вычислительную математику.

Чем приближать? Мы будем рассматривать приближение полиномами.

Как измерять близость функций? Равномерная близость.

Поясним, что это значит. Множество функций, заданных и непрерывных на промежутке $[a,b]$, обозначим через $C = C[a,b]$ и каждой функции $f \in C$ сопоставим число $\|f\| = \|f\|_C = \max_{x\in[a,b]} |f(x)|$, называемое нормой функции f (в пространстве C). Отметим свойства нормы:

$\langle 1\rangle$ $\|f\| \geq 0$ и $\|f\| = 0$ в том и только в том случае, если $f(x) \equiv 0$;

$\langle 2\rangle$ $\|\alpha f\| = |\alpha|\,\|f\|$;

$\langle 3\rangle$ $\|f+g\| \leq \|f\| + \|g\|$ (неравенство треугольника).

Из свойства $\langle 3\rangle$ следует, что для любых функций $f, g \in C$ выполняется неравенство $\big|\,\|f\| - \|g\|\,\big| \leq \|f-g\|$.

Пусть дана последовательность функций $\{\,f_n\,\} \subset C$. Соотношение $\|f_n - f\| \to 0$ (сходимость по норме в C) означает равномерную сходимость последовательности $\{\,f_n\,\}$ к f.

Введем обозначение: $\mathbb{P}_n = \{P_n\}$ — множество всех полиномов степени не выше n.

В терминах нормы известная из курса анализа 1-я теорема Вейерштрасса может быть сформулирована в виде

Теорема (1-я теорема Вейерштрасса). Для любой функции $f \in C$ по любому $\varepsilon > 0$ найдутся такое n и такой $P_n \in \mathbb{P}_n$, что $\|f - P_n\| < \varepsilon$.

Определение. *Наилучшим приближением* функции $f \in C$ полиномами степени n называется число

$$E_n(f) = \inf_{P_n\in\mathbb{P}_n} \|f - P_n\|.$$

Полином $P_n \in \mathbb{P}_n$ называется *полиномом наилучшего приближения* функции f, если $\|f - P_n\| = E_n(f)$.

Теорема Вейерштрасса означает, что для любой $f \in C$ $E_n(f) \to 0$.

Докажем существование полинома наилучшего приближения.

Лемма 1. $F(A) = F(a_0,\ldots,a_n) = \|f - P_n\|_C$, где $P_n(x) = P_n(A,x) = a_0 + \cdots + a_n x^n$, есть непрерывная функция аргументов a_k.

Доказательство. Положим $c = \max\{|a|, |b|\}$. Тогда в понятных обозначениях

$$|F(A+\Delta A) - F(A)| \leq \left\|\sum_{k=0}^{n} \Delta a_k x^k\right\| \leq \sum_{k=0}^{n} |\Delta a_k| c^k \leq \sqrt{\sum_{k=0}^{n} \Delta a_k^2}\sqrt{\sum_{k=0}^{n} c^k}. \quad \blacksquare$$

Лемма 2. Существует такая постоянная m, зависящая лишь от n и промежутка $[a,b]$, что для любого $P_n \in \mathbb{P}_n$ ($P_n = a_0 + \cdots + a_n x^n$) выполняется

неравенство

$$\|P_n\| \geq m \left(\sum_{k=0}^{n} a_k^2 \right)^{1/2}$$

Доказательство этой леммы будет получено позднее, в §1 главы 3, как следствие более общего утверждения.

Теорема. Для любой функции $f \in C[a,b]$ существует полином наилучшего приближения $P_n \in \mathbb{P}_n$.

Доказательство. Требуется доказать, что непрерывная функция $F(A)$, определенная в лемме 1, достигает своего наименьшего значения. Положим $R = 2\|f\|/m$ ($m > 0$ — число из леммы 2). В шаре $S_R = \{ A \mid \sum a_k^2 \leq R^2 \}$ функция $F(A)$ достигает своего наименьшего значения в некоторой точке $A^* \in S_R$ (т.к. S_R — замкнутое ограниченное множество). Если же $A \notin S_R$, то

$$F(A) = \|f - P_n(A,\cdot)\| \geq \|P_n(A,\cdot)\| - \|f\| > mR - \|f\| = \|f\| = F(0) \geq F(A^*),$$

так что A^* — точка глобального минимума. ■

Известно, что для любой непрерывной функции $f \in C$ ее полином наилучшего приближения в классе $\mathbb{P}_n$ единственный, но это утверждение оставим без доказательства.

Рассмотрим одну частную, но очень важную задачу. На промежутке $[-1,1]$ для функции $f(x) = x^n$ требуется построить ее полином наилучшего приближения степени $n-1$. Если $Q_{n-1} \in \mathbb{P}_{n-1}$ решает поставленную задачу, то $P_n(x) = x^n - Q_{n-1}(x)$ есть полином степени n со старшим коэффициентом 1, решающий задачу: среди всех полиномов степени n со старшим коэффициентом 1, найти тот, для которого $\|P_n\|_{C[-1,1]}$ минимальна (эти две задачи эквивалентны). Полином, решающий вторую задачу, называется *полиномом, наименее уклоняющимся от нуля*.

Определение. Полиномом Чебышева степени n называется функция, задаваемая на промежутке $[-1,1]$ формулой

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

Из формулы $\cos(n+2)\theta = 2cos\theta \cos(n+1)\theta - \cos n\theta$ легко получается *рекуррентная формула* для многочленов Чебышева:

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

которая (учитывая, что $T_0(x) = 1, T_1(x) = x$) позволяет легко доказать методом индукции, что $T_n(x)$ есть полином степени n со старшим коэффициентом (при $n \geq 1$) 2^{n-1}.

Непосредственно из определения легко находятся точки y_k максимума модуля и корни x_k полинома T_n:

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

Теорема. Наименее уклоняется от нуля *приведенный* многочлен Чебышева $\tilde{T}_n(x) = \frac{1}{2^{n-1}} T_n(x)$.

Доказательство. $\tilde{T}_n(y_k) = (-1)^k/2^{n-1}$. Если $P_n(x) = x^n + \dots$ таков, что $\|P_n\| < \|\tilde{T}_n\|$, то $\operatorname{sign}(\tilde{T}_n - P_n)(y_k) = (-1)^k$ и $\tilde{T}_n - P_n$ имеет n корней, так что $P_n \equiv \tilde{T}_n$, что невозможно. ■

Следствие. $E_{n-1}(x^n) = 1/2^{n-1}$.

Замечание. Как видно из доказательства теоремы, для функции x^n и ее полинома наилучшего приближения степени $n-1$ нашлись такие $n+1$ точки (это точки y_k), в которых разность между ними достигает максимального по величине значения с чередующимися знаками. Это — общее явление. Верна

Теорема П.Л.Чебышева (без доказательства). Пусть $f \in C$, $P_n \in \mathbb{P}_n$. Для того чтобы P_n был полиномом наилучшего приближения, необходимо и достаточно существование *чебышевского альтернанса*, т.е. таких точек $a \le x_1 < \dots < x_{n+2} \le b$, что

1) $|f(x_k) - P_n(x_k)| = \|f - P_n\|$,

2) $\operatorname{sign}\big(f(x_k) - P_n(x_k)\big) = -\operatorname{sign}\big(f(x_{k+1}) - P_n(x_{k+1})\big)$.

Гладкие функции хорошо приближаются полиномами. Без доказательства приведем теорему:

Теорема. (Д.Джексон, 1912). При каждом натуральном p найдется такая постоянная c_p, что для любой функции $f \in C^{(p)}[a,b]$ выполняются неравенства

$$E_n(f) \le \frac{c_p}{n^p}(b-a)^p \|f^{(p)}\| \quad (n \ge p-1).$$

Задача 1. Показать, что в теореме Джексона условие $n \ge p-1$ существенно.

Задача 2. Доказать теорему о единственности полинома наилучшего приближения, используя теорему об альтернансе. *Указание:* Показать, что если полиномов наилучшего приближения два, то их полусумма также полином наилучшего приближения и воспользоваться тем, что для него существует альтернанс.

Задача 3. Показать, что если для полинома $P_n \in \mathbb{P}_n$ при всех $x \in [-1,1]$ выполняется неравенство $|P_n(x)| \le 1$, то при $x > 1$ будет $|P_n(x)| \le T_n(x)$.

§2. Конечные и разделенные разности

Определение Дано $h > 0$. Конечной разностью с шагом h функции $f \in C[a,b]$ называется функция $\Delta f(x) = f(x+h) - f(x)$. Конечные разности высших порядков определяются рекурсивно: $\Delta^k f(x) = \Delta^{k-1} f(x+h) - \Delta^{k-1} f(x)$.

Конечная разность $\Delta^k f$ задана на промежутке $[a, b-kh]$.

Конечные разности — аппарат работы с функциями, заданными таблично в равноотстоящих узлах. Если нам известны значения функции f в точках $x_j = x_0 + jh$, $j = 0, \dots, N$, то в тех же узлах при $j = 0, \dots, N-k$ могут быть вычислены и значения $\Delta^k f$. В таблицах, которые наряду со значениями функции содержат и значения ее разностей, каждую следующую разность

обычно принято размещать на полстроки ниже, так что такая таблица выглядит так (если приводятся разности до 3-го порядка):

$$\begin{array}{ccccc}
x & f & \Delta f & \Delta^2 f & \Delta^3 f \\
\dots & \dots & \dots & \dots & \dots \\
x_{n-1} & f_{n-1} & & \Delta^2 f_{n-2} & \\
 & & \Delta f_{n-1} & & \Delta^3 f_{n-2} \\
x_n & f_n & & \Delta^2 f_{n-1} & \\
 & & \Delta f_n & & \Delta^3 f_{n-1} \\
x_{n+1} & f_{n+1} & & \Delta^2 f_n & \\
\dots & \dots & \dots & \dots & \dots
\end{array}$$

Отметим основные свойства конечных разностей.

⟨1⟩ Если $f = \alpha_1 g_1 + \alpha_2 g_2$, то $\Delta^k f = \alpha_1 \Delta^k g_1 + \alpha_2 \Delta^k g_2$.

⟨2⟩ Если p — полином степени n, то Δp — полином степени $n-1$, $\Delta^k p$ — полином степени $n-k$, в частности, $\Delta^n p$ — постоянная, а разности высших порядков тождественно равны нулю.

⟨3⟩ Непосредственно через значения самой функции конечные разности выражаются формулой:

$$\Delta^k f_0 = f_k - C_k^1 f_{k-1} + C_k^2 f_{k-2} + \dots + (-1)^k f_0.$$

Здесь C_k^j — биномиальные коэффициенты. Формула легко доказывается методом индукции с учетом известного равенства $C_{k-1}^j + C_{k-1}^{j-1} = C_k^j$.

Если ввести "оператор сдвига" $Ef(x) = f(x+h)$, то приведенная формула может быть записана в символической форме $\Delta^k = (E-1)^k$. Имеется в виду, что правая часть раскрывается по формуле бинома Ньютона.

⟨4⟩ Значение функции f в точке x_k может быть выражено через значения ее разностей в точке x_0:

$$f_n = f_0 + C_n^1 \Delta f_0 + C_n^2 \Delta^2 f_0 + \dots + \Delta^n f_0.$$

Формула также легко доказывается методом индукции. Для доказательства возможности индуктивного перехода следует воспользоваться тем, что $f_n = f_{n-1} + \Delta f_{n-1}$ и для обеих функций, стоящих справа, воспользоваться индуктивным предположением. Мнемоническая запись формулы: $E^n = (1+\Delta)^n$.

⟨5⟩ Если функция f r раз непрерывно дифференцируема ($f \in C^{(r)}$), то таковы же и ее конечные разности и $(\Delta^k f)^{(r)}(x) = (\Delta^k f^{(r)})(x)$.

⟨6⟩ Если функция f k раз непрерывно дифференцируема, то найдется такая точка $\xi \in (x_0, x_0 + kh)$, что $\Delta^k(x_0) = h^k f^{(k)}(\xi)$. При $k = 1$ доказываемое свойство есть формула Лагранжа. Возможность индуктивного перехода следует из цепочки равенств:

$$\Delta^k f(x_0) = \Delta^{k-1} f(x_0 + h) - \Delta^{k-1} f(x_0) = h \Delta^{k-1} f'(\eta) = h^k f^{(k)}(\xi).$$

Здесь $\eta \in (x_0, x_0 + h)$, $\xi \in (\eta, \eta + (k-1)h) \subset (x_0, x_0 + kh)$.

При работе с таблично заданными функциями при неравноотстоящих узлах конечные разности заменяются разделенными.

Определение. Разделенной разностью (разностным отношением) первого порядка функции $f(x)$ называется функция двух переменных

$$f(x_0, x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \qquad (x_1 \neq x_0).$$

Разделенные разности высших порядков определяются рекурсивно, причем разделенная разность k-го порядка есть функция $(k+1)$-го попарно не совпадающих аргументов:

$$f(x_0, x_1, \ldots, x_k) = \frac{f(x_1, \ldots, x_k) - f(x_0, x_1, \ldots, x_{k-1})}{x_k - x_0}.$$

Перечислим основные свойства разделенных разностей.
$\langle 1 \rangle$ Если $f = \alpha_1 g_1 + \alpha_2 g_2$, то

$$f(x_0, \ldots, x_k) = \alpha_1 g_1(x_0, \ldots, x_k) + \alpha_2 g_2(x_0, \ldots, x_k)$$

.

$\langle 2 \rangle$ Справедливо представление:

$$f(x_0, \ldots, x_k) = \frac{f(x_0)}{(x_0 - x_1)\ldots(x_0 - x_k)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)\ldots(x_1 - x_k)} + \\ + \cdots + \frac{f(x_k)}{(x_k - x_0)\ldots(x_k - x_{k-1})}.$$

Доказательство проводится методом индукции. При $k = 1$ формула очевидна. Возможность индуктивного перехода от k к $k+1$ покажем только для $k = 1$ — общий случай не сложнее в идейном отношении, но громоздок. Итак,

$$f(x_0, x_1, x_2) = \frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0} =$$
$$= \frac{1}{x_2 - x_0}\left[\frac{f(x_1)}{x_1 - x_2} + \frac{f(x_2)}{x_2 - x_1}\right] - \frac{1}{x_2 - x_0}\left[\frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0}\right] =$$
$$= \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{x_2 - x_0}\left[\frac{1}{x_1 - x_2} - \frac{1}{x_1 - x_0}\right] + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

и остается заметить, что

$$\frac{1}{x_1 - x_2} - \frac{1}{x_1 - x_0} = \frac{x_2 - x_0}{(x_1 - x_0)(x_1 - x_2)}. \quad \blacksquare$$

$\langle 3 \rangle$ Разделенная разность $f(x_0, \ldots, x_k)$ есть симметричная функция своих аргументов, т.е. от перестановки аргументов ее значение не меняется. Это свойство есть непосредственное следствие предыдущего.

Теперь можно сказать, что разделенная разность k-го порядка есть первая разделенная разность от $(k-1)$-й по любому ее аргументу.

⟨4⟩ Если p — полином степени n, то разделенная разность порядка k есть полином степени $n-k$ от $k+1$ аргументов.

⟨5⟩ Справедливо представление

$$f(x_n) = f(x_0) + (x_n - x_0)f(x_0, x_1) + (x_n - x_0)(x_n - x_1)f(x_0, x_1, x_2) + \\ + \cdots + (x_n - x_0)\ldots(x_n - x_{n-1})f(x_0, \ldots, x_n).$$

Доказательство проводится методом индукции. При $n = 1$ формула очевидна. Покажем возможность индуктивного перехода от $n-1$ к n. Используя индуктивное предположение для точек $x_0, \ldots, x_{n-2}, x_n$, имеем

$$f(x_n) = f(x_0) + (x_n - x_0)f(x_0, x_1) + \cdots + \\ + (x_n - x_0)\ldots(x_n - x_{n-2})f(x_0, \ldots, x_{n-2}, x_n)$$

и остается воспользоваться тем, что

$$f(x_0, \ldots, x_{n-2}, x_n) = f(x_0, \ldots, x_{n-2}, x_{n-1}) + f(x_0, \ldots, x_{n-1}, x_n)(x_n - x_{n-1}).$$

⟨6⟩ Пусть $\alpha = \min x_k$, $\beta = \max x_k$. Если на промежутке $[\alpha, \beta]$ функция f n раз непрерывно дифференцируема ($f \in C^{(n)}$), то найдется такая точка $\xi \in (\alpha, \beta)$, что

$$f(x_0, \ldots, x_n) = \frac{1}{n!} f^{(n)}(\xi).$$

Доказательство. Рассмотрим полином степени n

$$P_n(x) = f(x_0) + (x - x_0)f(x_0, x_1) + \cdots + (x - x_0)\ldots(x - x_{n-1})f(x_0, \ldots, x_n)$$

и функцию $\varphi(x) = f(x) - P_n(x)$. Очевидно, что $\varphi \in C^{(n)}$. Согласно предыдущему свойству $P_n(x_k) = f(x_k)$ $(k = 0, \ldots, n)$, так что функция φ имеет на $[\alpha, \beta]$ не менее чем $n+1$ различных корней. По теореме Ролля φ' имеет на (α, β) не менее n корней, φ'' — не менее, чем $n-1$, и $\varphi^{(n)}$ по меньшей мере один корень ξ. Но $\varphi^{(n)}(\xi) = f^{(n)}(\xi) - n! f(x_0, \ldots, x_n)$. ■

Доказанное свойство позволяет нам доопределить по непрерывности разделенную разность порядка n на случай, когда все или некоторые из ее аргументов совпадают.

⟨7⟩ Если функция f n раз непрерывно дифференцируема на промежутке $[a, b]$, то при $k \le n$ разделенная разность $f(x_0, \ldots, x_k)$ может быть продолжена по непрерывности на весь "куб" $[a, b]^{k+1}$, причем если $x_0 = x_1 = \cdots = x_k$, то

$$f(x_0, x_0, \ldots, x_0) = \frac{1}{k!}, \tag{1}$$

а если среди аргументов $x_0, \ldots, x_k$ имеются хоть два различных (для определенности $x_0 \ne x_k$), то[1]

$$f(x_0, \ldots, x_n) = \frac{f(x_1, \ldots x_n) - f(x_0, \ldots, x_{n-1})}{x_n - x_0}. \tag{2}$$

[1] Обратим внимание, что это *рекуррентное* (по k) определение.

Для доопределенных таким образом по непрерывности разделенных разностей сохраняются свойства ⟨1⟩, ⟨3⟩-⟨6⟩.

Доказательство. Заметим прежде всего, что если $f(x_0,\dots,x_k)$ доопределена по непрерывности на случай хотя бы нескольких совпадающих аргументов, то для нее выполняются свойства ⟨1⟩, ⟨3⟩-⟨6⟩. Это легко доказывается предельным переходом. Непрерывность доопределенной формулами (1)-(2) на случай совпадающих аргументов разделенной разности доказывается индукцией по k и следует из непрерывности производных f^k $(k \le n)$. ■

Аргументы разделенной разности часто называют узлами. Если некоторый узел встречается среди аргументов разделенной разности k раз, то его называют узлом кратности k. Для того чтобы вычислить $f(x_0,\dots,x_n)$, согласно формулам (1)-(2) достаточно знать в каждом узле x_k значение самой функции f и ее производных до порядка $m-1$ включительно, если кратность этого узла есть m. Если функция f n раз непрерывно дифференцируема, то ее разделенные разности порядка выше n определены и непрерывны для тех значений аргументов, когда кратность каждого узла не превышает n.

⟨8⟩ Если точки x_k равноотстоящи: $x_k = x_0 + kh$, то очевидна связь между конечными и разделенными разностями:

$$f(x_0,\dots,x_n) = \frac{1}{h^n n!}\Delta^n(x_0).$$

Таблица разделенных разностей обычно выглядит так:

x	$f(x)$	$f(x,y)$	$f(x,y,z)$	$f(x,y,z,t)$
x_0	$f(x_0)$			
		$f(x_0,x_1)$		
x_1	$f(x_1)$		$f(x_0,x_1,x_2)$	
		$f(x_1,x_2)$		$f(x_0,x_1,x_2,x_3)$
x_2	$f(x_2)$		$f(x_1,x_2,x_3)$	
		$f(x_2,x_3)$		$f(x_1,x_2,x_3,x_4)$
x_3	$f(x_3)$		$f(x_2,x_3,x_4)$	
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

Заметим, что если аргументы расположены в порядке возрастания и нам известны необходимые значения в узлах производных функции f, то вычисление всех находящихся в таблице значений разделенных разностей не составляет труда и в случае наличия кратных узлов.

Задача. Пусть $N > M \ge 0$ целые. Доказать:

$$\sum_{k=0}^{n}(-1)^{n-k}C_n^k\frac{(N+k)!}{(M+k)!} = \begin{cases} 0 & \text{при } N-M<n \\ n! & \text{при } N-M=n. \end{cases}$$

§3. Алгебраическая интерполяция

Общая постановка задачи интерполяции такова. На промежутке $[a,b]$ задана система непрерывных функций $\{\varphi_k(x)\}$ $(k=0,\dots,n)$. Линейные комбинации этих функций называются обобщенными полиномами (по системе $\{\varphi_k\}$). Заданы попарно различные точки $x_0,\dots,x_n$ промежутка $[a,b]$, называемые узлами[2]. Ставится задача: для произвольно заданной на $[a,b]$ функции $f(x)$ построить такой "полином" $q=\sum_{k=0}^{n} a_k\varphi_k$, который удовлетворял бы равенствам $q(x_k)=f(x_k)$ $(k=0,\dots,n)$.

Две ипостаси задачи: 1) приближение функции более простыми, 2) функция f известна нам лишь в конечном числе точек, а нас интересуют ее значения в других точках.

Будет ли поставленная задача разрешима?

Определение. Система функций $\{\varphi_k\}_{k=0}^{n}$ называется *чебышевской* на $[a,b]$, если любой нетривиальный "полином" $q=\sum_{k=0}^{n} a_k\varphi_k$ (т.е. такой, что хоть один из его коэффициентов a_k отличен от нуля) имеет на $[a,b]$ не более n корней.

Теорема 1. Для того чтобы система $\{\varphi_k\}_{k=0}^{n}$ была чебышевской, необходимо и достаточно, чтобы для любого набора попарно различных точек $x_0,\dots,x_n\in[a,b]$ определитель

$$\Delta(x_0,\dots,x_n)=\begin{vmatrix}\varphi_0(x_0) & \dots & \varphi_n(x_0)\\ \dots & \dots & \dots\\ \varphi_0(x_n) & \dots & \varphi_n(x_n)\end{vmatrix}$$

был отличен от нуля.

Доказательство. Докажем, что для того чтобы наша система *не была* чебышевской, необходимо и достаточно, чтобы нашлись такие попарно различные точки x_k, что $\Delta(x_0,\dots,x_n)=0$. Действительно, если система *не* чебышевская, то найдется нетривиальный "полином" q, который имеет по меньшей мере $n+1$ корень. Пусть $x_0,\dots,x_n$ — его корни. Тогда его коэффициенты удовлетворяют системе уравнений

$$\left.\begin{aligned} a_0\varphi_0(x_0)+\dots+a_n\varphi_n(x_0) &= 0\\ \cdot\quad\cdot\quad\cdot & \\ a_0\varphi_0(x_n)+\dots+a_n\varphi_n(x_n) &= 0\end{aligned}\right\}. \tag{1}$$

Итак, система однородных уравнений (1) имеет ненулевое решение и, значит, ее определитель $\Delta(x_0,\dots,x_n)=0$. Обратно, пусть нашлись такие попарно различные точки $x_0,\dots,x_n$, что $\Delta(x_0,\dots,x_n)=0$. Тогда система однородных уравнений (1) имеет ненулевое решение, и компоненты этого решения будут коэффициентами "полинома", который имеет все точки $x_0,\dots,x_n$ своими корнями. ■

Теорема 2. Для того чтобы для любой системы узлов $x_0,\dots,x_n\in[a,b]$ интерполяционная задача $q(x_k)=f_k$ была однозначно разрешима, необходимо и достаточно, чтобы система φ_k была чебышевской.

---

[2]Когда точки некоторой системы будут называться узлами, то всегда будет иметься в виду, что они попарно различны.

Доказательство. Если искать интерполяционный полином в форме $q(x) = \sum_{k=0}^{n} a_k \varphi_k(x)$, то требования $q(x_k) = f_k$ дадут систему $(n+1)$ линейных уравнений относительно его $(n+1)$ коэффициента с определителем $\Delta(x_0, \ldots, n)$. Необходимым и достаточным условием однозначной разрешимости этой системы является отличие от нуля определителя. Поэтому остается сослаться на предыдущую теорему. ■

Поскольку любой полином $P_n \in \mathbb{P}_n$ имеет не более n попарно различных корней, то система $\{1, x, \ldots, x^n\}$ является чебышевской на любом промежутке $[a, b]$, и из теоремы 2 немедленно вытекает

Следствие. Каковы бы ни были узлы $x_0, \ldots, x_n$ и числа $f_0, \ldots, f_n$ существует и притом единственный полином $P_n \in \mathbb{P}_n$, такой что при $k = 0, \ldots, n$ $P_n(x_k) = f_k$.

Если f_k — это значения в узлах некоторой функции $f(x)$, то P_n называется интерполяционным полиномом функции f.

Дальше будем рассматривать задачу построения алгебраического интерполяционного полинома.

Обозначим через $l_k(x)$ полином, решающий интерполяционную задачу[3]

$$l_k(x_j) = \delta_{kj}. \tag{2}$$

Легко видеть, что тогда полином $P_n(x) = \sum_{k=0}^{n} l_k(x) f_k$ удовлетворяет равенствам $P_n(x_j) = f_j$. Поэтому интерполяционный полином функции $f(x)$ может быть записан в виде

$$P_n(x) = \sum_{k=0}^{n} l_k(x) f(x_k). \tag{3}$$

Эта формула называется *интерполяционной формулой Лагранжа*, а полиномы $l_k(x)$ — фундаментальными полиномами интерполяции или полиномами влияния Лагранжа. Для этих полиномов нетрудно указать явное представление:

$$l_k(x) = \frac{(x - x_0) \ldots (x - x_{k-1})(x - x_{k+1}) \ldots (x - x_n)}{(x_k - x_0) \ldots (x_k - x_{k-1})(x_k - x_{k+1} \ldots (x_k - x_n)} = \frac{\omega(x)}{(x - x_k)\omega'(x_k)},$$

где $\omega(x) = (x - x_0) \ldots (x - x_n)$.

Другое представление интерполяционного полинома принадлежит Ньютону:

$$P_n(x) = f(x_0) + (x - x_0) f(x_0, x_1) + \cdots + (x - x_0) \ldots (x - x_{n-1}) f(x_0, \ldots, x_n).$$

Это — интерполяционная формула Ньютона. Полином P_n использовался при доказательстве свойства $\langle 6 \rangle$ разделенных разностей, там и было показано, что он удовлетворяет равенствам $P_n(x_k) = f(x_k)$.

Сравнение этих двух формул. Формула Ньютона удобнее для вычислений, в частности, тем, что легко добавлять новые узлы и вопрос о числе узлов

[3] δ_{kj} — символ Кронекера.

можно решать в процессе вычислений. Формула Лагранжа удобна в теоретических вопросах интерполяции. В практических применениях она удобнее, если нужно интерполировать много функций по одной и той же системе узлов.

Рассмотрим случай равноотстоящих узлов: $x_k = x_0 + kh$. Поскольку в основе формул будет лежать формула Ньютона, можно указывать лишь порядок привлечения узлов.

1. Пусть значения функции $f(x)$ известны в узлах $x_0, x_1, \ldots$ и точка x, в которой нам нужно найти ее значение, лежит вблизи x_0. Тогда привлекая узлы в порядке $x_0, x_1, \ldots$, делая замену $x = x_0 + th$ и учитывая, что $x - x_k = h(t-k)$ и $f(x_0, \ldots, x_k) = \frac{1}{h^k k!}\Delta^k f_0$, имеем

$$P(x_0 + th) = f_0 + t\Delta f_0 + \frac{t(t-1)}{2!}\Delta^2 f_0 + \frac{t(t-1)(t-2)}{3!}\Delta^3 f_0 + \ldots.$$

Это — формула Ньютона для начала таблицы.

2. Пусть значения функции $f(x)$ известны в точках $\ldots, x_{n-2}, x_{n-1}, x_n$ и точка интерполяции x лежит вблизи точки x_n. Естественный порядок привлечения узлов $x_n, x_{n-1}, x_{n-2}, \ldots$. Так же, как в предыдущем случае, учитывая при этом, что при замене разделенной разности конечной аргументом у конечной разности будет наименьший из аргументов разделенной, имеем:

$$P(x_n + th) = f_n + t\Delta f_{n-1} + \frac{t(t+1)}{2!}\Delta^2 f_{n-2} + \frac{t(t+1)(t+2)}{3!}\Delta^3 f_{n-3} + \ldots.$$

Это — формула Ньютона для конца таблицы.

3. Пусть значения функции $f(x)$ известны в узлах $\ldots, x_{-2}, x_{-1}, x_0, x_1, \ldots$ и точка интерполяции x лежит между x_0 и x_1. Будем привлекать узлы интерполяции в порядке $x_0, x_1, x_{-1}, x_2, x_{-2}, \ldots$. Тогда так же как в двух предыдущих случаях получим

$$P(x_0 + th) = f_0 + t\Delta f_0 + \frac{t(t-1)}{2!}\Delta^2 f_{-1} + \frac{(t+1)t(t-1)}{3!}\Delta^3 f_{-1} + \\ + \frac{(t+1)t(t-1)(t-2)}{4!}\Delta^4 f_{-2} + \frac{(t+2)(t+1)t(t-1)(t-2)}{5!}\Delta^5 f_{-2} + \ldots.$$

Это — формула Ньютона - Гаусса для середины таблицы.

Задача 1. Показать, что если функция $g(x)$ такова, что на промежутке $[a,b]$ $g^{(n)}(x) > 0$, то на этом промежутке система функций $1, x, x^2, \ldots, x^{n-1}, g(x)$ чебышевская.

Задача 2. Показать, что система функций $1, e^x, e^{2x}, \ldots, e^{nx}$ чебышевская на любом промежутке.

§4 Погрешность интерполяции

Пусть на $[a,b]$ заданы узлы $x_0, \ldots, x_n$. Для функции $f \in C[a,b]$ ее интерполяционный полином, построенный по этим узлам, условимся обозначать через $Q_n f$, а значение этого полинома в точке x через $Q_n(f;x)$. Отметим очевидные свойства:

1. $Q_n f \in \mathbb{P}_n$;

2. $Q_n(\alpha_1 f_1 + \alpha_1 f_2) = \alpha_1 Q_n f_1 + \alpha_2 Q_n f_2$ (линейность);
3. для любого $P_n \in \mathbb{P}_n \quad Q_n P_n = P_n$.

Разность $R_n(f;x) = f(x) - Q_n(f;x)$ есть погрешность (остаточный член) интерполяции.

Теорема 1. Если функция f $n+1$ раз непрерывно дифференцируема на $[a,b]$ ($f \in C^{(n+1)}[a,b]$), то для каждой точки $x \in [a,b]$ найдется такая точка $\xi \in (a,b)$, что

$$R_n(f;x) = \frac{\omega(x)}{(n+1)!} f^{(n+1)}(\xi), \quad \omega(x) = (x-x_0)\dots(x-x_n). \tag{1}$$

Доказательство. Если x совпадает с одним из узлов, то (1) очевидно — левая и правая части равны нулю. При x отличным от всех узлов воспользуемся свойством $\langle 5 \rangle$ разделенных разностей для узлов $x_0, \dots, x_n, x$. Имеем:

$$f(x) = Q_n(f;x) + \omega(x) f(x_0, \dots, x_n, x),$$

и для завершения доказательства остается воспользоваться свойством $\langle 6 \rangle$ разделенных разностей. ■

Остановимся на двух задачах выбора узлов интерполяции. Пусть $U \subset C[a,b]$ — некоторый класс непрерывных функций. При заданных узлах интерполяции введем обозначения:

$$R_n(U;x) = \sup_{f \in U} |R_n(f;x)|, \quad R_n(U) = \sup_{f \in U} \|R_n f\|_C -$$

погрешности (остатки) интерполяции на классе U. Для класса функций

$$KC^{(n+1)} = \{\, f \in C^{(n+1)} \;\mid\; \|f^{(n+1)}\|_C \le K \,\} \quad (K > 0)$$

эти погрешности легко вычисляются:

Теорема 2. Справедливы равенства:

$$R_n(KC^{(n+1)};x) = \frac{K|\omega(x)|}{(n+1)!}, \quad R_n(KC^{(n+1)}) = \frac{K\|\omega\|_C}{(n+1)!}.$$

Доказательство. То, что левая часть первого из этих равенств не превосходят правой, сразу же следует из теоремы 1. Докажем обратное неравенство. Рассмотрим функцию $f_0(x) = \frac{K}{(n+1)!}\omega(x)$. Так как $f_0^{(n+1)} \equiv K$, то $f_0 \in KC^{(n+1)}$. Очевидно, что $Q_n f_0 \equiv 0$, и потому $|R_n(f_0;x)| = |f_0(x)| = \frac{K|\omega(x)|}{(n+1)!}$, что и завершает доказательство первого равенства. Второе следует из первого ввиду очевидного тождества $R_n(KC^{(n+1)}) = \sup_x R_n(KC^{(n+1)};x)$. ■

При заданном классе U (или классе U и точке x) величина $R_n(U)$ (соответственно $R_n(U;x)$) есть функция узлов, и можно ставить задачу о минимизации этой функций. Те узлы, на которых функция $R_n(U)$ достигает минимального значения, называются *оптимальными* узлами для класса U.

Задача 1. Пусть на промежутке $[a,b]$ задано большое количество узлов $N > n+1$, в которых нам известны значения каких-то функций. Нас интересуют значения этих функций в некоторой точке x, отличной от всех узлов. Для вычисления этих значений мы хотим использовать интерполяцию по $n+1$ узлу. Задача состоит в таком выборе этих узлов из числа данных, чтобы для некоторого заданного класса функций U величина $R_n(U;x)$ была минимальной. Эта задача легко решается для класса $U = KC^{(n+1)}$. Действительно, в полученном в теореме 2 представлении остатка от узлов зависит только множитель $|\omega(x)|$, его-то и нужно минимизировать, а для этого следует выбрать из наших N узлов ближайшие к точке x. Заметим, что этот принцип учитывался для выбора порядка привлечения узлов при построении интерполяционных формул с равноотстоящими узлами в §3.

Задача 2 — это задача о выборе оптимальных узлов для класса U, т.е. таких узлов, для которых величина $R_n(U)$ минимальна. Решать эту задачу будем для класса функций $KC^{(n+1)}[-1,1]$.

Теорема 3. Оптимальными узлами для класса функций $KC^{(n+1)}[-1,1]$ являются корни полинома Чебышева $T_{n+1}(x)$: $x_k = \cos\frac{(2k+1)\pi}{2n+2}$ $(k = 0,\dots,n)$, называемые узлами Чебышева. Для этих узлов

$$R_n(KC^{(n+1)}[-1,1]) = \frac{K}{2^n(n+1)!}. \tag{2}$$

Доказательство. Для узлов Чебышева $\omega(x) = \tilde{T}_{n+1}(x)$, и (2) немедленно следует из теоремы 2 и равенства $\|\tilde{T}_{n+1}\|_C = \frac{1}{2^n}$. Поскольку полином Чебышева наименее уклоняется от нуля и для любых узлов $\omega(x)$ есть полином степени $n+1$ со старшим коэффициентом, равным 1, то всегда $\|\omega\|_C \geq \|\tilde{T}_{n+1}\|_C = \frac{1}{2^n}$, и потому для любых узлов $R_n(KC^{(n+1}) \geq \frac{K}{2^n(n+1)!}$ ■

Замечание. В случае произвольного промежутка $[a,b]$ оптимальные узлы для класса $KC^{(n+1)}[a,b]$ можно получить, если с помощью линейной замены переменной промежуток $[a,b]$ свести к промежутку $[-1,1]$ — образы узлов Чебышева и будут оптимальными узлами для этого класса, и для этих узлов

$$R_n(KC^{(n+1)}[a,b]) = \left(\frac{b-a)}{2}\right)^{n+1} \frac{K}{2^n(n+1)!}.$$

Другой подход к оценке погрешности интерполяции связан с понятием функций и постоянных Лебега.

