Оглавление

Глава 1. Введение

1. Основы теории погрешностей.

На погрешность результата приближенного решения задачи влияют следующие причины:

а) *неточность информации о решаемой задаче.* Ошибки в начальных данных дают ту часть погрешности в решении, которая не зависит от математической стороны решения задачи и называется *неустранимой погрешностью.* Информация о границах *неустранимой погрешности* используется:

- для упрощения самой задачи, при выборе метода вычислений, точность которого должна быть согласована с требуемой точностью решения задачи;
- для определения точности вычислений.

Непомерные требования к точности результата часто снимаются в процессе рассмотрения задачи на основе следующих соображений:

- столь высокая точность не требуется;
- математическая модель явления столь груба, что требовать столь высокую точность бессмысленно;
- параметры модели не могут быть определены с высокой точностью;
- в итоге нас интересует не количественный, а качественный результат.

б) *Погрешность аппроксимации* (**методическая погрешность**). При решении задачи численными методами необходимо считаться с тем, что неизбежно придётся иметь дело только с конечным количеством чисел, и с ними можно выполнить только конечное число операций. Граница для каждого из этих количеств определяется свойствами инструментария, используемого при решении, временем, целесообразностью, стоимостью и др.

Если количество чисел или операций превышает допустимые границы, то задачу приходится упрощать и заменять ее другой задачей, но уже удовлетворяющей нужным требованиям.

в) *Погрешность округления.* Всякое положительное число a может быть представлено в виде конечной или бесконечной десятичной дроби

$$a = \alpha_m 10^m + \alpha_{m-1}10^{m-1} + \alpha_{m-2}10^{m-2} + \ldots + \alpha_{m-n+1}10^{m-n+1} + \ldots \qquad (1)$$

где α_i – цифры числа a ($\alpha_i = 0, 1, 2, \ldots, 9$), причем старшая цифра $\alpha_m \neq 0$, а m – некоторое число (старший десятичный разряд числа a).
Например,

$$3141,59\ldots = 3 \cdot 10^3 + 1 \cdot 10^2 + 4 \cdot 10^1 + 1 \cdot 10^0 + 5 \cdot 10^{-1} + 9 \cdot 10^{-2} + \ldots .$$

На практике имеют дело с приближенными числами, представляющими собой конечные десятичные дроби

$$b^* = \beta_m 10^m + \beta_{m-1}10^{m-1} + \beta_{m-2}10^{m-2} + \ldots + \beta_{m-n+1}10^{m-n+1}, \qquad \beta_m \neq 0.$$

Определение 1. Цифра β_k в изображении числа b^* называется *верной,* если имеет место неравенство $|b - b^*| \leq \omega 10^k, \omega \leq 1$, чаще всего, $\omega = 0.5$. (Здесь b – точное значение величины, представленной приближённой записью через b^*.) ♣

Очевидно, что если цифра β_k верная, то и все цифры в записи числа b, расположенные левее неё, тоже верны.

Определение 2. *Значащей цифрой* числа называется всякая верная его цифра в десятичном изображении, кроме нулей, стоящих слева в записи числа до первой ненулевой цифры. ♣
Например, в числе 0.002080 первые три нуля не являются значащими цифрами, так как они служат только для установления десятичных разрядов других цифр. Остальные два нуля являются значащими. В случае, если в данном числе 0.002080 последняя цифра не является верной, то её не следует использовать в записи числа. Пусть a есть точное значение некоторой величины. a^* – приближенное ее значение.

Определение 3. *Разность $a - a^* = \varepsilon$ называется погрешностью приближённого значения a^*.* ♣

Точное значение a и ε, как правило, неизвестны, но очень часто известна верхняя граница $\triangle$ абсолютной величины погрешности:

$$|a - a^*| = |\varepsilon| \leq \triangle.$$

Её и будем называть границей погрешности ε. Точное значение a лежит в пределах

$$a^* - \Delta \leq a \leq a^* + \Delta$$

Определение 4. *Относительной погрешностью* величины a^* называют отношение

$$\frac{a^* - a}{a^*}. \quad ♣$$

Определение 5. *Границей относительной погрешности* a^* называют отношение

$$\left|\frac{a^* - a}{a^*}\right| \leq \frac{\triangle}{|a^*|} = \delta. \quad ♣$$

Погрешность суммы

Пусть $a = x_1 + x_2$, известны приближенные значения x_1^*, x_2^* слагаемых и границы Δ_1, Δ_2. Обозначим погрешности слагаемых соответственно ε_1,ε_2:
$a^* = x_1^* + x_2^*$, $\quad a = (x_1^* + \varepsilon_1) + (x_2^* + \varepsilon_2) = a^* + \varepsilon_1 + \varepsilon_2 = a^* + \varepsilon$. Поэтому $|\varepsilon| \leq |\varepsilon_1| + |\varepsilon_2| \leq \triangle_1 + \triangle_2$
Граница погрешности суммы не больше суммы границ погрешностей слагаемых. ♣

Погрешность произведения

Граница погрешности произведения по границам погрешности сомножителей определяется неравенством

$$|\varepsilon| \leq |\triangle| \leq |x_1^*| \triangle_2 + |x_2^*| \triangle_1 + \triangle_1 \triangle_2.$$

Более простым является правило оценки относительной погрешности произведения.

$$\frac{\varepsilon}{x_1^* x_2^*} = \frac{\varepsilon_1}{x_1^*} + \frac{\varepsilon_2}{x_2^*} + \frac{\varepsilon_2}{x_2^*}\frac{\varepsilon_1}{x_1^*}$$

Отсюда вытекает оценка границы относительной погрешности

$$\left|\frac{\varepsilon}{x_1^* x_2^*}\right| \le \delta \le \delta_1 + \delta_2 + \delta_1\delta_2. \quad \clubsuit$$

Погрешность частного

Пусть $x = \frac{x_1}{x_2}$. Сохраняя прежние обозначения для приближенных значений x_1^*, x_2^*, погрешностей и их границ, сделаем предположение: $|\Delta_2| < |x_2^*|$ (так как делитель x_2 может оказаться равным нулю).

Погрешность отношения может быть записана так:

$$\varepsilon = \frac{x_1}{x_2} - \frac{x_1^*}{x_2^*} = \frac{1}{x_2^*(x_2^* + \varepsilon_2)}(\varepsilon_1 x_2^* - \varepsilon_2 x_1^*),$$

что приводит к следующему правилу оценки границы погрешности:

$$|\varepsilon| \le \triangle \le \frac{1}{|x_2^*|(|x_2^*| - \triangle_2)}(\triangle_1|x_2^*| + \triangle_2|x_1^*|),$$

т.е. ***относительная погрешность дроби оценивается при помощи неравенства***

$$\left|\frac{\varepsilon}{x^*}\right| \le \delta \le \frac{1}{1-\delta_2}(\delta_1 + \delta_2). \quad \clubsuit$$

1.1. Погрешность вычисления функции. Оценка неустранимой погрешности

Пусть в выпуклой области $G \in R^n$ рассматривается непрерывно дифференцируемая функция $y = f(\cdot)$. Предположим, что в точке $x = (x_1, x_2, \ldots, x_n)$ области G нужно вычислить значение $y = f(x)$. Пусть нам известны лишь приближенные значения $x_1^*, x_2^*, \ldots, x_n^*$ такие, что точка $x^* = (x_1^*, x_2^*, \ldots, x_n^*) \in G$. Необходимо найти оценку погрешности приближённого значения функции $y^* = f(x^*)$. Через погрешности $\varepsilon_i = x_i - x_i^*$ аргументов она выражается следующим образом:

$$\varepsilon = f(x_1^* + \varepsilon_1, x_2^* + \varepsilon_2, \ldots, x_n^* + \varepsilon_n) - f(x_1^*, x_2^*, \ldots, x_n^*),$$

или, если воспользоваться формулой Лагранжа,

$$\varepsilon = \sum_{i=1}^{n} \frac{\partial}{\partial x_i} f(x_1^* + \theta\varepsilon_1, x_2^* + \theta\varepsilon_2, \ldots, x_n^* + \theta\varepsilon_n)\varepsilon_i, \qquad 0 \le \theta \le 1.$$

Отсюда получается оценка для границы погрешности вычисления функции

$$|\varepsilon| \le \Delta \le \sum_{i=1}^{n} B_i \Delta_i, \qquad (*)$$

где

$$|\varepsilon_i| \le \Delta_i, \qquad B_i = \max_{\theta\in[0,1]} \left|\frac{\partial}{\partial x_i} f(x_1^* + \theta\varepsilon_1, x_2^* + \theta\varepsilon_2, \ldots, x_n^* + \theta\varepsilon_n)\right|.$$

Таким образом решается **основная задача теории погрешности:** *известны погрешности некоторой системы величин, требуется определить погрешность данной функции от этих величин.* $\clubsuit$

1.2. Обратная задача теории погрешностей

каковы должны быть абсолютные погрешности аргументов функции, чтобы абсолютная погрешность функции не превышала заданной величины?

Эта задача математически неопределена, так как заданную предельную погрешность (верхнюю границу абсолютной погрешности) можно обеспечить, устанавливая по-разному предельные абсолютные погрешности ее аргументов. Простейшее решение обратной задачи дается так называемым принципом *равных влияний*. Предполагается, что все слагаемые $B_i\Delta_i \qquad i = 1, 2, \ldots, n$ в правой части формулы $(*)$ имеют одинаковую величину. Тогда

$$\Delta_i = \frac{|\varepsilon|}{nB_i}.$$

Другой столь же простой способ носит название принципа *равных погрешностей:* считается, что $\Delta_i = \Delta_j$, и тогда из той же формулы $(*)$ немедленно получаем:

$$\Delta_i = \frac{|\varepsilon|}{\sum_{i=1}^{n} B_i} \ .$$

Исходя из особенностей задачи и функции можно выставлять и другие требования к уровню погрешностей аргументов. ♣

Глава 2. Численное решение уравнений и систем уравнений

1. Решение скалярных уравнений

1.1. Метод Чебышева

Рассмотрим уравнение $f(x) = 0, x \in [a, b]$, и пусть на указанном интервале функция $f(\cdot)$ имеет обратную: $F(\cdot) = f^{-1}(\cdot)$. Очевидно, что тогда решение $\bar{x}$ уравнения $f(x) = 0$ находится тривиально: $\bar{x} = F(0)$. Следовательно, достаточно указать способ построения обратной функции или её приближения. В методе Чебышева функция $\tilde{F}(x) \approx F(x)$ строится в виде отрезка ряда Тейлора.

Итак, будем считать, что $f(\cdot) \in C^{m+1}[a, b],\ f'(x) \neq 0$ на всём интервале $[a, b]$. Следовательно, для $f(\cdot)$ существует обратная функция $F(\cdot) = f^{-1}(\cdot)$ ввиду монотонности $f(\cdot)$ на $[a, b]:\ F\big(f(x)\big) \equiv x$. Для определённости будем считать, что $f'(x) > 0$ на $[a, b]$, что обеспечивает монотонное возрастание функции $f(\cdot)$. Обозначим $c = f(a),\ d = f(b),\ c < d$, и пусть $\hat{x} \in [a, b]$ – некоторое приближение искомого корня $\bar{x}$ и $\hat{y} = f(\hat{x})$. Из анализа известно, что при сделанных предположениях обратная функция обладает той же гладкостью, что и сама функция: $F(\cdot) \in C^{m+1}[c, d]$. Следовательно, можем записать:

$$F(y) = \sum_{k=0}^{m} \frac{F^{(k)}(\hat{y})}{k!}(y - \hat{y})^k + R_m(\hat{y}, y),\ \hat{y} \in (c, d), \tag{1}$$

$$\text{где} \quad R_m(\hat{y}, y) = \frac{F^{(m+1)}(z)}{(m+1)!}(y - \hat{y})^{m+1}, \quad z \in (c, d). \tag{2}$$

Положим, далее

$$\tilde{F}(y) = \sum_{k=0}^{m} \frac{F^{(k)}(\hat{y})}{k!}(y - \hat{y})^k \tag{3}$$

и в соответствии с высказанными выше соображениями, положим

$$\bar{x} \approx \tilde{x} = \tilde{F}(0) = \sum_{k=1}^{m} \frac{F^{(k)}(\hat{y})}{k!}(0 - \hat{y})^k, \tag{4}$$

где $\hat{x}$ – начальное приближение к искомому корню. Получим такое представление для $\tilde{x}$:

$$\tilde{x} = \sum_{k=0}^{m} \frac{F^{(k)}(\hat{y})}{k!}\big(-f(\hat{x})\big)^k \equiv H_m(\hat{x}). \tag{5}$$

Теперь можно построить итеративный процесс, полагая

$$x_{k+1} = H_m(x_k). \tag{6}$$

Замечание. Уравнение $x = H_m(x)$ имеет корень $\bar{x}:\quad f(\bar{x}) = 0.$ ♣

Остаётся указать способ вычисления $F^{(k)}(\hat{y})$:

$$F\big(f(x)\big) \equiv x, \Longrightarrow F'_y \cdot f'_x \equiv 1, \Longrightarrow F'_y(\hat{y}) = \frac{1}{f'(\hat{x})}; \tag{7}$$

$$F''_{y^2}(f'_x)^2 + F'_y f''_{x^2} = 0, \Longrightarrow F''_{y^2}(\hat{y}) = -\frac{F'_y(\hat{y}) f''_{x^2}(\hat{x})}{[f'_x(\hat{x})]^2} = -\frac{f''_{x^2}(\hat{x})}{[f'_x(\hat{x})]^3} \quad \text{и т.д.} \tag{8}$$

Перейдём к оценке величины $\Delta x = |\bar{x} - x_{k+1}|$. Имеем:

$$\bar{x} - x_{k+1} = \bar{x} - \tilde{x} = F(0) - \tilde{F}(0) = \frac{F^{(m+1)}(z_k)}{(m+1)!}[-f(x_k)]^{m+1}. \tag{9}$$

Пусть известны оценки: $|f'(x)| \le q_1$, $|F^{(m+1)}| \le Q_{m+1}$, на $[a,b]$ и $[c,d]$ соответственно. Тогда

$$f(x_k) = f(x_k) - f(\bar{x}) = f'(\xi)(x_k - \bar{x}),$$
$$|f(x_k)| \le q_1 |\bar{x} - x_k|$$

и подставляя последнее выражение в формулу (9), получим:

$$|\bar{x} - x_{k+1}| \le \frac{Q_{m+1}}{(m+1)!} q_1^{m+1} |\bar{x} - x_k|^{m+1} \equiv p|\bar{x} - x_k|^{m+1}. \tag{10}$$

Применяя последовательно k раз полученную оценку, придём к следующему результату:

$$\begin{aligned}
&|\bar{x} - x_{k+1}| \le p\,|\bar{x} - x_k|^{m+1} \le p\,[\,p\,|\bar{x} - x_{k-1}|^{m+1}]^{m+1} = \\
&= p^{1+(m+1)} |\bar{x} - x_{k-1}|^{(m+1)^2} \le p^{1+(m+1)+(m+1)^2} |\bar{x} - x_{k-2}|^{(m+1)^3} \le \\
&\le p^t |\bar{x} - x_1|^{(m+1)^k},
\end{aligned}$$

где $t = 1 + (m+1) + \cdots + (m+1)^{k-1} = \frac{(m+1)^k - 1}{m}$. Или:

$$|\bar{x} - x_{k+1}| \le \frac{\big(\bar{p}\,|\bar{x} - x_1|\big)^{(m+1)^k}}{\bar{p}}, \qquad \text{где } \bar{p} = \sqrt[m]{p}. \tag{11}$$

Таким образом, в предположении

$$\bar{p}\,|\bar{x} - x_1| < 1 \tag{12}$$

имеет место сходимость: $x_k \to \bar{x}$ при $k \to \infty$. Необходимо лишь, чтобы все итерации $\{x_k\}_{k=1}^{\infty}$ оставались в $[a,b]$. Таким образом, доказана

Теорема 1. *Пусть* $f(\cdot) \in C^{m+1}[a,b]$, $|f'(x)| > 0$ *на* $[a,b]$ *и* $\exists\, \bar{x} : f(\bar{x}) = 0$. *Если* x_1 *удовлетворяет условию (12) и последовательность* $x_{k+1} = H_m(x_k) \in [a,b]$, *то* $\{x_k\}$ *сходится к* $\bar{x}$, *причём скорость сходимости характеризуется оценкой (11).* ♣

Замечание. Рассмотренный метод при $m = 1$ носит название метода Ньютона:

$$x_{k+1} = H_1(x_k) = F\big(f(x_k)\big) + \frac{F'\big(f(x_k)\big)}{1!}\big(-f(x_k)\big) = x_k - \frac{f(x_k)}{f'(x_k),}$$

причём оценка (10) даёт: $|\bar{x} - x_{k+1}| \le p\,|\bar{x} - x_k|^2$, т.е. порядок сходимости метода Ньютона равен двум (так, если $|x - x_1| \approx 10^{-1}$, то $|x - x_2| \approx 10^{-2}$). ♣

Выше использована терминология, требующая пояснения.

- Пусть для некоторого метода верна оценка: $|x_k - \bar{x}| \leq \omega q^k,\ q < 1,\ \omega = const$. Тогда говорят, что этот метод сходится со скоростью геометрической прогрессии со знаменателем q.

- Пусть существует окрестность корня $\bar{x}$ такая, что все приближения принадлежат ей и имеет место оценка: $|x_{k+1} - \bar{x}| \leq \omega |x_k - \bar{x}|^p$. Тогда число p называют порядком сходимости метода:

  – при $p = 1$ говорят о линейной сходимости;

  – при $p > 1$ говорят о сверхлинейной сходимости, в частности при $p = 2$ о квадратичной.

1.2. Метод итераций

Применение метода итераций (а таковым является, в частности, метод Ньютона) требует приведения уравнения к специальному (каноническому) виду

$$x = \varphi(x), \quad [a, b] \xrightarrow{\varphi} [a, b]. \tag{1}$$

Точки, удовлетворяющие (1), называются неподвижными точками преобразования $\varphi(\cdot)$. Геометрически неподвижная точка есть не что иное, как точка пересечения прямой $y = x$ и кривой $y = \varphi(x)$.

В методе итераций построение членов последовательности $\{x_k\}$ ведется по формуле

$$x_{k+1} = \varphi(x_k),\ k = 0, 1 \ldots \tag{2}$$

Рассмотрим поведение последовательности $\{x_k\}$, когда её члены находятся вблизи точки $\bar{x}$. Удобно ввести величину $\varepsilon_k = x_k - \bar{x}$ и учесть малость ε_k. Связь между ε_k и ε_{k+1} получим из (2), подставляя $x_k = \bar{x} + \varepsilon_k$ и $x_{k+1} = \bar{x} + \varepsilon_{k+1}$:

$$\bar{x} + \varepsilon_{k+1} = \varphi(\bar{x}) + \varepsilon_k \varphi'(\bar{x}) + o(\varepsilon_k).$$

Учитывая равенство $\bar{x} = \varphi(\bar{x})$, получаем:

$$\varepsilon_{k+1} \approx \varphi'(\bar{x}) \varepsilon_k,\ k = 0, 1 \ldots \tag{3}$$

Возможны случаи:

1. $|\varphi'(\bar{x})| > 1 \Longrightarrow |\varepsilon_{k+1}| > |\varepsilon_k| \Longrightarrow \bar{x}$ – точка "отталкивания";

2. $|\varphi'(\bar{x})| < 1$ – можно ожидать, что если x_0 близка к $\bar{x}$, то последовательность $\{x_k\}$ будет сходиться к $\bar{x}$ как геометрическая прогрессия со знаменателем $q = \varphi'(\bar{x})$.

   Если $\varphi'(\bar{x}) < 0$, то $\{x_k\}$ сходится к $\bar{x}$ с разных сторон, что облегчает оценку для $\bar{x}$: $\bar{x} \in (\min\{x_k,\ x_{k+1}\},\ \max\{x_k,\ x_{k+1}\})$.

3. $\varphi'(\bar{x}) = 0$. Пусть для определённости $\varphi(\cdot) \in C^{(m)}(a, b)$ и

$$\varphi'(\bar{x}) = 0,\ \varphi''(\bar{x}) = 0, \ldots,\ \varphi^{(m-1)}(\bar{x}) = 0,\ \varphi^{(m)}(\bar{x}) \neq 0.$$

В этом случае разложение для $\varphi(x_k) = \varphi(\bar{x} + \varepsilon_k)$ вблизи точки $\bar{x}$ имеет вид:

$$\varphi(x_k) = \varphi(\bar{x} + \varepsilon_k) = \varphi(\bar{x}) + \frac{1}{m!}\varphi^{(m)}(\xi)\varepsilon_k^m, \quad \xi \in [a, b]$$

и подстановка в (2) даёт:

$$\varepsilon_{k+1} = \frac{1}{m!}\varphi^{(m)}(\xi)\varepsilon_k^m \tag{4}$$

Если оценить $|\varphi^{(m)}(\xi)| \le M$, то $|\varepsilon_{k+1}| \le \dfrac{M}{m!}|\varepsilon_k|^m$, откуда получаем:

$$|\varepsilon_{k+1}| \le \left(\frac{M}{m!}|\varepsilon_0|\right)^{\frac{m^{k+1}-1}{m-1}} |\varepsilon_0|^{\frac{m^{k+2}-2m^{k+1}+1}{m-1}},$$

что указывает на быструю сходимость $\{x_k\}$ к $\bar{x}$ при $|\varepsilon_0| < 1$ и $\dfrac{M}{m!}|\varepsilon_0| < 1$.

Теперь сформулируем и докажем теорему:

Теорема 1. (достаточное условие сходимости метода итераций)
Пусть выполнены условия:

1. *функция $\varphi(\cdot)$ определена на $\Omega = \{x : |x - x_0| \le \delta\}$, непрерывна там и удовлетворяет условию Липшица: $|\varphi(x) - \varphi(x')| \le q|x - x'|, \quad 0 \le q < 1$;*
2. *для начального приближения выполнено условие: $|x_0 - \varphi(x_0)| \le m$;*
3. *числа δ, q, m удовлетворяют соотношению: $\dfrac{m}{1-q} \le \delta$.*

Тогда:

1. *Уравнение (1) имеет в Ω единственное решение $\bar{x}$;*
2. *Имеет место сходимость построенной последовательности:*

$$\{x_k\} \subset \Omega, \ \bar{x} = \lim_{k\to\infty} x_k;$$

3. *последовательность $\{x_k\}$ сходится к $\bar{x}$ со скоростью геометрической прогрессии со знаменателем q т.е. $|x_k - \bar{x}| \le \dfrac{m}{1-q}q^k, \ k = 1, 2, \dots$.* ♣

Доказательство. Покажем, прежде всего, что из условий теоремы следует

$$\{x_k\} \subset \Omega \text{ и } |x_{k+1} - x_k| \le mq^k, \quad k = 0, 1, \dots. \tag{5}$$

Доказательство проведём по индукции. При $k = 0$ можно построить $x_1 = \varphi(x_0)$, т.к. $x_0 \in \Omega$ и, кроме того, $|x_1 - x_0| = |\varphi(x_0) - x_0| \le m$ согласно условию 2 теоремы 1. Поэтому утверждение (5) верно для $k = 0$, а ввиду условия 3 той же теоремы $m \le \frac{m}{1-q} < \delta$, т.к. $q < 1$ (см. условие 1). Следовательно, $x_1 \in \Omega$. База для индукции создана.

Пусть теперь $\{x_0, \ x_1, \dots, x_n\} \subset \Omega$ и для членов вышеуказанной последовательности выполняется неравенство

$$|x_{k+1} - x_k| \le mq^k, \ k = 0, 1, \dots, n-1.$$

Поскольку $x_n \in \Omega$ согласно индуктивному предположению, то следующее приближение $x_{n+1} = \varphi(x_n)$ может быть построено. Далее, $|x_{n+1} - x_n| = |\varphi(x_n) - \varphi(x_{n-1})| \le q|x_n - x_{n-1}| \le q \cdot mq^{n-1} = mq^n$. Следовательно, неравенство (5) верно. Осталось проверить, что $x_{n+1} \in \Omega$: действительно,

$$|x_{n+1} - x_0| = |(x_{n+1} - x_n) + (x_n - x_{n-1}) + \cdots + (x_1 - x_0)| \le$$
$$\le m(q^n + q^{n-1} + \cdots + q^0) \le \frac{m}{1-q} \le \delta,$$

т.е. $x_{n+1} \in \Omega$.

Докажем теперь сходимость построенной последовательности. Проверим выполнение условия Больцано-Коши для неё:

$$|x_{n+p} - x_n| = |(x_{n+p} - x_{n+p-1}) + (x_{n+p-1} - x_{n+p-2}) + \cdots + (x_{n+1} - x_n)| \le$$
$$\le m(q^{n+p-1} + q^{n+p-2} + \cdots + q^n) \le \frac{m}{1-q} q^n \ < \varepsilon, \quad \text{при } n > N, \ \forall p > 0. \qquad (*)$$

Следовательно, построенная последовательность является фундаментальной (сходящейся в себе), а поскольку множество Ω замкнуто и $\{x_k\} \subset \Omega$, то

$$\exists\, x_* \in \Omega : \quad x_* = \lim x_k,$$

и переходя к пределу в (2), получим:

$$x_* = \varphi(x_*).$$

Если же допустить существование двух точек x_* и x_{**}, являющихся решением уравнения (1), то получим:

$$|x_* - x_{**}| = |\varphi(x_*) - \varphi(x_{**})| \le q|x_* - x_{**}|,$$

откуда ввиду $q < 1$ следует $|x_* - x_{**}| = 0$, т.е. $x_* = x_{**}$. Третье утверждение теоремы доказывается переходом к пределу в оценке $(*)$ при $p \to \infty$. Теорема доказана полностью. ♣

1.3. Ускорение сходимости. Преобразование Эйткена.

Рассмотрим показательную функцию целочисленного аргумента n :

$$s_n = s + Aq^n.$$

При $q < 1$ указанная последовательность сходится к s. В этом случае её предел можно выразить через три её последовательных члена:

$$q = \frac{s_n - s}{s_{n-1} - s} = \frac{s_{n+1} - s}{s_n - s},$$

откуда находим $(s_{n+1} - s)(s_{n-1} - s) = (s_n - s)^2$ и затем

$$s = \frac{s_{n+1}s_{n-1} - s_n^2}{s_{n+1} - 2s_n + s_{n-1}}.$$

Следуя Эйткену, рассмотрим преобразование произвольной последовательности $\{s_n\}$ в другую $\{\sigma_n\}$:

$$\sigma_n = \frac{s_{n+1}s_{n-1} - s_n^2}{s_{n+1} - 2s_n + s_{n-1}}.$$

Если это преобразование применить к последовательности $s_n = s + Aq^n$, то, очевидно, получим:

$$\sigma_n \equiv s = \lim s_k.$$

Естественно ожидать, что это преобразование приведёт к более быстрой сходимости последовательности $\{\sigma_n\}$ к *тому же пределу* s, если исходная последовательность будет меняться по закону, близкому к показательному.