Определение. Функцией Лебега узлов $x_0,\dots,x_n$ называется

$$\lambda_{n+1}(x) = \sum_{k=0}^{n} |l_k(x)|.$$

Здесь $l_k(x)$ — фундаментальные полиномы интерполяции. Постоянной Лебега узлов называется $\lambda_{n+1} = \max_{x\in[a,b]} \lambda_{n+1}(x)$.

Лемма. Для любой $f \in C[a,b]$ выполняются неравенства

$$|Q_n(f;x)| \le \lambda_{n+1}(x)\|f\|_C, \quad \|Q_n f\|_C \le \lambda_{n+1}\|f\|_C. \tag{3}$$

Доказательство. Первое из неравенств (3) есть очевидное следствие интерполяционной формулы Лагранжа. Для доказательства второго достаточно взять максимум по x от левой и правой части первого. ■

Теорема 4. Для любой $f \in C[a,b]$ выполняются неравенства

$$|R_n(f;x)| \le \big(\lambda_{n+1}(x)+1\big)E_n(f), \quad \|R_n(f)\| \le (\lambda_{n+1}+1)E_n(f).$$

Доказательство. Пусть $P_n \in \mathbb{P}_n$ — полином наилучшего приближения функции f. Тогда имеем

$$|R_n(f;x)| \le |f(x)-P_n(x)| + |P_n(x)-Q_n(f;x)| = |f(x)-P_n(x)| + \\ + |Q_n(P_n-f;x)| \le E_n(f) + \lambda_{n+1}(x)\|P_n-f\|_C = \big(\lambda_{n+1}(x)+1\big)E_n(f).$$

Этим доказано первое неравенство. Второе получается из первого, если в левой и правой его части перейти к максимумам по x. ■

С функцией и постоянной Лебега связаны вопросы сходимости интерполяционных полиномов к функции. Будем говорить, что для промежутка $[a,b]$ задан интерполяционный процесс, если при каждом n на этом промежутке заданы узлы $x_0^n, \dots, x_n^n$. Тогда $Q_n(f;x)$ — интерполяционный полином функции f, построенный по этим узлам. Говорят, что интерполяционный процесс для функции f сходится в точке x (сходится равномерно на $[a,b]$), если соответственно при $n \to \infty$ будет $Q_n(f;x) \to f(x)$ или $\|f - Q_n f\|_C \to 0$ (т.е. $Q_n f$ сходятся к f равномерно на $[a,b]$). Из теоремы 4 сразу же вытекает

Следствие. Если для некоторой непрерывной функции f выполняется соотношение $\lambda_{n+1}(x)E_n(f) \to 0$, то интерполяционный процесс для этой функции сходится в точке x. Если же $\lambda_{n+1}E_n(f) \to 0$, то интерполяционный процесс для нее сходится равномерно.

Известно, что для любого интерполяционного процесса $\lambda_{n+1} \to \infty$. С этим связана теорема Фабера (оба этих утверждения оставляем без доказательства):

Теорема (Фабер). Для любого интерполяционного процесса найдется такая непрерывная функция, для которой этот процесс не сходится равномерно.

С функцией и постоянной Лебега связана еще оценка погрешности в интерполяционном полиноме, возникающая вследствие неточного вычисления значений функции в узлах. Пусть при вычислении значений функции $f(x_k)$ мы допустили ошибки ε_k, для которых нам известны лишь оценки $|\varepsilon_k| \le \varepsilon$. Тогда вместо интерполяционного полинома $Q_n(f;x)$ мы получим

$$\overline{Q_n(f;x)} = \sum_{k=0}^{n} l_k(x)\big(f(x_k)+\varepsilon_k\big),$$

так что

$$|Q_n(f;x) - \overline{Q_n(f;x)}| = \left|\sum_{k=0}^{n} l_k(x)\varepsilon_k\right| \le \lambda_{n+1}(x)\varepsilon$$

и $\|Q_n f - \overline{Q_n f}\| \le \lambda_{n+1}\varepsilon$. Обе эти оценки являются точными в том смысле, что если при всех k $|\varepsilon_k| = \varepsilon$ и ε_k имеют соответствующим образом выбранные знаки, то это неравенства обращаются в равенства.

Простейшими узлами являются равноотстоящие: $x_k = a + kh$, где $h = (b-a)/n$, а $k = 0, \ldots, n$. Покажем, что эти узлы являются плохими в том отношении, что для них постоянная Лебега растет чрезвычайно быстро с ростом n.

Теорема 5. Для постоянной Лебега равноотстоящих узлов выполняется неравенство

$$\lambda_{n+1} > \frac{1}{3n}\left(\frac{3}{2}\right)^n.$$

Доказательство. Имеем $\lambda_{n+1} \ge \lambda_{n+1}(x^*)$, где $x^* = a + h/2$. Легко видеть, что

$$|l_k(x^*)| = \frac{(2n-1)!!}{2^n k!(n-k)!|2k-1|} > \frac{1}{2n}\frac{(2n-1)!!}{2^n k!(n-k)!}.$$

Отсюда

$$\lambda_{n+1} > \frac{1}{2n}\frac{(2n-1)!!}{2^n n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!} = \frac{1}{2n}\frac{(2n-1)!!}{n!}.$$

Произведя очень грубую оценку:

$$\frac{(2n-1)!!}{n!} = 1\cdot\frac{3}{2}\cdot\frac{5}{3}\cdots\frac{2n-1}{n} \ge \frac{2}{3}\left(\frac{3}{2}\right)^n,$$

придем к требуемому. ■

Доказанная в теореме оценка правильно показывает характер поведения λ_{n+1}, хотя и очень груба, что показывает следующая таблица:

n	оценка	$\lambda_{n+1}(x^*)$
10	1.92	24.6
20	55.4	7391
40	92144	$2.57\cdot 10^9$

Быстрый рост постоянной Лебега заставляет предполагать, что равномерная сходимость интерполяционного процесса по равноотстоящим узлам имеет место лишь для узкого класса функций. Действительно, как может быть показано, этот процесс в случае промежутка $[-1,1]$ не сходится равномерно для функций

$$g_p(x) = \begin{cases} x^p & \text{при } x \ge 0, \\ 0 & \text{при } x < 0 \end{cases}$$

при любом натуральном p, хотя функция g_p $p-1$ раз непрерывно дифференцируема.

Возникает вопрос, а существуют ли узлы, для которых постоянная Лебега существенно меньше, чем для равноотстоящих? Оказывается, что такими узлами являются узлы Чебышева. В случае узлов Чебышева удобнее оценивать не λ_{n+1}, а λ_n, так что узлы — корни полинома Чебышева $T_n(x)$:

$$x_k = \cos\theta_k, \qquad \theta_k = \frac{2k-1}{2n}\pi,$$

а фундаментальные полиномы интерполяции имеют вид:

$$l_k(x) = \frac{T_n(x)}{(x-x_k)T_n'(x_k)}, \quad k = 1, 2, \ldots, n.$$

Поскольку $|T_n'(x_k)| = n/\sin\theta_k$, то при $x = \cos\theta$

$$|l_k(x)| = \frac{|\cos n\theta|}{n|\cos\theta - \cos\theta_k|} \cdot \sin\theta_k.$$

Докажем несколько лемм.

Лемма 1. При $0 \le \alpha \le \pi/2$

$$\sin\alpha \ge \frac{2}{\pi}\alpha.$$

Доказательство. Ограничимся указанием, что это неравенство означает, что для промежутка $[0, \pi/2]$ график функции $\sin x$ лежит выше хорды, соединяющей начало координат с вершиной синусоиды. ■

Лемма 2. Если $0 \le x < x + h \le \pi$, то

$$\cos x - \cos(x+h) \ge \frac{2}{\pi^2}h^2.$$

Доказательство. На промежутке $[0, \pi - h]$ рассмотрим функцию $\varphi(x) = \cos x - \cos(x+h)$. Очевидно, что $\varphi(x) > 0$ и $\varphi''(x) = -\varphi(x) < 0$, так что φ не имеет точек локального минимума. В то же время

$$\varphi(0) = \varphi(\pi - h) = 1 - \cos h = 2\sin^2\frac{h}{2} \ge 2\left(\frac{h}{\pi}\right)^2.$$

Этим лемма доказана. ■

Лемма 3. $|l_k(x)| \le 2$.

Доказательство. Положим $x = \cos\theta$ ($\theta \in [0, \pi]$). Учитывая, что $\cos n\theta_k = 0$ и что при любом τ $|\sin n\tau| \le n|\sin\tau|$, имеем

$$\begin{aligned}
|l_k(x)| &= \left|\frac{\cos n\theta - \cos n\theta_k}{n(\cos\theta - \cos\theta_k)}\right| \sin\theta_k = \\
&= \left|\frac{\sin\frac{n}{2}(\theta-\theta_k)\cdot\sin\frac{n}{2}(\theta+\theta_k)}{n\sin\frac{1}{2}(\theta-\theta_k)\cdot\sin\frac{1}{2}(\theta+\theta_k)}\right| \sin\theta_k \le \\
&\le \frac{\sin\theta_k}{\sin\frac{1}{2}(\theta+\theta_k)} \le \frac{\sin\theta_k + \sin\theta}{\sin\frac{1}{2}(\theta+\theta_k)} = 2\cos\frac{1}{2}(\theta-\theta_k) \le 2,
\end{aligned}$$

что и требовалось доказать. ■

Теорема 6. Для постоянной Лебега узлов Чебышева верна оценка

$$\lambda_n \le 8 + \frac{4}{\pi}\ln n.$$

Доказательство. Для произвольной точки $x = \cos\theta \in [-1,1]$, считая $\theta_m < \theta < \theta_{m+1}$, имеем

$$\lambda_n(x) = \sum_{k=1}^{n} |l_k(x)| = \sum_{1}^{m-2} + \sum_{m-1}^{m+2} + \sum_{m+3}^{n} = S_1 + S_2 + S_3.$$

Из леммы 3 сразу же следует, что $S_2 \le 8$. Суммы S_1 и S_2 оцениваются одинаково. Оценим первую из них.

$$S_1 \le \frac{1}{n}\sum_{1}^{m-2} \frac{\sin\theta_k}{\cos\theta_k - \cos\theta}.$$

Функция $\sin u/(\cos u - \cos\theta)$ при $0 < u < \theta$ возрастает. Поэтому

$$\frac{\sin\theta_k}{\cos\theta_k - \cos\theta} \le \frac{\sin u}{\cos u - \cos\theta} \qquad u \in [\theta_k, \theta_{k+1}].$$

Интегрируя это неравенство по $[\theta_k, \theta_{k+1}]$, имеем

$$S_1 \le \frac{1}{\pi}\int_0^{\theta_{m-1}} \frac{\sin u}{\cos u - \cos\theta} du = \frac{1}{\pi}\ln\frac{1-\cos\theta}{\cos\theta_{m-1} - \cos\theta} \le$$

$$\le \frac{1}{\pi}\ln\frac{2}{\cos\theta_{m-1} - \cos\theta_m},$$

и по лемме 2 (при $h = \pi/n$)

$$S_1 \le \frac{1}{\pi}\ln n^2 = \frac{2}{\pi}\ln n.$$

Сумма S_3 допускает такую же оценку, и для завершения доказательства остается сложить полученные оценки для S_1, S_2 и S_3. ■

Полученная оценка правильно отражает характер роста постоянной Лебега для узлов Чебышева, хотя и немного завышена. Для сравнения с таблицей нижних оценок постоянных Лебега для равноотстоящих узлов (см. выше) приведем для того же числа узлов (11, 21, 41) значения постоянных Лебега узлов Чебышева:

$$\lambda_{11} = 2.489, \quad \lambda_{21} = 2.901, \quad \lambda_{41} = 3.327.$$

Из этой теоремы и теоремы Джексона (см. §1) легко получить, что интерполяционный процесс по узлам Чебышева равномерно сходится для всех

непрерывно дифференцируемых функций; в действительности такая сходимость имеет место для гораздо более широкого класса функций.

Задача 1. Доказать утверждение, содержащееся в замечании к теореме 3.

Задача 2. Показать, что в случае любых узлов при $n \geq 2$ для функции Лебега выполняется неравенство $\lambda_{n+1}(x) \geq 1$, причем знак равенства имеет место в том и только в том случае, если x совпадает с одним из узлов.

Задача 3. В тех же условиях показать, что между двумя соседними узлами функция Лебега имеет единственную точку максимума.

Задача 4. Доказать, что при $n \geq 3$ интерполяционный полином по узлам Чебышева на промежутке $[-1, 1]$ приближает функцию $\cos x$ лучше, чем отрезок ряда Тейлора той же степени.

§5 Эрмитовская интерполяция

Пусть на промежутке $[a, b]$ заданы узлы $x_0, \ldots, x_n$. Припишем каждому узлу некоторое натуральное число α_k — кратность узла. Положим $N = \sum \alpha_k - 1$. Пусть для некоторой функции f в каждом узле x_k нам известны значения ее самой и ее производных до порядка $\alpha_k - 1$ включительно. Задача эрмитовской интерполяции состоит в том, что требуется построить полином $P_N \in \mathbb{P}_N$, который при $k = 0, \ldots, n$ удовлетворял бы равенствам

$$P_N^{(j)}(x_k) = f^{(j)}(x_k) \quad j = 0, \ldots, \alpha_k - 1. \tag{1}$$

Заметим, что число условий, которые мы наложили на P_N, есть $N + 1$, т.е. столько же, сколько у него коэффициентов. Поэтому если искать этот полином с неопределенными коэффициентами, то условия (1) приведут к системе $(N + 1)$ линейных уравнений относительно $(N + 1)$ его коэффициентов.

Докажем однозначную разрешимость поставленной задачи.

Теорема 1. Каковы бы ни были числа b_k^j существует и притом единственный полином $P_N \in \mathbb{P}_N$, для которого выполняются равенства

$$P_N^{(j)}(x_k) = b_k^j \quad k = 0, \ldots, n, \quad j = 0, \ldots, \alpha_k - 1$$

Доказательство. Как уже отмечалось, задача построения такого полинома сводится к решению системы $(N + 1)$ линейных уравнений относительно $(N + 1)$ коэффициентов этого полинома. Требуется доказать, что эта система однозначно разрешима. Пусть $\tilde{P}_N$ — полином, коэффициенты которого удовлетворяют соответствующей однородной системе уравнений. Это означает выполнение равенств $\tilde{P}_N^{(j)}(x_k) = 0$ при $k = 0, \ldots, n$, $j = 0, \ldots, \alpha_k - 1$, т.е. x_k является корнем полинома $\tilde{P}_N$ кратности α_k, и с учетом кратностей этот полином степени не выше N имеет $(N + 1)$ корней. Но тогда он тождественно равен нулю, и равны нулю все его коэффициенты. Итак, наша однородная система уравнений имеет только нулевое решение. ■

Отметим частный случай поставленной задачи, когда имеется всего лишь один узел x_0 кратности α_0. Тогда $N = \alpha_0 - 1$ и, как легко видеть, P_N есть отрезок ряда Тейлора функции f:

$$P_N(x) = f(x_0) + (x - x_0)f'(x_0) + \cdots + \frac{(x - x_0)^N}{N!} f^{(N)}(x_0).$$

Эрмитовский интерполяционный полином может быть представлен в форме Ньютона. Пусть $y_0, \ldots, y_N$ — некоторая перестановка узлов $x_0, \ldots, x_n$ с повторениями, в которой каждый узел x_k встречается столько раз, какова его кратность. Располагая значениями $f^{(j)}(x_k)$ $(k = 0, 1, \ldots, n,\ j = 0, 1, \ldots, \alpha_k - 1)$, мы имеем возможность вычислить разделенную разность $f(y_0, \ldots, y_N)$. Строго говоря, когда вводились разделенные разности с кратными узлами, требовалось, чтобы производная $f^{(\alpha_k - 1)}$ была непрерывна по меньшей мере в окрестности точки x_k, а сейчас мы знаем только существование этой производной в самой точке x_k. Более того, если мы решаем интерполяционную задачу в постановке теоремы 1, то никакой функции f у нас вообще нет, хотя если считать b_k^j значением $f^{(j)}(x_k)$, обладающей непрерывными нужными производными, то $f(y_0, \ldots, y_N)$ мы можем вычислить. Чтобы разрешить эту коллизию, мы будем считать, что $f(y_0, \ldots, y_N)$ есть разделенная разность эрмитовского интерполяционного полинома, существование которого доказано в теореме 1. Для него выполняется равенство $P_N(y_0, \ldots, y_N) = f(y_0, \ldots, y_N)$, если только f достаточно гладкая функция, для которой $f^{(j)}(x_k) = b_k^j$. Так можно понимать $f(y_0, \ldots, y_N)$ (и разделенные разности низших порядков) в следующей теореме.

Теорема 2. Пусть $y_0, y_1, \ldots, y_N$ — произвольная перестановка узлов $x_0, \ldots, x_n$ с повторениями, в которой каждый узел x_k встречается столько раз, какова его кратность. Тогда эрмитовский интерполяционный полином имеет представление

$$P_N(x) = f(y_0) + (x - y_0)f(y_0, y_1) + \cdots + (x - y_0)\ldots(x - y_{N-1})f(y_0, \ldots, y_N).$$

Доказательство. Покажем, что выписанный полином P_N удовлетворяет интерполяционным условиям. Заметим, что если $y_0, \ldots, y_N$ различные узлы и $z_0, \ldots, z_N$ их произвольная перестановка, то выполняется равенство

$$\begin{aligned} f(y_0) + (x - y_0)f(y_0, y_1) + \cdots + (x - y_0)\ldots(x - y_{N-1})f(y_0, \ldots, y_N) = \\ = f(z_0) + (x - z_0)f(z_0, z_1) + \cdots + (x - z_0)\ldots(x - z_{N-1})f(z_0, \ldots, z_N), \end{aligned}$$

так как левая и правая части совпадают как интерполяционные полиномы функции f, построенные по одной и той же системе узлов. Поскольку разделенные разности суть непрерывные функции своих аргументов, то это же равенство соблюдается и при наличии кратных узлов. Поэтому при доказательстве равенства $P_N^{(j)}(x_k) = f^{(j)}(x_k)$ мы вправе считать, что $y_0 = \cdots = y_{\alpha_k - 1} = x_k$. Тогда

$$\begin{aligned} P_N(x) = f(x_k) + (x - x_k)f(x_k, x_k) + \cdots + (x - x_k)^{\alpha_k - 1}f(x_k, \ldots, x_k) + R(x) = \\ = f(x_k) + (x - x_k)f'(x_k) + \cdots + \frac{(x - x_k)^{\alpha_k - 1}}{(\alpha_k - 1)!}f^{(\alpha_k - 1)}(x_k) + R(x), \end{aligned}$$

где $R(x)$ — полином, содержащий множитель $(x - x_k)^{\alpha_k}$. Отсюда видно, что действительно при $j \le \alpha_k - 1$ будет $P_N^{(j)}(x_k) = f^{(j)}(x_k)$. ■

Погрешность эрмитовской интерполяции можно оценивать, используя следующую теорему.

Теорема 3. Пусть функция f $N+1$ раз непрерывно дифференцируема на некотором промежутке $[a,b]$, содержащем все узлы и точку интерполяции x. Тогда найдется такая точка $\xi \in (a,b)$, что

$$f(x) - P_N(x) = \frac{\Omega(x)}{(N+1)!} f^{(N+1)}(\xi).$$

Здесь $\Omega(x) = (x-x_0)^{\alpha_0} \dots (x-x_n)^{\alpha_n}$ — полином степени $N+1$.

Доказательство. Используя свойство $\langle 5 \rangle$ разделенных разностей, которое, как уже отмечалось, верно и в случае наличия кратных узлов, для узлов $y_0, \dots, y_N, x$ (узлы y_j те же, что в теореме 2), имеем

$$f(x) = P_N(x) + \Omega(x) f(y_0, \dots, y_N, x),$$

и остается воспользоваться свойством $\langle 6 \rangle$ разделенных разностей. ■

Задача 1. Показать однозначную разрешимость следующей интерполяционной задачи: $P_3(x_k) = a_k$, $P_3''(x_k) = b_k$ $(k = 0, 1)$.

Задача 2. Показать, что интерполяционная задача $P_2(-1) = a$, $P_2'(0) = b$, $P_2(1) = c$, вообще говоря, неразрешима и найти условие ее разрешимости, наложенное на числа a, b, c.

§6 Численное дифференцирование

Численное дифференцирование— это приближенное вычисление производных функции, заданной таблично.

Пусть нам известны значения функции f в узлах x_j, лежащих на промежутке $[a,b]$. Требуется найти значение производной этой функции $f^{(k)}(x)$ в некоторой точке x, которая может и совпадать с каким-нибудь из узлов. Способ решения — по узлам, в которых известно значение функции, (или части из них) строится интерполяционный полином, и за приближенное значение производной в точке x принимается значение в этой точке производной интерполяционного полинома. Если для интерполяции были выбраны узлы $x_0, \dots, x_n$ и P_n — соответствующий интерполяционный многочлен, то формула численного дифференцирования:

$$f^{(k)}(x) \approx P_n^{(k)}(x)$$

Займемся оценкой погрешности этой формулы.

Теорема. Пусть функция f $n+1$ раз непрерывно дифференцируема на промежутке $[a,b]$, содержащем узлы интерполяции и точку x, в которой вычисляется производная. Пусть $k \le n$ и пусть выполняется одно из условий: а) $x \notin [c,d]$, где $c = \min x_j$, $d = \max x_j$, б) $k = 1$ и x совпадает с одним из узлов x_j. Тогда найдется такая точка $\xi \in (a,b)$, что

$$f^{(k)}(x) - P_n^{(k)}(x) = \omega^{(k)}(x) \frac{f^{(n+1)}(\xi)}{(n+1)!}, \quad \omega(x) = (x-x_0)\dots(x-x_n).$$

Доказательство. Ввиду а) или б) $\omega^{(k)}(x) \neq 0$. Положим $A = [f^{(k)}(x) - P_n^{(k)}(x)]/\omega^{(k)}(x)$ и определим функцию $\varphi(z) = f(z) - P_n(z) - A\omega(z)$. Узлы x_j $(j = 0, \ldots, n)$ являются корнями этой функции, так что на $[c, d]$ она имеет $(n+1)$ различных корней. По теореме Ролля $\varphi^{(k)}$ имеет на (c, d) $n+1-k$ корней, не совпадающих с точкой x. Последнее следует из того, что в случае а) $x \notin [c, d]$, а в случае б) — из того, что по теореме Ролля существует корень первой производной функции, лежащий строго между корнями самой этой функции. Итак, точка x есть корень $\varphi^{(k)}$, отличный от сосчитанных ранее $n+1-k$ корней, так что на $[a, b]$ $\varphi^{(k)}$ имеет не менее $n+2-k$ корней. Продолжая применять теорему Ролля, получим, что $\varphi^{(n+1)}(z) = f^{(n+1)}(z) - A(n+1)!$ имеет на (a, b) хотя бы один корень ξ. Остается приравнять $\varphi^{(n+1)}(\xi)$ нулю. ■

Построим некоторые конкретные формулы численного дифференцирования в случае равноотстоящих узлов и точки дифференцирования, совпадающей с одним из узлов. При этом естественно использовать интерполяционные формулы с конечными разностями, построенные в §3. В этих формулах делалась замена переменной $x = x_0 + th$ (или $x = x_n + th$) и следует иметь в виду, что $\frac{d}{dx} = \frac{1}{h}\frac{d}{dt}$.

Дифференцируя формулу Ньютона для начала таблицы, имеем

$$P'(x) = \frac{1}{h}\frac{d}{dt}P(x_0 + th) = \frac{1}{h}\left[\Delta f_0 + \frac{2t-1}{2}\Delta^2 f_0 + \ldots\right].$$

Полагая в этой формуле $t = 0$, сохраняя в квадратных скобках лишь одно или два слагаемых и используя для остаточно члена R доказанную теорему, имеем для $f'(x_0)$ следующие формулы:

$$f'(x_0) = \frac{1}{h}\Delta f_0 + R = \frac{f_1 - f_0}{h} - \frac{h}{2}f''(\xi),$$
$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_0\right] + R = \frac{-3f_0 + 4f_1 - f_2}{2h} + \frac{h^2}{3}f'''(\xi). \quad (1)$$

Точно так же, дифференцируя формулу Ньютона для конца таблицы, можно получить:

$$f'(x_n) = \frac{f_n - f_{n-1}}{h} + \frac{h}{2}f''(\xi), \quad f'(x_n) = \frac{3f_n - 4f_{n-1} + f_{n-2}}{2h} + \frac{h^2}{3}f'''(\xi).$$

Дифференцирование формулы Ньютона - Гаусса дает:

$$P'(x) = \frac{1}{h}\frac{d}{dt}P_(x_0 + th) = \frac{1}{h}\left[\Delta f_0 + \frac{2t-1}{2}\Delta^2 f_{-1} + \ldots\right],$$

$$P''(x) = \frac{1}{h^2}\frac{d^2}{dt^2}P(x_0 + th) = \frac{1}{h^2}[\Delta^2 f_{-1} + \ldots].$$

Сохраняя в квадратных скобках выписанные члены и используя в случае первой производной теорему о представлении остаточного члена, имеем

$$f'(x_0) = \frac{1}{h}\left[\Delta f_0 - \frac{1}{2}\Delta^2 f_{-1}\right] + R = \frac{f_1 - f_{-1}}{2h} - \frac{h^2}{6}f'''(\xi), \quad (2)$$

$$f''(x_0) = \frac{1}{h^2}\Delta^2 f_{-1} + R = \frac{f_1 - 2f_0 + f_{-1}}{h^2} + R. \tag{3}$$

Интересно сравнить формулу (2) с первой из формул (1). В правых частях той и другой используются два значения функции f, но формула (2) имеет второй порядок точности относительно h, а первая из формул (1) лишь первый.

Условия теоремы об остаточном члене не выполнены в случае формулы (3), и для получения представления остатка в этом случае мы используем другой прием. Предполагая функцию f четырежды непрерывно дифференцируемой, напишем для нее формулы Тейлора, в которых значения всех производных, кроме последних, вычисляются в точке x_0:

$$f_1 = f_0 + hf' + \frac{h^2}{2}f'' + \frac{h^3}{6}f''' + \frac{h^4}{24}f^{IV}(\xi_1),$$

$$f_{-1} = f_0 - hf' + \frac{h^2}{2}f'' - \frac{h^3}{6}f''' + \frac{h^4}{24}f^{IV}(\xi_2).$$

Вычитая из суммы этих разложений $2f_0$ и поделив на h^2, придем к равенству

$$\frac{f_1 - 2f_0 + f_{-1}}{h^2} = f''(x_0) + \frac{h^2}{24}[f^{(IV)}(\xi_1) + f^{(IV)}(\xi_2)].$$

Заметив, что между точками ξ_1 и ξ_2 найдется такая точка ξ, что $f^{(IV)}(\xi_1) + f^{(IV)}(\xi_2) = 2f^{(IV)}(\xi)$, окончательно получим:

$$f''(x_0) = \frac{f_1 - 2f_0 + f_{-1}}{h^2} - \frac{h^2}{12}f^{(IV)}(\xi). \tag{4}$$

Остановимся теперь на влиянии ошибок, допущенных в значениях функции, на полученные в результате численного дифференцирования результаты на примере формулы (2).

Если для окрестности точки x_0 нам известна оценка $|f'''(x)| \le M$ и известно, что при использовании формулы (2) погрешности в значениях f не превосходят некоторого ε, то суммарная погрешность в значении $f'(x_0)$ оценивается величиной $\varepsilon/h + Mh^2/6$. При малых h влияние ошибок в значениях функции оказывается чрезвычайно большим. Если у нас есть возможность выбора шага h, то целесообразно находить его из условия минимума приведенной оценки погрешности. Таким образом нам следует найти точку минимума функции $\varphi(h) = \varepsilon/h + Mh^2/6$. Приравнивая нулю производную этой функции, легко находим эту точку и оценку E погрешности при таком выборе шага:

$$h_0 = \sqrt[3]{\frac{3\varepsilon}{M}}, \qquad E = \frac{1}{2}\sqrt[3]{9M\varepsilon^2}.$$

Заметим, что принципиально невозможно получить значение производной с погрешностью того же порядка, что в значениях функции.

Задача 1. Дифференцированием линейного интерполяционного полинома легко получается формула

$$f'(x) = \frac{f(x_0 + h) - f(x_0)}{h} + R(f; x). \tag{5}$$

Получить в случае $f \in C^{(2)}$ представление остатка $R(f;x)$:

$$R(f;x) = \frac{1}{h}\int_{x_0}^{x_0+h} K(x,t)f''(t)dt, \quad K(x,t) = \begin{cases} t - x_0 & \text{при } t < x, \\ t - x_0 - h & \text{при } t > x. \end{cases}$$

Задача 2. Показать, что при $x_0 < x < x_0 + h$ найдется такая функция $f \in C^{(2)}$, для которой не существует такой точки $\xi \in (x_0, x_0 + h)$, что в формуле (5)

$$R(f) = \frac{\omega'(x)}{2!}f''(\xi), \quad \omega(x) = (x - x_0)(x - x_0 - h).$$

Задача 3. Определить наилучший шаг h в формуле (4) при наличии ошибок округления в значениях функции.

§7 Тригонометрическая интерполяция. Дискретное преобразование Фурье.

Периодические функции естественно приближать периодическими. Простейшими 2π-периодическими функциями являются тригонометрические полиномы:

$$T_n(x) = a_0 + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx).$$

Если хоть один из коэффициентов a_n или b_n отличен от нуля, то n называется порядоком полинома T_n. Множество тригонометрических полиномов порядка не выше n обозначим через $\mathbb{T}_n$. Основные свойства тригонометрических полиномов:

1) Если $T_n, U_n \in \mathbb{T}_n$, то и $\alpha N_n + \beta U_n \in \mathbb{T}_n$ (линейность);

2) выражения для коэффициентов ($k = 1, 2, \dots$):

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} T_n(x)dx, a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} T_n(x)\cos kx\,dx, b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} T_n(x)\sin kx\,dx;$$

3) если $T_n \in \mathbb{T}_n$ есть четная функция, то все $b_k = 0$, а если нечетная, то все $a_k = 0$.

Если T_n имеет корень x^*, то он имеет бесконечно много корней, таковыми являются $x^* + 2j\pi$. Такие корни называются эквивалентными.

Теорема 1. Отличный от тождественного нуля $T_n \in \mathbb{T}_n$ имеет не более $2n$ попарно неэквивалентных корней.

Доказательство. Используя формулы Эйлера

$$\cos kx = \frac{e^{ikx} + e^{-ikx}}{2}, \quad \sin kx = \frac{e^{ikx} - e^{-ikx}}{2i},$$

приведем T_n к виду

$$T_n(x) = \sum_{k=-n}^{n} c_k e^{ikx} = e^{-inx}\sum_{k=0}^{2n} d_k e^{ikx} = e^{-inx}P_{2n}(z),$$

где c_k и $d_k = c_{k-n}$ — комплексные, вообще говоря, коэффициенты, $P_{2n}(z) = \sum_{k=0}^{2n} d_k z^k$ — полином степени не выше $2n$ и $z = e^{ix}$. Если x_0 — корень T_n, то $z_0 = e^{ix_0}$ — корень P_{2n}, если x_0 и x_1 — неэквивалентные корни T_n, то z_0 и z_1 — различные корни P_{2n}, так что P_{2n} имеет не менее различных корней, чем T_n попарно неэквивалентных. Но P_{2n} имеет не более $2n$ различных корней. ■

Следствие 1. Система функций $1, \cos x, \sin x, \ldots, \cos nx, \sin nx$ чебышевская на любом промежутке $[a, b]$, если только $b < a + 2\pi$.

Следствие 2. Каковы бы ни были точки $x_0 < x_1 < \cdots < x_{2n} < x_0 + 2\pi$ и числа $f_0, \ldots, f_{2n}$ существует и притом единственный $T_n \in \mathbb{T}_n$, такой что $T_n(x_k) = f_k$ при $k = 0, \ldots, 2n$.

Рассмотрим случай равноотстоящих узлов $x_j = jh$ $(j = 0, \ldots, 2n)$, $h = 2\pi/(2n+1)$.

Лемма 1. Справедливо равенство

$$\sigma_m = \sum_{j=0}^{2n} e^{imx_j} = \begin{cases} 2n+1 & \text{при } m = 0 \\ 0 & \text{при } m = 1, \ldots, 2n. \end{cases}$$

Доказательство. При $m = 0$ равенство очевидно. Если $1 \le m \le 2n$, то $\sigma_m = 1 + q + \cdots + q^{2n}$, где $q = e^{imh} \ne 1$, и потому $\sigma_m = (1 - q^{2n+1})/(1 - q) = 0$, так как $q^{2n+1} = e^{2m\pi i} = 1$. ■

Следствие . Справедливы равенства

$$C_m = \sum_{j=0}^{2n} \cos mx_j = \begin{cases} 2n+1 & \text{при } m = 0, \\ 0 & \text{при } m = 1, \ldots, 2n, \end{cases}$$

$$S_m = \sum_{j=0}^{2n} \sin mx_j = 0, \quad m = 0, \ldots, 2n.$$

Доказательство. Достаточно приравнять вещественную и мнимую части в равенстве, указанном в лемме 1. ■

Лемма 2. Пусть $0 \le k, l \le n$. Тогда

$$\sum_{j=0}^{2n} \cos kx_j \cos lx_j = \begin{cases} 2n+1 & \text{при } k = l = 0, \\ (2n+1)/2 & \text{при } k = l \ne 0, \\ 0 & \text{при } k \ne l, \end{cases}$$

$$\sum_{j=0}^{2n} \sin kx_j \sin lx_j = \begin{cases} (2n+1)/2 & \text{при } k = l \ne 0, \\ 0 & \text{при } k \ne l \text{ или } k = l = 0, \end{cases}$$

$$\sum_{j=0}^{2n} \cos kx_j \sin lx_j = 0.$$

Доказательство. Считая для определенности $k \geq l$ и используя формулу

$$\cos kx_j \cos lx_j = \frac{1}{2}[\cos(k+l)x_j + \cos(k-l)x_j],$$

имеем

$$\sum_{j=0}^{2n} \cos kx_j \cos lx_j = \frac{1}{2}[C_{k+l} + C_{k-l}].$$

Совершенно аналогично

$$\sum_{j=0}^{2n} \sin kx_j \sin lx_j = \frac{1}{2}[C_{k-l} - C_{k+l}],$$

$$\sum_{j=0}^{2n} \cos kx_j \sin lx_j = \frac{1}{2}[S_{k+l} \pm S_{|k-l|}].$$

Из этих равенств и следствия из леммы 1 легко следуют доказываемые. ■

Теорема 2. Коэффициенты полинома $T_n \in \mathbb{T}_n$, решающего интерполяционную задачу

$$T_n(x_j) = a_0 + \sum_{k=1}^{n}(a_k \cos kx_j + b_k \sin kx_j) = f_j, \quad j = 0, \ldots, 2n, \tag{1}$$

даются формулами ($k = 1, \ldots, n$)

$$a_0 = \frac{1}{2n+1}\sum_{j=0}^{2n} f_j, \qquad a_k = \frac{2}{2n+1}\sum_{j=0}^{2n} f_j \cos kx_j,$$

$$b_k = \frac{2}{2n+1}\sum_{j=0}^{2n} f_j \sin kx_j. \tag{2}$$

Доказательство. Достаточно умножить равенства (1) на $\cos lx_j$ или $\sin lx_j$, просуммировать по j от 0 до $2n$ и воспользоваться леммой 2. ■

На формулы (1) и (2) возможна несколько другая точка зрения. Пусть $F = (f_0, \ldots, f_{2n})$ произвольный вектор. Его компоненты можно рассматривать как значения в узлах x_j некоторой функции, и по формулам (2) поставить ему в соответствие вектор $A = (a_0, a_1, b_1, \ldots, a_n, b_n)$. Компоненты вектора F восстанавливаются по компонентам вектора A согласно формулам (1). Вектор A называют *дискретным преобразованием Фурье* вектора F. Процесс построения A можно рассматривать как переразложение вектора F по новому базису в пространстве $\mathbb{R}^{2n+1}$, компоненты нового базисного вектора являются значениями в узлах функции $\cos kx$ или $\sin kx$. Существо доказанной выше леммы 2 состоит в том, что этот базис является ортогональным.