Итак, пусть построены члены последовательности $x_1,\ x_2, \ldots,\ x_n$. Вычислим $x'_n = \varphi(x_n)$ и $x''_n = \varphi(x'_n)$. К трём значениям $x_n,\ x'_n,\ x''_n$ применяем преобразование Эйткена и его результат принимаем за новое приближение x_{n+1} :

$$x_{n+1} = \frac{x_n x''_n - [x'_n]^2}{x''_n - 2x'_n + x_n}.$$

Полученное равенство называют итерационной формулой Стеффенсена. Её можно истолковать как простой итерационный процесс для вспомогательного уравнения:

$$x = \Phi(x), \quad \Phi(x) = \frac{x\varphi(\varphi(x)) - \varphi^2(x)}{\varphi(\varphi(x)) - 2\varphi(x) + x} \qquad (*)$$

Для величин $\varepsilon_n = |x_n - \bar{x}|$ верна оценка:

$$\varepsilon_{n+1} \le B\varepsilon_n^2, \text{ где } B = \sup_{[a,b]} \left| \frac{\varphi'(x)\varphi''(x)}{2(\varphi'(x) - 1)} \right|,$$

что говорит о квадратичной сходимости преобразованной последовательности. Отметим, что в окрестности корня $\bar{x}$: $|\Phi'(x)| \le q < 1$, поскольку $\Phi'(\bar{x}) = 0$.

1.4. Границы расположения корней алгебраического уравнения

Пусть задано уравнение в комплексной области:

$$f(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_n = 0. \qquad (1)$$

Обозначим

$$a = \max_{i>0}\{|a_i|\}, \quad A = \max_{i<n}\{|a_i|\}.$$

Теорема 1. *Все корни уравнения (1) расположены в кольце:*

$$\frac{|a_n|}{A + |a_n|} < |z| < 1 + \frac{a}{|a_0|.} \qquad \clubsuit$$

Доказательство. Используя известное неравенство $|u + v| \ge |u| - |v|$, получаем:

$$\big|f(z)\big| \ge |a_0 z^n| - |a_1 z^{n-1} + \cdots + a_n|, \qquad (2)$$

Но при $|z| > 1$ имеет место оценка:

$$|a_1 z^{n-1} + \cdots + a_n| \le a\{|z|^{n-1} + |z|^{n-2} + \cdots + |z| + 1\} = a\frac{|z|^n - 1}{|z| - 1} < a\frac{|z|^n}{|z| - 1}.$$

Подставляя последнюю оценку в правую часть (2), получим:

$$\big|f(z)\big| > |a_0 z^n| - a\frac{|z|^n}{|z|-1}, \tag{3}$$

т.е. $|f(z)| > 0$ при $|a_0 z^n| - a\dfrac{|z|^n}{|z|-1} \geq 0$, что выполняется при $|a_0||z| - |a_0| - a \geq 0$, или

$$|z| \geq 1 + \frac{a}{|a_0|}. \tag{4}$$

Таким образом, в области комплексной плоскости, заданной неравенством (4) нет корней уравнения (1).

Далее, рассмотрим уравнение

$$\varphi(y) = a_0 + a_1 y + \cdots + a_n y^n = 0.$$

Оно, очевидно, имеет своими корнями числа, обратные к корням исходного уравнения. Поэтому на основании (4) введенное уравнение не имеет корней при

$$\frac{1}{|z|} \geq 1 + \frac{A}{|a_n|} \quad \text{или} \quad |z| \leq \frac{|a_n|}{A + |a_n|}\ , \tag{5}$$

что вместе с (4) и доказывает утверждение теоремы. ♣

Пусть теперь все коэффициенты уравнения (1) – действительные числа (и это будем подчеркивать заменой z на x) и $a_0 > 0$ (что никак не умаляет дальнейшего). Найдём границы *действительных* корней. Очевидно, однако, что достаточно указать верхнюю границу для положительных корней, т.к. заменой $x = -t$ получим уравнение, корни которого имеют знаки, противоположные знакам корней исходного уравнения.

Теорема 2. *Пусть* $a = \max_{a_i<0}\{|a_i|\}$ *и пусть* a_m *– первый отрицательный коэффициент в ряду* $a_0,\ a_1, \ldots,\ a_n$. *Тогда все положительные корни уравнения меньше числа* $p = 1 + (a/a_0)^{1/m}$. ♣

(Заметим, что если в уравнении нет отрицательных коэффициентов, то оно не имеет положительных корней!)

Доказательство. При $x > 1$ имеем:

$$f(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_n > a_0 x^n - a\big(x^{n-m} + x^{n-m-1} + \cdots + x + 1\big) =$$

$$= a_0 x^n - a\frac{x^{n-m+1} - 1}{x - 1} > a_0 x^n - a\frac{x^{n-m+1}}{x-1} =$$

$$= \frac{x^{n-m+1}}{x-1}\{a_0 x^{m-1}(x-1) - a\} \equiv u(x)v(x).$$

Здесь неотрицательные по условию теоремы коэффициенты $a_1,\ a_2, \ldots,\ a_{m-1}$ были заменены на нули, а остальные на $-a$ (см. условие). Далее, при $x > 1$ имеем:

$$u > 0, \quad v = a_0 x^{m-1}(x-1) - a > a_0(x-1)^m - a.$$

Правая часть полученной оценки для $f(x)$ обращается в нуль при $x = p$, т.к. $v(p) = 0$, а при $x > p$ положительна, т.е. при $x \geq p$ имеем $f(x) > 0$, чем и завершается доказательство. ♣

Замечание. Для нахождения двусторонних оценок действительных корней алгебраического уравнения достаточно уметь находить верхнюю границу положительных корней.

Действительно: обозначим

$$\varphi_0(x) = f(x),$$
$$\varphi_1(x) = x^n f\left(\frac{1}{x}\right),$$
$$\varphi_2(x) = f(-x),$$
$$\varphi_3(x) = x^n f\left(-\frac{1}{x}\right).$$

и пусть p_i есть верхняя оценка положительных корней для полинома $\varphi_i(x)$. Тогда число p_1^{-1} является нижней оценкой для положительных корней $f(x)$ т.к. если $f(\bar{x}) = 0$, то $\varphi_1(1/\bar{x}) = 0$, а поэтому $1/\bar{x} < p_1$, то есть $\bar{x} > p_1^{-1}$. Аналогично проводятся рассуждения для отрицательных корней.

Резюмируя, можно записать оценки:

$$p_1^{-1} < \bar{x} < p_0, \qquad \text{если } \bar{x} > 0;$$
$$-p_2 < \bar{x} < -p_3^{-1}, \qquad \text{если } \bar{x} < 0. \quad \clubsuit$$

1.5. О локализации корней в общем случае

Если в уравнении $f(x) = 0$ функция $f(\cdot)$ непрерывна, то основой для локализации корня обычно служит следствие из теоремы Коши: если $f(a)f(b) < 0$, то на интервале $[a, b]$ имеется по крайней мере один корень указанного уравнения (точнее нечётное число корней). Для локализации корня на интервале $[a, b]$ можно применять такие подходы:

- *Последовательный перебор.* Интервал $[a, b]$ разбивается на N равных отрезков и вычисляются значения функции $f(\cdot)$ в точках $x_k = a + kh,\ k = 0, 1, \ldots, N$, где $h = (b-a)/N$. Если при этом найдётся интервал $[x_k, x_{k+1}]$, для которого $f(x_k)f(x_{k+1}) < 0$, то тем самым корень функции будет локализован с точностью $h/2$. Может оказаться, что функция $f(\cdot)$ не меняет знака на последовательности $\{x_k\}$. Если корень на $[a, b]$ существует, то последнее означает, что шаг h слишком велик и его следует заменить на меньший, полагая, например, $N = 2N$.

- *Перебор с переменным шагом.* Если функция $f(x)$ является Липшицевой, т.е.

  $$|f(x') - f(x'')| \le L|x' - x''|, \quad x', x'' \in [a, b],$$

  то можно строить последовательность $\{x_k\}$ вида:

  $$x_0 = a,\ x_{k+1} = x_k + \frac{|f(x_k)|}{L}\ .$$

  Основанием к этому может служить то, что при $f(x) = cx+d$, можно принять $L = |c|$ и в этом случае значение x_1, полученное указанным способом, удовлетворяет уравнению $f(x) = 0$.

  Если L неизвестна, то можно её заменить через

  $$L_k = \frac{|f(x_k) - f(x_{k-1})|}{|x_k - x_{k-1}|}\ .$$

- *Использование мажорант.* Если известны оценки функции $f(\cdot)$ на $[a,b]$, т.е.

$$f^-(x) \le f(x) \le f^+(x),$$

и корни x^- и x^+ этих функций, то $\bar{x} \in [\min\{x^-, x^+\},\ \max\{x^-, x^+\}]$.

Пример. Пусть $f(x) = \sin x + x^3 - 2,\ x \in [0, \pi]$. На указанном интервале можно принять: $f^-(x) = x^3 - 2,\ \ f^+(x) = 1 + x^3 - 2 = x^3 - 1$. Следовательно, $\bar{x} \in \left[1, \sqrt[3]{2}\right]$. ♣

2. Решение систем линейных алгебраических уравнений

2.1. Нормы векторов и матриц

Напомним, что линейное пространство Ω элементов x называется нормированным, если в нём введена функция $\|\cdot\|_\Omega$, определённая для всех элементов пространства Ω и удовлетворяющая условиям:

1. $\|x\|_\Omega \ge 0$, причём $\|x\|_\Omega = 0 \Longleftrightarrow x = 0_\Omega$;
2. $\|\lambda x\|_\Omega = |\lambda| \cdot \|x\|_\Omega$;
3. $\|x + y\|_\Omega \le \|x\|_\Omega + \|y\|_\Omega$.

Договоримся в дальнейшем обозначать малыми латинскими буквами векторы, причём будем считать их вектор-столбцами, большими латинскими буквами обозначим матрицы, а греческими буквами станем обозначать скалярные величины (сохраняя за буквами i, j, k, l, m, n обозначения для целых чисел).

К числу наиболее употребительных норм векторов относятся следующие:

1. $\|x\|_1 = \sum\limits_{i=1}^{n} |x_i|$;
2. $\|x\|_2 = \sqrt{\sum\limits_{i=1}^{n} x_i^2}$;
3. $\|x\|_\infty = \max_i |x_i|$.

Отметим, что все нормы в пространстве R^n являются *эквивалентными*, т.е. любые две нормы $\|x\|_i$ и $\|x\|_j$ связаны соотношениями:

$$\alpha_{ij}\|x\|_j \le \|x\|_i \le \beta_{ij}\|x\|_j,$$

$$\tilde{\alpha}_{ij}\|x\|_i \le \|x\|_j \le \tilde{\beta}_{ij}\|x\|_i,$$

причём $\alpha_{ij},\ \beta_{ij},\ \tilde{\alpha}_{ij},\ \tilde{\beta}_{ij}$ не зависят от x. Более того, в *конечномерном* пространстве любые две нормы являются эквивалентными.

Пространство матриц с естественным образом введёнными операциями сложения и умножения на число образуют линейное пространство, в котором многими способами можно

ввести понятие нормы. Однако чаще всего рассматриваются так называемые *подчиненные* нормы, т.е. нормы, связанные с нормами векторов соотношениями:

$$\|A\| = \sup_{\|x\|=1} \frac{\|Ax\|}{\|x\|}.$$

Отмечая подчиненные нормы матриц теми же индексами, что и соответствующие нормы векторов, можно установить, что

$$\|A\|_1 = \max_j \sum_i |a_{ij}|; \quad \|A\|_2 = \sqrt{\max_i \lambda_i(A^T A)}; \quad \|A\|_\infty = \max_i \sum_j |a_{ij}|.$$

Здесь через $\lambda_i(A^T A)$ обозначено собственное число матрицы $A^T A$, где ^{T} – знак транспонирования. Кроме отмеченных выше трёх основных свойств нормы, отметим здесь ещё два:

- $\|AB\| \le \|A\| \cdot \|B\|$,
- $\|Ax\| \le \|A\| \cdot \|x\|$,

причём в последнем неравенстве матричная норма подчинена соответствующей векторной норме. Договоримся использовать в дальнейшем только нормы матриц, подчиненные нормам векторов. Отметим, что для таких норм справедливо равенство: $\|E\| = 1$, где E – единичная матрица.

2.2. Понятие обусловленности матриц и систем линейных алгебраических уравнений (СЛАУ)

Пусть требуется решить СЛАУ

$$Ax = b, \quad b \neq 0, \tag{1}$$

причём правая часть системы содержит погрешность, т.е. фактически требуется решить СЛАУ

$$A(x + \delta x) = b + \delta b. \tag{2}$$

Отсюда δx удовлетворяет СЛАУ $A\delta x = \delta b$ и, считая матрицу A неособой ($\det A \neq 0$), найдём:

$$\delta x = A^{-1}\delta b. \tag{3}$$

Поскольку из (1) следует

$$\|b\| = \|Ax\| \le \|A\| \cdot \|x\| \quad \text{ и отсюда } \frac{\|b\|}{\|A\|} \le \|x\|, \tag{4}$$

то из (3) с учетом (4) последовательно получаем:

$$\|\delta x\| \le \|A^{-1}\| \cdot \|\delta b\| = \|A^{-1}\| \cdot \|A\| \cdot \frac{\|b\|}{\|A\|} \cdot \frac{\|\delta b\|}{\|b\|} \le \nu(A)\|x\| \cdot \frac{\|\delta b\|}{\|b\|}, \tag{5}$$

где обозначено

$$\nu(A) = \|A\| \cdot \|A^{-1}\|.$$

Разделив обе части (5) на $\|x\| \neq 0$, получим:

$$\frac{\|\delta x\|}{\|x\|} \leq \nu(A) \cdot \frac{\|\delta b\|}{\|b\|}. \tag{6}$$

Введённая здесь величина $\nu(A)$ носит название числа обусловленности матрицы A или соответствующей СЛАУ (1). При $\nu(A) \gg 1$ СЛАУ называют плохо обусловленной. Отметим, что число обусловленности зависит от вида матричной нормы.

Свойства числа обусловленности

1. $\nu(E) = 1$;
2. $\nu(A) \geq 1: \quad 1 = \|E\| = \|A \cdot A^{-1}\| \leq \|A\|\|A^{-1}\| = \nu(A)$;
3. $\nu(\alpha A) = \nu(A), \quad \alpha \neq 0.$ ♣

Замечание 1. Получить оценку снизу для $\nu(A)$ можно с помощью формулы (6): из неё следует

$$\nu(A) \geq \frac{\|\delta x\|}{\|x\|} : \frac{\|\delta b\|}{\|b\|}.$$

Достаточно взять какой-либо вектор $x \neq 0$, и, умножив его на матрицу A, получить вектор b. Таким же образом взяв $\delta x \neq 0$ получим δb. Остаётся подставить эти значения в выписанную выше формулу. ♣

Замечание 2. Если $\|A\| = \|A\|_2$, то, как это следует из приведённых выше выражений для норм матрицы, получим:

$$\nu(A) = \sqrt{\frac{\max_i\{\lambda_i(A^T A)\}}{\min_i\{\lambda_i(A^T A)\}}}. \qquad ♣$$

Пример плохо обусловленной системы.

Рассмотрим СЛАУ (1), в которой

$$A = \begin{pmatrix} 1 & -1 & -1 & \dots & -1 \\ 0 & 1 & -1 & \dots & -1 \\ 0 & 0 & 1 & \dots & -1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ 1 \end{pmatrix}.$$

Данная система имеет единственное решение $x = (0, 0, \dots, 0, 1)^T$. Пусть правая часть системы содержит погрешность $\delta b = (0, 0, \dots, 0, \varepsilon),\ \varepsilon > 0$. Тогда

$$\delta x_n = \varepsilon,\ \delta x_{n-1} = \varepsilon,\ \delta x_{n-2} = 2\varepsilon,\ \delta x_{n-k} = 2^{k-1}\varepsilon, \dots, \delta x_1 = 2^{n-2}\varepsilon.$$

Отсюда

$$\|\delta x\|_\infty = \max_i\{|\delta x_i|\} = 2^{n-2}\varepsilon,\ \|x\|_\infty = 1;\ \|\delta b\|_\infty = \varepsilon,\ \|b\|_\infty = 1.$$

Следовательно,

$$\nu_\infty(A) \geq \frac{\|\delta x\|_\infty}{\|x\|_\infty} : \frac{\|\delta b\|_\infty}{\|b\|_\infty} = 2^{n-2}.$$

Поскольку $\|A\|_\infty = n$, то $\|A^{-1}\|_\infty \geq n^{-1}2^{n-2}$, хотя $\det A^{-1} = (\det A)^{-1} = 1$. Пусть, например, $n = 102$. Тогда $\nu(A) \geq 2^{100} > 10^{30}$. При этом если даже $\varepsilon = 10^{-15}$ получим $\|\delta x\|_\infty > 10^{15}$. И тем не менее $\|A\delta x\|_\infty = \varepsilon$.

2.3. Матрицы с диагональным преобладанием

Определение. Матрица $A = \{a_{ij}\}_{i,j=1}^n$ называется *матрицей с диагональным преобладанием* (величины $\delta > 0$), если

$$|a_{ii}| - \sum_{j \neq i} |a_{ij}| \geq \delta, \quad i = 1, 2, \ldots, n. \qquad \clubsuit$$

Теорема 1. *Пусть A – матрица с диагональным преобладанием величины $\delta > 0$. Тогда она неособая и $\|A^{-1}\|_\infty \leq 1/\delta$.* $\clubsuit$

Доказательство. Возьмём произвольный вектор x^0 и рассмотрим СЛАУ $Ax = b$, где $b = Ax^0$. Эта система, очевидно, имеет своим решением вектор x^0. Будем использовать ниже векторную норму $\|x\| = \|x\|_\infty = \max_i |x_i|$. Пусть число k определено из условия: $|x_k| = \max_i |x_i| = \|x\|$. Выпишем уравнение системы $Ax = b$ с этим номером и учитывая неравенство $|x_k| \geq |x_i|$, получим оценку:

$$|b_k| = \Big|\sum_j a_{kj}x_j\Big| \geq |a_{kk}| \cdot |x_k| - \sum_{j \neq k} |a_{kj}||x_j| \geq$$

$$\geq |a_{kk}| \cdot |x_k| - \Big(\sum_{j \neq k} |a_{kj}|\Big)|x_k| =$$

$$= \Big(|a_{kk}| - \sum_{j \neq k} |a_{kj}|\Big) \cdot |x_k| \geq \delta |x_k|.$$

Отсюда $|x_k| \leq |b_k|/\delta$. Но $|x_k| = \|x\|$, а $|b_k| \leq \max_i |b_i| = \|b\|$. Поэтому справедливо неравенство

$$\|x\| \leq \|b\|/\delta. \qquad (*)$$

Но последнее неравенство означает, что однородная система $Ax = 0$ имеет лишь нулевое (тривиальное) решение, а потому матрица A неособая и, стало быть, СЛАУ $Ax = b$ имеет единственное решение для любой правой части b. Используя оценку $(*)$ в форме $\|x\| = \|A^{-1}b\| \leq \|b\|/\delta$, получаем:

$$\|A^{-1}\| = \sup_{\|b\| \leq 1} \frac{\|A^{-1}b\|}{\|b\|} \leq \frac{1}{\delta}. \qquad \clubsuit$$

Следствие. Для выбранной нормы $\|\cdot\| = \|\cdot\|_\infty$ имеем:

$$\nu_\infty(A) = \nu(A) = \|A\|_\infty \cdot \|A^{-1}\|_\infty \leq \frac{1}{\delta}\|A\|_\infty. \quad \clubsuit$$

Замечание 1. Утверждение теоремы о неособости матрицы A не зависит от выбора нормы, использованной при доказательстве теоремы. $\clubsuit$

Замечание 2. Ввиду эквивалентности норм в конечномерных пространствах, нетрудно получить оценки для числа обусловленности матрицы A и при выборе норм $\|\cdot\|_1$, $\|\cdot\|_2$. Действительно: если известно, что $\|A\|_j \leq \lambda\|A\|_\infty$, то получаем:

$$\nu_j(A) = \|A\|_j \cdot \|A^{-1}\|_j \leq \lambda^2\|A\|_\infty \cdot \|A^{-1}\|_\infty \leq \frac{\lambda^2}{\delta}\|A\|_\infty. \quad \clubsuit$$

2.4. Точные методы. Методы Гаусса.

Определение. Метод решения СЛАУ будем называть точным, если при точном задании параметров задачи он позволяет (в принципе) найти точное решение за конечное число шагов. ♣

К точным методам относится, в частности, метод Гаусса, имеющий несколько разновидностей. Рассмотрим простейшую из них – простой метод Гаусса (метод исключения неизвестных).

Договоримся обозначать a_{*j} – j-ый столбец матрицы A, а через a_{k*} – k-ую строку матрицы A, через a_{ij}, как обычно, обозначим элемент матрицы A, стоящий на пересечении i-ой строки и j-го столбца, а саму СЛАУ записывать в матричной форме в виде:

$$Ax = b \tag{1}$$

Тогда схема простого метода Гаусса выглядит так: пусть $a_{11} \neq 0$; разделим первое уравнение СЛАУ на a_{11} (такой элемент будем называть ведущим) и умножая его последовательно на $a_{i1},\ i = 2, 3, \ldots, n$ вычтем результат умножения из соответствующих уравнений системы. СЛАУ примет вид:

$$\begin{aligned}
&1 \cdot x_1 + a^1_{12}x_2 + \cdots + a^1_{1n}x_n = b^1_1\\
&0 \cdot x_1 + a^1_{22}x_2 + \cdots + a^1_{2n}x_n = b^1_2\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&0 \cdot x_1 + a^1_{n2}x_2 + \cdots + a^1_{nn}x_n = b^1_n.
\end{aligned} \tag{2}$$

Если в системе (2) коэффициент a^1_{22} отличен от нуля, то повторяя указанные выше действия со второй строкой преобразованной матрицы и нижележащими строками, придём к матрице:

$$\begin{aligned}
&1 \cdot x_1 + a^1_{12}x_2 + \cdots + a^1_{1n}x_n = b^1_1\\
&0 \cdot x_1 + 1 \cdot x_2 + \cdots + a^2_{2n}x_n = b^2_2\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&0 \cdot x_1 + 0 \cdot x_2 + \cdots + a^2_{nn}x_n = b^2_n.
\end{aligned} \tag{3}$$

После n шагов в случае ненулевых ведущих элементов придём к треугольной матрице:

$$\begin{aligned}
&1 \cdot x_1 + a^1_{12}x_2 + \cdots + a^1_{1n}x_n = b^1_1\\
&0 \cdot x_1 + 1 \cdot x_2 + \cdots + a^2_{2n}x_n = b^2_2\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&0 \cdot x_1 + 0 \cdot x_2 + \cdots + 1 \cdot x_n = b^n_n.
\end{aligned} \tag{4}$$

Проделанные преобразования СЛАУ носят название прямого хода метода Гаусса, нахождение же компонент вектора x из полученной треугольной системы носит название обратного хода метода Гаусса.

Очевидным препятствием на пути применения этого метода может стать обнуление ведущего элемента на некотором шаге. Однако если исходная матрица системы есть матрица с диагональным преобладанием, то *этого не случится*. (В.С.Рябенький. Введение в ВМ.) Если же матрица не обладает этим свойством, то применяют метод Гаусса с выбором главного элемента по столбцу (реже по всей матрице). В данной модификации роль ведущего

элемента на k-ом шаге играет максимальный по модулю элемент, расположенный в k-ом столбце в строках с номерами, большими k. При этом строка, в которой находится этот элемент, переставляется с k-ой строкой. Очевидно, что если исходная матрица СЛАУ неособая, то данная модификация свободна от отмеченного недостатка простого метода Гаусса.

Ещё одной разновидностью метода Гаусса является метод Жордана, в котором совмещены прямой и обратный ход. Все отмеченные разновидности метода требуют $O(n^3)$ арифметических операций и являются в этом смысле оптимальными среди точных методов.

2.5. Метод квадратного корня

Будем считать, что матрица A симметричная и положительно определена. Поставим задачу представить её в виде произведения: $A = S^T S$, где S – правая треугольная матрица:

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ 0 & s_{22} & \dots & s_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_{nn} \end{pmatrix}.$$

Умножая матрицу S^T на S справа и приравнивая элементы результирующей матрицы соответствующим элементам матрицы A, получим:

$$\begin{aligned} & s_{11}^2 = a_{11}, \quad s_{1i} = a_{1i}/s_{11}, \quad i = \overline{2,n}, \\ & s_{kk}^2 = a_{kk} - \sum_{i<k} s_{ik}^2, \quad k \geq 2, \\ & s_{ij} = \Big(a_{ij} - \sum_{k<i} s_{ki} s_{kj}\Big)/s_{ii}, \quad j > i. \end{aligned} \tag{1}$$

2.6. Метод отражений

Пусть $w \in R^n$ – вектор-столбец единичной длины в евклидовой метрике: $w^T w = \sum_{j=1}^n w_i^2 = j$. С его помощью построим матрицу $U = E - 2ww^T$, где E – единичная матрица $n \times n$. Очевидно, что $U = U^T$ и, кроме того,

$$U^2 = U^T U = (E - 2ww^T)^T (E - 2ww^T) = E - 4ww^T + 4w(w^T w)w^T = E,$$

т.е. матрица U является симметричной и ортогональной. Последнее равенство означает, что все собственные числа построенной матрицы U удовлетворяют соотношению: $\lambda^2(U) = 1$. Проверим, что вектор w является собственным для матрицы U, отвечающим собственному значению $\lambda(U) = -1$:

$$Uw = (E - 2ww^T)w = w - 2w(w^T w) = -w.$$

Кроме того, любой вектор $v \in R^n$, $v \perp w$ (т.е. $w^T v = 0$) является собственным вектором матрицы U, отвечающим собственному значению $\lambda(U) = +1$:

$$Uv = (E - 2ww^T)v = v - 2w(w^T v) = v.$$

Если рассмотреть произвольный вектор $y \in R^n$ и разложить его по векторам $z, v:\ y = z + v$, где $z = \alpha w$, $v \perp w$, то после умножения его на матрицу U получим: $Uy = -z + v$, т.е. вектор Uy является зеркальным отражением вектора y относительно плоскости, перпендикулярной вектору w.