В вычислительной практике при работе с некоторым вектором часто оказывается удобнее иметь дело с его дискретным преобразованием Фурье, и это преобразование находит широкое применение.

Например, пусть требуется передать по каналу связи $2n+1$ число $f_0, \ldots, f_{2n}$. Эти числа можно рассматривать как значения в точках x_j тригонометрического полинома: $f_j = T_n(x_j)$. Коэффициенты этого полинома вычисляются по формулам (2), и иногда именно эти коэффиуциенты оказывается целесообразным передавать по каналу связи вместо чисел f_j (например, среди коэффициентов много очень маленьких, и их можно заменить нулями). Восстановление чисел f_j на другом конце канала связи по полученным коэффициентам также нетрудно.

Существуют и другие преобразования векторов, также называемые дискретным преобразованием Фурье. Остановимся на одном таком преобразовании векторов с комплексными компонентами. Для $y \in \mathbb{C}^N$ условимся писать $y = (y_0, \ldots, y_{N-1})$. В пространстве $\mathbb{C}^N$ таких векторов скалярное произведение задается формулой $(y, z) = \sum_{k=0}^{N} y_k \overline{z}_k$. Положим $h = 1/N$ и рассмотрим в $\mathbb{C}^N$ систему векторов

$$e_k = (1, e^{2\pi i(kh)}, e^{2\pi i(2kh)}, \ldots, e^{2\pi i(N-1)kh}) \quad k = 0, \ldots, N-1.$$

Эти векторы оказываются ортогональными:

$$(e_k, e_j) = N\delta_{kj},$$

(доказательство по существу совпадает с доказательством леммы 1) и потому образуют базис в $\mathbb{C}^N$, так что любой вектор y может быть разложен по этому базису:

$$y = \sum_{j=0}^{N-1} a_j e_j \qquad \left(y_k = \sum_{j=0}^{N-1} a_j e^{2\pi i(kjh)} \right).$$

Ввиду ортогональности базиса e_k коэффициенты a_j легко находятся:

$$a_j = \frac{(y, e_j)}{(e_j, e_j)} = \frac{1}{N} \sum_{k=0}^{N-1} y_k e^{-2\pi i(kjh)},$$

что можно переписать еще в виде

$$a_{N-j} = \frac{1}{N} \sum_{k=0}^{N-1} y_k e^{2\pi i(kjh)}, \quad j = 1, \ldots, N.$$

Вектор $(a_0, \ldots, a_{N-1})$ называется дискретным преобразованием Фурье (ДПФ) вектора y.

Вычисление преобразования Фурье вектора y и восстановление этого вектора по его преобразованию Фурье (нахождение компонент y_k) осуществляется по одинаковым (с точностью до множителя $1/N$) формулам и требует вычисления сумм вида

$$y_k = \sum_{j=0}^{N-1} a_j e^{2\pi i \frac{kj}{N}}.$$

Если мы располагаем таблицей соответствующих степеней $e^{2\pi i\frac{l}{N}}$, то прямое вычисление всех y_k требует произвести N^2 умножений. Число арифметических действий можно существенно сократить, если N есть произведение нескольких множителей. Ограничимся случаем, когда $N = 2^n$. Тогда полагая $N_1 = N/2 = 2^{n-1}$ и $z = e^{2\pi i/N}$, имеем

$$y_k = \sum_{m=0}^{N_1-1} a_{2m} z^{2mk} + \sum_{m=0}^{N_1-1} a_{2m+1} z^{(2m+1)k} = \\ = \sum_{m=0}^{N_1-1} a'_m z_1^{mk} + z^k \sum_{m=0}^{N_1-1} a''_m z_1^{mk} = y'_k + z^k y''_k.$$

Здесь $z_1 = z^2 = e^{2\pi i/N_1}$, $a'_m = a_{2m}$, $a''_m = a_{2m+1}$. Из равенства $z^{kN} = 1$ вытекает, что $y'_{N_1+k} = y'_k$ и $y''_{N_1+k} = y''_k$, так что реально требуется вычислить лишь компоненты векторов $y' = (y'_0, \ldots, y'_{N_1-1})$ и $y'' = (y''_0, \ldots, y''_{N_1-1})$. Из приведенных выше формул видно, что y' и y'' суть дискретные преобразования Фурье для $(a'_0, \ldots, a'_{N_1-1})$ и $(a''_0, \ldots, a''_{N_1-1})$ соответственно, и их можно вычислять пользуясь тем же приемом. Обозначим через q_n число умножений, которое потребно при применении такого процесса для вычисления ДПФ вектора размерности $N = 2^n$. Тогда $q_n = 2q_{n-1} + 2^n$. При $n = 0$ ($N = 1$) ДПФ "вектора" есть он сам, так что $q_0 = 0$. Это позволяет методом индукции легко доказать, что $q_n = nN = N\log_2 N$, так что число умножений по сравнению с применением прямых формул существенно сокращается. Например, при $N = 2^{10} = 1024$ будет $N^2 > 10^6$,а $Nn = 10240$, т.е. более чем в 100 раз меньше. Вычисление ДПФ с использованием приведенного приема называется быстрым преобразованием Фурье — БПФ.

Поясним формулы БПФ на примере $N = 4$. Тогда $z = e^{2\pi i/4} = i$, $z_1 = z^2 = -1$. Требуется вычислить коэффициенты разложения вектора $y = (y_0, y_1, y_2, y_3)$ по новому базису:

$$\begin{aligned} y_0 &= a_0 + a_1 + a_2 + a_3, \\ y_1 &= a_0 + a_1 z + a_2 z^2 + a_3 z^3, \\ y_2 &= a_0 + a_1 z^2 + a_2 z^4 + a_3 z^6, \\ y_3 &= a_0 + a_1 z^3 + a_2 z^6 + a_3 z^9. \end{aligned}$$

Тогда формулы БПФ принимают вид:

$$\left.\begin{aligned} y_0 &= (a_0 + a_2) + 1 \cdot (a_1 + a_3) = y'_0 + y''_0, \\ y_1 &= (a_0 + a_2 z_1) + z(a_1 + a_3 z_1) = y'_1 + z y''_1, \\ y_2 &= (a_0 + a_2) + z^2(a_1 + a_3) = y'_0 + z^2 y''_0, \\ y_3 &= (a_0 + a_2 z_1) + z^3(a_1 + a_3 z_1) = y'_1 + z^3 y''_1. \end{aligned}\right\} \tag{3}$$

$$\left.\begin{aligned} y'_0 &= a_0 + a_2, \quad & y''_0 &= a_1 + a_3, \\ y'_1 &= a_0 + a_2 z_1, \quad & y''_1 &= a_1 + a_3 z_1. \end{aligned}\right\} \tag{4}$$

При вычислениях сначала используются формулы (4), а затем (3).

Задача. Показать, что система функций $\sin x, \ldots, \sin nx$ является чебышевской на промежутке $[\varepsilon, \pi - \varepsilon]$ при любом $\varepsilon > 0$ и не является чебышевской на $[0, \pi - \varepsilon]$ и $[\varepsilon, \pi]$.

Глава 2
Приближенное вычисление интегралов

§1 Интерполяционные квадратурные формулы (ИКФ)

Основной способ приближенного вычисления интегралов — формулы механических квадратур.

Определение. *Формулой механических квадратур* или *квадратурной формулой* называется приближенная формула

$$\int_a^b f(x)dx \approx \sum_{k=1}^{n} A_k f(x_k). \tag{1}$$

Здесь *узлы* x_k и *коэффициенты (веса)* A_k не зависят от интегрируемой функции (формула — набор узлов и коэффициентов).

Все узлы считаются различными, и чаще всего $x_k \in [a,b]$, но делать такое предположение, если не оговорено противное, мы не будем.

Более общее понятие — формула механических квадратур с весом $w(x)$. Это

$$\int_a^b w(x)f(x)dx \approx \sum_{k=1}^{n} A_k f(x_k). \tag{2}$$

Применения — много интегралов от функций с одним множителем $w(x)$ и выделение стандартных особенностей. Формулу (1) можно считать частным случаем (2) при $w(x) \equiv 1$. Поэтому в этом параграфе рассматривается (2).

Будем обозначать квадратурную сумму формулы (2) через $Q_n(f)$, а погрешность формулы (остаток) через $R_n(f)$:

$$Q_n(f) = \sum_{k=1}^{n} A_k f(x_k), \quad R_n(f) = \int_a^b w(x)f(x)dx - Q_n(f).$$

Простейший способ построения (2): произвольно выбираются узлы и значение интеграла считается приближенно равным интегралу от интерполяционного полинома функции f, построенного по этим узлам. Так полученные формулы называются интерполяционно-квадратурными (ИКФ).

Определение. Формула (2) называется ИКФ, если

$$A_k = \int_a^b w(x)l_k(x)dx = \int_a^b w(x)\frac{\omega(x)dx}{(x-x_k)\omega'(x_k)}. \tag{3}$$

Здесь $\omega(x) = (x-x_1)\dots(x-x_n)$.

Таким образом, ИКФ полностью задается указанием ее узлов.

Определение. Говорят, что формула (2) имеет алгебраическую степень точности (АСТ) d, если она точна для всех алгебраических полиномов $p_d \in \mathbb{P}_d$ $(R_n(p_d) = 0)$ и существует хотя бы один полином $p_{d+1} \in \mathbb{P}_{d+1}$, для которого она не точна $(R_n(p_{d+1}) \neq 0)$.

Замечание. Очевидно, что для того чтобы (2) имела АСТ d, необходимо и достаточно, чтобы она была точна для x^j при $j = 0, \ldots, d$ и не была точна для x^{d+1}.

Теорема 1. Для того чтобы формула (2) была ИКФ необходимо и достаточно, чтобы она имела АСТ $d \geq n-1$.

Доказательство. 1) Необходимость. Произвольно взятый полином p_d представим в виде интерполчционного по узлам x_k и воспользуемся формулой (3).

2) Достаточность. Если АСТ формулы (2) $d \geq n-1$, то она точна, в частности, для $l_k(x)$, откуда формулы (3). ■

Замечание. Формула (2) может иметь АСТ $d > n-1$, но если вес $w(x)$ сохраняет на $[a,b]$ знак, то $d \leq 2n-1$ (формула не точна для $\omega^2(x)$).

Обозначим через $[A,B]$ наименьший промежуток, содержащий $[a,b]$ и все узлы x_k.

Определение. Говорят, что ФМК *имеет представление остатка в форме Лагранжа*, если существуют такое натуральное m и такая постоянная C, что для любой m раз непрерывно дифференцируемой на $[A,B]$ функции $f(x)$ найдется такая точка $\xi \in [A,B]$, что $R_n(f) = Cf^{(m)}(\xi)$.

Замечание. ФМК может и *не иметь* представления остатка в форме Лагранжа.

Теорема 2. Если ФМК имеет представление остатка в форме Лагранжа, то $m = d+1$.

Теорема 3. Если АСТ формулы (2) есть d, то для любой $f \in C[a,b]$ выполняется оценка

$$|R_n(f)| \leq \left[\int_a^b |w(x)|dx + \sum_{k=0}^{n} |A_k| \right] E_d(f), \tag{4}$$

где $E_d(f)$ — наилучшее приближение функции f полиномами степени n на промежутке $[A,B]$.

Доказательство. Вычесть из f и прибавить ее полином наилучшего приближения, для которого формула точна. ■

Будем считать, что дана последовательность квадратурных формул

$$\int_a^b w(x)f(x)dx \approx \sum_{k=1}^{n} A_k^n f(x_k^n).$$

(квадратурный процесс). АСТ формулы с номером n считаем равной d_n и сохраним для этих формул обозначения $Q_n(f)$ и $R_n(f)$.

Следствие. Если все узлы $x_k^n \in [a,b]$, $\sum_{k=1}^{n} |A_k^n| \leq C$ и $d_n \to \infty$, то для любой $f \in C[a,b]$

$$\sum_{k=1}^{n} A_k^n f(x_k^n) \to \int_a^b w(x)f(x)dx.$$

Замечание. Если $d \geq 0$, то

$$\sum_{k=0}^{n} A_k^n = \int_a^b w(x)dx,$$

и потому в случае $w(x) \geq 0$ и $A_k^n > 0$ оценку (4) можно переписать в виде

$$|R_n(f)| \leq 2 \int_a^b w(x)dx \, E_d(f).$$

Пусть $w(x) \geq 0$. Если среди коэффициентов A_k есть отрицательные, то

$$\sum |A_k| > \int_a^b w(x)dx,$$

и оценка (4) хуже, чем в случае положительных коэффициентов. Отсюда — требование положительности коэффициентов. Это существенно еще в одном отношении. Если значения $f(x_k)$ мы вычисляем с погрешностями ε_k, про которые известно лишь, что $|\varepsilon_k| \leq \varepsilon$, то вызванная этими погрешностями ошибка в квадратурной сумме оценивается через $\sum_{k=0}^{n} |A_k|\varepsilon$, причем эта оценка неулучшаема. Формулами, у которых среди коэффициентов имеются отрицательные, обычно не пользуются.

Задача 1. Пусть АСТ формулы (2) есть d. Найдется ли $p_{d+2} \in \mathbb{P}_{d+2}$, для которого она точна?

Задача 2. Пусть $w(x) \geq 0$ и k узлов формулы (2) принадлежат (a, b), а остальные лежат вне этого промежутка. Показать, что тогда АСТ $d \leq n + k - 1$.

Задача 3. Показать, что квадратурная формула

$$\int_0^1 f(x)dx \approx f(a),$$

где $a \in (0, 1)$ и $a \neq 1/2$, не имеет представления остатка в форме Лагранжа.

§2 Квадратурные формулы с постоянным весом. Формулы Котеса

Пусть квадратурную формулу

$$\int_a^b f(x)dx \approx \sum_{k=1}^{n} A_k f(x_k) = Q_n^1(f) \tag{1}$$

мы хотим использовать для вычисления интеграла по промежутку $[c, d]$. Сделав в интеграле по $y \in [c, d]$ линейную замену переменной интегрирования $y = \frac{d-c}{b-a}(x - a) + c$ и применив для вычисления полученного интеграла по $[a, b]$ квадратурную формулу (1), мы придем к приближенному равенству (квадратурной формуле)

$$\int_c^d g(y)dy \approx \sum_{k=1}^{n} B_k g(y_k) = Q_n^2(g), \tag{2}$$

где

$$B_k = \frac{d-c}{b-a} A_k, \qquad y_k = \frac{d-c}{b-a}(x_k - a) + c. \tag{3}$$

Определение. ФМК (2) называется *подобной* формуле (1), если узлы и коэффициенты этих формул связаны равенствами (3)

Отметим основные свойства ФМК с постоянным весом и, в частности, свойства подобных формул.

$\langle 1 \rangle$ Если формула (1) точна для постоянных (АСТ≥ 0), то $\sum A_k = b - a$.

$\langle 2 \rangle$ Если формула (2) подобна (1), то и (1) подобна (2).

$\langle 3 \rangle$ АСТ подобных формул совпадают.

Свойства $\langle 1 \rangle$ - $\langle 3 \rangle$ очевидны.

$\langle 4 \rangle$ Если одна из подобных формул есть ИКФ, то и другая тоже.

Это свойство немедленно следует из $\langle 3 \rangle$ и теоремы 1 предыдущего параграфа. Таким образом, если *узлы* интерполяционных квадратурных формул (1) и (2) связаны формулой (3), то эти формулы подобны — соотношения (3) для коэффициентов выполняются автоматически.

$\langle 5 \rangle$ Если (1) есть ИКФ и все ее узлы расположены симметрично (при всех k $x_k + x_{n+1-k} = a + b$), то $A_k = A_{n+1-k}$.

Предыдущее свойство позволяет доказывать это лишь для промежутка $[-1, 1]$, а в этом случае достаточно сослаться на очевидное равенство для фундаментальных полиномов интерполяции в случае симметрично расположенных узлов: $l_k(x) = l_{n+1-k}(-x)$.

$\langle 6 \rangle$ Если узлы ИКФ расположены симметрично, то ее АСТ есть нечетное число.

Действительно, сводя задачу опять к случаю промежутка $[-1, 1]$ и используя предыдущее свойство, легко заметить, что наша формула точна для всех нечетных полиномов.

$\langle 7 \rangle$ Если ФМК (1) имеет представление остатка в форме Лагранжа: $R_n^1(f) = C_1 f^{(m)}(\xi)$ ($f \in C^{(m)}[A, B]$, $\xi \in [A, B]$), то и (2) имеет такое представление: $R_n^2(g) = C_2 g^{(m)}(\eta)$, где $\eta \in [C, D]$ и

$$C_2 = \left(\frac{d-c}{b-a}\right)^{m+1} C_1.$$

Действительно, полагая $f(x) = g\left(\frac{d-c}{b-a}(x-a) + c\right)$, имеем

$$\int_c^d g(y)dy = \frac{d-c}{b-a}\int_a^b f(x)dx = \frac{d-c}{b-a}\left[Q_n^1(f) + R_n^1(f)\right] =$$
$$= Q_n^2(g) + \frac{d-c}{b-a} C_1 f^{(m)}(\xi) = Q_n^2(g) + \left(\frac{d-c}{b-a}\right)^{m+1} C_1 g^{(m)}(\eta),$$

где $\eta = \frac{d-c}{b-a}(\xi - a) + c \in [C, D]$.

$\langle 8 \rangle$ Если АСТ подобных формул (1) и (2) есть μ и $R_n^1(x^{\mu+1}) = r$, то

$$R_n^2(x^{\mu+1}) = \left(\frac{d-c}{b-a}\right)^{\mu+2} r.$$

Прежде чем переходить к конкретным ФМК с постоянным весом, докажем лемму, которая полезна при выводе представления остаточных членов.

Лемма. Пусть $q(x)$ – интегрируемая функция, причем $q(x) \geq 0$, функция $g(x)$ непрерывна на $[a,b]$ и $\xi(x)$ – произвольное (без каких-либо предположений о непрерывности) отображение промежутка $[a,b]$ в себя. Если написанный ниже интеграл I существует, то найдется такая точка $\eta \in [a,b]$, что

$$I = \int_a^b q(x)g\big(\xi(x)\big)dx = \int_a^b q(x)dx \cdot g(\eta).$$

Доказательство. Положим $M = \max g(x)$, $m = \min g(x)$. Тогда

$$m \leq I \Big/ \int_a^b q(x)dx = G \leq M.$$

Будучи непрерывной, функция g принимает в некоторой точке η значение, равное G. ■

Формулой средних прямоугольников называется ИКФ с единственным узлом — серединой промежутка интегрирования:

$$\int_a^b f(x)dx \approx (b-a)f\left(\frac{a+b}{2}\right) \tag{4}$$

Из самого определения видно, что формулы средних прямоугольников для всех промежутков подобны. Легко видеть, что АСТ этих формул $d \geq 1$ (они точны для постоянных, являются ИКФ и "узел расположен симметрично"). Представление остатка получим сначала для промежутка $[-1,1]$. Для $f \in C^{(2)}[-1,1]$, используя лемму, имеем:

$$\int_{-1}^1 f(x)dx = \int_{-1}^1 \big(f(0) + xf'(0) + \frac{1}{2}x^2 f''(\xi(x))\big)dx = 2f(0) + \frac{1}{3}f''(\eta).$$

Согласно свойству $\langle 7 \rangle$ в случае промежутка $[a,b]$ остаточный член (4) имеет представление $R(f) = \frac{(b-a)^3}{24}f''(\eta)$. Отсюда следует, что АСТ формулы прямоугольников есть 1.

Формулами Котеса называются ИКФ, узлами которых являются концы промежутка интегрирования и точки деления промежутка на $(n-1)$ равных частей (число узлов n). Формула полностью определяется промежутком и числом узлов. Все формулы Котеса с одним числом узлов подобны. Для АСТ d согласно теореме 1 из §1 и свойству $\langle 6 \rangle$ получаются оценки: при четном n $d \geq n-1$, при нечетном — $d \geq n$. В действительности в этих неравенствах можно поставить знак равенства (без доказательства).

Обратимся к частным случаям формул Котеса. Начнем с $n = 2$. Узлы этой формулы — a и b, и поскольку коэффициенты равны ($\langle 5 \rangle$) и в сумме дают $b-a$, то сама она имеет вид

$$\int_a^b f(x)dx \approx \frac{b-a}{2}\big[f(a) + f(b)\big]$$

и ввиду очевидного геометрического смысла называется *формулой трапеций*. Представление остатка этой формулы получим сначала для $[0,1]$. Для $f \in C^{(2)}[0,1]$, используя теорему о представлении остаточного члена интерполяции, имеем

$$f(x) = P_1(x) - x(1-x)f''(\xi)/2,$$

где P_1 — интерполяционный полином функции f, построенный по узлам 0 и 1. Поэтому, опять используя лемму и равенство

$$Q_2(f) = Q_2(P_1) = \int_a^b P_1(x)dx,$$

имеем

$$R_2(f) = \int_0^1 f(x)dx - Q_2(f) = -\int_0^1 \frac{x(1-x)}{2} f''(x(\xi))dx = -\frac{1}{12}f''(\eta),$$

откуда для произвольного промежутка

$$R_2(f) = -\frac{(b-a)^3}{12}f''(\eta).$$

Этим, в частности, доказано равенство $d = 1$ для формулы трапеций.

При $n = 3$ формула Котеса называется *формулой Симпсона*. Построим ее сначала для промежутка $[-1,1]$. В этом случае узлы формулы -1, 0 и 1, а коэффициенты удовлетворяют равенствам: $A_1 = A_3$, $A_1 + A_2 + A_3 = 2$. Поскольку

$$A_2 = \int_{-1}^1 l_2(x)dx = \int_{-1}^1 (1-x^2)dx = \frac{4}{3},$$

то $A_1 = A_3 = 1/3$ и для произвольного промежутка формула Симпсона выглядит так:

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

Выведем представление остаточного члена сначала для $[-1,1]$. Для функции $f \in C^{IV}[-1,1]$ построим эрмитовский интерполяционный полином $P_3(x)$ по узлам -1 и 1 первой кратности и 0 — второй. Тогда

$$f(x) = P_3(x) + \frac{1}{4!}\Omega(x)f^{IV}(\xi), \quad \Omega(x) = -x^2(1-x^2),$$

и так как

$$Q_3(f) = Q_3(P_3) = \int_{-1}^1 P_3(x)dx,$$

то

$$R_3(f) = -\frac{1}{4!}\int_{-1}^1 x^2(1-x^2)f^{IV}(\xi(x))dx = -\frac{1}{90}f^{IV}(\eta)$$

и в случае произвольного промежутка $[a,b]$

$$R_3(f) = -\frac{1}{90}\left(\frac{b-a}{2}\right)^5 f^{IV}(\eta).$$

Этим доказано и равенство $d=3$ для формулы Симпсона.

Приведем еще без вывода формулу Котеса при $n=4$, называемую *правилом 3/8 Ньютона*:

$$\int_a^b f(x)dx \approx \frac{b-a}{8}\left(f(a)+3f\left(\frac{2a+b}{3}\right)+3f\left(\frac{a+2b}{3}\right)+f(b)\right),$$

$$R_4(f) = -\frac{2}{405}\left(\frac{b-a}{2}\right)^5 f^{IV}(\eta).$$

Начиная с $n=9$ среди коэффициентов формул Котеса появляются отрицательные, и при $n\to\infty$ сумма абсолютных величин коэффициентов быстро стремится к бесконечности. Поэтому при больших n формулы Котеса не находят применения.

Задача. Показать, что если ИКФ имеет АСТ, равную $n-1$ (n — число узлов) и имеет представление остатка в форме Лагранжа, то в этом представлении

$$C = \frac{1}{n!}\int_a^b \omega(x)dx, \qquad \omega(x) = \prod(x-x_k).$$

§3 Составные формулы

Рассматривается ситуация, когда для вычисления интеграла

$$I = \int_a^b f(x)dx \tag{1}$$

мы хотим применить формулу

$$\int_0^1 g(x)dx = \sum_{j=1}^n A_j g(x_j) + R(g) = Q(g) + R(g), \tag{2}$$

узлы которой принадлежат промежутку интегрирования: $x_j\in[0,1]$, но формула, подобная (2), не дает нужной точности. Тогда можно разбить промежуток $[a,b]$ на N равных частей и к интегралу по каждой части применить формулу, подобную (2). Это приводит нас к формуле

$$I = h\sum_{k=0}^{N-1}\sum_{j=1}^n A_j f(y_k+hx_j) + R_N(f) = Q_N(f) + R_N(f), \tag{3}$$

где $h=(b-a)/N$, $y_k = a+kh$. Формула (3) называется *составной или большой*, а формула (2) по отношению к ней *исходной*. Составная формула

содержит Nn узлов, если хотя бы одна из точек 0, 1 не является узлом для формулы (2), и $N(n-1)+1$ узел в противном случае.

Основные свойства составных формул.

$\langle 1 \rangle$ Остаток $R_N(f)$ формулы (3) представим в виде

$$R_N(f) = \sum_{k=0}^{N-1} R_N^k(f), \quad R_N^k(f) = \int_{y_k}^{y_{k+1}} f(x)dx - h\sum_{j=1}^{n} A_j f(y_k + hx_j),$$

причем $R_N^k(f)$ — это остатки квадратурных формул, подобных (2).

$\langle 2 \rangle$ Если (2) имеет АСТ d и $R(x^{d+1}) = r$ (очевидно, что $r \neq 0$), то $R_N(x^{d+1}) = (b-a)h^{d+1}r$.

Доказательство. Воспользоваться предыдущим свойством и свойством $\langle 8 \rangle$ подобных формул (§2).

$\langle 3 \rangle$. АСТ формул (2) и (3) совпадает.

Доказательство. Очевидно, что АСТ формулы (3) не меньше, чем формулы (2). Обратное немедленно вытекает из предыдущего свойства. ■

$\langle 4 \rangle$. Если формула (2) имеет представление остатка в форме Лагранжа:

$$R(g) = Cg^{(m)}(\xi),$$

то и (3) имеет такое представление:

$$R_N(f) = C_N f^{(m)}(\eta),$$

где

$$C_N = C(b-a)h^m.$$

Доказательство. Применяя свойство $\langle 7 \rangle$ подобных формул, имеем

$$\begin{aligned} R_N^k(f) =& h^{m+1} C f^{(m)}(\eta_k), \quad \eta_k \in (y_k, y_{k+1},) \\ R_N(f) =& h^{m+1} C \sum_{k=0}^{N-1} f^{(m)}(\eta_k), \end{aligned}$$

и так как найдется такая точка η, что $\sum_{k=0}^{N-1} f^{(m)}(\eta_k) = Nf^{(m)}(\eta)$, (используется непрерывность функции $f^{(m)}$) и $Nh = b-a$, то этим свойство доказано. ■

Далее рассматривается *последовательность* составных формул: исходная формула считается фиксированной, а $N \to \infty$.

$\langle 5 \rangle$. Для того чтобы для любой непрерывной функции f последовательность квадратурных сумм сходилась к интегралу (т.е. $R_N(f) \to 0$), необходимо и достаточно, чтобы исходная формула (2) была точна для постоянных[4].

Доказательство. 1) Достаточность. Представим квадратурную сумму в виде

$$Q_N(f) = \sum_{j=1}^{n} A_j \sum_{k=0}^{N-1} hf(y_k + hx_j).$$

[4]Это может быть выражено также словами: АСТ формулы (2) ≥ 0 или $\sum A_j = 1$.

Каждая внутренняя сумма здесь есть сумма Римана для рассматриваемого интеграла и потому при $N \to \infty$ к нему сходится. Остается воспользоваться тем, что $\sum A_j = 1$.

2) Необходимость. Для функции $f(x) \equiv 1$ имеем

$$Q_N(1) = h \sum_{k=0}^{N-1} \sum_{j=1}^{n} A_j = (b-a) \sum_{j=1}^{n} A_j,$$

и так как $Q_N(1) \to (b-a)$, то $\sum A_j = 1$ ■

Определение Будем говорить, что для функции f последовательность квадратурных формул (3) сходится с порядком m, если найдется такая постоянная C, что при всех $N \quad |R_N(f)| \le Ch^m$.

⟨6⟩. Для того чтобы для любой m раз непрерывно дифференцируемой функции f ($f \in C^{(m)}$) последовательность (3) сходилась с порядком m, необходимо и достаточно чтобы для АСТ исходной формулы d выполнялось неравенство $d \ge m-1$.

Доказательство[5]. 1. Достаточность. Пусть $f \in C^{(m)}$ произвольна, $d \ge m-1$. Положим $M = \max_{[a,b]} |f^{(m)}(x)|$ и на каждом промежутке $[y_k, y_{k+1}]$ запишем $f(x)$ по формуле Тейлора:

$$f(x) = f(y_k) + (x-y_k)f'(y_k) + \cdots + \frac{(x-y_k)^{m-1}}{(m-1)!} f^{(m-1)}(y_k) + \frac{(x-y_k)^m}{m!} f^{(m)}(\xi_k) =$$
$$= p_{m-1}(x) + \frac{(x-y_k)^m}{m!} f^{(m)}(\xi_k).$$

Здесь $p_{m-1} \in \mathbb{P}_{m-1} \subseteq \mathbb{P}_d$, причем

$$\|f - p_{m-1}\|_{C[y_k, y_{k+1}]} \le \frac{M}{m!} h^m,$$

так что и наилучшее приближение функции f полиномами степени не выше d на промежутке $[y_k, y_{k+1}]$ удовлетворяет неравенству

$$E_d(f; [y_k, y_{k+1}]) \le \frac{M}{m!} h^m.$$

Используя оценку остатка квадратурной формулы через наилучшее приближение (теорема 3 из §1), теперь имеем

$$R_N^k(f) \le \big[h + h \sum |A_j|\big] E_d(f; [y_k, y_{k+1}]) \le \big[1 + \sum |A_j|\big] \frac{M h^{m+1}}{m!}.$$

Суммируя все эти оценки: $R_N(f) \le Ch^m$, где $C = M(b-a)[1 + \sum |A_j|]/m!$.

2. Необходимость. Согласно свойству ⟨2⟩ остаток $R_N(f)$ для функции $f(x) = x^{d+1}$ есть $(b-a)h^{d+1}r$ и в случае $d < m-1$ окажется $R_N(f)/h^m \to \infty$ ■

[5]Если исходная формула имеет представление остатка в форме Лагранжа, то свойство ⟨6⟩ легко следует из этого представления.

Как видно из доказательства, в части необходимости это утверждение допускает усиление: слова “для любой m раз непрерывно дифференцируемой функции” можно заменить на “для любого полинома”.

Наиболее употребительными являются составные формулы средних прямоугольников, трапеций и Симпсона. Выпишем эти формулы. При этом представление остаточного члена сразу же получается из представления остаточного члена исходной формулы применением свойства $\langle 4\rangle$. Конечно, приводимые формулы для остатка верны лишь в том случае, если подынтегральная функция нужное число раз непрерывно дифференцируема.

Составная формула средних прямоугольников:

$$\int_a^b f(x)dx \approx h\sum_{k=1}^{N} f\left(a+\frac{2k-1}{2}h\right), \quad h=\frac{b-a}{N}, \quad R(f)=\frac{b-a}{24}h^2f''(\xi).$$

Составная формула трапеций:

$$\int_a^b f(x)dx \approx \frac{h}{2}f(a)+h\sum_{k=1}^{N-1} f(a+kh)+\frac{h}{2}f(b), \ \ h=\frac{b-a}{N}, \ \ R(f)=-\frac{b-a}{12}h^2f''(\xi).$$

Как видно из представления остатков этих двух формул, если вторая производная функции f сохраняет на $[a,b]$ знак, то квадратурные суммы средних прямоугольников и трапеций дают двухсторонние приближения к интегралу.

Составная формула Симпсона:

$$\int_a^b f(x)dx \approx \frac{h}{3}\left[f(a)+2\sum_{k=1}^{N-1} f(a+2kh)+4\sum_{k=1}^{N} f\big(a+(2k-1)h\big)+f(b)\right],$$
$$h=\frac{b-a}{2N}, \quad R(f)=-\frac{b-a}{180}h^4f^{IV}(\xi).$$

Задача. Показать, что свойства $\langle 3\rangle$ и $\langle 4\rangle$ составных формул (второе из них с изменением представления для C_N) сохраняются и в том случае, если составная формула строится на основе *неравномерного* разбиения отрезка $[a,b]$ на части.

§4 Квадратурные формулы гауссова типа

Вернемся к рассмотрению формул с весом $w(x)$. На всем протяжении параграфа функцию w будем считать суммируемой, неотрицательной и отличной от нуля на множестве положительной меры. ФМК с n узлами содержит $2n$ параметров — кроме узлов еще и коэффициенты. Требование, чтобы ее АСТ была не ниже d, налагает на эти параметры $d+1$ условие (точность для x^j, $j=0,\dots,d$). Поэтому можно рассчитывать на существование ФМК с n узлами и АСТ $2n-1$. О таких формулах и идет речь в этом параграфе. Но сначала некоторые вспомогательные сведения.

Для полиномов f, g определим скалярное произведение:

$$(f,g) = \int_a^b w(x)f(x)g(x)dx.$$

Оно обладает обычными свойствами скалярного произведения: 1) линейность по каждому аргументу, 2) $(g,f) = (f,g)$, 3) $(f,f) \geq 0$, и $(f,f) = 0$ тогда и только тогда, когда $f(x) \equiv 0$. Полиномы f и g назовем ортогональными, если $(f,g) = 0$. Если f ортогонален g_1 и g_2, то он ортогонален и $\alpha_1 g_1 + \alpha_2 g_2$.

Отметим очевидный факт: если q_k — полиномы степени в точности k, то любой полином $p_n \in \mathbb{P}_n$ допускает представление $p_n(x) = \sum_{k=0}^n c_k q_k(x)$.

Теорема 1. Существует и притом единственный с точностью до постоянного множителя полином $\omega_n(x)$, ортогональный всем $q_{n-1} \in \mathbb{P}_{n-1}$. Полиномы ω_n называются ортогональными полиномами.

Доказательство. Положим $\omega_0 = 1$.

1) Существование. Метод индукции. При $n = 1$ достаточно положить $\omega_1 = x - a$, где $a = (x,1)/(1,1)$. Пусть для $k = 1, \ldots, n-1$ существование ω_k уже доказано. Полином

$$\omega_n(x) = x^n - a_{n-1}\omega_{n-1}(x) - \cdots - a_0\omega_0(x), \quad a_k = (x^n, \omega_k)/(\omega_k, \omega_k),$$

ортогонален полиномам ω_k, а значит, и всем $q_{n-1} \in \mathbb{P}_{n-1}$.

2) Единственность. От противного. Пусть ω_n и ω_n' — два ортогональных полинома со старшими коэффициентами a_n и a_n'. Тогда $\omega_n/a_n - \omega_n'/a_n'$ — многочлен степени не выше $n-1$, ортогональный сам себе и потому тождественно равный нулю, так что $\omega_n' = (a_n'/a_n)\omega_n$. ■

Теорема 2. Все корни ортогонального многочлена $\omega_n(x)$ вещественны, различны и принадлежат (a,b).

Доказательство. Убедимся, что ω_n имеет на (a,b) n точек перемены знака. Пусть их $m < n$ и это $x_1, \ldots, x_m$. Тогда положим $q_m(x) = (x - x_1)\ldots(x - x_m) \in \mathbb{P}_{n-1}$. Функция $w(x)\omega_(x)q_m(x)$ сохраняет на (a,b) знак, и потому $(\omega_n, q_m) \neq 0$, что противоречит ортогональности. ■

Теорема 3. Для того чтобы квадратурная формула

$$\int_a^b w(x)f(x)dx \approx \sum_{k=1}^n A_k f(x_k) = Q_n(f) \tag{1}$$

имела АСТ $2n-1$, необходимо и достаточно:

1) (1) есть ИКФ,

2) узлы x_k суть корни ортогонального полинома ω_n.