Пусть y, z – произвольные векторы из R^n. Построим вектор w таким образом, чтобы $Uy = \alpha z$. Поскольку матрица U является ортогональной и, следовательно, вектор Uy имеет в евклидовой метрике ту же длину, что и y, то α определится из условия $||Uy|| = ||y||$, т.е. $\alpha = \|y\|/\|z\|$. Следовательно, положив $w = (y - \alpha z)/\rho$, где $\rho = \|y - \alpha z\|$, мы получим искомый вектор.

Используем полученные результаты для упрощения СЛАУ. Возьмём на первом шаге преобразования СЛАУ $Ax = b$ в качестве вектора y первый столбец матрицы A, а в качестве z – орт $e^1 = (1, 0, \ldots, 0)^T$ и построим как указано выше вектор $w^1 = y - \alpha e^1$. Умножив обе части исходной системы на матрицу $U_1 = (E - 2w^1(w^1)^T)$, получим СЛАУ $A_1 x = b^1,\ A_1 = U_1 A,\ b^1 = U_1 b$, у которой первый столбец имеет нули во всех строках, кроме первой (если он уже имел такой вид, то никаких преобразований с ним производить не требуется).

На втором шаге положим $y = (a^1_{22}, a^1_{32}, \ldots, a^1_{n2})^T \in R^{n-1}$ (если второй столбец матрицы A_1 не коллинеарен вектору $e^2 = (0, 1, 0, \ldots, 0)$), $z = e^1 \in R^{n-1}$. Построив матрицу $U_2 = \{u^2_{ij}\}$ (размерности $(n-1) \times (n-1)$) и умножив обе части СЛАУ $A_1 x = b_1$ на матрицу

$$\hat{U}_2 = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & u^2_{11} & \ldots & u^2_{1,n-1} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & u^2_{n-1,1} & \ldots & u^2_{n-1,n-1} \end{pmatrix},$$

получим СЛАУ $A_2 x = b^2$, у которой в первых двух столбцах под главной диагональю стоят нули. Дальнейшее очевидно. Получив систему с треугольной матрицей, решаем её как и в методе Гаусса.

Отметим, что преобразования подобного типа (с помощью ортогональных матриц) позволяют минимизировать влияние неустранимых погрешностей в исходных данных на получаемое решение.

2.7. Метод окаймления

Отметим, что все методы, применяемые для решения СЛАУ применимы и для построения обратной матрицы, ибо последняя задача эквивалентна задаче решения совокупности n систем вида $Ax = e^i$, где e^i – i-ый орт пространства R^n. Излагаемый ниже метод предназначен для построения обратных матриц для последовательности матриц увеличивающихся размерностей.

Представим исходную матрицу $A_n = A$ и искомую обратную к ней матрицу в блочном виде

$$A_n = \begin{pmatrix} A_{n-1} & v \\ u^T & a_n \end{pmatrix}, \quad A_n^{-1} = \begin{pmatrix} B_{n-1} & w \\ s^T & \alpha \end{pmatrix},$$

причём считаем, что матрица A^{-1}_{n-1} уже построена. Производя поблочное умножение $A_n A_n^{-1}$ и $A_n^{-1} A_n$ и приравнивая результат единичной матрице, получим соотношения, из которых определятся блочные элементы матрицы A_n^{-1} (далее обозначено для краткости $c = A^{-1}_{n-1} v$):

$$\alpha = (a_n - u^T c)^{-1}, \quad s^T = -\alpha u^T A^{-1}_{n-1}, \quad w = -\alpha c, \quad B_{n-1} = A^{-1}_{n-1} - c s^T.$$

2.8. Итерационные методы. Метод простой итерации

Пусть СЛАУ

$$Ax = b \tag{1}$$

тем или иным способом записана в виде:

$$x = Bx + c. \tag{2}$$

Метод простой итерации (МПИ) состоит в следующем: берётся некоторый вектор $x^0 \in R^n$ и строится последовательность векторов $\{x^k\}$ по формуле

$$x^{k+1} = Bx^k + c. \tag{3}$$

Теорема 1. (достаточное условие сходимости МПИ) *Если $\|B\| < 1$, то СЛАУ (2) имеет единственное решение и последовательность (3) сходится к нему со скоростью геометрической прогрессии.* ♣

Доказательство. Рассмотрим однородную систему

$$x = Bx \tag{4}$$

и пусть $\hat{x} \neq 0$ – ненулевое решение этой системы. Тогда $\|\hat{x}\| \leq \|B\|\|\hat{x}\|$ или $\|\hat{x}\| \cdot (1 - \|B\|) \leq 0$, откуда немедленно $\|\hat{x}\| = 0$, т.е. $\hat{x} = 0$. Поскольку однородная система (4) имеет лишь тривиальное решение, то соответствующая неоднородная система (2) при любом векторе c имеет единственное решение. Далее, пусть $\bar{x} = B\bar{x} + c,\ x^k = Bx^{k-1} + c$. Тогда

$$\|\bar{x} - x^k\| = \|B(\bar{x} - x^{k-1})\| \leq \|B\|^k \|\bar{x} - x^0\|,$$

что и означает сходимость последовательности (3). ♣

Замечание. При выполнении условий теоремы МПИ сходится к решению системы для любого начального вектора. ♣

Более точное условие сходимости содержится в следующей теореме.

Теорема 2. *МПИ сходится при любом начальном векторе и любом векторе c тогда и только тогда, когда все собственные значения матрицы B лежат в единичном круге.* ♣

Интересно и для практики важно получить оценку уклонения вновь построенного члена последовательности от решения через последние члены этой последовательности. Проделаем это.

$$\begin{aligned}
&\bar{x} = B\bar{x} + c,\\
&x^{k+1} = Bx^k + c,\\
&\bar{x} - x^k = B(\bar{x} - x^{k-1}),\\
&\bar{x} - x^{k-1} = x^k - x^{k-1} + B(\bar{x} - x^{k-1}),\\
&\|\bar{x} - x^{k-1}\| \leq \|x^k - x^{k-1}\| + \|B\|\|\bar{x} - x^{k-1}\|,\\
&\|\bar{x} - x^{k-1}\| \leq \frac{\|x^k - x^{k-1}\|}{1 - \|B\|}.
\end{aligned}$$

Последнее соотношение уже может быть использовано для оценки уклонения члена x^{k-1} от искомого решения $\bar{x}$, однако это можно сделать только после построения следующего члена последовательности x^k, хотя, возможно, желаемая точность уже достигнута. Устраним этот недостаток.

Используя третье и последнее из выписанных здесь соотношений, получим желаемый результат:

$$\|\bar{x} - x^k\| \le \|B\|\|\bar{x} - x^{k-1}\| \le \frac{\|B\|}{1 - \|B\|} \cdot \|x^k - x^{k-1}\|.$$

Замечание. Для сходимости построенной последовательности достаточно, чтобы нашлась *любая подчиненная* норма матрицы, в которой выполнено условие теоремы 2. При этом сходимость $\{x^k\}$ к $\bar{x}$ будет иметь место в *любой норме пространства* R^n ввиду эквивалентности норм в конечномерных пространствах. ♣

Определение. Если для системы (2) имеет место сходимость последовательности $\{x^k\}$, то будем говорить, что система (1) приведена к виду, пригодному для применения МПИ. ♣

Покажем, что любую СЛАУ с неособой матрицей можно привести к виду, пригодному для применения МПИ.

Рассмотрим сначала СЛАУ (1), в которой матрица A является симметричной и положительно определённой. Пусть в (2) матрица B и вектор c имеют вид:

$$B = E - \mu A, \quad c = \mu b, \qquad \text{где } \mu = \frac{2}{\|A\| + \varepsilon}, \quad \varepsilon > 0. \tag{5}$$

Очевидно, что система (2) в таком случае эквивалентна системе (1). Отметим здесь же, что матрица B является симметричной (хотя, возможно, и не положительно определённой). Симметрия её влечет за собой тот факт, что $\|B\|_2 = \max|\lambda_B|$, где λ_B – собственное число матрицы B. Поскольку матрица B имеет вид (5), то $\lambda_B = 1 - \mu\lambda_A$. Согласно сделанному предположению о положительной определённости матрицы A верно неравенство: $0 < \lambda_A \le \|A\|$, причём здесь под $\|\cdot\|$ можно понимать *любую (но подчиненную)* норму матрицы. Из последнего неравенства последовательно имеем:

$$0 < \mu\lambda_A < 2, \quad -1 < \mu\lambda_A - 1 < 1, \text{ т.е.} \quad |\lambda_B| < 1,$$

что в соответствии с теоремой 3 означает сходимость МПИ для системы (2) с матрицей (5). Если же в исходной системе A не является симметричной и положительно определённой, то умножая обе части (1) на транспонированную матрицу A^T, придём к системе

$$\hat{A}x = \hat{b}, \ \hat{A} = A^T A, \ \hat{b} = A^T b,$$

которая эквивалентна исходной системе ввиду предположенной неособости A и в которой матрица $\hat{A}$ обладает требуемыми свойствами симметричности и положительной определённости. Тем самым установлено, что *любая* СЛАУ с неособой матрицей указанным выше способом может быть приведена к виду, пригодному для применения МПИ.

Отметим в заключение, что в МПИ на выполнение одной итерации требуется $\approx n^2$ арифметических операций, поэтому если число итераций для достижения заданной точности меньше n, то МПИ по числу операций оказывается экономичнее метода Гаусса.

2.9. Метод Зейделя.

Пусть матрица СЛАУ $Ax = b$ такова, что $\forall i \ a_{ii} \neq 0$. Если в i-ом уравнения произвести деление на a_{ii} и все неизвестные кроме x_i перенести направо, то получим систему вида

$$x = Cx + d, \quad c_{ii} = 0, \ c_{ij} = -a_{ij}/a_{ii}.$$

Представим матрицу C в виде суммы двух треугольных матриц L и U, причём у матрицы L все диагональные и наддиагональные элементы равны нулю, а у матрицы U – диагональные и поддиагональные. Тогда система примет вид

$$x = Lx + Ux + d.$$

Задавшись начальным вектором x^0, будем строить последовательность $\{x^k\}$ по формуле:

$$x^{k+1} = Lx^{k+1} + Ux^k + d.$$

Нетрудно видеть, что хотя в правую часть этого соотношения входит вектор x^{k+1}, формально который ещё не построен, фактически для вычисления i-ой компоненты вектора x^{k+1} в выписанной формуле используются лишь компоненты вектора x^{k+1} с номерами, меньшими i и к этому моменту уже вычисленные.

Данный метод построения итеративной последовательности носит название метода Зейделя. Этот метод можно рассматривать как МПИ. Действительно: из последней формулы выразим x^{k+1}:

$$x^{k+1} = (E - L)^{-1}Ux^k + (E - L)^{-1}d.$$

Стало быть, $B = (E - L)^{-1}U$. Отметим, что $\det(E - L) = 1$, поэтому существует матрица $(E - L)^{-1}$. Выпишем уравнение для собственных чисел матрицы B. Имеем:

$$\det\{(E - L)^{-1}U - \lambda E\} = 0, \text{ или } \det\{\lambda E - (U + \lambda L)\} = 0,$$

$$\begin{vmatrix} \lambda a_{11} & a_{12} & \dots & a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ \lambda a_{n1} & \lambda a_{n2} & \dots & \lambda a_{nn} \end{vmatrix} = 0.$$

Таким образом в соответствии с теоремой 2 необходимым и достаточным условием сходимости метода Зейделя является попадание всех корней последнего уравнения в круг единичного радиуса.

Приведём без доказательства ещё одну теорему:

Теорема 3. *Для сходимости метода Зейделя достаточно, чтобы выполнялось хотя бы одно из условий:*

1. *матрица A исходной системы обладает свойством диагонального преобладания;*
2. *матрица A симметрична и положительно определена.* ♣

2.10. Метод прогонки

Рассмотрим СЛАУ следующего вида:

$$z_0 = k_0 z_1 + n_0, \tag{1}$$

$$a_j z_{j-1} + b_j z_j + c_j z_{j+1} = f_j, \quad j = 1, 2, \dots, N-1 \tag{2}$$

$$z_N = k_N z_{N-1} + n_N, \tag{3}$$

где $z_0,\ z_1, \dots\ z_N$ – неизвестные, а $a_j,\ b_j,\ c_j,\ f_j,\ k_i,\ n_i$ – заданные числа, причём

$$|b_j| \geq |a_j| + |c_j| \geq |a_j| > 0,\ |k_0| < 1,\ |k_N| \leq 1 \tag{4}$$

В матрично-векторной форме систему запишем в виде $Ax = b$, где:

$$A = \begin{pmatrix} 1 & -k_0 & 0 & 0 & 0 & \dots & 0 & 0 \\ a_1 & b_1 & c_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & a_2 & b_2 & c_2 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & -k_N & 1 \end{pmatrix}, \quad b = \begin{pmatrix} n_0 \\ f_1 \\ f_2 \\ \vdots \\ n_N \end{pmatrix}.$$

Матрицы подобной структуры называют *трёхдиагональными.* СЛАУ указанного вида часто возникают при построении сплайнов и решении краевых задач для дифференциальных уравнений методом сеток.

Выражая последовательно неизвестные и подставляя в следующие уравнения, получаем:

$$z_0 = k_0 z_1 + n_0, \quad a_1(k_0 z_1 + n_0) + b_1 z_1 + c_1 z_2 = f_1$$

или

$$z_1 = k_1 z_2 + n_1, \qquad \text{где } k_1 = \frac{-c_1}{b_1 + a_1 k_0},\ n_1 = \frac{-a_1 n_0 + f_1}{b_1 + a_1 k_0}.$$

Пусть уже найдены $z_{m-1} = k_{m-1} z_m + n_{m-1}, \quad m < N - 1$. Тогда повторяя проделанное с первыми двумя уравнениями, получим:

$$a_m(k_{m-1} z_m + n_{m-1}) + b_m z_m + c_m z_{m+1} = f_m.$$

Разрешив последнее равенство относительно z_m, получим:

$$z_m = k_m z_{m+1} + n_m, \tag{5}$$

где

$$k_m = \frac{-c_m}{b_m + a_m k_{m-1}},\ n_m = \frac{-a_m n_{m-1} + f_m}{b_m + a_m k_{m-1}}. \tag{6}$$

Поскольку $k_0,\ n_0$ заданы, то все коэффициенты $\{k_m,\ n_m\}$ формулы (5) определяются по формуле (6). Наконец, подставив выражение (5) при $m = N - 1$ в (3), получаем:

$$z_N = k_N(k_{N-1} z_N + n_{N-1}) + n_N, \tag{7}$$

откуда

$$z_N = \frac{n_N + k_N n_{N-1}}{1 - k_N k_{N-1}} \tag{8}$$

Затем по формуле (6) находим в обратном порядке неизвестные $z_{N-1},\ \dots,\ z_1,\ z_0$. Получение коэффициентов $\{k_j,\ n_j\}$ по формулам (6) называется прямым ходом метода прогонки, а вычисление неизвестных – обратным ходом.

Покажем, что все знаменатели полученных формул отличны от нуля. Прежде всего установим, что $\forall m \le N - 1$ верно неравенство $|k_m| < 1$. Поскольку по условию $|k_0| < 1$, то база для индукции есть. Допустим, что $|k_{m-1}| < 1$ и покажем,что $|k_m| < 1$. Ввиду условия (4) получаем:

$$|b_m + a_m k_{m-1}| \ge |b_m| - |a_m| \cdot |k_{m-1}| > |b_m| - |a_m| \ge 0. \tag{9}$$

При этом для k_m имеем с учетом (9) и индуктивного предположения:

$$|k_m| = \frac{|c_m|}{|b_m + a_m k_{m-1}|} < \frac{|b_m| - |a_m|}{|b_m| - |a_m| \cdot |k_{m-1}|} < 1. \tag{10}$$

В силу неравенства (9) знаменатели в формулах (6) отличны от нуля. Кроме того $1 - k_N k_{N-1} > 0$ в силу установленного выше свойства (10) для параметров $k_1,\ k_2, \ldots, k_{N-1}$, поэтому z_N можно найти по формуле (8), а вслед за этим и все неизвестные $\{z_j\}$.

Таким образом при выполнении ограничений, наложенных на параметры системы, задача имеет единственное решение. Отметим, что количество арифметических операций для нахождения неизвестных указанным способом составляет $8(N+1) - 9$.

3. Метод Ньютона решения систем нелинейных уравнений

Рассмотрим систему нелинейных уравнений

$$F(x) = 0,\ F(x), \quad x \in R^n, \tag{1}$$

и предположим, что существует вектор $\bar{x} \in D \subset R^n$, являющийся решением системы (1). Будем считать, что $F(x) = (f_1(x), f_2(x), \ldots, f_n(x))^T$, причём $f_i(\cdot) \in C^1(D)\ \forall i$.

Разложим $F(x)$ в окрестности точки $\bar{x}$: $F(x) = F(x^0) + F'(x^0)(x - x^0) + o(\|x - x^0\|)$. Здесь

$$F'(x) = \frac{\partial F(x)}{\partial x} = \begin{bmatrix} \dfrac{\partial f_1(x)}{\partial x_1}, & \dfrac{\partial f_1(x)}{\partial x_2}, & \cdots & \dfrac{\partial f_1(x)}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1}, & \dfrac{\partial f_2(x)}{\partial x_2}, & \cdots & \dfrac{\partial f_2(x)}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial f_n(x)}{\partial x_1}, & \dfrac{\partial f_n(x)}{\partial x_2}, & \cdots & \dfrac{\partial f_n(x)}{\partial x_n} \end{bmatrix}$$

называется матрицей Якоби, а её определитель – якобианом системы (1). Исходное уравнение заменим следующим: $F(x^0) + F'(x^0)(x - x^0) = 0$. Считая матрицу Якоби $F'(x^0)$ неособой, разрешим это уравнение относительно x: $\hat{x} = x^0 - [F'(x)]^{-1} F(x^0)$. И вообще положим

$$x^{k+1} = x^k - [F'(x^k)]^{-1} F(x^k). \tag{2}$$

Оценим уклонение

$$\|x^{k+1} - \bar{x}\| = \|(x^k - \bar{x}) - [F'(x^k)]^{-1}(F(x^k) - F(\bar{x}))\|. \tag{3}$$

Преобразуем последнюю формулу:

$$F(z) - F(y) = \int_0^1 F'_t(y + t(z-y))dt = \left(\int_0^1 \frac{\partial F}{\partial x} dt \right)(z - y) \equiv H(z, y)(z - y).$$

По теореме о среднем

$$\int_a^b f(t)g(t)\,dt = f(\xi) \int_a^b g(t)\,dt, \qquad \text{при } g(x) > 0$$

получим:

$$\int_0^1 \frac{\partial f_i(y+t(z-y))}{\partial x_j}\,dt = \frac{\partial f_i(y+\nu_{ij}(z-y))}{\partial x_j} \to \frac{\partial f_i(y)}{\partial x_j} \text{ при } z \to y. \tag{4}$$

Здесь $\nu_{ij} \in (0,1)$. Теперь формула (3) примет вид:

$$\|x^{k+1} - \bar{x}\| = \|\big(E - [F'(x^k)]^{-1}H(x^k,\bar{x})\big)(x^k - \bar{x})\| \equiv \|U(x^k)(x^k - \bar{x})\|. \tag{5}$$

Ввиду (4) имеет место $U(x^k) \to 0$ при $x^k \to \bar{x}$, т.е.

$$\forall \varepsilon:\ 0 < \varepsilon < 1\ \exists \rho:\ \|U(x^k)\| \le \varepsilon < 1 \quad \text{при } x^k \in S_\rho(\bar{x}),$$

где $S_\rho(\bar{x}) = \{x:\ \|x - \bar{x}\| \le \rho\}$.

Таким образом из $x^k \in S_\rho(\bar{x})$ следует

$$\|x^{k+1} - \bar{x}\| \le \|U(x^k)\| \cdot \|x^k - \bar{x}\| \le \varepsilon\|x^k - \bar{x}\| \le \varepsilon\rho < \rho,$$

т.е. $x^{k+1} \in S_\rho(\bar{x})$. Если считать, что $x^0 \in S_\rho(\bar{x})$, то $\{x^k\} \subset S_\rho(\bar{x})$ и имеет место оценка $\|x^k - \bar{x}\| \le \varepsilon^k\|x^0 - \bar{x}\|$ откуда и следует сходимость последовательности $\{x^k\}$, причём со скоростью геометрической прогрессии.

При дополнительном предположении $F(\cdot) \in C^2$ имеет место квадратичная сходимость метода, т.е.

$$\|x^{k+1} - \bar{x}\| \le \omega\|x^k - \bar{x}\|^2.$$

Сформулируем теорему.

Теорема. *Пусть в некоторой окрестности решения $\bar{x}$ системы (1) функции $f_i(\cdot) \in C^2$ и якобиан системы отличен от нуля в этой окрестности. Тогда существует δ-окрестность точки $\bar{x}$ такая, что при любом выборе начального приближения x^0 из этой окрестности последовательность $\{x^k\}$ не выходит из неё и имеет место квадратичная сходимость этой последовательности.* ♣

Замечание 1. В качестве критерия окончания процесса итераций обычно берут условие: $\|x^{k+1} - x^k\| < \varepsilon$. ♣

Замечание 2. Сложность метода Ньютона – в обращении матрицы Якоби. Вводя обозначение $\delta x^k = x^{k+1} - x^k$ получаем для вычисления δx^k СЛАУ

$$\frac{\partial F(x^k)}{\partial x} \cdot \delta x^k = -F(x^k),$$

откуда и находим искомую поправку δx^k, а затем и следующее приближение $x^{k+1} = x^k + \delta x$ к решению $\bar{x}$. Очевидно, что это значительно сокращает количество арифметических операций для построения очередного приближения. ♣

Замечание 3. Начиная с некоторого шага k_0 решают стационарную СЛАУ

$$\frac{\partial F(x^{k_0})}{\partial x} \cdot \delta x^k = -F(x^k).$$

Данное видоизменение носит название *модифицированный метод Ньютона.* ♣

Замечание 4. (О выборе начального приближения). Пусть вектор-функция $\Phi(\lambda, x)$ такова, что $\Phi(1, x) = F(x)$, а система $\Phi(0, x) = 0$ может быть решена. Тогда разбивая $[0, 1]$ на N частей решают методом Ньютона набор из N систем

$$\Phi(i/N, x) = 0,\ i = \overline{1, N},$$

принимая для каждой следующей системы в качестве начального приближения решение предыдущей системы. ♣

Задачи, предлагаемые для лучшего усвоения материала данной главы.

1. Для полинома $P(x) = x^4 - x^3 - 7x^3 + x + 6$ оценить интервалы, которым могут принадлежать положительные и отрицательные корни этого полинома.

2. Выписать формулу Ньютона (метод касательных) для извлечения квадратного корня из числа a.

Пусть заданы матрица и вектор:

$$A = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 4 \end{pmatrix}, \qquad b = \begin{pmatrix} -1 \\ 0 \\ 3 \end{pmatrix}.$$

3. Вычислить числа обусловленности матрицы A с использованием трех рассмотренных в лекциях матричных норм $\|\cdot\|_1,\ \|\cdot\|_2,\ \|\cdot\|_\infty$.

4. Решить СЛАУ $Ax = b$ методами:

- квадратного корня;
- отражений;
- простой итерации;
- Зейделя.

Глава 3. Методы поиска экстремума

Пусть X – нормированное векторное пространство. Рассмотрим множество $Q \subset X$ и пусть на Q определена функция $f(\cdot)$. Будем в дальнейшем рассматривать две задачи:

Задача 1. Найти элемент $\bar{x} \in Q: \ f(\bar{x}) = \min\{f(x), \ x \in Q\}$. ♣

Задача 2. Найти число $f_* = \inf\{f(x), \ x \in Q\}$. ♣

Что понимается под решением задач 1, 2 ?

Определение 1. Решением задачи 1 называется элемент $\bar{x} \in Q: \ f(\bar{x}) \leq f(x), \ \forall x \in Q$ ♣.

Определение 2. Решением задачи 2 будем называть последовательность (и называть её *минимизирующей*) $\{x^k\}_{k=1}^{\infty} \subset Q$, обладающую свойством: $f(x^k) \to f_*$ при $k \to \infty$. ♣

Отметим, что решение задачи 2 всегда существует, в то время как решение задачи 1 может и не существовать. Поэтому важно и интересно выяснить условия, при выполнении которых $x^k \to \bar{x}$ при $k \to \infty$. Очевидно эти условия определяются свойствами функции $f(x\cdot)$ и множества Q .

1. Задача минимизации квадратичной функции

1.1. Постановка задачи.

Пусть X – евклидово n-мерное пространство, обозначаемое далее R^n, $Q = R^n$. Векторы пространства R^n будем считать записанными в столбец, обозначая их строчными латинскими буквами; скалярное произведение векторов x и y будем обозначать в зависимости от удобства использования трояко: $x^T y$, либо (x, y), либо $x \cdot y$.

В пространстве R^n рассмотрим квадратичную функцию

$$f(x) = \frac{1}{2} x^T A x + x^T b, \tag{1}$$

где A – положительно определенная симметричная матрица, т.е. имеют место неравенства:

$$m\|x\|^2 \leq (x, Ax) \leq M\|x\|^2, \quad m > 0. \tag{2}$$

Для функции (1) рассмотрим задачу 1.

1.2. Существование и единственность точки минимума.