Доказательство. 1) Необходимость первого условия следует из теоремы об АСТ ИКФ. Докажем необходимость второго. Пусть (1) точна для всех $p \in \mathbb{P}_{2n-1}$. Положим $\omega_n(x) = (x - x_1)\ldots(x - x_n)$ и пусть $q_{n-1} \in \mathbb{P}_{n-1}$ произволен. Тогда $\omega_n q_{n-1} \in \mathbb{P}_{2n-1}$ и потому

$$(\omega_n, q_{n-1}) = \sum_{k=1}^n A_k \omega_n(x_k) q_{n-1}(x_k) = 0,$$

так что ω_n — ортогональный полином.

2) Достаточность. Пусть выполнены условия 1) и 2). Покажем, что (1) точна для всех полиномов степени $2n-1$. Возьмем любой такой полином $Q_{2n-1}(x)$ и поделим его на ω_n. Тогда получим представление $Q_{2n-1}(x) = \omega(x)q_{n-1}(x) + r_{n-1}(x)$, где q_{n-1} и r_{n-1} — частное и остаток деления — полиномы степени не выше $n-1$. Учитывая, что (1) как ИКФ точна для всех полиномов степени не выше $n-1$, имеем

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

$$= \int_a^b w(x)r_{n-1}(x) = \sum_{k=1}^n A_k r_{n-1}(x_k) = \sum_{k=1}^n A_k\big(\omega_n(x_k)q_{n-1}(x_k) + r_{n-1}(x_k)\big) =$$

$$= \sum_{k=1}^n A_k Q_{2n-1}(x_k),$$

Поскольку квадратурная формула с неотрицательным весом и n узлами не может иметь АСТ больше $2n-1$, то этим теорема доказана. ■

Определение. Формула (1), имеющая при n узлах АСТ $2n-1$, называется *формулой гауссова типа* или *формулой наивысшей степени точности.*

Следствие. При наложенных на вес $w(x)$ условиях при каждом n формула гауссова типа существует и единственна.

Отметим свойства формул гауссова типа.

⟨1⟩ Коэффициенты такой формулы положительны: $A_k > 0$.

Доказательство. Для полиномов $l_k^2(x)$, где $l_k(x)$ — фундаментальные полиномы интерполяции по узлам x_k ($l_k(x_j) = \delta_{kj}$), имеющие степень $n-1$, формула точна. Поэтому

$$\int_a^b w(x)l_k^2(x)dx = \sum_{j=1}^n A_j l_k^2(x_j) = A_k. \quad ■$$

⟨2⟩ Формула гауссова типа имеет представление в форме Лагранжа:

$$R_n(f) = C_n f^{(2n)}(\xi), \quad C_n = \frac{1}{(2n)!}\int_a^b w(x)\omega_n^2(x)dx,$$

где $\omega_n(x)$ — ортогональный полином со старшим коэффициентом, равным единице.

Доказательство. Для $f \in C^{(2n)}[a,b]$ построим эрмитовский интерполяционный полином $P_{2n-1} \in \mathbb{P}_{2n-1}$ по узлам x_k кратности 2. Тогда

$$f(x) - P_{2n-1}(x) = \frac{\Omega_{2n}(x)}{(2n)!}f^{(2n)}(\eta(x)), \quad \big(\Omega_{2n}(x) = \omega_n^2(x)\big)$$

и т.к. $R_n(P_{2n-1}) = 0$, так что $R_n(f) = R_n(f - P_{2n-1})$, и $Q_n(f) = Q_n(P_{2n-1})$, то

$$R_n(f) = \int_a^b w(x)\frac{\Omega_{2n}(x)}{(2n)!}f^{(2n)}(\eta(x))dx = \frac{1}{(2n)!}\int_a^b w(x)\omega_n^2(x)dx\, f^{(2n)}(\xi)$$

(мы воспользовались леммой из §2). ■

⟨3⟩ Для остатка справедлива оценка

$$R_n(f) \le 2 \int_a^b w(x)dx\, E_{2n}(f).$$

Это — непосредственное следствие теоремы 3 из §1.

⟨4⟩ Для любой непрерывной функции $f \in C[a,b]$

$$Q_n(f) \to \int_a^b w(x)f(x)dx.$$

Это немедленно следует из⟨3⟩.

Определение. Квадратурная формула гауссова типа для промежутка $[-1,1]$ с весом $w(x) \equiv 1$ называется *квадратурной формулой Гаусса.*

Замечание. Для любого промежутка $[a,b]$ формула гауссова типа с весом $w(x) \equiv 1$ подобна формуле Гаусса. Иногда формулы, подобные формуле Гаусса, также называют формулами Гаусса.

Определение. *Многочленом Лежандра* степени n называется

$$P_n(x) = \frac{n!}{(2n)!}\frac{d^n}{dx^n}(x^2-1)^n.$$

Теорема 4. Многочлен Лежандра есть ортогональный на промежутке $[-1,1]$ с весом $w(x) \equiv 1$ полином. Его старший коэффициент равен единице.

Доказательство. То, что P_n есть полином степени n со старшим коэффициентом единица, очевидно. Покажем ортогональность. Учитывая, что при $k < n$

$$\frac{d^k}{dx^k}(x^2-1)^n \Big|_{x=\pm 1} = 0,$$

и интегрируя по частям, для любого полинома $q_{n-1} \in \mathbb{P}_{n-1}$ имеем

$$\int_{-1}^1 P_n(x)q_{n-1}(x)dx = \frac{n!}{(2n)!}\int_{-1}^1 \frac{d^n}{dx^n}(x^2-1)^n q_{n-1}(x)dx =$$
$$= (-1)^n \frac{n!}{(2n)!}\int_{-1}^1 P_n(x)\frac{d^n}{dx^n}q_{n-1}(x)dx = 0,$$

т.к. $\frac{d^n}{dx^n}q_{n-1}(x) \equiv 0$. ■

Корни многочлена Лежандра являются узлами квадратурной формулы Гаусса. Сам этот многочлен в зависимости от четности или нечетности n является четной или нечетной функцией. Поэтому (см. также свойство ⟨4⟩) верна

Теорема 5. Узлы квадратурной формулы Гаусса симметричны (при нумерации в порядке возрастания $x_k = -x_{n+1-k}$), а коэффициенты при симметричных узлах равны ($A_k = A_{n+1-k}$).

Лемма. Справедливо равенство

$$I_n = \int_{-1}^{1} (1-x^2)^n dx = 2\frac{(2n)!!}{(2n+1)!!}.$$

Доказательство. Применяя интегрирование по частям, имеем

$$I_n = \int_{-1}^{1} (1-x^2)^{n-1}(1-x^2)dx = I_{n-1} - \frac{1}{2}\int_{-1}^{1} x[2x(1-x^2)^{n-1}]dx =$$
$$I_{n-1} + \frac{1}{2n}\int_{-1}^{1} x[(1-x^2)^n]'dx = I_{n-1} - \frac{1}{2n}I_n,$$

так что $I_n = \frac{2n}{2n+1}I_{n-1}$ и остается применить метод математической индукции, учитывая, что $I_o = 2$. ■

Следствие. Справедливо равенство

$$J_n = \int_{-1}^{1} P_n^2(x)dx = 2\frac{n!}{(2n)!}\frac{(2n)!!}{(2n+1)!!}n!.$$

Доказательство. Интегрируя по частям:

$$J_n = \frac{n!}{(2n)!}\int_{-1}^{1} P_n(x)[(x^2-1)^n]^{(n)}dx =$$
$$= (-1)^n\frac{n!}{(2n)!}\int_{-1}^{1} P_n^{(n)}(x)(x^2-1)^n dx = \frac{n!}{(2n)!}n!I_n$$

(здесь учтено, что $P_n^{(n)}(x) \equiv n!$), и остается воспользоваться леммой. ■

Теорема 6. Для квадратурной формулы Гаусса при $f \in C^{(2n)}[-1,1]$ справедливо представление остатка

$$R_n(f) = C_n f^{(2n)}(\xi), \quad C_n = \frac{2^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3}.$$

Доказательство. Согласно свойству $\langle 2 \rangle$ доказываемое представление имеет место при $C_n = \frac{1}{(2n)!}J_n$. Остается воспользоваться доказанным следствием и равенствами

$$(2n)!! = 2^n n!, \quad (2n+1)!! = \frac{(2n+1)!}{(2n)!!} = \frac{(2n+1)(2n)!}{2^n n!} \quad ■$$

Замечание 1. Постоянные C_n очень быстро убывают. Приведем несколько первых из них:

$$C_2 = \frac{1}{135}, \quad C_3 = \frac{1}{15750} \quad C_4 = \frac{1}{3472875}.$$

Замечание 2. При $n = 1$ формула Гаусса совпадает с формулой средних прямоугольников.

Замечание 3. При $n = 2$ в представление остаточного члена формулы Гаусса, как и для формулы Симпсона, входит 4-ая производная, но коэффициенты при них имеют противоположные знаки. Поэтому если четвертая производная функции сохраняет знак на промежутке интегрирования, то квадратурные суммы Гаусса (при $n = 2$) и Симпсона дают двусторонние приближения к интегралу. То же относится и к построенным на основе этих формул составным квадратурным формулам.

Задача 1. Доказать ортогональность многочленов Чебышева на промежутке $[-1, 1]$ с весом $w(x) = 1/\sqrt{1 - x^2}$.

Задача 2. Пусть x_k и A_k $(k = 1, \dots, n)$ — узлы и коэффициенты формулы гауссова типа, $\widehat{\omega}_k(x)$ $(k = 0, 1, \dots)$ — ортогональные полиномы, нормированные условием $(\widehat{\omega}_k, \widehat{\omega}_k) = 1$. Доказать, что

$$\mathcal{A} = \begin{pmatrix} \sqrt{A_1}\widehat{\omega}_0(x_1) & \sqrt{A_2}\widehat{\omega}_0(x_2) & \dots & \sqrt{A_n}\widehat{\omega}_0(x_n) \\ \sqrt{A_1}\widehat{\omega}_1(x_1) & \sqrt{A_2}\widehat{\omega}_1(x_2) & \dots & \sqrt{A_n}\widehat{\omega}_1(x_n) \\ \dots & \dots & \dots & \dots \\ \sqrt{A_1}\widehat{\omega}_{n-1}(x_1) & \sqrt{A_2}\widehat{\omega}_{n-1}(x_2) & \dots & \sqrt{A_n}\widehat{\omega}_{n-1}(x_n) \end{pmatrix} \quad -$$

ортогональная матрица $(\mathcal{A}\mathcal{A}^T = E)$.

Задача 3. Используя предыдущую задачу, показать, что коэффициенты квадратурной формулы гауссова типа можно вычислять по формулам:

$$A_k = \left(\sum_{j=0}^{n-1} \widehat{\omega}_j^2(x_k) \right)^{-1} .$$

Глава 3
Решение задач линейной алгебры

§1 Нормы векторов и матриц

Обозначение: $\mathbb{C}^n$ — пространство n-мерных векторов с комплексными компонентами. Естественный базис в $\mathbb{C}^n$ — векторы $e_k = \{\delta_{kj}\}_1^n$. Компоненты векторов x, y будем обычно обозначать: $x = (\xi_1, \ldots, \xi_n)$, $y = (\eta_1, \ldots, \eta_n)$, их скалярное произведение — $(x, y) = \sum_{k=1}^n \xi_k \overline{\eta_k}$.

Определение. Заданная на $\mathbb{C}^n$ вещественная функция $\nu(x)$, обозначаемая обычно $\nu(x) = \|x\|$, обладающая свойствами:

1) $\|x\| \geq 0$, $\|x\| = 0$ тогда и только тогда, когда $x = 0$ (нулевой вектор),

2) для любых $x \in \mathbb{C}^n$ и $\lambda \in \mathbb{C}$ $\|\lambda x\| = |\lambda| \, \|x\|$,

3) для любых $x, y \in \mathbb{C}^n$ $\|x+y\| \leq \|x\|+\|y\|$ (неравенство треугольника),

называется *нормой*.

Сразу же отметим очевидное следствие 3):

4) Для любых $x, y \in \mathbb{C}^n$ $\big|\|x\| - \|y\|\big| \leq \|x - y\|$.

Наиболее употребительными являются следующие нормы, имеющие специальные обозначения[6]:

$$\|x\|_2 = \sqrt{\sum_{k=1}^n |\xi_k|^2} = \sqrt{(x,x)} \ - \text{ евклидова норма, длина вектора,}$$

$$\|x\|_\infty = \max |\xi_k|,$$

$$\|x\|_1 = \sum_{k=1}^n |\xi_k|.$$

Проверка аксиом 1)-3) для этих норм элементарна.

В силу неравенства Коши - Буняковского $|(x, y)| \leq \|x\|_2 \|y\|_2$. Кроме того, очевидно неравенство $|(x, y)| \leq \|x\|_1 \|y\|_\infty$.

Приведем еще один важный пример нормы. Пусть φ_k $(k = 1, \ldots, n)$ комплекснозначные непрерывные на $[a, b]$ линейно независимые функции. Тогда

$$\|x\| = \left\| \sum_{k=1}^n \xi_k \varphi_k \right\|_C = \max_{t \in [a,b]} \left| \sum_{k=1}^n \xi_k \varphi_k(t) \right|$$

есть норма в $\mathbb{C}^n$. Доказательство очевидно.

Теорема 1 (об эквивалентности норм). Все нормы в $\mathbb{C}^n$ эквивалентны. Это значит, что для любых двух норм $\|\cdot\|'$ и $\|\cdot\|''$ найдутся такие постоянные c' и c'', что для всех $x \in \mathbb{C}^n$ выполняются неравенства

$$\|x\|' \leq c' \|x\|'', \quad \|x\|'' \leq c'' \|x\|'.$$

---

[6] Эти обозначения связаны с тем, что перечисляемые ниже нормы являются частными случаями *норм Гельдера*: $\|x\|_p = \left(\sum_{k=1}^n |\xi_k|^p\right)^{1/p}$ $(p \geq 1)$.

Доказательство. Можно считать, что одна из норм есть $\|\cdot\|_2$, а другая $\|\cdot\|$ произвольна. Тогда

$$\|x\| = \Big\|\sum \xi_k e_k\Big\| \le \sum |\xi_k|\,\|e_k\| \le c'\|x\|_2, \quad c' = \sqrt{\sum \|e_k\|^2}.$$

В частности, $\big|\|x\| - \|y\|\big| \le \|x-y\| \le c'\|x-y\|_2$, так что $\|\cdot\|$ — непрерывная функция. На сфере $S = \{\, x \;\big|\; \|x\|_2 = 1 \,\}$ она достигает своего минимального значения, отличного от нуля, т.к. на S в нуль не обращается: $\min_{x\in S}\|x\| = m > 0$. Для любого $x \ne 0$ положим $y = x/\|x\|_2$; тогда $\|x\| = \|x\|_2\,\|y\| \ge m\|x\|_2$, так что для любого x $\|x\|_2 \le \frac{1}{m}\|x\|$. ■

Если использовать приведенный выше пример нормы, то из теоремы 1 немедленно получим

Следствие. При заданных n и $[a,b]$ найдется такая постоянная $m > 0$, что для любого полинома $P_n \in \mathbb{P}_n$: $P_n(t) = a_0 + \cdots + a_n t^n$ выполняется неравенство $\|P_n\|_{C[a,b]} \ge m\sqrt{\sum a_k^2}$.

Это следствие формулировалось в §1 главы 1 в виде леммы.

Пусть даны последовательность векторов $x_s = (\xi_1^s, \ldots, \xi_n^s)$ и вектор x.

Теорема 2. Эквивалентны утверждения:

А. При всех k $\xi_k^s \to \xi_k$,

Б. $\|x_s - x\| \to 0$,

В. для всех $y \in \mathbb{C}^n$ $(x_s, y) \to (x, y)$.

Доказательство. Если Б. выполняется для какой-то одной нормы, то в силу теоремы 1 и для всех остальных.

1) А⇒Б. Очевидно, если $\|\cdot\| = \|\cdot\|_2$.

2) Б⇒В. $|(x_s, y) - (x, y)| \le \|x_s - x\|_2\|y\|_2$.

3) В⇒А Достаточно взять $y = e_k$. ■

Если выполнено хоть одно из требований А-В, то говорят, что последовательность x_s сходится к x ($x_s \to x$).

Множество (комплексных) квадратных матриц порядка n обозначим $\mathbb{M}_n$. Элементы матрицы A будем обозначать a_{kj}:

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

Единичную матрицу будем обозначать I. Заданная на $\mathbb{M}_n$ вещественная функция $\|A\|$ называется *нормой*, если она удовлетворяет требованиям:

1) $\|A\| \ge 0$, $\|A\| = 0 \Leftrightarrow A = 0$,

2) $\|\lambda A\| = |\lambda|\,\|A\|$,

3) $\|A + B\| \le \|A\| + \|B\|$.

Как и для векторов, из 3) следует

4) $\big|\|A\| - \|B\|\big| \le \|A - B\|$.

Таким образом, норму в $\mathbb{M}_n$ можно рассматривать как норму в $\mathbb{C}^{n^2}$. Поэтому в силу теоремы 1 все нормы в $\mathbb{M}_n$ эквивалентны.

Определение. Пусть $\|\cdot\|$ некоторая норма в $\mathbb{C}^n$. Матричная норма

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\| \tag{1}$$

называется *операторной* нормой, порожденной соответствующей векторной.

То, что определяемая формулой (1) функция действительно есть норма, проверяется элементарно. Если $\|\cdot\|$ — операторная норма, то помимо 1)-4) она обладает еще легко проверяемыми свойствами:

5) для всех векторов x $\|Ax\| \le \|A\|\,\|x\|$ и существует такой вектор x_0, что $\|x_0\| = 1$, $\|Ax_0\| = \|A\|$,

6) $\|I\| = 1$,

7) $\|AB\| \le \|A\|\,\|B\|$,

8) для любого собственного числа матрицы A $|\lambda| \le \|A\|$.

Рассмотрим последовательность матриц

$$A_s = \begin{pmatrix} a^s_{11} & \dots & a^s_{1n} \\ \dots & \dots & \dots \\ a^s_{n1} & \dots & a^s_{nn} \end{pmatrix}$$

и матрицу A.

Теорема 3. Эквивалентны утверждения:

А. при всех k, j $a^s_{kj} \to a_{kj}$,

Б. $\|A_s - A\| \to 0$,

В. для всех x $A_s x \to Ax$,

Г. для всех x и y $(A_s x, y) \to (Ax, y)$.

Доказательство. Как и в теореме 2, норма в Б произвольна.

А$\Leftrightarrow$Б — из теоремы 2.

Б$\Rightarrow$В. Для операторной нормы $\|A_s x - Ax\| \le \|A_s - A\|\,\|x\|$.

В$\Rightarrow$Г. $\big|(A_s x, y) - (Ax, y)\big| \le \|A_s x - Ax\|_2 \|y\|_2$,

Г$\Rightarrow$А. Достаточно рассмотреть $x = e_j$, $y = e_k$. ■

Определение сходимости последовательности матриц дается так же, как для векторов.

Операторные нормы матрицы, порожденные введенными выше векторными $\|\cdot\|_2$, $\|\cdot\|_\infty$ и $\|\cdot\|_1$, помечаются теми же значками.

Теорема 4. Справедливы равенства:

$$\|A\|_\infty = \max_k \sum_j |a_{kj}|, \tag{2}$$

$$\|A\|_1 = \max_j \sum_k |a_{kj}|, \tag{3}$$

$$\|A\|_2 = \sqrt{\Lambda}, \tag{4}$$

где Λ — максимальное собственное число матрицы[7] A^*A.

[7]Корни из собственных чисел матрицы A^*A называются *сингулярными числами* матрицы A.

Доказательство. 1) Обозначим правую часть (2) $\varkappa$ и положим $y = Ax$. Тогда

$$|\eta_k| = \left|\sum_j a_{kj}\xi_j\right| \le \varkappa\|x\|_\infty,$$

так что $\|Ax\|_\infty = \|y\|_\infty \le \varkappa\|x\|_\infty$ и $\|A\|_\infty \le \varkappa$. Пусть $\varkappa = \sum_j |a_{k_0 j}|$. Положим $\xi_j = \operatorname{sign} a_{k_0 j}$, $y = Ax$. Тогда $\|x\|_\infty = 1$, $\eta_{k_0} = \varkappa$, $\|Ax\|_\infty = \|y\|_\infty \ge \varkappa$ и потому $\|A\|_\infty \ge \varkappa$.

2) Пусть $\varkappa$ — правая часть (3), $Ax = y$.

$$\|y\|_1 = \sum_k \left|\sum_j a_{kj}\xi_j\right| \le \sum_j |\xi_j| \sum_k |a_{kj}| \le \varkappa\|x\|_1$$

и $\|A\|_1 \le \varkappa$. Пусть $\varkappa = \sum_k |a_{kj_0}|$. Тогда для вектора $x = e_{j_0}$ имеем $\|x\|_1 = 1$, $\|Ax\|_1 = \varkappa$, и $\|A\|_1 \ge \varkappa$.

3) Матрица A^*A эрмитова. Все ее собственные числа вещественны, неотрицательны, и она имеет полную ортогональную систему собственных векторов. Пусть $\Lambda_1 \ge \Lambda_2 \ge \cdots \ge \Lambda_n$ ($\Lambda = \Lambda_1$) все ее собственные числа и $z_1, \ldots, z_n$ — соответствующие собственные векторы, такие что $(z_k, z_j) = \delta_{kj}$. Произвольный вектор x разложим по этим векторам: $x = \sum \alpha_k z_k$. Тогда $\|x\|_2^2 = (x, x) = \sum |\alpha_k|^2$ и

$$\|Ax\|_2^2 = (Ax, Ax) = (x, A^*Ax) = \sum \Lambda_k |\alpha_k|^2 \le \Lambda\|x\|_2^2.$$

Так что $\|A\|_2 \le \sqrt{\Lambda}$. В то же время $\|Az_1\|_2^2 = (Az_1, Az_1) = \Lambda(z_1, z_1) = \Lambda\|z_1\|_2^2$ и $\|A\|_2 \ge \sqrt{\Lambda}$. ■

Следствие. Если матрица A эрмитова, то $\|A\|_2 = \max|\lambda_k|$, где λ_k — собственные числа самой матрицы A.

Доказательство. В эрмитовом случае $A^*A = A^2$, и собственные числа этой матрицы суть квадраты собственных чисел матрицы A. ■

Теорема 5. Для $\|A\|_2$ справедливы оценки:

$$\|A\|_2 \le \sqrt{\sum_k \sum_j a_{kj}^2}, \quad \|A\|_2 \le \sqrt{\|A\|_\infty \|A\|_1}.$$

Доказательство. 1) Для $y = Ax$ имеем

$$\|y\|_2^2 = \sum_k \left|\sum_j a_{kj}\xi_j\right|^2 \le \sum_k \left(\sum_j |a_{kj}|^2\right)\left(\sum_j |\xi_j|^2\right) = \|x\|_2^2 \sum_k \sum_j |a_{kj}|^2.$$

2) $\|A\|_2^2 = \Lambda \le \|A^*A\|_1 \le \|A\|_1\|A^*\|_1 = \|A\|_1\|A\|_\infty$. ■

Задача 1. Пусть B неособенная матрица и $\|.\|$ — векторная норма. Положим $\|x\|' = \|Bx\|$. Показать, что $\|\cdot\|'$ также норма и найти выражение для порожденной этой нормой операторной матричной.

Задача 2. Найти наилучшие значения (зависящих от n !) постоянных в неравенствах, связывающих $\|x\|_p$ с $\|x\|_q$ и $\|A\|_p$ с $\|A\|_q$ при $p, q = 1, 2, \infty$.

§2 Матричная геометрическая прогрессия и некоторые оценки

В этом §, если не оговорено противное, мы всегда считаем, что задана произвольная векторная норма, а норма матрицы — всегда порожденная этой векторной операторная норма.

Пусть $A \in \mathbb{M}_n$ — некоторая матрица. Матричная геометрическая прогрессия — последовательность A^s. Вопрос: когда $A^s \to 0$?

Лемма 1. Если $|\lambda| < 1$, $\nu \in \mathbb{N}$, то $C_s^\nu \lambda^{s-\nu} \to_0$ при $s \to \infty$.

Доказательство следует из того, что $C_s^\nu \le s^\nu$. ■

Лемма 2. Пусть

$$D = \begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ 0 & 0 & \lambda & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \lambda \end{pmatrix} \in \mathbb{M}_n. \tag{1}$$

Если $|\lambda| < 1$, то $D^s \to 0$.

Доказательство. Достаточно показать, что при всех k $D^s e_k \to 0$. Введем обозначение: $e_k = 0$ при $k \le 0$. Тогда $De_k = \lambda e_k + e_{k-1}$ и методом индукции легко показывается, что

$$D^s e_k = \sum_{\nu=0}^{s} C_s^\nu \lambda^{s-\nu} e_{k-\nu}.$$

При $s > k - 1$

$$D^s e_k = \sum_{\nu=0}^{k-1} s C_s^\nu \lambda^{s-\nu} e_{k-\nu}.$$

Используя лемму 1, получаем требуемое. ■

Определение. Спектральным радиусом матрицы A (обозначение $\rho(A)$) называется максимальный из модулей ее собственных чисел.

Теорема 1. Для того чтобы $A^s \to 0$, необходимо и достаточно выполнение неравенства $\rho(A) < 1$.

Доказательство. 1) Необходимость. От противного. Пусть матрица A имеет такое собственное число, что $|\lambda| \ge 1$ и пусть z — соответствующий собственный вектор. Тогда $\|A^s z\| = |\lambda|^s \|z\| \nrightarrow 0$ и $A^s \nrightarrow 0$.

2) Достаточность легко показывается, если A имеет полную систему собственных векторов, но предполагать это мы не будем и воспользуемся для A канонической формой Жордана: $A = BDB^{-1}$, где

$$D = \begin{pmatrix} D_{n_1} & 0 & \dots & 0 \\ 0 & D_{n_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & D_{n_l} \end{pmatrix},$$

D_{n_j} — матрицы вида (1) с собственными числами λ_j матрицы A на главной диагонали. Тогда $A^s = BD^sB^{-1}$, причем

$$D^s = \begin{pmatrix} D^s_{n_1} & 0 & \dots & 0 \\ 0 & D^s_{n_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & D^s_{n_l} \end{pmatrix},$$

и в случае $\rho(A) < 1$ $D^s \to 0$. ■

Рассмотрим теперь сумму членов матричной геометрической прогрессии (ряд)

$$I + A + A^2 + A^3 + \dots. \tag{2}$$

Теорема 2. Для того чтобы ряд (2) сходился, необходимо и достаточно $\rho(A) < 1$. В случае сходимости его сумма $S = (I - A)^{-1}$.

Доказательство. Необходимость следует из того, что общий член сходящегося ряда обязан стремиться к нулю, и теоремы 1.

Достаточность. Пусть S_m — частная сумма ряда. Тогда $(I - A)S_m = I - A^{m+1}$. Т.к. 1 не есть собственное число матрицы A ($\rho(A) < 1$), то существует $(I - A)^{-1}$ и $S_m = (I - A)^{-1}(I - A^{m+1})$. Остается перейти к пределу при $m \to \infty$. ■

Следствие. Если $\|A\| < 1$[8], то ряд (2) сходится, матрица $I - A$ неособенная и

$$\|(I - A)^{-1}\| \le \frac{1}{1 - \|A\|}. \tag{3}$$

Доказательство требуется лишь для (3):

$$\|S_m\| \le 1 + \|A\| + \|A\|^2 + \dots + \|A\|^m \le \frac{1}{1 - \|A\|}. \quad ■$$

Определение. Говорят, что A матрица с диагональным преобладанием, если при всех k

$$|a_{kk}| > \sum_{j \ne k} |a_{kj}|.$$

Теорема 3 (признак Адамара). Матрица с диагональным преобладанием неособенная.

Доказательство. Диагональные элементы матрицы A отличны от нуля и она допускает представление $A = D(I + R)$, где

$$D = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix}, R = \begin{pmatrix} 0 & a_{12}/a_{11} & \dots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 0 & \dots & a_{2n}/a_{22} \\ \dots & \dots & \dots & \dots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & \dots & 0 \end{pmatrix}.$$

Здесь $\|R\|_\infty < 1$, поэтому матрица $I + R$ обратима и A представлена в виде произведения двух обратимых матриц. ■

[8]Напомним, что $\|A\|$ — операторная норма.

Замечание. Так как неособенность матрицы влечет и неособенность транспонированной, то для неособенности матрицы достаточно и "диагонального преобладания в столбцах".

Определение. Кругами Гершгорина матрицы A называются круги на комплексной плоскости

$$\Lambda_k = \{\, \lambda \in \mathbb{C} \;\big|\; |\lambda - a_{kk}| \le \sum_{j \ne k} |a_{kj}| \,\}.$$

Теорема 4. Все собственные числа матрицы A содержатся в объединении ее кругов Гершгорина.

Доказательство. Если $\lambda \notin \Lambda_k$, то $|a_{kk} - \lambda| > \sum_{j \ne k} |a_{kj}|$, так что если $\lambda \notin \cup \Lambda_k$, то $A - \lambda I$ матрица с диагональным преобладанием и потому неособенная. ■

Замечание. Точно так же все собственные числа содержатся в кругах Гершгорина транспонированной матрицы.

Пусть исходные данные в какой-то задаче вычисления известны нам неточно. Ошибка в решении, вызванная неточностью исходных данных, называется *неустранимой*. Займемся оценкой неустранимой ошибки в задачах обращения матриц и решения систем линейных уравнений.

Теорема 5. Пусть матрица A неособенная и матрица ΔA такова, что $\|A^{-1}\| \, \|\Delta A\| < 1$. Тогда матрица $A + \Delta A$ также неособенная и выполняются оценки

$$\|(A + \Delta A)^{-1}\| \le \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \, \|\Delta A\|}, \quad \|(A + \Delta A)^{-1} - A^{-1}\| \le \frac{\|A^{-1}\|^2 \|\Delta A\|}{1 - \|A^{-1}\| \, \|\Delta A\|}.$$

Доказательство. По следствию из теоремы 2 матрица $I + A^{-1}\Delta A$ обратима и

$$\|(I + A^{-1}\Delta A)^{-1}\| \le \frac{1}{1 - \|A\| \, \|\Delta A\|}.$$

Т.к. $A + \Delta A = A(I + A^{-1}\Delta A)$, то существует $(A+\Delta A)^{-1} = (I + A^{-1}\Delta A)^{-1} A^{-1}$, откуда сразу же следует первая из доказываемых оценок. Докажем вторую:

$$(A + \Delta A)^{-1} - A^{-1} = [I - A^{-1}(A + \Delta A)](A + \Delta A)^{-1} = -A^{-1}\Delta A (A + \Delta A)^{-1}$$

и остается применить первую, уже доказанную, оценку. ■

Обратимся к задаче решения систем линейных уравнений. Пусть x^* — решение системы линейных уравнений

$$Ax = y, \tag{4}$$

а $x^* + \Delta x$ — системы уравнений

$$(A + \Delta A)x = y + \Delta y.$$

Теорема 6. Пусть матрица A обратима и $\|A^{-1}\|\,\|\Delta A\| < 1$[9]. Тогда

$$\|\Delta x\| \le \frac{\|A^{-1}\|}{1-\|A^{-1}\|\,\|\Delta A\|}\big[\|\Delta A\|\,\|x^*\| + \|\Delta y\|\big].$$

Доказательство. Введем еще $x^* + \Delta_1 x$ — решение системы уравнений $Ax = y + \Delta y$. Тогда

$$\Delta_1 x = A^{-1}\Delta y, \quad \|\Delta_1 x\| \le \|A^{-1}\|\,\|\Delta y\|,$$

$$(A+\Delta A)(x^* + \Delta_1 x) = y + \Delta y + \Delta A(x^* + \Delta_1 x),$$
$$(A+\Delta A)(\Delta x - \Delta_1 x) = -\Delta A(x^* + \Delta_x).$$

Используя теорему 5:

$$\|\Delta x - \Delta_1 x\| \le \frac{\|A^{-1}\|\,\|\Delta A\|}{1-\|A^{-1}\|\,\|\Delta A\|}(\|x^*\| + \|\Delta_1 x\|) \le$$
$$\le \frac{\|A^{-1}\|\,\|\Delta A\|}{1-\|A^{-1}\|\,\|\Delta A\|}(\|x^*\| + \|A^{-1}\|\,\|\Delta y\|).$$

Остается, используя полученную оценку для $\|\Delta_1 x\|$, применить неравенство треугольника: $\|\Delta x\| \le \|\Delta x - \Delta_1 x\| + \|\Delta_1 x\|$. ■

Замечание. Для того чтобы избежать в правой части оценки неизвестной нам величины $\|x^*\|$, можно воспользоваться неравенством $\|x^*\| \le \|A^{-1}\|\,\|y\|$.

В теоремах 5 и 6 речь шла об *абсолютных* погрешностях. Введем *относительные* погрешности:

$$\delta A = \frac{\|\Delta A\|}{\|A\|}, \quad \delta y = \frac{\|\Delta y\|}{\|y\|}, \quad \delta A^{-1} = \frac{\|A+\Delta A)^{-1} - A^{-1}\|}{\|A^{-1}\|}, \quad \delta x = \frac{\|\Delta x\|}{\|x^*\|}.$$

Определение. Пусть A неособенная матрица. *Числом обусловленности матрицы* A называется[10]

$$\mu(A) = \|A\|\,\|A^{-1}\|.$$

Теорема 7. Пусть A обратимая матрица. Если $\mu(A)\delta A < 1$, то

$$\delta A^{-1} \le \frac{\mu(A)\delta A}{1-\mu(A)\delta A}, \quad \delta x \le \frac{\mu(A)}{1-\mu(A)\delta A}.$$

---

[9] Заметим, что в силу теоремы 5 тогда и $A + \Delta A$ обратима.

[10] Заметим, что число обусловленности, как и операторная матричная норма, связано с введенной в $\mathbb{R}^n$ векторной нормой. Если последняя помечена каким-нибудь значком, то тем же значком помечается и $\mu(A)$. Например, евклидовой векторной норме соответствует $\mu_2(A)$.

Доказательство. Поскольку $\mu(A)\delta A = \|A^{-1}\|\,\|\Delta A\|$, то в силу теоремы 5 $A + \Delta A$ обратима. По той же теореме

$$\delta A^{-1} \le \frac{\|A^{-1}\|\,\|\Delta A\|}{1 - \|A^{-1}\|\,\|\Delta A\|} = \frac{\mu(A)\delta A}{1 - \mu(A)\delta A}.$$

Перейдем к доказательству второй оценки. По теореме 6 (используя равенство $\|A^{-1}\|\,\|\Delta A\| = \mu(A)\delta A$)

$$\delta x \le \frac{1}{1 - \mu(A)\delta A}\left(\mu(A)\delta A + \|A^{-1}\|\frac{\|\Delta y\|}{\|x^*\|}\right).$$

Остается оценить $\|x^*\|$ снизу из неравенства $\|y\| = \|Ax^*\| \le \|A\|\,\|x^*\|$. ■

Задача 1. Показать, что в случае эрмитовой матрицы A

$$\mu_2(A) = \frac{\max|\lambda_k|}{\min|\lambda_k|},$$

где λ_k — собственные числа матрицы A.

Задача 2. Показать, что для любой матрицы A и любой операторной нормы выполняется равенство

$$\lim_{s\to\infty} \sqrt[s]{\|A^s\|} = \rho(A).$$

§3 Вопросы устойчивости в задаче на собственные значения

Рассмотрим вопрос, насколько могут изменяться собственные числа и векторы матрицы при малом ее возмущении. Начнем с примеров, которые иллюстрируют те трудности, которые могут здесь возникнуть.

Пример 1. Пусть $A_\varepsilon \in \mathbb{M}_n$ отличается от "канонического ящика Жордана" с нулями на главной диагонали лишь одним элементом: "в левом нижнем углу" стоит $(-1)^n\varepsilon$ $(\varepsilon \ge 0)$. У матрицы A_0 $\lambda = 0$ является собственным числом кратности n. При $\varepsilon > 0$ все собственные числа матрицы A_ε различны: $\lambda_k = \sqrt[n]{\varepsilon}e^{i2\pi k/n}$ $(k = 0, \dots, n-1)$. Отсюда два вывода. Во-первых, кратность собственного числа неустойчива по отношению к малым возмущениям матрицы. Во-вторых, в случае кратного собственного числа его возмущение может иметь меньший порядок малости, чем возмущение самой матрицы. Так, например, при $n = 10$ и $\varepsilon = 10^{-10}$ окажется $|\lambda_k| = 0.1$.