Теорема 1. *Квадратичная функция (1) при выполнении условия (2) имеет в R^n единственную точку минимума, удовлетворяющую системе линейных алгебраических уравнений $Ax + b = 0$.* ♣

Доказательство. Пусть точка минимума функции $f(x)$ существует. Рассмотрим окрестность точки минимума, т.е. точки вида $x = \bar{x} + \alpha q$, где α – произвольное действительное

число, а q – произвольный, но фиксированный вектор пространства R^n. Подставим данное выражение в формулу (1):

$$f(x) = f(\bar{x} + \alpha q) = f(\bar{x}) + \alpha(A\bar{x} + b)^T q + \frac{1}{2}\alpha^2 q^T A q. \tag{1}$$

Введем обозначение $\bar{\varphi}(\alpha) = f(\bar{x} + \alpha q)$. Функция $\bar{\varphi}(\alpha)$ является, очевидно, квадратным трехчленом относительно переменной α и этот трехчлен имеет минимум ввиду положительности старшего коэффициента $\frac{1}{2}q^T Aq$, причем этот минимум достигается при $\alpha = 0$. Поэтому необходимое условие минимума функции (обращение в нуль ее производной) является в данном случае и достаточным. Выписывая это условие, получаем ($\dot{\varphi}(\alpha) = \frac{d}{d\alpha}$):

$$\dot{\bar{\varphi}}(0) = (A\bar{x} + b)^T q = 0, \tag{2}$$

откуда ввиду произвольности вектора q следует:

$$A\bar{x} + b = 0, \tag{3}$$

стоит только в соотношении (2) положить $q = A\bar{x} + b$ и воспользоваться свойствами нормы. Таким образом, точка минимума функции (1) в предположении ее существования удовлетворяет системе линейных алгебраических уравнений (3) (в дальнейшем – СЛАУ). Покажем, что точка $\bar{x}$, найденная как решение системы уравнений (3), является точкой минимума функции $f(\cdot)$.

Действительно,

$$f(x) = f\big(\bar{x} + (x - \bar{x})\big) = f(\bar{x}) + q^T(A\bar{x} + b) + \frac{1}{2}(x - \bar{x})^T A(x - \bar{x}) \geq f(\bar{x}) + \frac{1}{2}m\|x - \bar{x}\|^2 \geq f(\bar{x})$$

ввиду неравенства (2) и с учетом того, что вектор $\bar{x}$ удовлетворяет системе (3). Более того, при $x \neq \bar{x}$ имеет место строгое неравенство: $f(x) > f(\bar{x})$, на чём и заканчивается доказательство. ♣

2. Одношаговые градиентные методы

Поскольку сходящаяся минимизирующая последовательность позволяет найти точку минимума рассматриваемой функции с произвольной наперед заданной точностью, то следует подробнее рассмотреть вопрос о построении такой последовательности. Поставим следующую задачу: имея точку $x^k \in R^n$ построить точку $x^{k+1} \in R^n$ такую, чтобы выполнялось соотношение:

$$f(x^{k+1}) < f(x^k). \tag{1}$$

Будем искать точку x^{k+1} в следующем виде:

$$x^{k+1} = x^k + \mu q, \quad q \neq 0, \tag{2}$$

где q – заданный вектор из R^n, называемый направлением спуска, а μ – искомый параметр, называемый обычно шагом метода в направлении спуска. Имеем:

$$f(x^{k+1}) = f(x^k + \mu q) \equiv \varphi(\mu) = \varphi(0) + \mu\dot{\varphi}(0) + \frac{1}{2}\mu^2\ddot{\varphi}(0), \quad \dot{\varphi} = \frac{d\varphi}{d\mu}, \tag{3}$$

где

$$\varphi(0) = f(x^k),\ \dot{\varphi}(0) = q^T(Ax^k + b),\ \ddot{\varphi}(0) = q^T A q. \tag{4}$$

Поскольку $\varphi(\mu)$ является квадратным трехчленом с положительным старшим коэффициентом, то у этого трехчлена существует точка минимума $\bar{\mu}_k$, которая может быть найдена из необходимого условия экстремума $\dot{\varphi}(\mu) = 0$, откуда и получаем:

$$\bar{\mu}_k = -\frac{\dot{\varphi}(0)}{\ddot{\varphi}(0)} = -\frac{q^T(Ax^k + b)}{q \cdot Aq}. \tag{5}$$

Очевидно, что если в формуле (2) положить $\mu = \bar{\mu}_k$, то построенная по формуле (2) точка x^{k+1} будет удовлетворять условию (1), причем там будет строгое неравенство в случае $x^k \neq \bar{x}$ и $q^T(Ax^k + b) \neq 0$. Продолжая указанные построения, получим последовательность $\{x^k\}$, которую естественно назвать последовательностью убывания для функции $f(x)$. Если в приведенных выше формулах считать, что $q = Ax^k + b$ (то есть вектор q является вектором градиента функции $f(x)$ в точке x^k), то соответствующий метод построения последовательности называют *градиентным методом*. Если к тому же шаг метода μ выбирается по формуле (3), то такой метод называют (одношаговым) *методом наискорейшего градиентного спуска*.

Замечание. В случае выбора направлений спуска на каждом шаге в виде $q = e^i$, где $e^i = (\underbrace{0, 0, \ldots, 0, 1}_{i}, 0 \ldots, 0)^T$ – i-ый орт пространства R^n, описанный выше метод носит название *метода покоординатного спуска.* ♣

Пусть последовательность убывания $\{x^k\}$ для функции $f(x)$ строится по формулам вида (2), т.е. $x^{k+1} = x^k + \mu_k q$. Построение этой последовательности сопровождается построением ещё двух последовательностей $\{q^k\}$ и $\{\mu_k\}$.

Теорема 2. *Пусть для последовательностей $\{x^k\}$, $\{q^k\}$, $\{\mu_k\}$ выполнены следующие условия:*

1. $\exists\, \gamma \in (0, 1]$ *такое, что* $(Ax^k + b)^T q^k \leq -\gamma \|q^k\| \cdot \|Ax^k + b\|$;
2. $\mu_k = \tau_k \bar{\mu}_k,\ \tau_k \in (\varepsilon, 2 - \varepsilon),\ 0 < \varepsilon < 1,\ \bar{\mu}_k = -\dfrac{q^k \cdot (Ax^k + b)}{q^k \cdot Aq^k}$.

Тогда $x^k \to \bar{x}$ при $k \to \infty$. ♣

Лемма. (Свойства минимизирующей последовательности) Пусть $\{x^k\}_{k=1}^{\infty}$ – минимизирующая последовательность в задаче 2.

Следующие три утверждения выполняются одновременно:

а) $\lim\limits_{k\to\infty} x^k = \bar{x}$;

б) $\lim\limits_{k\to\infty} f(x^k) = f(\bar{x})$;

в) $\|Ax^k + b\| \to 0$ *при* $k \to \infty$. ♣

Доказательство леммы. Очевидно, что из а) следует б), ибо рассматриваемая функция непрерывна. Покажем, что из б) следует в), для чего оценим норму градиента функции $f(\cdot)$

через разность значений функции в точках x^k и $\bar{x}$:

$$|f(x^k) - f(\bar{x})| = (x^k - \bar{x})^T A(x^k - \bar{x}) = \left[A(x^k - \bar{x})\right]^T A^{-1}\left[A(x^k - \bar{x})\right] =$$
$$= (Ax^k + b)^T A^{-1}(Ax^k + b) \geq \frac{1}{M}\|Ax^k + b\|^2. \tag{6}$$

В оценке (6) использовано то обстоятельство, что матрица A^{-1} так же как и A является положительно определённой, причём для квадратичной формы с обратной матрицей в формуле (2) роль m играет M^{-1}, а роль $M - m^{-1}$. Остаётся показать, что из в) следует а). Действительно:

$$m\|x^k - \bar{x}\|^2 \leq (x^k - \bar{x})^T A(x^k - \bar{x}) \leq (x^k - \bar{x})^T [A(x^k - \bar{x})] =$$
$$= (x^k - \bar{x})^T (Ax^k + b) \leq \|x^k - \bar{x}\| \cdot \|Ax^k + b\|, \tag{7}$$

откуда, сокращая обе части на $\|x^k - \bar{x}\| \neq 0$, получаем оценку:

$$\|x^k - \bar{x}\| \leq \frac{1}{m}\|Ax^k + b\|, \tag{8}$$

откуда и следует доказываемое утверждение. ♣

Замечание. При $x^k = \bar{x}$ неравенство (8), очевидно, верно. ♣

Доказательство теоремы. Оценим разность $-\Delta f_k = f(x^k) - f(x^{k+1})$:

$$-\Delta f_k = -\mu_k \dot{\varphi}(0) - \frac{1}{2}\mu_k^2 \ddot{\varphi}(0) = -\left(-\tau_k \frac{\dot{\varphi}(0)}{\ddot{\varphi}(0)}\right)\dot{\varphi}(0) - \frac{1}{2}\left(-\tau_k \frac{\dot{\varphi}(0)}{\ddot{\varphi}(0)}\right)^2 \ddot{\varphi}(0) =$$

$$= \tau_k \frac{[\dot{\varphi}(0)]^2}{\ddot{\varphi}(0)} - \frac{1}{2}\tau_k^2 \frac{[\dot{\varphi}(0)]^2}{\ddot{\varphi}(0)} = \frac{\tau_k(2 - \tau_k)[(Ax^k + b)^T q^k]^2}{2q^k \cdot Aq^k} \geq$$
$$\geq \frac{\varepsilon(2 - \varepsilon)\gamma^2}{2M} \cdot \frac{\|q^k\|^2 \cdot \|Ax^k + b\|^2}{\|q^k\|^2} \equiv \sigma^2 \|Ax^k + b\|^2,$$

где обозначено

$$\sigma^2 = \frac{\varepsilon(2 - \varepsilon)\gamma^2}{2M} > 0.$$

Следовательно, имеет место оценка:

$$\Delta f_k \leq -\sigma^2 \|Ax^k + b\|^2, \tag{9}$$

т.е. на построенной последовательности происходит монотонное убывание функции $f(x)$. Пусть, однако, вопреки утверждению теоремы последовательность $\{x^k\}$ не сходится к точке $\bar{x}$, другими словами, в соответствии с леммой существует число $\rho > 0$ и подпоследовательность натурального ряда $\{k_s\}_{s=1}^{\infty}$ такая, что на ней $\|Ax^{k_s} + b\| \geq \rho > 0$. Используя полученную выше оценку (9), на указанной подпоследовательности имеем:

$$\Delta f_{k_s} \leq -\sigma^2 \rho^2, \quad s = 1, 2, \ldots \ .$$

Рассмотрим последовательность $\{f(x^k)\}$: очевидно, что

$$f(x^{k+1}) = f(x^1) + \sum_{i=1}^{k} \Delta f_i,$$

а поскольку $\Delta f_i \le 0 \ \ \forall i$ то, оставляя в последней сумме только слагаемые, номера которых входят в $\{k_s\}$ и не превосходят k, получаем оценку:

$$f(x^{k+1}) \le f(x^1) + \sum_{k_s \le k} \Delta f_i \le f(x^1) - \sigma^2 \rho^2 N(k),$$

где $N(k)$ – количество слагаемых в последней сумме, причём ввиду неограниченности членов в подпоследовательности $\{x^{k_s}\}$, $N(k)$ также неограниченно возрастает при $k \to \infty$, иными словами, имеем: $N(k) \to \infty$ при $k \to \infty$. Но тогда из оценки (7) следует: $f(x^{k+1}) \to -\infty$ при $k \to \infty$, что противоречит теореме о существовании точки минимума (см. с.30). ♣

Следствие. Метод наискорейшего градиентного спуска сходится.

Замечание 1. Формула (8) позволяет оценить расстояние от произвольной точки $x^k \in R^n$ до точки минимума $\bar{x}$ функции $f(\cdot)$. ♣

Замечание 2. Формула (8) для матриц с диагональным преобладанием величины δ может быть представлена в виде:

$$\|x^k - \bar{x}\| \le \frac{1}{\delta} \|Ax^k + b\|. \quad ♣$$

3. Многошаговые градиентные методы

Пусть из точки x^k методом наискорейшего градиентного спуска (МНГС) сделано s шагов и при этом получены промежуточные векторы $x^{k+1}, x^{k+2}, \ldots, x^{k+s}$. Выпишем выражение каждого из этих векторов через вектор x^k и направление спуска в этой точке q^k, используя формулы МНГС:

$$x^{k+1} = x^k + \bar{\mu}_k q^k, \quad q^k = Ax^k + b, \quad \mu_k = -\frac{q^k \cdot (Ax^k + b)}{q^k \cdot Aq^k}.$$

Получаем последовательно:

$$\begin{array}{rcl} x^{k+1} & = & x^k + \bar{\mu}_k q^k; \\ x^{k+2} & = & x^{k+1} + \bar{\mu}_{k+1} q^{k+1} = x^k + \bar{\mu}_k q^k + \bar{\mu}_{k+1} q^{k+1} = \\ & = & x^k + \bar{\mu}_k q^k + \bar{\mu}_{k+1} \big(A(x^k + \bar{\mu}_k q^k) + b\big) = \\ & = & x^k + (\bar{\mu}_k + \bar{\mu}_{k+1}) q^k + \bar{\mu}_k \bar{\mu}_{k+1} A q^k. \end{array}$$

Таким образом, очевидно, для x^{k+s} получим представление:

$$x^{k+s} = x^k + \left(\sum_{i=1}^{s} \alpha_i A^{i-1} \right) q^k,$$

где $\alpha_i = \alpha_i(\bar{\mu}_k, \bar{\mu}_{k+1}, \ldots, \bar{\mu}_{k+s-1})$ – функции указанных аргументов.