Пример 2. Рассмотрим матрицы

$$A_1 = \begin{pmatrix} 1+\varepsilon & 0 \\ 0 & 1-\varepsilon \end{pmatrix}, \qquad A_2 = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix};$$

у обеих этих близких при малом ε матрицах одни и те же собственные числа $1+\varepsilon$ и $1-\varepsilon$, но собственные векторы, нормированные условием $\|z\|_2 = 1$, у первой из них $(1,0)$ и $(0,1)$, а у второй $\frac{1}{\sqrt{2}}(1,1)$ и $\frac{1}{\sqrt{2}}(1,-1)$). Таким образом,

в случае наличия близких собственных чисел малые возмущения матрицы могут кардинально менять систему ее собственных векторов.

Дальше будем рассматривать случай простого собственного числа.

Пусть $A_0 \in \mathbb{M}$, λ_0 — простое собственное число этой матрицы и z_0 — соответствующий собственный вектор. Устойчивость λ_0 и z_0 означает, что малые изменения матрицы A_0 вызывают малые изменения λ_0 и z_0, т.е. при малой матрице ΔA матрица $A_0 + \Delta A$ будет иметь собственное число $\lambda_0 + \Delta\lambda$ и соответствующий собственный вектор $z_0 + \Delta z$, где $\Delta\lambda$ и Δz малы. Поскольку собственный вектор определен с точностью до множителя, естественно считать его нормированным каким-либо условием и тем же условием нормированным вектор $z_0 + \Delta z$. Для определенности будем предполагать, что для некоторого вектора y_0 отлично от нуля скалярное произведение (z_0, y_0) и в качестве нормирующего условия для собственного вектора выберем

$$(z, y_0) = 1. \tag{1}$$

Теорема 1. Если λ_0 есть простое собственное число матрицы A_0, то в некоторой окрестности этой матрицы собственные число и вектор (нормированный условием (1)) суть непрерывно дифференцируемые функции ее элементов[11].

Доказательство начнем с λ. Пусть $P(A, \lambda)$ — характеристический полином матрицы A. Его коэффициенты суть непрерывно дифференцируемые (и даже полиномиальные) функции элементов матрицы A. Поскольку λ_0 — простое собственное число, в точке (A_0, λ_0) будет $\frac{\partial}{\partial\lambda}P(A, \lambda) \neq 0$, и по теореме о неявных функциях в некоторой окрестности A_0 корень $\lambda(A)$ характеристического полинома (т.е. собственное число матрицы A) есть непрерывно дифференцируемая функция элементов матрицы A.

Матрица $A_0 - \lambda_0 I$ имеет нулевой определитель, и не умаляя общности будем считать, что ее первая строка есть линейная комбинация остальных строк. Каждой близкой к A_0 матрице A и числу λ, близкому к λ_0, сопоставим матрицу $B(A, \lambda)$, которая получается из матрицы $A - \lambda I$ заменой ее первой строки на вектор y_0. Тогда матрица $B(A_0, \lambda_0)$ неособенная, поскольку вектор z_0 есть *единственное* решение системы уравнений $B(A_0, \lambda_0)z = e_1$. Если возмущения матрицы A_0 и числа λ_0 достаточно малы, то матрица $B(A, \lambda)$ также будет неособенной, причем элементы обратной матрицы $B^{-1}(A, \lambda)$ суть непрерывно дифференцируемые функции матрицы A и λ. Именно только такие A и λ мы будем теперь рассматривать.

Система уравнений $B(A, \lambda)z = e_1$ определяет z как неявную непрерывно дифференцируемую функцию аргументов A и λ: $z(A, \lambda)$. Тогда $z(A) =$

[11]Это означает, что в некоторой области $\|A - A_0\| < \varepsilon$ ($\varepsilon > 0$) существуют непрерывно дифференцируемые функции $\lambda(A)$ и $z(A)$ элементов матрицы A, такие что $\lambda(A_0) = \lambda_0$, $z(A_0) = z_0$ и при всех A из этой области $\lambda(A)$ есть собственное число матрицы A, а $z(A)$ — соответствующий ему собственный вектор, нормированный условием (1). Как будет видно из доказательства, можно утверждать, что $\lambda(A)$ и $z(A)$ являются бесконечно дифференцируемыми функциями.

$z\big(A, \lambda(A)\big)$ есть непрерывно дифференцируемая функция элементов матрицы A. Остается показать, что $z(A)$ есть собственный вектор матрицы A, соответствующий собственному числу $\lambda(A)$. Действительно, строки матрицы $A - \lambda(A)I$, начиная со второй по последнюю, линейно независимы, т.к. они же — строки неособенной матрицы $B(A, \lambda(A))$. Но сама матрица $A - \lambda(A)I$ особенная, и поэтому ее первая строка есть линейная комбинация остальных. Поэтому, будучи ортогональным строкам со второй по n-ю этой матрицы, вектор $z(A)$ ортогонален и первой, т.е. есть собственный вектор матрицы A. ■

Степень устойчивости простого собственного числа и его собственного вектора зависит от величины их производных по элементам матрицы A. Получать соответствующие оценки мы будем в предположении, что *все* собственные числа матрицы A_0 простые.

Итак, пусть $\lambda_1, \dots, \lambda_n$ — все и притом различные собственные числа матрицы A_0, $z_1, \dots, z_n$ — соответствующие им собственные векторы. Комплексно сопряженные $\overline{\lambda}_1, \dots, \overline{\lambda}_n$ — собственные числа сопряженной матрицы A_0^*, и пусть $y_1, \dots, y_n$ — соответствующие собственные векторы. Эти последние мы считаем фиксированными и в качестве нормирующего условия для вектора z_k и соответствующего вектора возмущенной матрицы выберем $(z, y_k) = 1$, так что $(z_k, y_j) = \delta_{kj}$.

Пусть ΔA — малое возмущение матрицы A_0, а $\Delta\lambda_k$ и Δz_k — вызванные им возмущения собственных чисел и векторов (в силу нормирующего условия $(\Delta z_k, y_k) = 0$). Тогда

$$(A_0 + \Delta A)(z_k + \Delta z_k) = (\lambda_k + \Delta\lambda_k)(z_k + \Delta z_k).$$

Переходя к дифференциалам:

$$(dA)z_k + A_0\, dz_k = \lambda_k\, dz_k + d\lambda_k z_k. \tag{2}$$

Умножим это равенство скалярно на y_j:

$$\big((dA)z_k, y_j\big) + \big(A_0(dz_k), y_j\big) = \lambda_k(dz_k, y_j) + d\lambda_k(z_k, y_j). \tag{3}$$

Здесь $\big(A_0(dz_k), y_j\big) = (dz_k, A_0^* y_j) = \lambda_j(dz_k, y_j)$, так что (3) может быть переписано в виде

$$\big((dA)z_k, y_j\big) = (\lambda_k - \lambda_j)(dz_k, y_j) + d\lambda_k(z_k, y_j). \tag{4}$$

В силу нормирующего условия $(z_k, y_k) = 1$, поэтому при $j = k$ равенство (4) дает

$$|d\lambda_k| = \big|\big((dA)z_k, y_k\big)\big| \le \|dA\|_2 \|z_k\|_2 \|y_k\|_2 = c_k \|dA\|_2.$$

Здесь

$$c_k = \|z_k\|_2 \|y_k\|_2 = \frac{\|z_k\|_2 \|y_k\|_2}{|(z_k, y_k)|}.$$

Последняя формула лучше в том отношении, что она не связана с нормирующим условием — написанное отношение инвариантно относительно *любой*

нормировки векторов z_k и y_k. Число c_k называется *коэффициентом перекоса* матрицы A_0, соответствующим собственному числу λ_k. Очевидно, что всегда $c_k \geq 1$, а для эрмитовых матриц $c_k = 1$. Итак, задача нахождения простого собственного числа матрицы хорошо обусловлена, если только соответствующий этому числу коэффициент перекоса невелик.

Обратимся к собственным векторам. При $j \neq k$ $(z_k, y_j) = 0$ и из формулы (4) следует

$$(dz_k, y_j) = \big((dA)z_k, y_j\big)/(\lambda_k - \lambda_j).$$

Разложим вектор dz_k по векторам z_j: $dz_k = \sum_{j=1}^{n} \alpha_{kj} z_j$, $\alpha_{kj} = (dz_k, y_j)$. Тогда в силу нормирующего условия $\alpha_{kk} = 0$ и при $j \neq k$

$$|\alpha_{kj}| = \frac{|\big((dA)z_k, y_j\big)|}{|\lambda_k - \lambda_j|} \leq \frac{\|dA\|_2 \|z_k\|_2 \|y_j\|_2}{|\lambda_k - \lambda|},$$

откуда

$$\frac{\|dz_k\|_2}{\|z_k\|_2} \leq \|dA\|_2 {\sum_{j=1}^{n}}' \frac{\|z_j\|_2 \|y_j\|_2}{|\lambda_k - \lambda_j|} = \|dA\|_2 {\sum_{j=1}^{n}}' \frac{c_j}{|\lambda_k - \lambda_j|}.$$

Штрих у суммы означает, что слагаемое с номером k отсутствует — $\alpha_{kk} = 0$ ввиду нормирующего условия. Итак, задача отыскания собственного вектора, соответствующего простому собственному числу, хорошо обусловлена, если это собственное число достаточно далеко отстоит от остальных собственных чисел и все коэффициенты перекоса невелики. Впрочем, некоторую роль здесь играет еще порядок n матрицы.

В случае эрмитовой матрицы A можно указать оценку возмущения собственных чисел, вызванного возмущением ΔA, если ΔA также эрмитова матрица. Эта оценка связана с экстремальными свойствами собственных чисел эрмитовых матриц.

Пусть матрица $A \in \mathbb{M}_n$ эрмитова, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ — ее собственные числа, $z_1, \ldots, z_n$ — соответствующие собственные векторы, выбранные так, что $(z_k, z_j) = \delta_{kj}$.

Теорема 2. Справедливы равенства:

$$\lambda_k = \max\{\, (Ax, x) \;\big|\; \|x\|_2 = 1, (x, z_j) = 0 \;\text{ при } j = 1, \ldots, k-1 \,\}.$$

Доказательство. Вектор x, удовлетворяющий выписанным условиям, разложим по векторам z_ν. Коэффициенты при z_ν для $\nu = 1, \ldots, k-1$ окажутся равными нулю, и $x = \sum_{\nu=k}^{n} \alpha_\nu z_\nu$, причем $\sum_{\nu=k}^{n} |\alpha_\nu|^2 = \|x\|_2^2 = 1$. Поэтому

$$(Ax, x) = \sum_{\nu=k}^{n} \lambda_\nu |\alpha_\nu|^2 \leq \lambda_k \sum_{\nu=k}^{n} |\alpha_\nu|^2 = \lambda_k.$$

В то же время $(Az_k, z_k) = \lambda_k$. ■

Введем функцию аргументов $h_j \in \mathbb{C}^n$:

$$\varphi_A(h_1, \ldots, h_{k-1}) = \max\{\, (Ax, x) \;\big|\; \|x\|_2 = 1, (x, h_j) = 0, j = 1, \ldots, k-1 \,\}.$$

Теорема 3 (минимально-максимальный принцип Куранта). Справедливо равенство

$$\min_{h_j} \varphi_A(h_1, \ldots, h_{k-1}) = \lambda_k = \varphi_A(z_1, \ldots, z_{k-1}).$$

Доказательство. Второе из доказываемых равенств следует из теоремы 2. Поэтому $\min \varphi_A \le \lambda_k$. Докажем обратное неравенство. Рассмотрим для произвольных h_j систему однородных линейных уравнений

$$\sum_{\nu=1}^{k} \alpha_\nu (z_\nu, h_j) = 0, \quad j = 1, \ldots, k-1.$$

Пусть $\{\alpha_\nu\}$ ее ненулевое решение, нормированное условием $\sum |\alpha_\nu|^2 = 1$. Положим $x = \sum_{\nu=1}^{k} \alpha_\nu z_\nu$. Тогда $\|x\|_2 = 1$, при $j = 1, \ldots, k-1 \quad (x, h_j) = 0$ и $(Ax, x) = \sum_{\nu=1}^{k} |\alpha_\nu|^2 \lambda_\nu \ge \lambda_k$. Итак, при любых $h_j \quad \varphi_A(h_1, \ldots, h_{k-1}) \ge \lambda_k$. ■

Лемма. Пусть $A, B \in \mathbb{M}_n$ эрмитовы, λ_k^A и λ_k^B их собственные числа, расположенные в порядке убывания, $\mu_1 \ge \cdots \ge \mu_n$ — собственные числа матрицы $A + B$. Тогда $\mu_k \le \lambda_k^A + \lambda_1^B$.

Доказательство. Пусть $\{z_k\}$ собственные векторы матрицы A. Рассмотрим множество

$$\mathcal{M} = \{\, x \in \mathbb{C}^n \;\mid\; \|x\|_2 = 1, (x, z_j) = 0 \;\text{ при } j = 1, \ldots, k-1 \,\}.$$

По теореме 2 для $x \in \mathcal{M} \;\; (Ax, x) \le \lambda_k^A$ и $(Bx, x) \le \lambda_1^B$ и потому $\big((A+B)x, x\big) \le \lambda_k^A + \lambda_1^B$. Теперь имеем

$$\mu_k = \min \varphi_{A+B}(h_1, \ldots, h_{k-1}) \le \varphi_{A+B}(z_1, \ldots, z_{k-1}) = \\ = \max\{\, \big((A+B)x, x\big) \;\mid\; x \in \mathcal{M} \,\} \le \lambda_k^A + \lambda_1^B. \quad \blacksquare$$

Следствие. В условиях леммы $\mu_k \le \lambda_k^A + \|B\|$ ($\|B\|$ — любая операторная норма матрицы B).

Теорема 4. Пусть матрицы A и ΔA эрмитовы и λ_k и $\lambda_k + \Delta\lambda_k$ собственные числа матриц A и $A + \Delta A$, расположенные в порядке убывания. Тогда для любой операторной нормы матриц $|\Delta\lambda_k| \le \|\Delta A\|$.

Доказательство. Согласно следствию $\lambda_k + \Delta\lambda_k \le \lambda_k + \|\Delta A\|$, так что $\Delta\lambda_k \le \|\Delta A\|$. Применяя то же следствие к матрице $A = (A + \Delta A) - \Delta A$: $\lambda_k \le \lambda_k + \Delta\lambda_k + \|\Delta A\|$ и $-\|\Delta A\| \le \Delta\lambda_k$. ■

§4 Метод исключений Гаусса

Метод исключений для решения системы n линейных уравнений заключается в следующем. Из первого уравнения первая неизвестная выражается через остальные, это выражение подставляется в остальные уравнения, и мы получаем систему $(n-1)$-го порядка относительно оставшихся неизвестных, с которой поступаем так же, и т.д. Это равносильно (на первом шаге) к переходу от системы уравнений с расширенной матрицей

$$A[1:n, 1:n+1] = \{a_{kj}\} = \begin{pmatrix} a_{11} & \ldots & a_{1n} & a_{1n+1} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} & a_{nn+1} \end{pmatrix}$$

к равносильной ей системе с расширенной матрицей

$$A_1 = \begin{pmatrix} 1 & b_2^1 & \dots & b_n^1 & b_{n+1}^1 \\ 0 & a_{22}^1 & \dots & a_{2n}^1 & a_{2n+1}^1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & a_{n2}^1 & \dots & a_{nn}^2 & a_{nn+1}^2 \end{pmatrix},$$

элементы которой вычисляются по формулам

$$b_j^1 = \frac{a_{1j}}{a_{11}}, \qquad a_{kj}^1 = a_{kj} - b_j^1 a_{k1},$$

т.е. первая строка матрицы A делится на a_{11}, а из остальных строк вычитается эта новая первая, умноженная на коэффициент при первой неизвестной. Полученную первую строку мы больше не трогаем, и на следующем шаге вторую строку делим на a_{22}^1 и , домножая на соответствующие коэффициенты, вычитаем из следующих строк. В результате после n таких шагов мы получим эквивалентную исходной систему уравнений с расширенной матрицей

$$A_n = \begin{pmatrix} 1 & b_2^1 & b_3^1 & \dots & b_{n-1}^1 & b_n^1 & b_{n+1}^1 \\ 0 & 1 & b_3^2 & \dots & b_{n-1}^2 & b_n^2 & b_{n+1}^2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & 1 & b_{n+1}^n \end{pmatrix},$$

из которой легко находятся последовательно все неизвестные:

$$\left.\begin{array}{lcllll} \xi_n & = & b_{n+1}^n & & & \\ \xi_{n-1} & = & b_{n+1}^{n-1} & -b_n^{n-1}\xi_n & & \\ \dots & & \dots & \dots & \dots & \dots \\ \xi_1 & = & b_{n+1}^1 & -b_2^1\xi_2 & -\dots & -b_n^1\xi_n \end{array}\right\}.$$

Весь процесс решения требует порядка $\frac{2}{3}n^3$ арифметических операций.

Если требуется решить несколько (m) систем уравнений с одной и той же матрицей коэффициентов и различными правыми частями, то удобно это делать одновременно, преобразуя так же, как и раньше, расширенную матрицу коэффициентов, которая будет теперь содержать $n + m$ столбцов.

Метод в изложенной форме окажется неприменим, если $a_{11} = 0$ или подобное обстоятельство встретится на одном из последующих шагов, например, $a_{22}^1 = 0$. Но плохо и в том случае, если, например, элемент $a_{22}^1 \neq 0$, но мал по абсолютной величине. Он получен путем вычитания, и его малость означает обычно, что велика его относительная погрешность (мы ведем счет с округлением результатов арифметических действий), а тогда и все последующие результаты будут иметь малую точность. Избежать этого можно, если на каждом шаге совершать перенумерацию уравнений или неизвестных, или того и другого. В связи с этим различают схемы исключения с выбором "ведущего элемента" по строке, по столбцу или по всей матрице. На первом шаге в первом случае это означает такое изменение порядка уравнений, чтобы после перестановки строк у новой матрицы оказалось $|a_{11}| \geq |a_{k1}|$ $(k = 2, \dots, n)$, во втором — такую перенумерацию неизвестных, чтобы было

$|a_{11}| \geq |a_{1j}|$ $(j = 2, \ldots, n)$, а в последнем — такую перестановку уравнений и перенумерацию неизвестных, чтобы оказалось $|a_{11}| \geq |a_{kj}|$ при $k, j = 1, \ldots, n$. Подобная перенумерация осуществляется при переходе к каждому следующему шагу.

§5 Итеративные методы решения систем

Метод простой итерации.

Пусть система линейных алгебраических уравнений записана в виде

$$x = Ax + y. \tag{1}$$

Метод итерации заключается в том, что, взяв произвольный вектор (начальное приближение) x_0, строят последовательность векторов

$$x_s = Ax_{s-1} + y.$$

Очевидно, что если $x_s \to x^*$, то x^* — решение системы (1). Условимся решение системы всегда обозначать x^*.

Теорема 1. Для того чтобы при любом начальном приближении x_0 было $x_s \to x^*$, необходимо и достаточно условие $\rho(A) < 1$.

Доказательство. Из равенств $x^* = Ax^* + y$, $x_s = Ax_{s-1} + y$ следует, что $x_s - x^* = A(x_{s-1} - x^*) = A^s(x_0 - x^*)$. Поэтому для сходимости метода при любом начальном приближении необходимо и достаточно $A^s \to 0$, и остается воспользоваться теоремой 1 из §2 и добавить, что при $\rho(A) < 1$ решение x^* существует и единственно. ■

Замечание 1. Отметим связь метода итерации с рассмотренным в §2 рядом:

$$x_s = A^s x_0 + (I + A + A^2 + \cdots + A^{s-1})y.$$

Замечание 2. В случае $\rho(A) \geq 1$ метод расходится почти при любом начальном приближении x_0 — если только при разложении $x_0 - x^*$ по собственным векторам матрицы A хоть один из коэффициентов при собственных векторах, отвечающих собственным числам, таким что $|\lambda| \geq 1$, отличен от нуля.

Пусть теперь в $\mathbb{C}^n$ задана некоторая норма $\|\cdot\|$, и нормы матриц ниже — это операторные нормы, порожденные введеной векторной.

Теорема 2. Если $\|A\| < 1$, то $x_s \to x^*$ и выполняются оценки:

$$\|x_s - x^*\| \leq \frac{\|A\|^s}{1 - \|A\|}\|x_1 - x_0\|, \tag{2}$$

$$\|x_s - x^*\| \leq \frac{\|A\|}{1 - \|A\|}\|x_s - x_{s-1}\|. \tag{3}$$

Доказательство. Сходимость следует из теоремы 1 и неравенства $\rho(A) \leq \|A\|$. Докажем оценку (2) Учитывая равенства $x_0 - x_1 = (I - A)x_0 - y$ и $x^* = (I - A)^{-1}y$, а также оценку $\|(I - A)^{-1}\| \leq 1/(1 - \|A\|)$, имеем

$$\|x_0 - x^*\| \leq \|(I - A)^{-1}\|\,\|(I - A)x_0 - y\| \leq \frac{1}{1 - \|A\|}\|x_1 - x_0\|$$

и остается воспользоваться тем, что $x_s - x^* = A^s(x_0 - x^*)$.

Неравенство (3) немедленно следует из (2), если принять $\tilde{x}_0 = x_{s-1}$, и тогда $\tilde{x}_1 = x_s$. ■

Замечание. Оценка (2) является *априорной*. Это значит, что ее правая часть может быть вычислена *до того*, как построено само приближение x_s и она годится для того, чтобы заранее оценить, сколько шагов метода следует сделать, чтобы добиться необходимой точности. Оценка (3) *апостериорная* — ее правая часть вычисляется, когда x_s уже известно, и может служить для принятия решения о прекращении итераций, когда необходимая точность уже достигнута. Заметим одно преимущество оценки (3) — она остается верной, какие бы ошибки ни были допущены при вычислении x_j при $j = 1, \ldots, s-1$.

Займемся вопросом о приведении системы уравнений к виду (1), если первоначально она задана в виде $Bx = z$. Остановимся на двух случаях.

1) Пусть $B = \{b_{kj}\}$ — матрица с диагональным преобладанием. Тогда систему $Bx = z$ можно переписать в виде $x = Ax + y$, где

$$A = \begin{pmatrix} 0 & -b_{12}/b_{11} & \ldots & -b_{1n}/b_{11} \\ \ldots & \ldots & \ldots & \ldots \\ -b_{n1}/bnn & -bn2/b_{nn} & \ldots & 0 \end{pmatrix}, y = \begin{pmatrix} \zeta_1/b_{11} \\ \ldots \\ \zeta_n/b_{nn} \end{pmatrix}. \qquad (4)$$

Очевидно, что $\|A\|_\infty < 1$, и потому метод итерации будет сходиться.

2) Пусть B эрмитова матрица. Взяв $\alpha \neq 0$, приведем систему к эквивалентной: $x = x - \alpha(Bx - z)$, так что $A = I - \alpha B$, и займемся вопросом о выборе α. Пусть μ_k — собственные числа матрицы B. Тогда собственные числа матрицы A суть $\lambda_k = 1 - \alpha\mu_k$. Если собственные числа матрицы B имеют разные знаки, то среди λ_k найдется хоть одно, большее единицы. Поэтому выбор числа α, гарантирующий сходимость метода итерации, возможен лишь при условии, что все μ_k имеют один знак. Для определенности будем считать, что они положительны (тогда эрмитова матрица B называется положительной), так что следует выбрать $\alpha > 0$. Мы заинтересованы в том, чтобы $\|A\|_2 = \rho(A) = \max\{\alpha M - 1, 1 - \alpha m\}$, где $m = \min \mu_k$, $M = \max \mu_k$, была минимальной. Этому соответствует выбор $\alpha = \frac{2}{M+m}$, тогда $\rho(A) = \frac{M-m}{M+m} < 1$. Заметим, что если собственные числа B имеют сильный разброс (M во много раз больше, чем m), то $\rho(A)$ близко к 1, и сходимость метода итерации может оказаться довольно медленной. Собственные числа матрицы B обычно неизвестны. Легко видеть, что если взять $0 < \alpha < \frac{1}{M'}$, где M' — любая верхняя граница для μ_k (например, $M' = \|B\|$), то окажется $\rho(A) < 1$, и метод итерации будет сходиться.

Метод Зайделя.

Это — другой итеративный способ решения системы (1). Предполагается заданным некоторое начальное приближение к решению x_0. Введем обозначения для компонент последующих приближений: $x_s = (\xi_1^s, \ldots, x_n^s)$. Вычислительные формулы метода:

$$\left.\begin{array}{rl} \xi_1^{s+1} = & a_{11}\xi_1^s + a_{12}\xi_2^s + \cdots + a_{1n}\xi_n^s + \eta_1 \\ \xi_2^{s+1} = & a_{21}\xi_1^{s+1} + a_{22}\xi_2^s + \cdots + a_{2n}\xi_n^s + \eta_2 \\ & \cdots\cdots\cdots \\ \xi_n^{s+1} = & a_{n1}\xi_1^{s+1} + \cdots + a_{nn-1}\xi_{n-1}^{s+1} + a_{nn}\xi_n^s + \eta_n \end{array}\right\}.$$

Эти формулы можно записать так: $x_{s+1} = Rx_s + Lx_{s+1} + y$, где

$$R = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix}, L = \begin{pmatrix} 0 & 0 & \dots & 0 \\ a_{21} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn-1} & 0 \end{pmatrix}, \quad A = R + L,$$

или $x_{s+1} = (I - L)^{-1}Rx_s + (I - L)^{-1}y$. Из теоремы 1 сразу же вытекает, что для сходимости метода Зайделя при любом начальном приближении x_0 необходимо и достаточно неравенство $\rho\big((I-L)^{-1}R\big) < 1$. Это условие трудно проверять, и мы докажем некоторый достаточный признак сходимости, но сначала отметим, что известны примеры, когда метод простой итерации сходится, а метод Зайделя нет, и наоборот, когда сходится метод Зайделя, но не метод простой итерации.

Теорема 3. Если $\|A\|_\infty < 1$, то метод Зайделя сходится при любом начальном приближении.

Доказательство. Положим

$$\beta_i = \sum_{j=1}^{i-1} |a_{ij}|, \quad \gamma_i = \sum_{j=i}^{n} |a_{ij}|; \quad \max_i(\beta_i + \gamma_i) = \|A\|_\infty < 1.$$

Тогда

$$\beta_i + \gamma_i - \frac{\gamma_i}{1-\beta_i} = \frac{\beta_i(1-\beta_i-\gamma_i)}{1-\beta_i} \geq 0, \qquad \frac{\gamma_i}{1-\beta_i} \leq \beta_i + \gamma_i \leq \|A\|_\infty.$$

Так как $x^* = Rx^* + Lx^* + y$, то $x^* - x_{s+1} = L(x^* - x_{s+1}) + R(x^* - x_s)$. Пусть $\|x^* - x_{s+1}\|_\infty = |\xi^*_{i_0} - \xi^{s+1}_{i_0}|$. При этом

$$\xi^*_{i_0} - \xi^{s+1}_{i_0} = \sum_{j=1}^{i_0-1} a_{i_0 j}(\xi^*_j - \xi^{s+1}_j) + \sum_{j=i_0}^{n} a_{i_0 j}(\xi^*_j - \xi^s_j),$$

так что

$$\|x^* - x_{s+1}\|_\infty \leq \beta_{i_0}\|x^* - x_{s+1}\|_\infty + \gamma_{i_0}\|x^* - x_s\|_\infty,$$

$$\|x^* - x_{s+1}\|_\infty \leq \frac{\gamma_{i_0}}{1-\beta_{i_0}}\|x^* - x_s\|_\infty \leq \|A\|_\infty\|x^* - x_s\|_\infty.$$

Так как последнее неравенство выполняется при всех s, то дальнейшее очевидно. ■

Метод Некрасова

Если первоначально система уравнений задана в виде $Bx = z$, причем все диагональные элементы матрицы B отличны от нуля, то она приводится к виду (1), если для A и y воспользоваться формулами (4). Метод Зайделя, примененный к полученной таким путем системе (1), называется *методом Некрасова* для исходной системы. Таким образом, компоненты следующего

приближения в методе Некрасова находятся по компонентам предыдущего из уравнений

$$\left.\begin{array}{cccccc} b_{11}\xi_1^{s+1} & +b_{12}\xi_2^{s} & +b_{13}\xi_3^{s} & +\ldots & +b_{1n}\xi_n^{s} & =\zeta_1 \\ b_{21}\xi_1^{s+1} & +b_{22}\xi_2^{s+1} & +b_{23}\xi_3^{s} & +\ldots & +b_{2n}\xi_n^{s} & =\zeta_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ b_{n1}\xi_1^{s+1} & +b_{n2}\xi_2^{s+1} & +b_{n3}\xi_3^{s+1} & +\ldots & +b_{nn}\xi_n^{s+1} & =\zeta_n \end{array}\right\}.$$

Теорема 4. Если выполнено хотя бы одно из условий

а) матрица B имеет диагональное преобладание,

б) матрица B положительно определена,

то метод Некрасова сходится при любом начальном приближении.

Доказательство. В случае а) утверждение теоремы непосредственно следует из теоремы 3.

Случай б). Положительно определенную матрицу B представим в виде суммы $B = R + D + R^*$, где

$$D = \begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{nn} \end{pmatrix}, R = \begin{pmatrix} 0 & a_{12} & \ldots & a_{1n} \\ 0 & 0 & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 \end{pmatrix}.$$

Тогда формулы метода Некрасова приводятся к виду

$$(D + R^*)x_{s+1} + Rx_s = y, \qquad x_{s+1} = -(D + R^*)^{-1}Rx_s + (D + R^*)^{-1}y,$$

и для сходимости метода достаточно показать, что $\rho\big((D + R^*)^{-1}R\big) < 1$. Пусть λ и z собственное число и соответствующий собственный вектор этой матрицы, так что

$$Rz = \lambda Dz + \lambda R^* z = \lambda Bz - \lambda Rz.$$

Введем обозначения:

$$(Bz, z) = p > 0, \quad (Dz, z) = d > 0, \quad (Rz, z) = a + ib, \quad (R^*z, z) = a - ib.$$

Тогда $p = d + 2a$, и т.к. $(Rz, z) = \lambda(Bz, z) - \lambda(Rz, z)$, то $a + ib = \lambda(p - a - ib)$. Далее

$$\lambda = \frac{a + ib}{p - a - ib}, \quad |\lambda|^2 = \frac{a^2 + b^2}{(p - a)^2 + b^2} = \frac{a^2 + b^2}{a^2 + b^2 + p(p - 2a)} = \frac{a^2 + b^2}{a^2 + b^2 + pd} < 1.$$

Поскольку λ любое собственное число, нужное неравенство для $\rho\big((D+R^*)^{-1}R\big)$ доказано. ■

Задача 1. Доказать, что правая часть оценки (3) всегда не больше правой части (2).

Задача 2. Показать, что в случае метода простой итерации по любым $\varepsilon > 0$ и начальному приближению x_0 найдется такая постоянная C_ε, что

$$\|x_s - x^*\| \le C_\varepsilon[\rho(A) + \varepsilon]^s.$$

§6 Обращение матриц

Построить матрицу $D = A^{-1}$, обратную данной, можно используя метод исключений Гаусса, поскольку j-й столбец матрицы D есть решение системы уравнений $Ax = e_j$. В §4 уже обращалось внимание, что решать несколько систем уравнений с одной и той же матрицей коэффициентов, но различными правыми частями удобно одновременно, в единой схеме. Именно так и следует поступать при обращении матриц, решая все n систем уравнений с правыми частями e_j одновременно.

При наличии не слишком плохого приближения D_0 к обратной матрице можно для построения обратной матрицы использовать итеративный процесс

$$D_{s+1} = D_s(2I - AD_s). \tag{1}$$

Стоит обратить внимание, что эта вычислительная формула вполне аналогична формуле метода Ньютона (касательных) для решения числового уравнения $\frac{1}{t} - a = 0$.

Формулу (1) удобно переписать в виде

$$R_s = I - AD_s, \quad D_{s+1} = D_s(I + R_s) \quad s = 0, 1, \ldots \tag{2}$$

Теорема. Для выполнения соотношения $D_s \to A^{-1}$ необходимо и достаточно выполнение неравенства $\rho(R_0) < 1$. Если при этом[12] $\|R_0\| = q < 1$, то выполняется оценка

$$\|D_s - A^{-1}\| \le \|D_0\| \frac{q^{2^s}}{1-q}. \tag{3}$$

Доказательство. Заметим прежде всего, что если $\rho(R_0) < 1$, то матрица $I - R_0 = AD_0$ обратима, а значит, обратима и A. Преобразуем формулу для R_{s+1}:

$$R_{s+1} = I - AD_{s+1} = I - AD_s(I + R_s) = I - AD_s - AD_sR_s = R_s - AD_sR_s = R_s^2.$$

Докажем необходимость. Если $D_s \to A^{-1}$, то $R_s \to 0$. Поскольку $R_s = R_0^{2^s}$, то $\rho(R_0) < 1$ является для этого необходимым условием.

Докажем достаточность и оценку (3). Если $\rho(R_0) < 1$, то $R_s = R_0^{2^s} \to 0$, т.е. $AD_s \to I$ и $D_s \to A^{-1}$. Если $\|R_0\| = q < 1$, то $\|R_s\| = \|R_0^{2^s}\| \le \|R_o\|^{2^s} = q^{2^s}$. Далее

$$A^{-1} - D_s = A^{-1}R_s, \quad \|A^{-1} - D_s\| \le \|A^{-1}\| q^{2^s}$$

и остается заметить, что

$$AD_0 = I - R_0, \quad A = (I - R_0)D_0^{-1}, \quad A^{-1} = D_0(I - R_0)^{-1},$$
$$\|A^{-1}\| \le \|D_0\| \, \|(I - R_0)^{-1}\|,$$

[12]Как обычно считаем заданной некоторую векторную норму, и матричная — операторная, порожденная этой векторной.

причем $\|(I-R_0)^{-1} \le \frac{1}{1-q}$ ■

Замечание. Следует подчеркнуть, что оценка (3) означает очень быструю сходимость процесса (2).

Следствие. Если матрица A эрмитова и обратима, то при $D_0 = \alpha A$, где $0 < \alpha < \frac{2}{(\rho(A))^2}$, процесс (2) сходится.

Доказательство. Если λ_k — собственные числа матрицы A (они вещественны), то при указанном выборе D_0 собственными числами матрицы R_0 будут $1 - \alpha\lambda_k^2$. Все они меньше 1, а неравенство $\alpha < \frac{2}{((\rho(A))^2}$ гарантирует, что они больше -1. ■

Следствие гарантирует сходимость процесса для эрмитовой матрицы, если взять, например, $D_0 = \frac{1}{\|A\|^2}A$ при любой операторной норме матрицы.

§7 Степенной метод

В задаче на отыскание собственных чисел и векторов матрицы A различают полную и частичную проблему собственных чисел. В первом случае требуется найти все собственные числа и векторы, а во втором — одно собственное число (чаще всего максимальное по модулю) и соответствующий собственный вектор. Степенной метод — это метод решения частичной проблемы собственных чисел.

Идея метода основана на анализе поведения последовательности векторов

$$y_k = A^k y_0 = A y_{k-1}, \tag{1}$$

где y_0 произвольно взятый ненулевой вектор. Будем считать, что матрица A имеет полную систему собственных векторов, λ_j — ее собственные числа, а z_j — соответствующие собственные векторы, причем $|\lambda_1| \ge \cdots \ge |\lambda_n|$. Пусть $y_0 = \sum_{j=1}^n \alpha_j z_j$ — разложение вектора y_0 по собственным векторам. Тогда

$$y_k = \sum_{j=1}^n \alpha_j \lambda_j^k z_j = \lambda_1^k \left[\alpha_1 z_1 + \sum_{j=2}^n \left(\frac{\lambda_j}{\lambda_1}\right)^k \alpha_j z_j \right]. \tag{2}$$

Основной случай

Будем сейчас считать, что $|\lambda_1| > |\lambda_2|$ (основной случай) и что $\alpha_1 \ne 0$ (это условие обычно выполнено для произвольно взятого y_0). Тогда суммой, стоящей в правой части (2), при больших k можно пренебречь, и мы имеем приближенное равенство $y_k \approx \lambda_1^k \alpha_1 z_1$. Пусть u — некоторый вектор, такой что $(z_1, u) \ne 0$. Положим

$$\varphi_k = (y_k, u) = \lambda_1^k \left[\alpha_1 (z_1, u) + \sum_{j=2}^n \left(\frac{\lambda_j}{\lambda_1}\right)^k \alpha_j (z_j, u) \right].$$

Тогда, как видно из этой формулы, $\varphi_k/\varphi_{k-1} \to \lambda_1$ и, более того,

$$\frac{\varphi_k}{\varphi_{k-1}} = \lambda_1 + \mathcal{O}\left(\left(\frac{\lambda_2}{\lambda_1}\right)^k\right).$$

Если $|\lambda_1| < 1$, то векторы y_k стремятся к нулевому, а при $|\lambda| > 1$ — к бесконечности (по норме). Это обычно неудобно, и поэтому на каждом шаге процесса часто осуществляют некоторую нормировку этих векторов, так что процесс выглядит так:

$$\tilde{x}_0 = y_0, \quad x_k = \frac{\tilde{x}_k}{(\tilde{x}_k, u)}, \quad \tilde{x}_{k+1} = Ax_k. \tag{3}$$

Векторы $\tilde{x}_k$ и x_k лишь числовыми множителями отличаются от вектора y_k: $x_k = \beta_k y_k$, и для введенных выше чисел φ_k имеем

$$\frac{\varphi_k}{\varphi_{k-1}} = \frac{(y_k, u)}{(y_{k-1}, u)} = \frac{(Ay_{k-1}, u)}{(y_{k-1}, u)} = \frac{(Ax_{k-1}, u)}{(x_{k-1}, u)} = (\tilde{x}_k, u),$$

поскольку $(x_{k-1}, u) = 1$, так что приближения к λ_1 и z_1 даются формулами

$$\lambda_1^{(k)} = (\tilde{x}_k, u), \quad z_1^k = x_k.$$

Вектор x_k, будучи нормирован условием $(x_k, u) = 1$, с точностью до вектора порядка малости $\mathcal{O}\big((\lambda_2/\lambda_1)^k\big)$ совпадает с собственным вектором z_1, если последний нормирован тем же условием.

В качестве нормирующего вектора u удобно выбирать один из ортов e_j. После нескольких шагов процесса, когда наметилась некоторая стабилизация векторов x_k, целесообразно в качестве j взять номер максимальной по модулю компоненты x_k. Фактически доказана теорема о сходимости метода, при формулировке которой будут использоваться введенные выше обозначения; кроме того, мы будем предполагать, что при всех k $(y_k, u) \neq 0$.

Теорема. Пусть выполнены условия:
1) $|\lambda_1| > |\lambda_2|$;
2) $(z_1, u) = 1$;
3) $\alpha_1 \neq 0$.
Тогда описанный формулами (3) процесс сходится: $\lambda_1^{(k)} \to \lambda_1$, $z_1^k \to z_1$, причем

$$\lambda_1^{(k)} = \lambda_1 + \mathcal{O}\left(\left(\frac{\lambda_2}{\lambda_1}\right)^k\right), \quad z_1^k = z_1 + \Delta z^k, \quad \|\Delta z^k\| = \mathcal{O}\left(\left(\frac{\lambda_2}{\lambda_1}\right)^k\right).$$

Замечание. Если в условиях теоремы $|\lambda_2| > |\lambda_3|$, $(u, z_2) \neq 0$ и $\alpha_2 \neq 0$, то характер сходимости $\lambda_1^{(k)}$ к λ_1 можно охарактеризовать более точно:

$$\lambda_1^{(k)} = \lambda_1 + c\left(\frac{\lambda_2}{\lambda_1}\right)^k + \mathcal{O}\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2k} + \left(\frac{\lambda_3}{\lambda_1}\right)^k\right),$$

где

$$c = \frac{\lambda_1}{\lambda_2}\frac{\alpha_2(z_2, u)}{\alpha_1(z_1, u)}(\lambda_2 - \lambda_1).$$

При сделанных предположениях такое же уточнение можно получить и по отношению к сходимости векторов $z_1^k = (\zeta_1^k, \ldots, \zeta_n^k)$ к собственному вектору $z_1 = (\zeta_1, \ldots, \zeta_n)$. Действительно,

$$y_k = \lambda_1^k \left[\alpha_1 z_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k z_2 + w_k \right], \quad \|w_k\| = \mathcal{O}\left(\left(\frac{\lambda_3}{\lambda_1} \right)^k \right),$$

и так как $z_1^k = x_k = y_k/(y_k, u)$, то для любого вектора v

$$(z_1^k, v) = \frac{\lambda_1^k \left[\alpha_1 (z_1, v) + \alpha_2 (\lambda_2/\lambda_1)^k (z_2, v) + \varepsilon_{k1} \right]}{\lambda_1^k \left[\alpha_1 + \alpha_2 (\lambda_2/\lambda_1)^k (z_2, u) + \varepsilon_{k2} \right]}, \quad \varepsilon_{kj} = \mathcal{O}\left(\left(\frac{\lambda_1}{\lambda_3} \right)^k \right).$$

Отсюда нетрудно усмотреть, что

$$(z_1^k, v) = (z_1, v) + c(v) \left(\frac{\lambda_2}{\lambda_1} \right)^k + \mathcal{O}\left(\left(\frac{\lambda_2}{\lambda_1} \right)^{2k} + \left(\frac{\lambda_3}{\lambda_1} \right)^k \right),$$

Взяв в качестве v орты: $v = e_j$, имеем

$$\zeta_j^k = \zeta_j + c_j \left(\frac{\lambda_2}{\lambda_1} \right)^k + \mathcal{O}\left(\left(\frac{\lambda_2}{\lambda_1} \right)^{2k} + \left(\frac{\lambda_3}{\lambda_1} \right)^k \right),$$

Метод скалярных произведений.

Если матрица A эрмитова, то располагая векторами x_k, построенными согласно (3), можно получить существенно лучшее приближение к λ_1. Действительно, поскольку собственные векторы z_j попарно ортогональны, то

$$(y_k, y_k) = \alpha_1^2 \lambda_1^{2k} (z_1, z_1) + \sum_{j=2}^{n} \alpha_j^2 \lambda_j^{2k} (z_j, z_j),$$

аналогичный вид имеет и (y_k, y_{k-1}), так что

$$\frac{(y_k, y_k)}{(y_k, y_{k-1})} = \lambda_1 \frac{1 + \sum_{j=2}^{n} \gamma_j \left(\frac{\lambda_j}{\lambda_1} \right)^{2k}}{1 + \sum_{j=2}^{n} \gamma_j \left(\frac{\lambda_j}{\lambda_1} \right)^{2k-1}}, \quad \gamma_j = \left(\frac{\alpha_j}{\alpha_1} \right)^2 \frac{(z_j, z_j)}{(z_1, z_1)}.$$

Таким образом

$$\tilde{\lambda}_1^{(k)} = \frac{(y_k, y_k)}{(y_k, y_{k-1})} = \frac{(\tilde{x}_k, \tilde{x}_k)}{(\tilde{x}_k, x_{k-1})} = \lambda_1 + \mathcal{O}\left(\left(\frac{\lambda_2}{\lambda_1} \right)^{2k} \right)$$

есть, вообще говоря, существенно лучшее приближение к λ_1, чем $\lambda_1^{(k)}$. В этом и состоит метод скалярных произведений. Улучшить этим методом приближение к собственному вектору эрмитовой матрицы не удается.

Случай $|\lambda_2| = |\lambda_1|$.

Остановимся кратко на случае $|\lambda_2| = |\lambda_1|$, считая для простоты матрицу A вещественной; тогда и векторы y_0 и u естественно считать вещественным. Из всех возможностей разберем три наиболее характерные.

1) λ_1 — кратное собственное число: $\lambda_1 = \cdots = \lambda_m$, причем $|\lambda_1| > |\lambda_{m+1}|$. Тогда

$$y_k = \lambda_1^k \left[\sum_{j=1}^{m} \alpha_j z_j + \sum_{j=m+1}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k z_j \right],$$

откуда видно, что $\lambda_1^{(k)}$ и z_1^k, построенные по формулам (3), сходятся соответственно к λ_1 и некоторому из собственных векторов матрицы A, соответствующих λ_1, с той же быстротой, что указана в теореме. То обстоятельство, что λ_1 кратное собственное число, в процессе счета замечено не будет.

2) $\lambda_2 = -\lambda_1$, $|\lambda_1| > |\lambda_3|$. В этом случае

$$y_k = \lambda_1^k \left[\alpha_1 z_1 + (-1)^k \alpha_2 z_2 + \sum_{j=3}^{n} \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^k z_j \right].$$

Поэтому характерным признаком, по которому мы можем судить, что имеем дело с этим случаем, будет близость при больших k векторов x_k и x_{k-2} при существенном отличии x_{k-1} от x_k. Приближения к собственным числам и векторам можно вычислить по формулам

$$\left(\lambda_1^{(k)}\right)^2 = \left(\lambda_2^{(k)}\right)^2 = \frac{(y_k, u)}{(y_{k-2}, u)}, \quad z_m^k = c_m(y_k + \lambda_m^{(k)} y_{k-1}), \quad m = 1, 2.$$

Заниматься представлением правых частей этих формул через векторы $\tilde{x}_k$ и x_k не будем.

3) $\lambda_1 = \rho e^{i\theta}$ и $\lambda_2 = \rho e^{-i\theta}$ — комплексно сопряженные собственные числа, $\rho > |\lambda_3|$. Векторы z_1 и z_2 имеют комплексно сопряженные компоненты. Поскольку векторы y_0 и u вещественные, то числа $\alpha_{1-2}(z_{1-2}, u) = re^{\pm i\varphi}$ также комплексно сопряжены. Поэтому числа

$$(y_k, u) = \rho^k \left[2r \cos(k\theta + \varphi) + \mathcal{O}\left((\lambda_3/\rho)^k \right) \right]$$

ведут себя "неправильно", что и служит признаком комплексно сопряженных наибольших по модулю собственных чисел. Модуль первых двух собственных чисел может быть приближенно найден по формуле

$$\rho^2 \approx \frac{(y_k, u)(y_{k-2}, u) - (y_{k-1}, u)^2}{(y_{k-1}, u)(y_{k-3}, u) - (y_{k-2}, u)^2},$$

так как

$$\begin{aligned}(y_k, u)(y_{k-2}, u) - (y_{k-1}, u)^2 \approx \\ \approx \rho^{2k-2} 4r^2 \left[\cos(k\theta + \varphi) \cos((k-2)\theta + \varphi) - \cos^2((k-1)\theta + \varphi) \right] = \\ = -4r^2 \sin^2\theta \, \rho^{2k-2}.\end{aligned}$$

Приводить формулы для вычисления θ не будем.

Метод обратных итераций.

Пусть нам требуется найти *наименьшее по модулю* собственное число λ_n матрицы A. Поскольку $\mu = \frac{1}{\lambda_n}$ — наибольшее по модулю собственное число матрицы $B = A^{-1}$, то можно найти λ_n, применяя степенной метод к матрице B. Обращать матрицу A нерационально, более выгодно вычислять на каждом шаге вектор y_k, решая систему уравнений $Ay_k = y_{k-1}$. При решении этих систем методом исключений следует учитывать, что все они имеют одну и ту же матрицу коэффициентов, так что значительная часть вычислений при решении этих систем проделывается только один раз. В этом и состоит метод обратных итераций.

Метод обратных итераций можно применять для уточнения произвольного собственного числа λ матрицы A, если для него известно достаточно хорошее приближение $\lambda^{(0)}$. Следует воспользоваться тем, что $\lambda - \lambda^{(0)}$ есть собственное число матрицы $B = A - \lambda^{(0)}I$, минимальное по модулю, если $\lambda^{(0)}$ — хорошее приближение. Более того, сходимость метода обратных итераций для нахождения $\lambda - \lambda^{(0)}$ будет в этом случае очень быстрой.

Задача. Пусть для матрицы A $\lambda_1 = \rho e^{i\theta}$, $\lambda_2 = \rho e^{-i\theta}$, $\rho > |\lambda_3|$. Рассмотрев последовательность

$$p_k = \frac{(y_{k-1}, u)(y_{k+2}, u) - (y_k, u)(y_{k+1}, u)}{(y_{k-1}, u)(y_{k+1}, u) - (y_k, u)^2},$$

указать способ нахождения $\mathrm{Re}\lambda_1$ степенным методом в этом случае.

§8. Метод Крылова

Метод А.Н.Крылова — это метод решения полной проблемы собственных чисел, основанный на вычислении коэффициентов характеристического полинома матрицы A. С самого начала будем предполагать, что все собственные числа этой матрицы различны — в противном случае метод имеет некоторые осложнения.

Метод основан на равенстве Кели–Гамильтона. Пусть

$$P_n(\lambda) = \det(A - \lambda I) = (-1)^n\lambda^n + p_1\lambda^{n-1} + \cdots + p_n \quad -$$

характеристический полином матрицы A. Тогда

$$P_n(A) = (-1)^n A^n + p_1 A^{n-1} + \cdots + p_n I = 0.$$

В случае, когда матрица A имеет полную систему собственных векторов, это равенство доказывается совсем просто. Действительно, пусть z — собственный вектор: $Az = \lambda z$. Тогда $P_n(A)z = P_n(\lambda)z = 0$ (т.к. λ — собственное число). Ввиду полноты системы собственных векторов и для любого вектора x будет $P_n(A)x = 0$, и значит $P_n(A)$ нулевая матрица.

Обратимся к методу. Возьмем произвольный ненулевой вектор y_0 и вычислим векторы

$$y_k = (\eta_1^k, \ldots, \eta_n^k), \quad y_k = A^k y_0 = Ay_{k-1}, \quad k = 1, \ldots, n.$$

Векторное равенство $P_n(A)y_0$ перепишем покомпонентно, перенеся первое слагаемое в правую часть:

$$\left.\begin{array}{cccc} \eta_1^{n-1}p_1 & +\cdots+ & \eta_1^0 p_n & = (-1)^{n+1}\eta_1^n \\ \cdots & \cdots & \cdots & \cdots \\ \eta_n^{n-1}p_1 & +\cdots+ & \eta_n^0 p_n & = (-1)^{n+1}\eta_n^n \end{array}\right\}. \qquad (1)$$

Будем рассматривать (1) как систему линейных алгебраических уравнений относительно коэффициентов характеристического полинома и кратко запишем ее в виде $Yp = -y_n$ (матрица Y и вектор y_n известны, ищется вектор $p = (p_1, \ldots, p_n)$). Пусть z_k собственные векторы матрицы A: $Az_k = \lambda_k z_k$ (напомним, что все λ_k различны), и пусть $y_0 = \sum_{j=1}^n \alpha_j z_j$.

Теорема. для того чтобы матрица Y системы уравнений (1) была неособенной, необходимо и достаточно чтобы при всех j было $\alpha_j \neq 0$[13].

Доказательство. Введем обозначения для компонент векторов $z_j = (\zeta_1^j, \ldots, \zeta_n^j)$ и определим матрицы Z, Λ и диагональную матрицу D:

$$Z = \begin{pmatrix} \zeta_1^1 & \ldots & \zeta_1^n \\ \ldots & \ldots & \ldots \\ \zeta_n^1 & \ldots & \zeta_n^n \end{pmatrix}, \Lambda = \begin{pmatrix} \lambda_1^{n-1} & \ldots & 1 \\ \ldots & \ldots & \ldots \\ \lambda_n^{n-1} & \ldots & 1 \end{pmatrix}, D = \begin{pmatrix} \alpha_1 & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & \alpha_n \end{pmatrix}.$$

Поскольку элемент k-й строки и j-го столбца матрицы Y есть

$$\eta_k^{n-j} = \sum_{\nu=1}^n \alpha_\nu \lambda_\nu^{n-j} \zeta_k^\nu,$$

то $Y = Z\Lambda D$. Матрица Z неособенная, поскольку собственные векторы z_k линейно независимы, матрица Λ неособенная как матрица Вандермонда, а для того чтобы D была неособенной, необходимо и достаточно, чтобы все коэффициенты α_j были отличны от нуля. ■

Замечание. Если среди собственных чисел матрицы A имеются равные, то $\det\Lambda = 0$, так что и $\det Y = 0$. В этом и состоит упоминавшееся в начале параграфа затруднение.

Пусть коэффициенты полинома P_n уже вычислены и его корни λ_k — собственные числа матрицы A найдены. Обратимся в нахождению собственных векторов. Построим полиномы

$$P_{n-1}^k(\lambda) = \frac{P_n(\lambda)}{(\lambda - \lambda_k)} = (-1)^n\lambda^{n-1} + p_1^k\lambda^{n-2} + \cdots + p_{n-1}^k.$$

Очевидно, что $P_{n-1}^k(\lambda_j) = 0$ при $j \neq k$ и $P_{n-1}^k(\lambda_k) = P_n'(\lambda_k) \neq 0$. Поэтому $P_{n-1}^k(A)z_j = P_{n-1}^k(\lambda_j)z_j = 0$ $(j \neq k)$ и $P_{n-1}^k(A)z_k = P_n'(\lambda_k)z_k$. Таким образом

$$P_{n-1}^k(A)y_0 = \sum_{j=1}^n \alpha_j P_{n-1}^k(A)z_j = \alpha_k P_n'(\lambda_k)z_k = \tilde{z}_k$$

[13]Про такой вектор y_0 иногда говорят, что он “находится в общем положении”.

есть собственный вектор матрицы A, соответствующий собственному числу λ_k. Остается еще заметить, что

$$\tilde{z}_k = (-1)^n y_{n-1} + p_1^k y_{n-2} + \cdots + p_{n-1}^k y_0. \qquad (2)$$

Итак, алгоритм метода Крылова состоит в следующем:

1) выбирается произвольный ненулевой вектор y_0 и строятся векторы $y_k = Ay_{k-1}$, $k = 1, \ldots, n$;

2) путем решения системы уравнений (1) находятся коэффициенты характеристического полинома $P_n(\lambda)$ матрицы A;

3) собственные числа λ_k матрицы A находятся как корни полинома $P_n(\lambda)$;

4) вычисляются коэффициенты полиномов $P_{n-1}^k(\lambda) = P_n(\lambda)/(\lambda - \lambda_k)$;

5) собственные векторы $\tilde{z}_k$ вычисляются по формулам (2).

Если определитель системы уравнений (1) оказался нулевым, то это означает, что либо у матрицы A имеется кратное собственное число, либо вектор y_0 выбран неудачно, и тогда его следует заменить.

§9 Метод Якоби

В этом параграфе матрицы будем считать вещественными.

Метод Якоби — это метод решения полной проблемы собственных чисел для симметричных матриц.

Напомним некоторые сведения из линейной алгебры. Матрица T называется ортогональной, если ее столбцы попарно ортогональны и сумма квадратов элементов каждого столбца равна единице, т.е. если[14] $T'T = I$. Произведение ортогональных матриц есть ортогональная матрица. Если A — симметричная матрица, а T — ортогональная, то $T'AT$ также симметричная и собственные числа этих матриц совпадают. Симметричная матрица A может быть ортогональным преобразованием приведена к диагональному виду, т.е. существует такая ортогональная матрица T, что $T'AT = \Lambda$, где Λ — диагональная матрица. При этом диагональные элементы матрицы Λ — собственные числа матрицы A, а столбцы матрицы T — соответствующие собственные векторы.

Элементарной ортогональной матрицей назовем ортогональную матрицу $T_{ij}(\varphi)$ $(i < j)$, которая лишь четырьмя элементами отличается от единичной: $t_{ii} = t_{jj} = \cos\varphi$, $-t_{ij} = t_{ji} = \sin\varphi$.

Лемма 1. Пусть $B = T'_{ij}(\varphi)AT_{ij}(\varphi)$. Тогда матрица B лишь двумя столбцами и двумя строками с номерами i и j отличается от матрицы A (т.е. при $\mu \neq i, j$ и $\nu \neq i, j$ будет $b_{\mu\nu} = a_{\mu\nu}$). Кроме того $\hat{B} = T'(\varphi)\hat{A}T(\varphi)$. Здесь

$$\hat{A} = \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix}, \quad T(\varphi) = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \quad -$$

матрицы второго порядка.

---

[14] Штрих означает транспонирование матрицы.

Лемма 2. Для симметричной матрицы A и индексов $i < j$ найдется такой угол φ, что для матрицы $B = T'_{ij}(\varphi)AT_{ij}(\varphi)$ окажется $b_{ij} = b_{ji} = 0$. При таком выборе φ

$$b_{ii}^2 + b_{jj}^2 = a_{ii}^2 + a_{jj}^2 + 2a_{ij}^2. \tag{1}$$

Доказательство. Используя лемму 1, нетрудно получить, что при произвольном φ

$$b_{ij} = \frac{1}{2}(a_{jj} - a_{ii}) \sin 2\varphi + a_{ij} \cos 2\varphi.$$

Если $a_{ii} = a_{jj}$, то достаточно положить $\varphi = \pi/4$, а при $a_{ii} \neq a_{jj}$

$$\varphi = \frac{1}{2} \arctan \frac{2a_{ij}}{a_{ii} - a_{jj}}.$$

Доказательство формулы (1) элементарно. ■

Метод Якоби состоит в построении последовательности матриц A_s ($A_0 = A$) с таким расчетом, чтобы при больших s матрицы A_s были близки к диагональным, так что их диагональные элементы близки к собственным числам матрицы A. Матрица A_{s+1} строится в виде $A_{s+1} = T'_s A_s T_s$, где $T_s = T_{ij}(\varphi)$, индексы $i = i_s$ и $j = j_s$ находятся из условия $|a_{ij}^{(s)}| = \max_{\mu \neq \nu} |a_{\mu\nu}^{(s)}|$, а $\varphi = \varphi_s$ вычисляется (согласно лемме 2) так, чтобы оказалось $a_{ij}^{(s+1)} = 0$. Напомним, что каждая следующая матрица отличается от предыдущей лишь двумя строками и столбцами.

Введем обозначения для сумм квадратов всех элементов матрицы A и ее диагональных элементов[15]:

$$t^2(A) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2, \qquad d^2(A) = \sum_{i=1}^{n} a_{ii}^2.$$

Тогда $r^2(A) = t^2(A) - d^2(A)$ — сумма квадратов недиагональных элементов.

Лемма 3. Если T — ортогональная матрица и $B = T'AT$, то $t^2(B) = t^2(A)$.

Доказательство. Очевидно равенство[16] $t^2(A) = \mathrm{tr}(AA')$. Поэтому

$$t^2(B) = \mathrm{tr}(B'B) = \mathrm{tr}(T'ATT'A'T) = \mathrm{tr}(T'AA'T) = \mathrm{tr}(AA') = t^2(A) \quad ■$$

Теорема. При $s \to \infty$ все недиагональные элементы матриц A_s стремятся к нулю.

Доказательство. Достаточно доказать, что $r^2(A_s) \to 0$. Согласно лемме 3 $t^2(A_{s+1}) = t^2(A_s)$. Все диагональные элементы матриц A_{s+1} и A_s, кроме двух, совпадают, и ввиду равенства (1) $d^2(A_{s+1}) = d^2(A_s) + 2\left(a_{ij}^{(s)}\right)^2$, причем согласно выбору индексов i_s и j_s $\left(a_{ij}^{(s)}\right)^2 \geq r^2(A_s)/[n(n-1)]$ и потому

$$d^2(A_{s+1}) \geq d^2(A_s) + 2r^2(A_s)/[n(n-1)],$$

---

[15] $t(A)$ называется нормой Фробениуса матрицы A.

[16] tr(A) — след матрицы A, сумма ее диагональных элементов, равная сумме собственных чисел.

$r^2(A_{s+1}) = t^2(A_{s+1}) - d^2(A_{s+1}) \le t^2(A_s) - d^2(A_s) - 2r^2(A_s)/[n(n-1)] = qr^2(A_s)$,

где $q = 1 - 2/[n(n-1)] < 1$. Методом индукции теперь легко показывается, что $r^2(A_s) \le q^s r^2(A)$, откуда и следует сходимость. ■

Покажем, что диагональные элементы матриц A_s сходятся к собственным числам матрицы A. Пусть $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ — собственные числа A (а значит, и A_s), $\mu_1^s \ge \mu_2^s \ge \cdots \ge \mu_n^s$ — диагональные элементы матрицы A_s, расположенные в порядке убывания.

Следствие. При $s \to \infty$ выполняются соотношения $\mu_k^s \to \lambda_k$.

Доказательство. Матрицу A_s представим в виде $A_s = \Lambda_s + R_s$, где Λ_s — диагональная матрица, а у R_s на главной диагонали нули, так что μ_k^s — собственные числа матрицы Λ_s. По доказанной теореме $R_s \to 0$, и по теореме 4 из §3 при всех k $|\lambda_k - \mu_k^s| \le \|R_s\|$ ■

Замечание. Из теоремы следует сходимость метода с быстротой геометрической прогрессии со знаменателем q, близким при большом n к единице. В действительности сходимость метода существенно быстрее.

Итак, за приближения к собственным числам матрицы A принимаются при большом s диагональные элементы матрицы A_s. За приближения к собственным векторам можно принять столбцы матрицы $T_1 T_2 \dots T_s$.

§10 Об ускорении сходимости

.

Пусть имеется сходящаяся числовая последовательность $a_s \to a^*$. Ускорением сходимости называется построение другой последовательности $\{b_s\}$, которая сходится к тому же пределу a^*, но быстрее, чем исходная. Для того, чтобы это было возможно, мы должны располагать некоторой дополнительной информацией о последовательности $\{a_s\}$. В этом параграфе будем предполагать известным, что $\{a_s\}$ сходится к пределу с быстротой общего члена геометрической прогрессии, т.е. что

$$a_s = a^* + q^s(c + \varepsilon_s), \tag{1}$$

где $c \ne 0$, $0 < |q| < 1$ и $\varepsilon_s \to 0$. Числа a^*, c и последовательность ε_s нам неизвестны (получение лучших, чем a_s, приближений к a^* и составляет цель ускорения сходимости), известен только факт существования такого представления. Что касается q, то следует различать два случая, когда q известно хотя бы приближенно, и неизвестно. Первый случай связан с методом Л.А.Люстерника ускорения сходимости метода итерации решения систем линейных уравнений.

Метод Люстерника.

Итак, нам известно, что a_s имеют представление (1) и известно число q. Из (1) следует, что $a_s - qa_{s-1} = (1-q)a^* + q^s(\varepsilon_s - \varepsilon_{s-1})$, и если мы положим $b_s = (a_s - qa_{s-1})/(1-q)$, то окажется $b_s = a^* + \frac{1}{1-q} q^s(\varepsilon_s - \varepsilon_{s-1})$ и $(b_s - a^*)/(a_s - a^*) = \frac{1}{1-q}(\varepsilon_s - \varepsilon_{s-1})/(c + \varepsilon_s) \to 0$. Это и означает, что b_s сходится к a^* быстрее, чем a_s.

Положение не изменится, если под a^*, a_s, c, ε_s мы будем понимать векторы. Тогда (1) — векторное равенство, и если векторы b_s строить по

формуле $b_s = (a_s - qa_{s-1})/(1-q)$, то т.к. $\|b_s - a^*\| \le \frac{1}{1-q} q^s(\|\varepsilon_s\| + \|\varepsilon_{s-1}\|)$ и $\|a_s - a^*\| \ge q^s(\|c\| - \|\varepsilon_s\|)$, будет $\|b_s - a^*\|/\|a_s - a^*\| \to 0$.

Обратимся теперь к методу итерации для решения системы линейных уравнений $x = Ax + y$: $x_{s+1} = Ax_s + y$. Сделаем предположения: 1) максимальное по модулю собственное число матрицы A, имеющей полную систему собственных векторов, единственно, так что $1 > |\lambda_1| > |\lambda_2| \ge |\lambda_3| \ge \cdots \ge |\lambda_n|$; 2) при разложении вектора $x_0 - x^* = \sum_{j=1}^n \alpha_j z_j$ по собственным векторам матрицы A $\alpha_1 \ne 0$. Тогда

$$x_s - x^* = A^s(x_0 - x^*) = \sum_{j=1}^n \alpha_j \lambda_j^s z_j = \lambda_1^s \left(\alpha_1 z_1 + \sum_{j=2}^n \alpha_j \left(\frac{\lambda_j}{\lambda_1}\right)^s z_j \right) = \\ = \lambda_1^s(\alpha_1 z_1 + \varepsilon_s), \quad \|\varepsilon_s\| \to 0,$$

т.е. последовательность $\{x_s\}$ имеет представление (1) при $q = \lambda_1$. В то же время

$$x_s - x_{s-1} = x_s - x^* - (x_{s-1} - x^*) = \lambda_1^{s-1}\big((\lambda_1 - 1)\alpha_1 z_1 + \lambda_1 \varepsilon_s - \varepsilon_{s-1}\big),$$

откуда видно, что если $(z_1, u) \ne 0$, то

$$\lambda_1^{(s)} = \frac{(x_s - x_{s-1}, u)}{(x_{s-1} - x_{s-2}, u)} \to \lambda_1,$$

и можно при больших s заменить $q = \lambda_1$ на $\lambda_1^{(s)}$ (в сущности мы применили для нахождения приближения к λ_1 степенной метод). В этом и состоит метод Люстерника, имеющий вычислительные формулы:

$$\lambda_1^{(s)} = \frac{(x_s - x_{s-1}, u)}{(x_{s-1} - x_{s-2}, u)}, \quad \tilde{x}_s = \frac{x_s - \lambda_1^{(s)} x_{s-1}}{1 - \lambda_1^{(s)}} = \frac{1}{1 - \lambda_1^{(s)}} x_s - \frac{\lambda_1^{(s)}}{1 - \lambda_1^{(s)}} x_{s-1}.$$

Вектор $\tilde{x}_s$, вообще говоря, ближе к x^*, чем x_s. Метод особенно эффективен, если $|\lambda_1|$ существенно больше, чем $|\lambda_2|$.

δ^2-процесс Эйткена.

Рассмотрим второй случай, когда в представлении (1) числовой последовательности коэффициент q нам неизвестен. Пренебрегая малыми слагаемыми ε_j, будем находить b_s из системы уравнений относительно b_s, c и q

$$\left.\begin{array}{rl} a_{s-1} = & b_s + cq^{s-1} \\ a_s = & b_s + cq^s \\ a_{s+1} = & b_s + cq^{s+1} \end{array}\right\}.$$

Система легко решается:

$$q = \frac{a_{s+1} - a_s}{a_s - a_{s-1}}, \qquad b_s = \frac{a_s - qa_{s-1}}{1-q} = \frac{\begin{vmatrix} a_{s+1} & a_s \\ a_s & a_{s-1} \end{vmatrix}}{a_{s+1} - 2a_s + a_{s-1}}.$$

Последняя формула и применяется для вычисления b_s. Простые вычисления показывают. что при любом числе a выполняются также равенства

$$b_s = a + \frac{\begin{vmatrix} a_{s+1}-a & a_s-a \\ a_s-a & a_{s-1}-a \end{vmatrix}}{a_{s+1}-2a_s+a_{s-1}}. \tag{2}$$

При вычислениях используют именно эту формулу, выбирая a близким к a_s, чтобы избежать существенного влияния ошибок округления в произведениях при вычислении определителя. В частности, если положить $a = a_s$, то

$$b_s = a_s + \frac{(a_{s+1}-a_s)(a_{s-1}-a_s)}{a_{s+1}-2a_s+a_{s-1}}.$$

Для последовательностей рассматриваемого вида последовательность $\{b_s\}$ сходится к a^* быстрее, чем $\{a_s\}$:

Теорема. Если $0 < |q| < 1$ и $c \neq 0$, то

$$\frac{b_s - a^*}{a_s - a^*} \to 0.$$

Доказательство. Используя для b_s формулу (2) при $a = a^*$, имеем

$$b_s - a^* = \frac{\begin{vmatrix} q^{s-1}(c+\varepsilon_{s-1}) & q^s(c+\varepsilon_s) \\ q^s(c+\varepsilon_s) & q^{s+1}(c+\varepsilon_{s+1}) \end{vmatrix}}{q^{s+1}(c+\varepsilon_{s+1}) - 2q^s(c+\varepsilon_s) + q^{s-1}(c+\varepsilon_{s-1})} =$$

$$= q^s \frac{c(\varepsilon_{s+1}+\varepsilon_{s-1}-2\varepsilon_s) + \varepsilon_{s+1}\varepsilon_{s-1} - \varepsilon_s^2}{q(c+\varepsilon_{s+1}) - 2(c+\varepsilon_s) + \frac{1}{q}(c+\varepsilon_{s-1})} = \frac{e_s}{Q_s} q^s.$$

Здесь $e_s \to 0$ и $Q_s \to c\left(q - 2 + \frac{1}{q}\right) = \frac{c}{q}(1-q)^2 \neq 0$. Поэтому

$$\frac{b_s - a^*}{a_s - a^*} = \frac{e_s}{(c+\varepsilon_s)Q_s} \to 0. \qquad \blacksquare$$

Замечание. Как видно из доказательства, порядок сходимости b_s к a^* определяется формулой $b_s - a^* = \mathcal{O}\big(q^s(|\varepsilon_s| + |\varepsilon_{s+1}| + |\varepsilon_{s-1}|)\big)$.

Вспомним степенной метод нахождения максимального собственного числа матрицы A. Как было показано в §7, при соответствующих предположениях

$$\lambda_1^{(s)} = \lambda_1 + c\left(\frac{\lambda_2}{\lambda_1}\right)^s + \mathcal{O}\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2s} + \left(\frac{\lambda_3}{\lambda_1}\right)^s\right).$$

Итак, при тех условиях, которые указаны в §7, последовательность $\{\lambda_1^{(s)}\}$ принадлежит к тому классу последовательностей, к которым применим δ^2-процесс Эйткена. Последовательность $\{\mu_1^{(s)}\}$, полученная из $\{\lambda_1^s\}$ этим методом, сходится к λ_1 быстрее, причем, согласно замечанию,

$$\mu_1^{(s)} - \lambda_1 = \mathcal{O}\left(\left(\frac{\lambda_2}{\lambda_1}\right)^{2s} + \left(\frac{\lambda_3}{\lambda_1}\right)^s\right).$$

δ^2-процесс применим и к методу скалярных произведений в случае эрмитовой матрицы A, а также к уточнению собственных векторов x_s, полученных степенным методом; в последнем случае речь идет о покомпонентном применении этого метода. Останавливаться на подробностях не будем.

Задача. Определить быстроту сходимости полученной с помощью δ^2-процесса последовательности приближений к λ_1 в случае метода скалярных произведений.

Глава 4
Приближенное решение нелинейных уравнений м систем

§1. Метод итерации

Пусть дано уравнение

$$t = \varphi(t). \tag{1}$$

Метод итерации для решения этого уравнения состоит в том, что, начиная с некоторого начального приближения t_0, строится последовательность

$$t_{s+1} = \varphi(t_s).$$

Очевидно, что если функция $\varphi(t)$ непрерывна и $t_s \to t^*$, то t^* — корень уравнения (1). Графически t^* это абсцисса точки пересечения графика функции $\varphi(t)$ с биссектрисой первого координатного угла. Из рассмотрения графика можно сделать следующие выводы.

1) Если $|\varphi'(t)| < 1$, то следует ожидать сходимость метода.

2) Если $|\varphi'(t)| > 1$, то следует ожидать расходимость метода.

3) Если $0 < \varphi'(t) < 1$, то приближения t_s лежат по одну сторону от t^*.

4) Если $-1 < \varphi'(t) < 0$, то имеет место альтернирующая сходимость, т.е. t^* лежит между последовательными приближениями t_s и t_{s+1}.

Однако делать строгие выводы мы будем сразу для системы нелинейных уравнений.

Пусть дана система нелинейных уравнений

$$\xi_k = \varphi_k(\xi_1, \ldots, \xi_n), \qquad k = 1, \ldots, n,$$

которую будем записывать в векторной форме

$$x = \Phi(x), \tag{2}$$

где $x = (\xi, \ldots, \xi_n) \in \mathbb{R}^n$, $\Phi(x) = (\varphi_1(x), \ldots, \varphi_n(x))$. Метод итерации состоит в том, что, выбрав начальное приближение x_0, мы строим последовательные приближения по формуле $x_{s+1} = \Phi(x_s)$. Дальше мы считаем, что в $\mathbb{R}^n$ введена некоторая норма $\|.\|$, отображение Φ задано в некоторой области $\Omega \subset \mathbb{R}^n$. Кроме того будем пользоваться обозначением $S_r(y) = \{\, x \in \mathbb{R}^n \mid \|x - y\| \le r \,\}$ – шар (соответствующий введенной норме!) радиуса r с центром в точке y.

Теорема 1. Пусть $x_0 \in \Omega$ и выполнены следующие условия:

1^0. $\|\Phi(x_0) - x_0\| \le m$,

2^0. существует $q < 1$, такое что для любых точек $x', x'' \in \Omega$ выполняется неравенство $\|\Phi(x') - \Phi(x'')\| \le q\|x' - x''\|$,

3^0. $S_r(x_0) \subseteq \Omega$, где $r = m/(1-q)$.

Тогда

а) в области Ω существует и притом единственное решение x^* уравнения (1),

б) $x^* \in S_r(x_0)$,

в) $x_s \to x^*$,

г) выполняются оценки погрешности:

$$\|x_s - x^*\| \le \frac{mq^s}{1-q}, \qquad \|x_s - x^*\| \le \frac{q}{1-q}\|x_s - x_{s-1}\|.$$

Доказательство. Докажем сначала методом индукции, что при всех $s \ge 1$ $\alpha)$ $x_s \in S_r(x_0)$ и $\beta)$ $\|x_{s+1} - x_s\| \le mq^s$. Действительно, при $s = 0$ $\alpha)$ очевидно и по условию 1^0 $\|x_1 - x_0\| \le m$, так что $\beta)$ также выполнено. Докажем возможность индуктивного перехода от s к $s+1$:
$\alpha)$ $\|x_{s+1} - x_0\| \le \|x_{s+1} - x_s\| + \cdots + \|x_1 - x_0\| \le mq^s + mq^{s-1} + \cdots + m \le m/(1-q) = r$;
$\beta)$ $\|x_{s+2} - x_{s+1}\| = \|\Phi(x_{s+1}) - \Phi(x_s)\| \le q\|x_{s+1} - x_s\| \le mq^{s+1}$.

Из $\beta)$ следует, что при любом натуральном p

$$\|x_{s+p} - x_s\| \le \|x_{s+p} - x_{s+p-1}\| + \cdots + \|x_{s+1} - x_s\| \le$$
$$\le mq^{s+p-1} + \cdots + mq^s \le \frac{mq^s}{1-q} \to 0, \quad (3)$$

так что последовательность $\{x_s\}$ сходится в себе, и потому сходится: $x_s \to x^*$. Из $x_s \in S_r(x_0)$ вытекает $x^* \in S_r(x_0)$, а из непрерывности Φ — $x^* = \Phi(x^*)$. Этим существование решения и утверждения б) и в) теоремы доказаны. Первая из оценок г) получается предельным переходом при $p \to \infty$ из (3), а вторая также предельным переходом по p из следующего неравенства:

$$\|x_{s+p} - x_s\| \le \|x_{s+p} - x_{s+p-1}\| + \cdots + \|x_{s+1} - x_s\| \le$$
$$\le (q^p + q^{p-1} + \cdots + q)\|x_s - x_{s-1}\| \le \frac{q}{1-q}\|x_s - x_{s-1}\|,$$

которое следует из того, что $\|x_{j+1} - x_j\| = \|\Phi(x_j) - \Phi(x_{j-1})\| \le q\|x_j - x_{j-1}\|$. Единственность решения доказывается от противного. Пусть $x^*, x^{**} \in \Omega$ — два решения. Тогда

$$\|x^* - x^{**}\| = \|\Phi(x^*) - \Phi(x^{**})\| \le q\|x^* - x^{**}\|,$$

откуда следует, что $x^* = x^{**}$. ■

Замечание 1. Решение уравнения (2) обычно называют *неподвижной точкой* отображения Φ. Отображение Φ, удовлетворяющее условию 2^0 теоремы называют *сжатым* или *сжимающим*. В связи с этим теорему 1 (или некоторые ее модификации) обычно называют *принципом сжатых отображений*. Это один из *принципов неподвижной точки*.

Замечание 2. Первая из оценок г) является *априорной*, а вторая — *апостериорной*. Эти понятия уже были введены ранее в §5 главы 3.

Введем понятие интеграла от вектор-функции. Пусть вектор $x(t)$ непрерывно зависит от $t \in [a, b]$. Тогда

$$\int_a^b x(t)dt = y = (\eta_1, \ldots, \eta_n), \text{ где } \eta_k = \int_a^b \xi_k(t)dt.$$

Лемма. Выполняется неравенство

$$\left\| \int_a^b x(t)dt \right\| \le \int_a^b \|x(t)\|dt.$$

Доказательство. Подобное неравенство очевидно для римановых сумм, и остается совершить предельный переход. ■

Отображение Φ называется дифференцируемым в точке x_0, если в этой точке дифференцируемы все его составляющие φ_k. В этом случае $\Phi'(x_0)$ – матрица Якоби:

$$\Phi'(x_0) = \left\{ \frac{\partial \varphi_k}{\partial \xi_j}(x_0) \right\}.$$

Отображение Φ непрерывно дифференцируемо в Ω, если таковы все φ_k.

Мы по-прежнему считаем, что в $\mathbb{R}^n$ введена некоторая векторная норма, и норма матрицы — это всегда операторная норма, порожденная введенной векторной.

Теорема 2. Пусть область Ω выпукла, Φ непрерывно дифференцируемо в Ω и при всех $x \in \Omega \quad \|\Phi'(x)\| \le q$. Тогда для всех $x', x'' \in \Omega$ $\|\Phi(x') - \Phi(x'')\| \le q\|x' - x''\|$.

Доказательство. При $t \in [0,1]$ рассмотрим

$$y(t) = \big(\eta_1(t), \dots, \eta_n(t)\big) = \Phi\big((1-t)x' + tx''\big) = \Phi\big(x' + t(x'' - x')\big)$$

(ввиду выпуклости $\Omega \quad (1-t)x' + tx'' \in \Omega$). Очевидно

$$y(0) = \Phi(x'), \quad y(1) = \Phi(x''), \quad \eta_k(t) = \varphi_k\big(x' + t(x'' - x')\big).$$

Теперь имеем:

$$\eta_k(1) - \eta_k(0) = \int_0^1 \eta_k'(t)dt,$$

$$\eta_k'(t) = \sum_{j=1}^n \frac{\partial \varphi_k}{\partial \xi_j}\big(x' + t(x'' - x')\big) \cdot (\xi_j'' - \xi_j'),$$

$$y(1) - y(0) = \int_0^1 \Phi'\big(x' + t(x'' - x')\big) \cdot (x'' - x')dt,$$

$$\|\Phi(x'') - \Phi(x')\| = \|y(1) - y(0)\| \le$$

$$\int_0^1 \|\Phi'\big(x' + t(x'' - x')\big)\| \cdot \|x'' - x'\|dt \le q\|x' - x''\|. \quad ■$$

Заметим, что при применении этой теоремы удобно использовать нормы $\|\cdot\|_\infty$ и $\|\cdot\|_1$, поскольку именно для этих векторных норм порожденные ими матричные вычисляются по простым формулам.

В принципе сжатых отображений (теорема 1) в случае выпуклой области Ω условие 2^0 можно заменить на: $\|\Phi'(x)\| \le q < 1 \quad \forall x \in \Omega$.

Задача 1. Сформулировать принцип сжатых отображений в той форме, как это предлагается в последнем абзаце. Обратить внимание, что предположения о том, что область Ω выпукла не требуется (за исключением утверждения о единственности) — достаточно воспользоваться выпуклостью $S_r(x_0)$.

Задача 2. Показать, что правая часть во второй из оценок г) в теореме 1 всегда не больше правой части в первой.

Задача 3. Показать, что теорема 2 (§5 главы 3) может быть получена как следствие теоремы 1 этого параграфа.

§2. Метод итерации (продолжение)

В этом параграфе сосредоточены результаты, имеющие локальный характер.

Напомним из курса анализа: функция n переменных $\varphi(x)$ $(x \in \mathbb{R}^n)$ называется дифференцируемой в точке x_0, если по всякому $\varepsilon > 0$ найдется такое $\delta > 0$, что при $\|x - x_0\|_2 < \delta$

$$\left| \varphi(x) - \varphi(x_0) - \sum_{j=1}^{n} \frac{\partial \varphi}{\partial \xi_j}(\xi_j - \xi_j^0) \right| < \varepsilon \|x - x_0\|_2.$$

Ввиду эквивалентности всех заданных в $\mathbb{R}^n$ норм норму $\|\cdot\|_2$ можно заменить на любую другую – получится эквивалентное определение.

Отсюда легко следует, что данному в предыдущем параграфе определению дифференцируемости отображения $\Phi(x)$ в точке x_0 эквивалентно следующее: по всякому $\varepsilon > 0$ найдется такое $\delta > 0$, что при $\|x - x_0\| < \delta$

$$\|\Phi(x) - \Phi(x_0) - \Phi'(x_0)(x - x_0)\| < \varepsilon \|x - x_0\|.$$

Последнее определение инвариантно относительно выбранной в $\mathbb{R}^n$ нормы.

Ниже, как и в предыдущем параграфе, норма матрицы — это всегда операторная норма, порожденная введенной в $\mathbb{R}^n$ векторной.

Пусть $\Phi : \Omega \mapsto \mathbb{R}^n$ $(\Omega \subset \mathbb{R}^n)$ – отображение, имеющее неподвижную точку $x^* \in \Omega$: $x^* = \Phi(x^*)$. Пусть x^* принадлежит Ω вместе с некоторой окрестностью. Рассматривается итеративная последовательность: $x_{s+1} = \Phi(x_s)$ при некотором начальном приближении x_0.

Теорема 1. Если отображение Φ дифференцируемо в точке x^* и выполняется неравенство $\|\Phi'(x^*)\| < 1$, то найдется такое $\delta > 0$, что при любом начальном приближении, удовлетворяющем неравенству $\|x_0 - x^*\| < \delta$, последовательность $\{x_s\}$ сходится к x^*.

Доказательство. Выберем число q так, что $\|\Phi'(x^*)\| < q < 1$ и положим $\varepsilon = q - \|\Phi'(x^*)\|$. По этому ε найдется такое $\delta > 0$, что из неравенства $\|x - x^*\| < \delta$ следует

$$\|\Phi(x) - \Phi(x^*) - \Phi'(x^*)(x - x^*)\| < \varepsilon \|x - x^*\|.$$

Если $\|x_0 - x^*\| < \delta$, то

$$\|x_1 - x^*\| = \|\Phi(x_0) - \Phi(x^*)\| \leq$$
$$\leq \|\Phi'(x^*)\| \cdot \|x_0 - x^*\| + \varepsilon \|x_0 - x^*\| = q \|x_0 - x^*\|.$$

Методом индукции отсюда легко можно получить, что при всех $s = 1, 2, \ldots$ $\|x_s - x^*\| < \delta$ и $\|x_s - x^*\| \le q^s \|x_0 - x^*\|$, откуда и следует сходимость. ■

При применении этой теоремы в распоряжении исследователя находится параметр q, лежащий в указанных пределах. Увеличение этого параметра ослабляет напрашивающееся утверждение о скорости сходимости, но увеличивает возможную область выбора начального приближения. Кроме того, формулировка теоремы *не* инвариантна относительно введенной в $\mathbb{R}^n$ нормы, так что эта норма – еще один инструмент, находящийся в руках исследователя. В действительности верно несколько более сильное утверждение: условие $\|\Phi'(x^*)\| < 1$ может быть заменено на $\rho\big(\Phi'(x^*)\big) < 1$. Здесь $\rho(A)$ – спектральный радиус матрицы A.

Определение. Пусть $\alpha > 1$ и дана последовательность векторов $\{a_s\}$. Говорят, что эта последовательность сходится к вектору a^* *с порядком* α, если $a_s \to a^*$ и существует такая постоянная c, что при всех s $\|a_{s+1} - a^*\| \le c\|a_s - a^*\|^\alpha$.

Пусть имеется некоторый метод, который исходя из произвольного вектора x_0 строит последовательность приближений x_s к вектору x^*. Говорят, что этот метод сходится *с порядком* α, если существует такое $\delta > 0$, что при выполнении условия $\|x_0 - x^*\| < \delta$ последовательность $\{x_s\}$ сходится к x^* с порядком α.

Замечание 1. Сходимость с порядком $\alpha = 2$ называется квадратичной. Существуют еще и такие термины. Последовательность $\{a_s\}$ сходится к a^* *линейно*, если существует такое $q < 1$, что $\|a_{s+1} - a^*\| \le q\|a_s - a^*\|$, и *сверхлинейно*, если $\|a_{s+1} - a^*\| / \|a_s - a^*\| \to 0$. Понятие линейной и сверхлинейной сходимости переносится и на методы построения последовательностей.

Замечание 2. Легко видеть, что понятие сходимости с порядком α и сверхлинейной сходимости инвариантно относительно выбранной в $\mathbb{R}^n$ нормы. К линейной сходимости это не относится.

Лемма. Пусть x^* – неподвижная точка отображения Φ: $x^* = \Phi(x^*)$, и пусть нашлись такие $r > 0$, $c > 0$ и $\alpha > 1$, что $S_r(x^*) \subset \Omega$ и для любого $x \in S_r(x^*)$ выполняется неравенство $\|\Phi(x) - \Phi(x^*)\| \le c\|x - x^*\|^\alpha$. Тогда метод итерации для нахождения x^* сходится с порядком α.

Доказательство. Возьмем произвольное $q < 1$ и найдем $\delta > 0$ из условий $\delta < r$ и $c\delta^{\alpha-1} \le q$. Тогда если $\|x_0 - x^*\| < \delta$, то

$$\|x_1 - x^*\| = \|\Phi(x_0) - \Phi(x^*)\| \le c\|x_0 - x^*\|^\alpha \le q\|x_0 - x^*\| \le \delta.$$

Отсюда методом индукции легко заключить, что при всех s

$$\|x_s - x^*\| \le \|x_0 - x^*\| q^s \to 0,$$

причем $\|x_{s+1} - x^*\| = \|\Phi(x_s) - \Phi(x^*)\| \le c\|x_s - x^*\|^\alpha$. ■

Теорема 2. Пусть отображение Φ дважды непрерывно дифференцируемо в некоторой окрестности неподвижной точки x^* и $\Phi'(x^*) = \mathbb{O}$. Тогда метод итерации для нахождения x^* сходится квадратически.

Доказательство. Используя инвариантность понятия квадратичной сходимости относительно выбранной в $\mathbb{R}^n$ нормы, проверим выполнение

условий леммы при $\alpha = 2$ для нормы $\|\cdot\|_\infty$. Пусть отображение Φ дважды непрерывно дифференцируемо в шаре $S_r(x_0)$ ($\|x - x^*\| \le r$). Положим

$$M = \max\left\{ \left| \frac{\partial^2 \varphi_k}{\partial \xi_j \partial \xi_l} \right| \;\middle|\; 1 \le k, j, l \le n,\ x \in S_r(x^*) \right\}.$$

Пусть $x \in S_r(x^*)$. Определим функции

$$u_k(t) = \varphi_k\big((1-t)x^* + tx\big) = \varphi_k\big(x^* + t(x - x^*)\big).$$

Тогда имеем

$$u_k'(t) = \sum_{j=1}^n \frac{\partial \varphi_k}{\partial \xi_j} \cdot (\xi_j - \xi_j^*),$$
$$u_k''(t) = \sum_{j,l=1}^n \frac{\partial^2 \varphi_k}{\partial \xi_j \partial \xi_l} \cdot (\xi_j - \xi_j^*)(\xi_l - \xi_l^*),$$

где все производные вычисляются в точке $x^* + t(x - x^*)$. Отсюда

$$u_k'(0) = 0;$$
$$|u_k''(t)| \le M n^2 \|x - x^*\|_\infty^2 \quad \forall t \in [0,1];$$
$$\varphi_k(x) - \varphi_k(x^*) = u_k(1) - u_k(0) = u_k'(0) + \frac{1}{2} u_k''(\tau_k) = \frac{1}{2} u_k''(\tau_k);$$
$$|\varphi_k(x) - \varphi_k(x^*)| \le \frac{1}{2} M n^2 \|x - x^*\|_\infty^2;$$
$$\|\Phi(x) - \Phi(x^*)\|_\infty \le c \|x - x^*\|_\infty^2$$

при $c = \frac{1}{2} M n^2$. ■

Задача. Показать, что в случае одного уравнения $t = \varphi(t)$ с трижды непрерывно дифференцируемой функцией φ условие $\varphi(t^*) = \varphi'(t^*) = \varphi''(t^*) = 0$ гарантирует сходимость третьего порядка метода итерации для корня t^*.

§3. Метод Ньютона

Метод Ньютона сначала рассмотрим для одного уравнения $f(t) = 0$. Пусть функция f дважды непрерывно дифференцируема в окрестности корня t^* этого уравнения и пусть нам известно достаточно близкое приближение t_0 к этому корню. Тогда

$$0 = f(t^*) = f(t_0) + (t^* - t_0) f'(t_0) + \frac{1}{2}(t^* - t_0)^2 f''(\tau).$$

Последнее слагаемое в правой части мало, и им можно пренебречь, так что t^* с хорошей точностью удовлетворяет уравнению $f(t_0) + (t - t_0) f'(t_0) = 0$. Решение этого уравнения $t_1 = t_0 - f(t_0)/f'(t_0)$ принимается за следующее

приближение к решению. Итак, метод Ньютона состоит в следующем. Выбирается некоторое начальное приближение t_0 и строится последовательность

$$t_{s+1} = t_s - \frac{f(t_s)}{f'(t_s)}. \tag{1}$$

Метод Ньютона имеет простой геометрический смысл: t_{s+1} есть абсцисса точки пересечения касательной к графику функции f, построенной в точке $\big(t_s, f(t_s)\big)$, с осью абсцисс.

Иногда оказывается удобным не пересчитывать на каждом шаге производную и пользоваться упрощенной формулой

$$\tilde{t}_{s+1} = \tilde{t}_s - \frac{f(\tilde{t}_s)}{f'(t_0)}.$$

Этот метод называется *модифицированным методом Ньютона*. Очевидно, что если $\tilde{t}_0 = t_0$, то и $\tilde{t}_1 = t_1$. Модифицированный метод Ньютона также имеет очевидный геометрический смысл.

В случае трудностей с вычислением производной ее значения можно заменять, используя численное дифференцирование. Пусть взяты два близких начальных приближения t_0 и t_1. Тогда можно построить следующее приближение по формуле $t_2 = t_1 - f(t_1)/f(t_0, t_1)$, заменив в формуле (1) при $s = 1$ производную $f'(t_1)$ на разделенную разность $f(t_0, t_1)$. Это приводит к последовательности

$$t_{s+1} = t_s - \frac{f(t_s)}{f(t_{s-1}, t_s)} = t_s - \frac{t_s - t_{s-1}}{f(t_s) - f(t_{s-1})} f(t_s). \tag{2}$$

Этот метод называется *методом секущих*.

Иногда употребляют также *метод хорд*. Предполагается, что известны начальные приближения t_0 и t_1, такие что $f(t_0)$ и $f(t_1)$ имеют противоположные знаки. Тогда построенное по формуле (2) (при $s = 1$) приближение t_2 лежит между t_0 и t_1. Из промежутков $[t_0, t_2]$ и $[t_2, t_1]$ выбирается тот, на концах которого функция f принимает значения разных знаков, и делается новый шаг, аналогичный предыдущему.

Вернемся к методу Ньютона. Формулу (1) можно рассматривать как формулу метода итераций для уравнения $t = \varphi(t)$, где $\varphi(t) = t - f(t)/f'(t)$. Легко проверить, что $\varphi'(t^*) = 0$. Поэтому, в связи с результатами предыдущего параграфа, следует ожидать квадратичную сходимость метода.

Аналогично модифицированный метод Ньютона можно рассматривать как метод итерации для уравнения $t = \tilde{\varphi}(t)$, где $\tilde{\varphi}(t) = t - f(t)/f(t_0)$ Здесь $\tilde{\varphi}'(t_0) = 0$, и при достаточно хорошем начальном приближении можно ожидать сходимость этого метода с быстротой геометрической прогрессии с малым знаменателем.

Исследование методов Ньютона и модифицированного будем проводить для случая системы уравнений

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

$(x = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n,\ F(x) = \big(f_1(x), \ldots, f_n(x)\big))$. Согласно формуле Тейлора

$$\mathbb{O} = F(x^*) \approx F(x_0) + F'(x_0)(x^* - x_0).$$

Так что исходя из начального приближения x_0 последующие по методу Ньютона строятся по формуле, вполне аналогичной (1):

$$x_{s+1} = x_s - [F'(x_s)]^{-1} F(x_s). \tag{4}$$

Формула модифицированного метода:

$$\tilde{x}_{s+1} = \tilde{x}_s - [F'(x_0)]^{-1} F(\tilde{x}_s). \tag{5}$$

Реально в методе Ньютона нахождение x_{s+1} требует не обращения матрицы (что нерационально), а решения системы уравнений:

$$x_{s+1} = x_s + \Delta x_s, \qquad F'(x_s)\Delta x_s = -F(x_s).$$

В модифицированном методе обращение матрицы может оказаться оправданным.

Теорема 1 (о методе Ньютона). Пусть x^* – решение системы уравнений (1). Пусть в некоторой окрестности точки x^* отображение $F(x)$ трижды непрерывно дифференцируемо и $\det F'(x^*) \neq 0$. Тогда метод Ньютона сходится для x^* квадратически.

Доказательство. $F'(x) \to F(x^*)$ при $x \to x^*$. Но при условии $\|F'(x) - F'(x^*)\| \cdot \|[F'(x^*)]^{-1}\| < 1$ матрица Якоби $F'(x)$ также обратима (мы используем теорему 5 §2 главы 3). Поэтому для некоторого $r > 0$ при всех тех x, для которых $\|x - x^*\| \leq r$, существует матрица $\Gamma(x) = [F'(x)]^{-1} = \{\gamma_{kj}(x)\}$. Легко видеть, что функции $\gamma_{kj}(x)$ дважды дифференцируемы. Метод Ньютона теперь запишется так: $x_{s+1} = x_s - \Gamma(x_s)F(x_s) = \Phi(x_s)$. Отображение $\Phi(x)$ дважды непрерывно дифференцируемо в окрестности точки x^*, и чтобы воспользоваться теоремой 2 из предыдущего параграфа, достаточно убедиться, что $\Phi'(x^*) = \mathbb{O}$. Для компонент отображения Φ имеем

$$\varphi_k(x) = \xi_k - \sum_{j=1}^{n} \gamma_{kj}(x) \cdot f_j(x),$$

$$\frac{\partial \varphi_k}{\partial \xi_l} = \delta_{kl} - \sum_{j=1}^{n} \left[\frac{\partial \gamma_{kj}}{\partial \xi_l}(x) f_j(x) + \gamma_{kj}(x) \frac{\partial f_j}{\partial \xi_l}(x) \right].$$

При $x = x^*$ первое слагаемое в квадратных скобках обращается в ноль, а второе есть (k,l)-ый элемент матрицы $\Gamma(x^*)F'(x^*) = E$, т.е. δ_{kj}. Итак, все элементы матрицы $\Phi'(x^*)$ равны нулю, $\Phi'(x^*) = \mathbb{O}$, и для завершения доказательства остается воспользоваться теоремой 2 из §2. ■

Замечание 1. В условиях этой теоремы требование на гладкость отображения F завышено — в действительности достаточно двукратной дифференцируемости.

Замечание 2. Условие $\det F'(x^*) \neq 0$ теоремы существенно – без него квадратичной сходимости может и не быть. Приведем соответствующий пример в одномерном случае. Пусть $f(t) = t^2$. Корень $t^* = 0$ уравнения $f(t) = 0$ таков, что $f'(t^*) = 0$. Для метода Ньютона легко получаем $t_{s+1} = \frac{1}{2}t_s$, и при $t_0 \neq 0$ сходимость метода всего лишь линейная.

Теорема 2 (о модифицированном методе Ньютона). Пусть $F(x^*) = 0$ и в некоторой окрестности x^* отображение F дважды непрерывно дифференцируемо. Пусть $\det F'(x^*) \neq 0$. Тогда найдутся такие $\delta > 0$ и $c > 0$, что при $x_0 \in S_\delta(x^*)$ модифицированный метод Ньютона сходится с быстротой геометрической прогрессии: $\|\tilde{x}_s - x^*\| \leq q^s \|x_0 - x^*\|$, где $q \leq c\|x_0 - x^*\|$.

Доказательство. Обозначим r радиус того шара $S_r(x^*)$, в котором отображение F дважды дифференцируемо. Найдется такая постоянная c_1, что для $x', x'' \in S_r(x^*)$ выполняется неравенство $\|F'(x') - F'(x'')\| \leq c_1\|x' - x''\|$. Выберем $r_0 < r$ так, что $\|[F'(x^*)]^{-1}\|c_1 r_0 < \frac{1}{2}$. По теореме 5 из §2 главы 4 для всех $x \in S_{r_0}(x^*)$ существует обратная матрица $\Gamma(x) = [F'(x)]^{-1}$ и $\|\Gamma(x)\| \leq 2\|\Gamma(x^*)\| = c_2$. Положим $c = 2c_1c_2$ и выберем δ так, что $\delta \leq r_0$ и $c\delta < 1$. Итак, c и δ выбраны. Покажем, что они требуемые. Пусть $x_0 \in S_\delta(x^*)$. В модифицированном методе Ньютона $\tilde{x}_0 = x_0$ и $\tilde{x}_{s+1} = \Phi(\tilde{x}_s)$, где $\Phi(x) = x - \Gamma_0 F(x)$, $\Gamma_0 = \Gamma(x_0)$. Так как $\Phi'(x) = E - \Gamma_0 F'(x) = \Gamma_0\big(F'(x_0) - F'(x)\big)$,то при $\|x - x^*\| \leq \|x_0 - x^*\|$ будет

$$\|\Phi'(x)\| \leq c_2c_1\|x_0 - x\| \leq c_1c_2(\|x_0 - x^*\| + \|x - x^*\|) \leq c\|x_0 - x^*\| = q < 1.$$

Поэтому по теореме 2 из §1 для любых точек x', x'', таких что $\|x' - x^*\|$, $\|x'' - x^*\| \leq \|x_0 - x^*\|$ имеем

$$\|\Phi(x') - \Phi(x'')\| \leq q\|x' - x''\|.$$

Отсюда

$$\|\tilde{x}_s - x^*\| = \|\Phi(\tilde{x}_{s-1}) - \Phi(x^*)\| \leq q\|\tilde{x}_{s-1} - x^*\| \leq \cdots \leq q^s\|x_0 - x^*\|.$$

Этим теорема доказана. ■

Известны методы и более высокого, чем второй, порядка сходимости. Например, методами третьего порядка являются метод касательных гипербол и Чебышева, вычислительные формулы которых в случае одного вещественного уравнения $f(t) = 0$ выглядят соответственно так:

$$t_{s+1} = t_s - \frac{1}{1 - \frac{f''(t_s)f(t_s)}{2[f'(t_s)]^2}} \frac{f(t_s)}{f'(t_s)},$$

$$t_{s+1} = t_s - \left(1 + \frac{f(t_s)f''(t_s)}{2[f'(t_s)]^2}\right) \frac{f(t_s)}{f'(t_s)}.$$

Методы высших порядков обычно менее эффективны, чем метод Ньютона, особенно в случае систем уравнений. Например, при решении систем указанные методы третьего порядка на одном шаге требуют вычисления самих функций φ_k (их n), их первых производных (их n^2) и вторых (их n^3) и

решения двух систем линейных уравнений порядка n. Так что один шаг такого метода более трудоемок, чем два шага метода Ньютона. В то же время, грубо говоря, один шаг такого метода возводит погрешность в третью степень, в то время как два шага метода Ньютона – в четвертую.

Но бывают случаи, когда начальное приближение чем-то особенно просто – в этой точке уже известны все производные, среди них много нулей и т.п. Если одного шага метода Ньютона здесь не хватает для достижения нужной точности, то применение методов высшего порядка оправдано. Возможно, методы высших порядков имеют некоторое преимущество при распараллеливании вычислительных процессов.

Задача. Показать, что в случае одного уравнения для четырежды непрерывно дифференцируемой функции f методы Чебышева и касательных гипербол имеют третий порядок сходимости.

Глава 5
Численное решение задачи Коши

§1 Простейшие методы

Будем рассматривать задачу Коши для обыкновенного дифференциального уравнения (ОДУ)

$$y' = f(x, y), \quad y(x_0) = y_0. \tag{1}$$

Численное решение этой задачи — построение таблицы приближенного решения, чаще всего для равноотстоящих узлов $x_k = x_0 + kh$ ($h > 0$ — шаг): $y_k \approx y(x_k)$.

Метод Эйлера.

В точке x_0 легко найти производную решения: $y'(x_0) = f(x_0, y_0)$. Поскольку (при малом шаге h)

$$y(x_1) = y(x_0 + h) \approx y_0 + hy'(x_0) = y_0 + hf(x_0, y_0),$$

то можно положить $y_1 = y_0 + hf(x_0, y_0)$. По аналогичной формуле можно найти y_2 и т.д. В этом и состоит метод Эйлера:

$$y(x_{k+1}) \approx y_{k+1} = y_k + hf(x_k, y_k). \tag{2}$$

В методе Эйлера не обязательно считать узлы x_k равноотстоящими. Если $x_{k+1} = x_k + h_k$, то в правой части (2) h следует заменить на h_k.

Метод Эйлера имеет простой геометрический смысл. Если каждую точку (x_k, y_k) соединить отрезком со следующей, то мы получим ломаную линию — график приближенного решения. Поэтому метод Эйлера называют иногда *методом ломаных*. Звенья этой ломаной — отрезки касательных, проведенных в точке (x_k, y_k) к проходящей через эту точку интегральной кривой.

Если функцию f считать дифференцируемой, то погрешность метода Эйлера на одном шаге имеет порядок $\mathcal{O}(h^2)$. Это означает, что в предположении, что $y_k = y(x_k)$ будет $y(x_{k+1}) - y_{k+1} = \mathcal{O}(h^2)$. В действительности, кроме ошибки, вызванной тем, что в формуле Тейлора мы ограничились лишь двумя членами, в y_{k+1} войдет еще и "наследственная" ошибка — y_{k+1} не совпадает с $y(x_{k+1})$, в частности, и по той причине, что все предыдущие значения $y_k, \ldots, y_1$ были найдены неточно. Для получения решения нашей задачи на промежутке $[x_0, X]$ нам придется сделать примерно $(X - x_0)/h$ шагов, и так как на каждом шаге мы допускаем ошибку порядка h^2, то можно ожидать, что ошибка, допущенная на всем промежутке, будет порядка h. Влияние "наследственных ошибок" зависит от свойств решаемого уравнения. В частности, если интегральные линии с ростом x сближаются (это так, если $\partial f/\partial y < 0$), то влияние "наследственных ошибок" убывает, а если расходятся ($\partial f/\partial y > 0$), то возрастает.

Метод Эйлера прост. Главный его недостаток — малая точность. В связи с этим рассматриваются некоторые усовершенствования этого метода. Сам метод Эйлера допускает такую трактовку: в равенстве

$$y(x_{k+1}) = y(x_k) + \int_{x_k}^{x_{k+1}} y'(t)dt = y(x_k) + \int_{x_k}^{x_{k+1}} f\big(t, y(t)\big)dt \tag{3}$$

стоящий в правой части интеграл приближенно заменяется на $hf(x_k, y_k)$. Усовершенствования связаны с более точным приближением этого интеграла.

1-й усовершенствованный метод.

Упомянутый выше интеграл будем вычислять по формуле средних прямоугольников, причем неизвестное нам значение подынтегральной функции в середине промежутка будем находить с помощью метода Эйлера. Это приводит с следующему алгоритму:

$$y_{k+1} = y_k + hf\left(x_k + \frac{h}{2}, y_{k+1/2}\right), \quad \text{где} \quad y_{k+1/2} = y_k + \frac{h}{2}f(x_k, y_k).$$

Оценим погрешность этой формулы на одном шаге (для краткости обозначений — на первом), считая при этом функцию f нужное число раз дифференцируемой. В понятных обозначениях для точного решения имеем

$$y(x_1) = y_0 + hy_0' + \frac{h^2}{2}y_0'' + \mathcal{O}(h^3), \quad y_0' = f(x_0, y_0), \quad y_0'' = f_x' + y_0' f_y'.$$

Для приближенного решения:

$$y_1 = y_0 + hf\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2}y_0'\right) =$$

$$= y_0 + hf(x_0, y_0) + \frac{h^2}{2}f_x' + \frac{h^2}{2}y_0' f_y' + \mathcal{O}(h^3) = y(x_1) + \mathcal{O}(h^3).$$

Итак, погрешность метода на одном шаге имеет порядок малости h^3.

2-ой усовершенствованный метод.

Интеграл, стоящий в правой части (3), будем вычислять по формуле трапеций, опять используя для приближения неизвестного значения подынтегральной функции на правом конце промежутка метод Эйлера:

$$y_{k+1} = y_k + \frac{h}{2}\big(f(x_k, y_k) + f(x_{k+1}, \tilde{y}_{k+1})\big), \text{ где } \tilde{y}_{k+1} = y_k + hf(x_k, y_k).$$

При оценке ошибки этого метода на одном шаге для точного решения будем использовать то же разложение, что и выше, а для приближенного:

$$y_1 = y_0 + \frac{h}{2}\big(f(x_0, y_0) + f(x_0 + h, y_0 + hy_0')\big) =$$

$$= y_0 + hf(x_0, y_0) + \frac{h^2}{2}f_x' + \frac{h^2}{2}y_0' f_y' + \mathcal{O}(h^3) = y(x_1) + \mathcal{O}(h^3).$$

Итак, ошибка этого метода на одном шаге также имеет порядок малости h^3.

При решении задачи (1) на заданном промежутке обоими этими усовершенствованными методами можно ожидать погрешность порядка $\mathcal{O}(h^2)$.

Задача. Найти порядок погрешности на одном шаге метода, определяемого (на первом шаге) формулами:

$$\overline{y}_1 = y_0 + hf(x_0, y_0), \quad y_{1/2} = y_0 + \frac{h}{2}f(x_0, y_0), \quad \tilde{y}_1 = y_{1/2} + \frac{h}{2}f\left(x_0 + \frac{h}{2}, y_{1/2}\right),$$

$$y_1 = 2\tilde{y}_1 - \overline{y}_1.$$

§2 Методы Адамса.

Квадратурные формулы Адамса.

При рассмотрении во второй главе интерполяционных квадратурных формулах предположение о том, что узлы этих формул принадлежали промежутку интегрирования, не делалось, хотя это было так в приведенных там конкретных формулах. Сейчас мы рассмотрим формулы, в которых часть узлов не принадлежит этому промежутку.

Определение 1. *Экстраполяционными квадратурными формулами Адамса* называются интерполяционные квадратурные формулы для промежутка $[0,1]$ с узлами $0,-1,\dots,-p$

$$\int_0^1 g(x)dx = \sum_{j=0}^{p} \alpha_j^p g(-j) + R_p(g), \tag{1}$$

а также подобные им формулы.

Согласно определению интерполяционных квадратурных формул коэффициенты α_j^p вычисляются по правилу

$$\alpha_j^p = \int_0^1 \left(\prod_{k=0,k\neq j}^{p} (k+x) \right) \Big/ \left(\prod_{k=0,k\neq j}^{p} (k-j) \right) dx.$$

Найдем представление остаточного члена формулы (1) в предположении должной дифференцируемости функции g. Пусть $P_p(x)$ — интерполяционный полином функции g, построенный по узлам $0,-1,\dots,-p$, и $\omega(x)=\prod_{k=0}^{p}(x+k)$. Поскольку формула (1) интерполяционная. то

$$R_p(g) = \int_0^1 \big(g(x)-P_p(x)\big)dx = \frac{1}{(p+1)!}\int_0^1 \omega(x) g^{(p+1)}\big(\eta(x)\big)dx =$$
$$= \frac{1}{(p+1)!}\int_0^1 \omega(x)dx\, g^{(p+1)}(\xi) = A_p g^{(p+1)}(\xi) \quad \xi\in(-p,1).$$

Мы воспользовались представлением остатка интерполяции, леммой из §1 главы 2 и неотрицательностью полинома $\omega(x)$ на промежутке $[0,1]$.

Приведем таблицу коэффициентов формулы (1) для первых значений p, а также значений коэффициентов A_p остаточного члена.

	$j=0$	$j=1$	$j=2$	$j=3$	A_p
$p=1$	$3/2$	$-1/2$			$5/12$
$p=2$	$23/12$	$-4/3$	$5/12$		$3/8$
$p=3$	$55/24$	$-59/24$	$37/24$	$-3/8$	$251/720$

Определение 2. *Интерполяционными квадратурными формулами Адамса* называются интерполяционные квадратурные формулы для промежутка $[0,1]$ с узлами $1,0,-1,\dots,-(p-1)$

$$\int_0^1 g(x)dx = \sum_{j=-1}^{p-1} \beta_j^p g(-j) + R_p(g), \tag{2}$$

а также подобные им формулы.

При $p = 1$ эта формула совпадает с формулой трапеций. Коэффициенты формулы (2):

$$\beta_j^p = \int_0^1 \left(\prod_{k=-1, k\neq j}^{p-1} (k+x) \right) \Big/ \left(\prod_{k=-1, k\neq j}^{p-1} (k-j) \right) dx.$$

Поскольку полином $\omega(x) = (x-1)x(x+1)\dots(x+p-1)$ не меняет знака на $[0,1]$, то как и в случае экстраполяционной формулы Адамса легко получить представление остаточного члена в форме Лагранжа:

$$R_p(g) = B_p g^{(p+1)}(\xi), \quad B_p = \frac{1}{(p+1)!} \int_0^1 \omega(x)dx, \quad \xi \in (-p+1, 1).$$

Приведем таблицу коэффициентов β_j^p и B_p.

	$j=-1$	$j=0$	$j=1$	$j=2$	B_p
$p=1$	$1/2$	$1/2$			$-1/12$
$p=2$	$5/12$	$2/3$	$-1/12$		$-1/24$
$p=3$	$3/8$	$19/24$	$-5/24$	$1/24$	$-19/720$

Формулы Адамса обычно применяются, когда функция f задана таблицей своих значений в равноотстоящих узлах $x_k = x_0 + kh$: $f(x_k) = f_k$ и требуется вычислить интеграл от этой функции по промежутку между двумя соседними узлами. Применяя формулы, подобные формулам Адамса, тогда получим в случае экстраполяционной формулы

$$\int_{x_k}^{x_{k+1}} f(x)dx \approx h \sum_{j=0}^{p} \alpha_j^p f_{k-j}, \qquad R_p(f) = A_p h^{p+2} f^{(p+1)}(\xi),$$

а в случае интерполяционной

$$\int_{x_k}^{x_{k+1}} f(x)dx \approx h \sum_{j=-1}^{p-1} \beta_j^p f_{k-j}, \qquad R_p(F) = B_p h^{p+2} f^{(p+1)}(\xi).$$

Часто эти формулы записывают в другом виде, с использованием конечных разностей. Действительно, квадратурная сумма есть интеграл от интерполяционного полинома функции f. В случае экстраполяционной формулы воспользуемся формулой Ньютона для конца таблицы

$$P_p(x_k + th) = f_k + t\Delta f_{k-1} + \frac{t(t+1)}{2!}\Delta^2 f_{k-2} + \dots$$

и тем, что

$$\int_{x_k}^{x_{k+1}} P_p(x)dx = h \int_0^1 P_p(x_k + th)dt.$$

Тогда получим экстраполяционную формулу Адамса в виде:

$$\int_{x_k}^{x_{k+1}} f(x)dx \approx h\Big[f_k + \frac{1}{2}\Delta f_{k-1} + \frac{5}{12}\Delta^2 f_{k-2} + \frac{3}{8}\Delta^3 f_{k-3} + \\ + \frac{251}{720}\Delta^4 f_{k-4} + \cdots + a_p\Delta^p f_{k-p}\Big], \quad R_p(f) = A_p h^{p+2} f^{(p+1)}(\xi). \quad (3)$$

В случае интерполяционной формулы

$$P_p(x_{k+1} + th) = f_{k+1} + t\Delta f_k + \frac{t(t+1)}{2!}\Delta^2 f_{k-1} + \ldots,$$

$$\int_{x_k}^{x_{k+1}} P_p(x)dx = h\int_{-1}^{0} P_p(x_{k+1} + th)dt,$$

так что интерполяционная формула Адамса принимает вид

$$\int_{x_k}^{x_{k+1}} f(x)dx \approx h\Big[f_{k+1} - \frac{1}{2}\Delta f_k - \frac{1}{12}\Delta^2 f_{k-1} - \frac{1}{24}\Delta^3 f_{k-2} - \\ - \frac{19}{720}\Delta^4 f_{k-3} - \cdots - b_p\Delta^p f_{k-p+1}\Big], \quad R_p(f) = B_p h^{p+2} f^{(p+1)}(\xi). \quad (4)$$

Экстраполяционный метод Адамса.
Вернемся к задаче Коши:

$$y' = f(x, y), \quad y(x_0) = y_0.$$

Мы ищем значения приближенного решения в равноотстоящих узлах $x_j = x_0 + jh$. Пусть при $j = 0, \ldots, k$, где $k \geq p - 1$, эти приближенные значения $y_j \approx y(x_j)$ уже найдены. Будем искать y_{k+1}. По формуле Ньютона - Лейбница

$$y(x_{k+1}) = y(x_k) = \int_{x+k}^{x_{k+1}} y'(x)dx.$$

Для вычисления интеграла воспользуемся экстраполяционной формулой Адамса, заменив значения y' в узлах приближенными значениями $y'(x_j) \approx f(x_j, y_j)$, и заменим $y(x_k)$ на y_k. Тогда, полагая $\eta_j = hf(x_j, y_j)$, получим

$$y_{k+1} = y_k + \eta_k + \Delta\eta_{k-1} + \cdots + a_p\Delta^p\eta_{k-p} \quad (5)$$

или в безразностной форме

$$y_{k+1} = y_k + \sum_{j=0}^{p} \alpha_j^p \eta_{k-p}.$$

Метод Адамса есть способ *продолжения* таблицы — первые значения искомой функции должны быть найдены другим способом. Если метод применяется в разностной форме, то к тому моменту, когда мы вычисляем y_{k+1},

мы должны располагать таблицей (в случае $p = 3$)

$$\begin{array}{cccccc}
x_j & y_j & \eta_j & \Delta\eta_j & \Delta^2\eta_j & \Delta^3\eta_j \\
\dots & \dots & \dots & & \dots & \\
 & & & \dots & & \dots \\
x_{k-2} & y_{k-2} & \eta_{k-2} & & \Delta^2\eta_{k-3} & \\
 & & & \Delta\eta_{k-2} & & \Delta^3\eta_{k-3} \\
x_{k-1} & y_{k-1} & \eta_{k-1} & & \Delta^2\eta_{k-2} & \\
 & & & \Delta\eta_{k-1} & & \\
x_k & y_k & \eta_k & & &
\end{array}$$

Значение y_{k+1} вычисляется по формуле (5). В эту формулу входят разности, находящиеся в нижней косой строке. После этого вычисляется $\eta_{k+1} = hf(x_{k+1}, y_{k+1})$, заполняется следующая косая строка разностей, и все готово для следующего шага. Безразностная форма метода позволяет избежать вычисления разностей.

Счет с разностями обычно применяется при ручных вычислениях. Он имеет то преимущество, что позволяет контролировать разумность выбранного порядка метода p и шага h — первый отброшенный в формуле (5) член не должен превосходить принятую точность вычисления решения. Впрочем, при $p > 5$ метод обычно не применяется, поскольку в разностях высокого порядка сильно влияние ошибок округления.

При счете на компьютере обычно применяется безразностная формула. Тогда требуются другие способы контроля за шагом h.

Интерполяционный метод Адамса.

Отличие этого метода от экстраполяционного состоит в том, что для вычисления интеграла по $[x_k, x_{k+1}]$ от производной решения используется интерпроляционная квадратурная формула Адамса. Тогда мы приходим к формуле

$$y_{k+1} = y_k + \eta_{k+1} - \frac{1}{2}\Delta\eta_k - \frac{1}{12}\Delta^2\eta_{k-1} - \dots \tag{6}$$

или в безразностной форме

$$y_{k+1} = y_k + \sum_{j=-1}^{p-1} \beta_j^p \eta_{k-j}. \tag{7}$$

Когда нам следует вычислить y_{k+1}, величина η_{k+1} (и ее разности, входящие в (6), если мы используем эту формулу) нам неизвестна. Поэтому (6) и (7) следует рассматривать как *уравнение* вида $y_{k+1} = \varphi(y_{k+1})$, решив которое, мы и найдем искомое значение y_{k+1}. Уравнение легко решается методом итерации: выбрав начальное приближение y_{k+1}^0, мы строим последующие $y_{k+1}^\nu = \varphi(y_{k+1}^{\nu-1})$. Легко видеть, что $\varphi'(y) = h\beta_{-1}^p f_y'(x_{k+1}, y)$. Поскольку вычисления обычно ведутся с малым шагом h, эта производная мала по абсолютной величине, и метод итерации сходится очень быстро. Хорошее начальное приближение y_{k+1}^0 получают обычно, используя экстраполяционный метод Адамса. При использовании конечных разностей есть и другой способ построения начального приближения. Обычно шаг h выбирается так, что последние используемые

разности порядка p (для определенности пусть $p = 3$) почти постоянны с принятой точностью вычислений. Поэтому полагают $\Delta^3\eta^0_{k-2} = \Delta^3\eta_{k-1}$ и путем сложения заполняют косую строку разностей вплоть до η^0_{k+1}, после чего y^0_{k+1} находят по формуле (6).

При вычислениях на компьютере без разностей часто используется комбинация экстраполяционного и интерполяционного методов Адамса. Сначала вычисляется по экстраполяционному методу Адамса значение y^0_{k+1}, и оно уточняется по формуле (7). По разности этих двух значений можно судить о пригодности выбранного шага вычислений.

Системы уравнений и уравнения высшего порядка.

Методы Адамса применимы и для решения задачи Коши для систем дифференциальных уравнений. Для определенности ограничимся рассмотрением системы второго порядка (обобщение на случай любого порядка очевидно):

$$y' = f(x, y, z), \quad z' = g(x, y, z), \quad y(x_0) = y_0, \quad z(x_0) = z_0.$$

Формулы методов Адамса выводились при единственном предположении, что $\eta_k \approx hy'(x_k)$. Пусть приближенные значения $y_j \approx y(x_j)$ и $z_j \approx z(x_j)$ при $j = 0, \ldots, k$ уже известны. Учитывая, что $hy'(x_j) \approx hf(x_j, y_j, z_j) = \eta_j$ и $hz'(x_j) \approx hg(x_j, y_j, z_j) = \zeta_j$, мы можем для нахождения y_{k+1} и z_{k+1} использовать формулы Адамса

$$y_{k+1} = y_k + \sum_{j=0}^{p} \alpha^p_j \eta_j, \quad z_{k+1} = z_k + \sum_{j=0}^{p} \alpha^p_j \zeta_j$$

в случае экстраполяционного метода, а в случае интерполяционного

$$y_{k+1} = y_k + \sum_{j=-1}^{p-1} \beta^p_j \eta_j, \quad z_{k+1} = z_k + \sum_{j=-1}^{p-1} \beta^p_j \zeta_j.$$

Эти формулы можно было бы записать и с конечными разностями. В случае интерполяционного метода на каждом шаге нам, естественно, придется решать *систему* двух уравнений методом итерации.

Методы Адамса применимы и для решения задачи Коши для уравнений высшего порядка:

$$y^{(m)} = f(x, y, y', \ldots, y^{(m-1)}), \quad y(x_0) = y_0, \ \ldots, \ y^{(m-1)}(x_0) = y_0^{(m-1)},$$

поскольку эта задача легко сводится к задаче Коши для системы введением новых неизвестных функций $z_l = y^{(l)}$ $(l = 1, \ldots, m-1$:

$$\left.\begin{array}{c} y' = z_1 \\ z_1' = z_2 \\ \cdots \\ z'_{m-2} = z_{m-1} \\ z'_{m-1} = f(x, y, z_1, \ldots, z_{m-1}) \end{array}\right\}$$

$$y(x_0) = y_0, \quad z_1(x_0) = y_0', \quad \ldots, \quad z_{m-1}(x_0) = y_0^{(m-1)}.$$

Два заключительных замечания. Во-первых, еще раз напомним, что оба метода Адамса требуют, чтобы начало таблицы было построено каким-либо другим способом. Во-вторых, если в процессе счета нам потребовалось изменить шаг интегрирования, то это связано с необходимостью построения нового начала таблицы, что может быть сделано, например, с помощью интерполяции.

§3 Способы построения начала таблицы

В этом параграфе излагаются некоторые способы построения начала таблицы, т.е. вычисления приближенных значений $y_j \approx y(x_j)$ $(x_j = x_0 + jh)$ решения задачи Коши

$$y' = f(x, y), \quad y(x_0) = y_0 \tag{1}$$

при малых j.

1. Разложение решения в ряд Тейлора.

Пусть $y(x)$ — решение нашей задачи: $y'(x) = f\big(x, y(x)\big)$. Дифференцируя это тождество, мы получим:

$$y''(x) = f'_x\big(x, y(x)\big) + f'_y\big(x, y(x)\big),$$
$$y'''(x) = f''_{xx}\big(x, y(x)\big) + 2f''_{xy}\big(x, y(x)\big)y'(x) + f''_{yy}\big(x, y(x)\big)[y'(x)]^2 + f'_y\big(x, y(x)\big)y''(x)$$

и, продолжая дифференцирование, представление следующих производных. Учитывая, что $y'(x_0) = f(x_0, y_0)$, и подставляя x_0 в приведенные формулы, мы можем найти в этой точке значения производных решения $y^{(\nu)}(x_0) = y_0^{(\nu)}$. После этого приближенные значения решения в интересующих нас точках находятся из ряда Тейлора:

$$y(x_j) \approx y_j = y_0 + (jh)y'_0 + \frac{(jh)^2}{2!}y''_0 + \cdots + \frac{(jh)^n}{n!}y_0^{(n)}.$$

При построении начала таблицы начальные точки выгодно брать как правее, так и левее точки x_0. Например, если мы собираемся применять метод Адамса 4-го порядка и нам требуется знать пять первых значений, то в качестве начальных точек берут $x_{-2}, x_{-1}, x_0, x_1, x_2$. Это лучше, чем x_0, x_2, x_2, x_3, x_4, поскольку позволяет брать меньше членов в разложении Тейлора.

2. Итеративный метод (метод А.Н.Крылова)

Для определенности будем считать, что требуется построить (кроме известного x_0) еще три первых значения решения задачи (1) y_1, y_2, y_3. Нетрудно построить интерполяционные квадратурные формулы

$$\int_k^{k+1} g(x)dx \approx \sum_{j=0}^{3} A_j^k g(j), \quad k = 0, 1, 2. \tag{2}$$

Приведем таблицу их коэффициентов

	$j=0$	$j=1$	$j=2$	$j=3$
$k=0$	$3/8$	$19/24$	$-5/24$	$1/24$
$k=1$	$-1/24$	$13/24$	$13/24$	$-1/24$
$k=2$	$1/24$	$-5/24$	$19/24$	$3/8$

Тогда, используя квадратурные формулы, подобные (2), для решения $y(x)$ имеем

$$y(x_{k+1})-y(x_k)=\int_{x_k}^{x_{k+1}} y'(x)dx=\int_{x_k}^{x_{k+1}} f\big(x,y(x)\big)dx\approx h\sum_{j=0}^{3}A_j^k f(x_j,y_j)=\Delta y_k.$$

Таким образом для неизвестных Δy_k $(k=0,1,2)$ мы получаем (учитывая, что y_0 известно) систему уравнений

$$\Delta y_k=\varphi_k(\Delta y_0,\Delta y_1,\Delta y_2),\quad k=0,1,2,$$

где

$$\varphi_k(\Delta y_0,\Delta y_1,\Delta y_2)=h\sum_{j=0}^{3}A_j^k f(x_j,y_j),\quad y_j=y_0+\sum_{l=0}^{j-1}\Delta y_l.$$

При малых h производные функций φ_k малы, и для решения системы быстро сходится метод итерации. Начальное приближение может быть построено, например, методом Эйлера.

Оба изложенных метода применимы и в случае задачи Коши для системы дифференциальных уравнений, а значит, и для уравнений высшего порядка.

Для построения начала таблицы можно применять также излагаемый в следующем параграфе метод Рунге - Кутта.

Задача. Объяснить связь при $k=0$ и $k=2$ коэффициентов A_j^k, приведенных в таблице, с коэффициентами квадратурной интерполяционной формулы Адамса.

§4 Метод Рунге - Кутта

Начнем сразу с вычислительных формул. Пусть для задачи Коши

$$y'=f(x,y),\qquad y(x_0)=y_0 \tag{1}$$

уже найдено приближенное значение решения в точке x_n: $y(x_n)\approx y_n$, и требуется найти решение в точке $x_{n+1}=x_n+h$. Согласно методу Рунге - Кутта это делается по формулам:

$$k_1=hf(x_n,y_n),\quad k_2=hf\left(x_n+\frac{h}{2},y_n+\frac{k_1}{2}\right),$$

$$k_3=hf\left(x_n+\frac{h}{2},y_n+\frac{k_2}{2}\right),\quad k_4=hf(x_n+h,y_n+k_3),$$

$$y_{n+1}=y_n+\frac{1}{6}(k_1+2k_2+2k_3+k_4). \tag{2}$$

Это — некоторый аналог применения для вычисления интеграла в равенстве

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} y'(t)dt$$

квадратурной формулы Симпсона. Действительно, $k_1 \approx hy'(x_n)$, k_2 и k_3 — некоторые приближения к $hy'(x_n + h/2)$, а k_4 — к $hy'(x_{n+1})$.

Основным и неочевидным свойством метода Рунге - Кутта является то, что его погрешность на одном шаге имеет порядок малости $\mathcal{O}(h^5)$. Это означает, что если функция f имеет нужное число производных, а y_n точно совпадает с $y(x_n)$ и мы разложим $y(x_{n+1})$ и правую часть равенства (2) по степеням h, то расхождения в этих разложениях начнутся лишь с членов, содержащих h^5. Связанные с этим выкладки чрезвычайно громоздки[17], и мы их приводить не будем.

Метод Рунге - Кутта применим и для систем дифференциальных уравнений, а значит, и для уравнений высшего порядка. Для системы двух уравнений

$$y' = f(x, y, z), \quad z' = g(x, y, z).$$

формулы метода таковы

$$\begin{array}{ll} k_1 = hf(x_n, y_n, z_n) & l_1 = hg(x_n, y_n, z_n) \\ k_2 = hf\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}, z_n + \frac{l_1}{2}\right) & l_2 = hg\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}, z_n + \frac{l_1}{2}\right) \\ k_3 = hf\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}, z_n + \frac{l_2}{2}\right) & l_3 = hg\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}, z_n + \frac{l_2}{2}\right) \\ k_4 = hf(x_n + h, y_n + k_3, z_n + l_3) & l_4 = hg(x_n + h, y_n + k_3, z_n + l_3), \end{array}$$

$$y_{n+1} = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad z_{n+1} = \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4).$$

Отметим некоторые свойства метода Рунге - Кутта в сравнении с методами Адамса. На одном шаге метод Рунге - Кутта имеет погрешность того же порядка, что и методы Адамса при $p = 3$. В то же время в экстраполяционном методе Адамса на одном шаге требуется вычислять лишь одно значение правой части f, в интерполяционном — в зависимости от числа итераций, в среднем можно считать 2-3 значения, а в методе Рунге - Кутта 4 значения. Если использовать методы Адамса при $p = 4, 5$, то и шаг в этих методах можно выбирать больше. Так что метод Рунге - Кутта существенно более трудоемок, чем методы Адамса. По этой причине до появления вычислительных машин метод Рунге - Кутта рассматривался в основном как способ построения начала таблицы. Но с появлением машин на первый план вышли некоторые преимущества метода Рунге - Кутта, о которых будет сказано ниже, и этот метод применяется для решения задач Коши на больших промежутках. Впрочем, если решение задач Коши является составной частью больших вычислений, и таких задач требуется решить очень много, то фактор трудоемкости может оказаться очень существенным.

[17] Их можно найти в книге И.П.Мысовских "Лекции по методам вычислений", СПбГУ 1998

Метод Рунге - Кутта имеет два важных преимущества перед методами Адамса. Во-первых, он не требует предварительного построения начала таблицы и сам пригоден для такого построения. Во-вторых, шаг в этом методе не обязательно постоянный, и его изменение в любой момент не связано с дополнительными трудностями. Напомним, что в методах Адамса переход к новому шагу связан с необходимостью строить новое начало таблицы.

Для контроля за выбором шага в методе Рунге - Кутта иногда применяется просчет одной задачи с разными шагами.

Задача. Доказать основное свойство метода Рунге - Кутта (порядок точности на одном шаге) для случая простейшего дифференциального уравнения $y' = f(x)$.

§5 О граничных задачах

.

Существуют приближенные методы решения граничных задач, общие для обыкновенных дифференциальных уравнений и уравнений в частных производных (метод сеток, проекционные методы). Эти методы будут изучаться во второй части курса (7-й семестр). В этом параграфе речь пойдет лишь о методах, специфичных для ОДУ. Это методы сведения граничной задачи к задачам Коши. Первый из них так и называется:

Метод сведения к задачам Коши.

Будем рассматривать граничную задачу для *линейного* ОДУ второго порядка:

$$\begin{aligned} L(y) =& y'' + p(x)y' + q(x)y = f(x), \\ & A_0y'(a) + B_0y(a) = D_0, \quad A_1y'(b) + B_1y(b) = D_1. \end{aligned} \tag{1}$$

Общее решение нашего дифференциального уравнения имеет вид:

$$y(x) = C_1z_1(x) + C_2z_2(x) + y_0(x), \tag{2}$$

где y_0 — какое-либо частное решение ($L(y_0) = f$), а z_1 и z_2 — линейно независимые решения однородного уравнения $L(z_k) = 0$. Если мы построим y_0, z_1 и z_2 как решения некоторых задач Коши, то нам останется только подобрать постоянные C_1 и C_2 так, чтобы удовлетворить граничным условиям. Это требование даст нам систему двух линейных уравнений относительно искомых постоянных.

Сказанное выше позволяет свести решение задачи (1) к решению трех задач Коши. В действительности можно обойтись решением двух задач Коши. Выберем какие-нибудь числа y_0, y_0', z_0 и z_0', которые удовлетворяют равенствам $A_0y_0' + B_0y_0 = D_0$, $A_0z_0' + B_0z_0 = 0$ и найдем решения двух задач Коши:

$$L(y_0) = f, \quad y_0(a) = y_0, \quad y'(a) = y_0',$$

$$L(z_1) = 0, \quad z_1(a) = z_0, \quad z'(a) = z_0'.$$

Пусть z_2 какое-то решение уравнения $L(z_2) = 0$, линейно независимое с z_1, так что $A_0z_2(a) + B_0z_2(a) \neq 0$. Решение задачи (1), если оно существует,

должно иметь вид (2) при некоторых постоянных C_1 и C_2, для определения которых мы имеем систему уравнений

$$\left.\begin{array}{l}\big(A_0z_1'(a)+B_0z_1(a)\big)C_1+\big(A_0z_2'(a)+B_0z_2(a)\big)C_2+\big(A_0y_0'(a)+B_0y_0(a)\big)=D_0\\ \big(A_1z_1'(b)+B_1z_1(b)\big)C_1+\big(A_1z_2'(b)+B_1z_2(b)\big)C_2+\big(A_1y_0'(b)+B_1y_0(b)\big)=D_1\end{array}\right\}$$

В силу выбора начальных условий для y_0 и z_1 первое из этих уравнений есть $\big(A_0z_2'(a)+B_0z_2(a)\big)C_2=0$, т.е. $C_2=0$, и из второго уравнения мы сразу же находим C_1. Нам остается построить решение задачи (1) как линейную комбинацию y_0 и z_1: $y^*(x)=y_0(x)+C_1z_1(x)$.

Отметим одну трудность, которая может встретиться при применении этого метода. Если однородное уравнение $L(z)=0$ имеет быстро- и медленнорастущие решения, то может оказаться, что вблизи правого конца b промежутка решения y_0 и z_1 "почти линейно зависимы", и нам придется строить "небольшое" решение нашей задачи как линейную комбинацию "больших" "почти линейно зависимых" функций, что связано с пропаданием знаков. Приведем

Пример Рассмотрим граничную задачу

$$L(y)=y''-y'-6y=6,\quad y(-1)=y(1)=0.$$

Мы будем пользоваться тем, что общее решение соответствующего однородного уравнение известно: $C_1e^{3x}+C_2e^{-2x}$, так что нетрудно найти решение поставленной задачи в явном виде, но тем не менее будем считать, что это решение мы ищем численно рассмотренным методом (можно считать, что коэффициенты дифференциального уравнения переменные, но близки к написанным постоянным — качественная картина при этом будет такая же). Найдем y_0 как решение нашего дифференциального уравнения при условиях Коши: $y(-1)=0$, $y'(-1)=1$:

$$y_0(x)=\frac{3}{5}e^{3(1+x)}+\frac{2}{5}e^{-2(1+x)}-1,$$

а z_1 как решение задачи Коши $L(z_1)=0$, $z(-1)=0$, $z'(-1)=1$:

$$z_1(x)=\frac{1}{5}e^{3(1+x)}-\frac{1}{5}e^{-2(1+x)}.$$

Будем считать, что мы нашли эти функции численно с 5 верными значащими цифрами, так что $y_0(1)=241.05$, $z_1(1)=80.682$ и решение нашей задачи $y^*(x)=y_0(x)+C_1z_1(x)$, где $C_1=-y_0(1)/z_1(1)=-2.9816$. При $x=1/2$ получаем $y^*(1/2)=53.030+17.993C_1=-0.619$, и это значение получилось всего лишь с 3 верными знаками.

2. Метод дифференциальной прогонки.

Для простоты ограничимся рассмотрением задачи

$$L(y)=y''+p(x)y'+q(x)y=f(x), \tag{3}$$

$$y'(a)+\alpha_0y(a)=\beta_0,\quad y'(b)+\alpha_1y(b)=\beta_1. \tag{4}$$

Лемма. Пусть $\alpha(x)$ и $\beta(x)$ — решения дифференциальных уравнений

$$\alpha' + p\alpha - \alpha^2 = q, \qquad \beta' + (p-\alpha)\beta = f. \tag{5}$$

Тогда любое решение дифференциального уравнения

$$y' + \alpha y = \beta \tag{6}$$

удовлетворяет уравнению (3).

Доказательство. Пусть y — решение уравнения (6). Тогда $y'' = -\alpha' y - \alpha y' + \beta'$ и потому

$$y'' + py' + qy = (p-\alpha)y' + (q-\alpha')y + \beta' = (p-\alpha)(\beta - \alpha y) + (q-\alpha')y + \beta' = \\ = (\alpha^2 - p\alpha - \alpha' + q)y + (p\beta - \alpha\beta + \beta') = f,$$

и этим лемма доказана. ■

Изложим теперь алгоритм метода.

1) Решаем задачу Коши

$$\alpha' + p\alpha - \alpha^2 = q, \quad \alpha(a) = \alpha_0;$$

2) решаем задачу Коши

$$\beta' + p\beta - \alpha\beta = f, \quad \beta(a) = \beta_0;$$

3) находим числа y_1 и y_1' из системы уравнений

$$\left.\begin{array}{c} y_1' + \alpha(b)y_1 = \beta(b) \\ y_1' + \alpha_1 y_1 = \beta_1 \end{array}\right\};$$

4) находим функцию $y_*(x)$ как решение задачи Коши[18]

$$y' + \alpha(x)y = \beta(x), \quad y(b) = y_1.$$

Теорема. Построенная по указанному алгоритму функция $y_*(x)$ есть решение задачи (3)-(4).

Доказательство. Дифференциальному уравнению (3) функция y_* удовлетворяет ввиду леммы. Первому из условий (4) — поскольку оно эквивалентно $y'(a) + \alpha(a) = \beta(a)$. Наконец, второму из условий (4):

$$y'(b) = \beta(b) - \alpha(b)y(\beta) = \beta(b) - \alpha(b)y_1 = y_1' = \beta_1 - \alpha_1 y_1. \quad ■$$

Заметим, что дифференциальное уравнение (5) относительно α нелинейно, и его решение может не существовать на промежутке $[a,b]$. Тогда изложенный метод окажется неприменимым. Существуют варианты метода

[18]Эту задачу Коши нам придется решать “в обратном направлении”, двигаясь от правого конца промежутка b к левому.

прогонки, которые позволяют избежать этой неприятности[19]. Сам термин "метод прогонки" связан с тем, что пункты 1)-2) алгоритма можно трактовать так, что мы "прогоняем" граничное условие, заданное на левом конце промежутка на правый его конец (равенство $y'(b)+\alpha y(b)=\beta(b)$ для решения уравнения (3) есть следствие граничного условия на левом конце). Можно показать, что если определитель системы уравнений 4) равен нулю, то это означает отсутствие однозначной разрешимости задачи (3)-(4) — решение либо не существует, либо их бесконечно много (в соответствии с тем, имеет ли решение система уравнений этого пункта).

В методе сведения к задачам Коши для (3)-(4) нам пришлось бы решать две задачи Коши, а в методе дифференциальной прогонки — три. Но в первом случае это задачи Коши для уравнений второго порядка, а во втором — для уравнений первого порядка.

Задача. Показать, что при решении граничной задачи для системы линейных дифференциальных уравнений

$$y_k' + \sum_{j=1}^{n} a_{kj}(x) y_j = f_j(x), \quad k = 1, \ldots, n,$$

$$y_k(a) = y_k,\ k = 1, \ldots, m, \quad y_k(b) = y_k,\ k = m+1, \ldots, n$$

методом сведения к задачам Коши можно обойтись решением $l+1$ задачи Коши, где $l = \min\{\, m,\, n-m \,\}$.

[19] См., например, В.И.Крылов, В.В.Бобков, П.И.Монастырный "Вычислительные методы высшей математики", т.2, Минск, 1975.

Вопросы по курсу "Методы вычислений - 1".

1. Наилучшее равномерное приближение функций. Существование полинома наилучшего приближения.

2. Задача о полиноме, наименее уклоняющемся от нуля. Многочлены Чебышева.

3. Конечные разности: определение и основные свойства.

4. Разделенные разности: определение и основные свойства.

5. Понятие чебышевской системы функций. Признак такой системы. Разрешимость интерполяционной задачи.

6. Интерполяционные формулы Лагранжа и Ньютона.

7. Интерполяционные формулы с равноотстоящими узлами.

8. Представление остатка алгебраического интерполирования в форме Лагранжа.

9. Задача о минимуме остатка интерполирования на классе функций $KC^{(n+1)}$.

10. Понятие функции и постоянной Лебега интерполяционного процесса. Оценка погрешности интерполяции через наилучшее приближение. Признак сходимости интерполяционного процесса.

11. Оценка снизу постоянной Лебега для равноотстоящих узлов.

12. Оценка постоянной Лебега для узлов Чебышева.

13. Эрмитовская интерполяция: разрешимость задачи, представление интерполяционного полинома через разделенные разности.

14. Представление остатка эрмитовской интерполяции.

15. Численное дифференцирование: постановка задачи, теорема о представлении остатка.

16. Простейшие формулы численного дифференцирования.

17. Влияние ошибок в значениях функции на результат численного дифференцирования. Задача о наилучшем выборе шага на примере простейшей формулы.

18. Тригонометрическая интерполяция. Существование и единственность интерполяционного полинома.

19. Тригонометрическая интерполяция по равноотстоящим узлам.

20. Дискретное преобразование Фурье. Быстрое преобразование Фурье.

21. Формулы механических квадратур: понятие алгебраической степени точности (АСТ); верхняя оценка АСТ в случае положительного веса; оценка погрешности через наилучшее приближение функции.

22. Интерполяционные квадратурные формулы и их АСТ.

23. Квадратурные формулы в случае постоянного веса. Подобные квадратурные формулы и их свойства.

24. Квадратурные формулы Котеса. Формулы средних прямоугольников, трапеций и Симпсона и представление остатков этих формул.

25. Составные квадратурные формулы и их свойства. Теорема о сходимости.

26. Теорема о порядке сходимости составных квадратурных формул.

27. Составные формулы средних прямоугольников, трапеций и Симпсона.

28. Ортогональные полиномы: существование и единственность,.

29. Задача о построении квадратурной формулы наивысшей АСТ. Необходимый и достаточный признак такой формулы.

30. Основные свойства квадратурных формул гауссова типа: положительность коэффициентов, представление остатка, сходимость.

31. Многочлены Лежандра: формула Родрига, интеграл от квадрата.

32. Квадратурная формула Гаусса: симметрия узлов и коэффициентов, представление остатка.

33. Нормы векторов: определение, примеры, теорема об эквивалентности норм, сходимость последовательности векторов.

34. Нормы матриц. Операторные нормы и их свойства. Сходимость последовательности матриц.

35. Представление трех матричных норм. Оценки для $\|A\|_2$.

36. Необходимое и достаточное условие сходимости последовательности матриц D^s к нулевой матрице.

37. Теорема о сходимости суммы членов матричной геометрической прогрессии. Оценка $\|(E-A)^{-1}\|$ в случае $\|A\| < 1$.

38. Обратимость матриц с диагональным преобладанием. Теорема о кругах Гершгорина.

39. Обратимость матрицы, близкой к обратимой, и оценка близости обратных матриц.

40. Оценка абсолютной погрешности при замене системы линейных уравнений близкой.

41. Число обусловленности матрицы. Оценка относительной погрешности обратной матрицы и решения при замене системы линейных уравнений близкой.

42. Дифференцируемость простого собственного числа матрицы и соответствующего собственного вектора. Оценка дифференциалов.

43. Экстремальные свойства собственных чисел эрмитовой матрицы.

44. Минимально-максимальный принцип Куранта для собственных чисел эрмитовой матрицы.

45. Оценка погрешности собственных чисел при возмущении эрмитовой матрицы.

46. Метод итерации для решения систем линейных уравнений. Необходимый и достаточный признак сходимости. Оценки погрешности в случае $\|A\| < 1$.

47. Приведение системы уравнений к виду, удобному для применения метода итерации.

48. Метод Зейделя: алгоритм, сходимость в случае $\|A\|_\infty < 1$.

49. Метод Некрасова. Достаточные признаки сходимости.

50. Итеративный процесс уточнения обратной матрицы.

51. Степенной метод нахождения первого собственного числа матрицы (основные случаи). Метод скалярных произведений.

52. Метод А.Н.Крылова нахождения собственных чисел и векторов матрицы.

53. Метод Якоби нахождения собственных чисел и векторов симметричной матрицы.

54. Метод Люстерника ускорения сходимости метода итерации для решения линей- ных систем уравнений.

55. Метод Эйткена ускорения сходимости последовательностей и его применение к степенному методу.

55. Метод итерации для решения систем нелинейных уравнений. Теорема о сходимости.

57. Признак отображения сжатия в случае дифференцируемости.

58. Теорема о сходимости метода итерации для нелинейных уравнений при достаточно близком начальном приближении.

59. Понятие порядка сходимости итеративного процесса. Случай квадратичной сходимости метода итерации.

60. Метод Ньютона: алгоритм, теорема о квадратичной сходимости.

61. Теорема о сходимости модифицированного метода Ньютона.

62. Три простейших метода решения задачи Коши для обыкновенного дифференциального уравнения. Порядок ошибки на одном шаге.

63. Квадратурные формула Адамса. Две формы записи.

64. Экстраполяционный и интерполяционный методы Адамса: алгоритмы.

65. Построение начала таблицы при численном интегрировании обыкновенных дифференциальных уравнений.

66. Метод Рунге-Кутта.

67. Решение линейных граничных задач сведением к задачам Коши.

68. Метод дифференциальной прогонки.