Введём линейное многообразие

$$L_k = \left\{ x(c) : \ x(c) = x^k + \left(\sum_{i=1}^{s} c_i A^{i-1} \right) q^k \right\}, \tag{1}$$

причём векторы $\{q^k, Aq^k, \ldots, A^{s-1}q^k\}$ мы *вправе считать линейно независимыми*. Поставим задачу: найти $\min\{f(x), \ x \in L_k\}$, или другими словами: найти точку $x^{k+1} \in L_k$ такую, что

$$f(x^{k+1}) = \min_{x \in L_k} f(x). \tag{2}$$

Выпишем необходимые (а в рассматриваемой задаче и достаточные) условия минимума в задаче (2). Поскольку функция

$$\psi(c) = f\big(x^k + \sum_{i=1}^{s} c_i A^{i-1} q^k\big), \quad c = (c_1, c_2, \ldots, c_s)^T$$

является, очевидно, квадратичной по $\{c_i\}$ (и, следовательно, дифференцируемой), упомянутые условия минимума имеют вид:

$$\frac{\partial \psi}{\partial c_j} = 0, \text{ или } \left.\frac{\partial f}{\partial x}\right|_{x=x^{k+1}} \cdot \frac{\partial x(c)}{\partial c_j} = 0, \quad j = \overline{1,s}.$$

С учетом вида функции $f(x)$ эти условия принимают вид:

$$\Big(Ax^{k+1} + b\Big) \cdot A^{j-1} q^k = 0. \tag{3}$$

Поскольку $x^{k+1} \in L_k$, то подставляя сюда представление векторов из L_k по формуле (1), получаем следующую СЛАУ относительно $\{c_i\}$:

$$\sum_{i=1}^{s} c_i q^k \cdot A^{i+j-1} q^k = -q^k \cdot A^{j-1} q^k, \quad j = \overline{1,s}. \tag{4}$$

Матрица полученной системы неособая как матрица Грама и ввиду линейной независимости векторов $\{q^k, Aq^k, \ldots, A^{s-1}q^k\}$. Рассмотренный метод носит название *s-шагового метода наискорейшего градиентного спуска.*

Замечание. Пусть в линейном пространстве H со скалярным произведением (u, v), $u, v \in H$ выбрано s элементов $\{u^i\}_{i=1}^s$. Рассмотрим линейную комбинацию этих элементов $u = \sum_{i=1}^{s} c_i u^i$ и её норму:

$$\|u\|^2 = (u, u) = \sum_{i,j=1}^{s} c_i c_j (u^i, u^j) \geq 0,$$

Таким образом данная квадратичная форма (её матрица и носит название матрицы Грама) является неотрицательно определённой, а при условии линейной независимости элементов $\{u^i\}_{i=1}^s$ – положительно определённой, а следовательно её матрица $C = \{(u^i, u^j)\}$ является неособой.
Матрица СЛАУ (4) является матрицей Грама, поскольку построена по линейно независимым векторам $q^k, Aq^k, \ldots, A^{s-1}q^k$ указанным выше способом, если скалярное произведение в R^n ввести так: $[x, y] = (x, Ay)$, где $(\cdot, \cdot)$ – введенное ранее скалярное произведение: $(x, y) = \sum x_i y_i$. Поэтому СЛАУ (4) имеет единственное решение при любой правой части. ♣

4. Стационарный s-шаговый метод спуска

Рассмотренный в предыдущем разделе s-шаговый МНГС называется *нестационарным* ввиду того, что на каждом шаге приходится заново пересчитывать параметры метода $\{c_i\}$, решая для этого каждый раз СЛАУ с новой матрицей, что является трудоёмкой процедурой.

Построим *стационарный* метод s-шагового спуска, т.е. метод, в котором в отличие от предыдущего параметры не изменяются от шага к шагу.

Пусть $\bar{x}$ – точка минимума функции $f(\cdot)$, а очередное приближение строится как в s-шаговом градиентном методе:

$$x^{k+1} = x^k + \sum_{i=1}^{s} c_i A^{i-1} q^k. \tag{1}$$

Оценим уклонение x^{k+1} от $\bar{x}$:

$$\begin{aligned} x^{k+1} - \bar{x} &= x^k + \sum_{i=1}^{s} c_i A^{i-1} q^k - \bar{x} = \\ &= (x^k - \bar{x}) + \sum_{i=1}^{s} c_i A^{i-1}(Ax^k + b) = \big(E + \sum_{i=1}^{s} c_i A^i\big)(x^k - \bar{x}). \end{aligned} \tag{2}$$

Здесь учтено то обстоятельство, что $q^k = Ax^k + b = Ax^k - A\bar{x} = A(x^k - \bar{x})$.

Введём в рассмотрение полином

$$\Phi(\lambda) = 1 + c_1\lambda + c_2\lambda^2 + \cdots + c_s\lambda^s. \tag{3}$$

Тогда формула (2) может быть записана в виде $x^{k+1} - \bar{x} = \Phi(A)(x^k - \bar{x})$ или, оценивая по норме обе части равенства, получим:

$$\|x^{k+1} - \bar{x}\| \le \|\Phi(A)\| \cdot \|(x^k - \bar{x})\|. \tag{4}$$

Рассмотрим последовательность x^k, члены которой определяются по формуле

$$x^{k+1} = x^k + \sum_{i=1}^{s} c_i A^{i-1} q^k, \qquad \text{где } x^1 \in R^n \text{ – произвольный вектор.} \tag{5}$$

Теорема. *Пусть числа $\{c_i\}$ в (3) таковы, что $\gamma_s = \|\Phi(A)\| < 1$. Тогда $x^k \to \bar{x}$ при $k \to \infty$, причём скорость сходимости характеризуется оценкой*

$$\|x^{k+1} - \bar{x}\| \le \gamma_s^k \|x^1 - \bar{x}\|. \quad \clubsuit \tag{6}$$

Доказательство. Из (4) методом математической индукции следует оценка (6) в которой $\gamma_s^k \to 0$ при $k \to \infty$. $\clubsuit$

Чтобы перейти к вопросу о фактическом выборе коэффициентов стационарного метода и притом оптимальным образом, рассмотрим вспомогательный вопрос.

5. Полиномы Чебышева

Рассмотрим функции вида:

$$T_n(x) = \cos(n \arccos x), \quad |x| \le 1. \tag{1}$$

Имеет место легко проверяемое тождество:

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

Подставляя сюда $\varphi = \arccos x$, получим:

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

а поскольку $T_0(x) = 1,\ T_1(x) = x$, то в соответствии с полученным рекуррентным соотношением функции $T_n(x)$ являются полиномами и носят название полиномов Чебышева. Например,

$$T_2(x) = 2x^2 - 1,\ T_3(x) = 4x^3 - 3x,\ T_4(x) = 8x^4 - 8x^2 + 1.$$

Свойства полиномов Чебышева.

1. Полином $T_n(x)$ имеет n различных действительных корней, расположенных на интервале $(-1, 1)$:

$$\cos(n \arccos x) = 0;\ n \arccos x = \frac{\pi}{2}(2m+1);$$

$$\arccos x = \frac{\pi}{2n}(2m+1),\ m = \overline{0, n-1};\ x = \cos\frac{\pi(2m+1)}{2n},\ m = \overline{0, n-1}.$$

2. Экстремальные значения $T_n(x)$ суть $+1$ и -1, чередуются и достигаются в $(n+1)$ точках интервала $[-1, 1]$:

$$n \arccos(x) = \pi m,\ x = \cos\frac{\pi m}{n}, \quad m = \overline{0, n}.$$

3. Полином Чебышева степени n является четной или нечетной функцией вместе с четностью или нечетность его степени n, что легко следует из формулы (2).

4. Пусть $\mathcal{P}_n^1$ – множество полиномов степени n с коэффициентом 1 при старшей степени. Тогда $\bar{T}_n = \dfrac{1}{2^{n-1}} T_n(x)$ является решением задачи:

$$\max_{x\in[-1,1]} |\bar{T}_n(x)| = \min_{P_n(x)\in\mathcal{P}_n^1} \max_{x\in[-1,1]} |P_n(x)|.$$

Чтобы убедиться в этом, учтём, что

$$\max_{x\in[-1,1]} |\bar{T}_n(x)| = \frac{1}{2^{n-1}}$$

и покажем, что для произвольного полинома $P_n(\cdot) \in \mathcal{P}^1$ верно неравенство

$$\min_{P_n(x)\in\mathcal{P}_n^1} \max_{x\in[-1,1]} |P_n(x)| \geq \frac{1}{2^{n-1}}.$$

Действительно, если это не так, то найдётся полином $P_n(\cdot)$, для которого

$$\max_{x\in[-1,1]} |P_n(x)| < \frac{1}{2^{n-1}},$$

Тогда полином степени $(n-1)$

$$Q_{n-1}(x) = \bar{T}_n(x) - P_n(x),$$

принимающий в $(n+1)$ точках

$$x = \cos\frac{\pi m}{n}, \quad m = \overline{0, n},$$

являющихся точками экстремума полинома Чебышева, попеременно положительные и отрицательные значения, должен иметь n корней, что невозможно. ♣

Ввиду последнего свойства многочлены $\bar{T}_n(x)$ получили название *многочленов, наименее отклоняющихся от нуля.*

Замечание. Очевидно, что многочлен $\alpha\bar{T}_n(x)$ является многочленом, наименее отклоняющимся от нуля на интервале $[-1,1]$ среди всех многочленов со старшим коэффициентом α. В частности, если положить $\alpha = \eta\dfrac{1}{\bar{T}_n(\xi)}$, $\xi \notin [-1,1]$, то такой полином является полиномом, наименее отклоняющимся от нуля на интервале $[-1,1]$ и принимающим значение η в точке ξ, *расположенной вне этого интервала.* ♣

6. Стационарный оптимальный s-шаговый метод спуска

Как было установлено ранее, для обеспечения сходимости стационарного s-шагового метода спуска достаточно обеспечить выполнение неравенства $\gamma_s = \|\Phi(A)\| < 1$. Оценим норму матрицы $\|\Phi(A)\|$, считая, что используется вторая норма матрицы: $\|A\| = \sqrt{\max\lambda_i(A^TA)}$. Поскольку матрица A симметричная и положительно определённая, то $\|A\| = \max\lambda_i(A)$. Очевидно, что матрица $\Phi(A)$ также симметрична и $\lambda(\Phi(A)) = \Phi(\lambda(A))$. Тем самым имеем:

$$\|\Phi(A)\| = \max_i |\Phi(\lambda_i(A))|.$$

Поскольку же собственные числа матрицы A находятся на интервале $[m,M]$ (см. с. 30, формула (2)), то

$$\max_i |\Phi(\lambda_i(A))| \le \max_{\tau\in[m,M]} |\Phi(\tau)|,$$

причём оценка является точной, т.е. не может быть улучшена. Скорость сходимости последовательности

$$x^{k+1} = x^k + \sum_{i=1}^{s} c_i A^{i-1} q^k, \quad x^0 \in R^n, \quad q^k = Ax^k + b,$$

как это следует из полученной ранее на стр.36 оценки (6) $\|x^{k+1}-\bar{x}\| \le \gamma_s^k\|x^1-\bar{x}\|$ возрастает вместе с уменьшением величины $\gamma_s = \|\Phi(A)\|$, поэтому имеет смысл искать параметры метода $\{c_i\}$ (они же коэффициенты полинома $\Phi(\tau)$) из условия

$$\bar{\gamma}_s = \min_{\{c_i\}} \max_{\tau\in[m,M]} |\Phi(\tau)|. \tag{1}$$

Сделав в последней задаче замену переменных

$$z = a\tau + b, \quad \text{где } a = \frac{2}{M-m}, \quad b = -\frac{M+m}{M-m} \tag{2}$$

и обозначив $\Psi(z) = \Phi\Big(\dfrac{z-b}{a}\Big)$ перепишем (1) в виде:

$$\min_{\Phi(0)=1} \max_{\tau\in[m,M]} |\Phi(\tau)| = \min_{\Psi(b)=1} \max_{z\in[-1,1]} |\Psi(z)|, \quad b < -1. \tag{3}$$

Решением же последней задачи, как это следует из замечания на с.38, является полином

$$\bar{\Phi}(\tau) = \frac{T_s(a\tau+b)}{T_s(b).} \tag{4}$$

При этом $\bar{\gamma}_s = \dfrac{1}{|T_s(b)|} < 1$, поскольку вне отрезка $[-1,1]$ полиномы Чебышева являются монотонными функциями, принимая значения -1 либо $+1$ на концах.

Рассмотренный метод выбора параметров многошагового метода будем называть *методом оптимального стационарного (градиентного) спуска.* Он же известен в литературе как *метод Ричардсона.*

7. Методы сопряженных направлений

Поскольку матрица A квадратичной формы, участвующая в записи квадратичной функции $f(x)$, симметрична и положительно определена, то она имеет ровно n собственных векторов, образующих базис пространства R^n, которые можно считать ортонормированными, т.е. $x^i \cdot x^j = \delta_{ij}$, где δ_{ij} – символ Кронекера. Пусть $P = \{x^1, x^2, \ldots, x^n\}$ – матрица, составленная из собственных векторов матрицы A. Она является ортогональной, поскольку ввиду сделанных предположений $P^T P = E$, т.е. $P^T = P^{-1}$. В выражении для функции $f(x)$ положим $x = Pz$:

$$f(x) = f(Pz) = \frac{1}{2} z^T (P^T A P) z + b^T P z = \frac{1}{2} \sum_{i=1}^{n} \lambda_i z_i^2 + \sum_{i=1}^{n} \mu_i z_i. \tag{1}$$

Ясно, что

$$\min_{x \in R^n} f(x) = \min_{z \in R^n} f(Pz), \quad \bar{x} = P\bar{z}.$$

Минимум же функции $g(z) = f(Pz)$ находится из (1) с учетом использования необходимых (и достаточных в данном случае) условий:

$$\lambda_i \bar{z}_i + \mu_i = 0, \quad \bar{z}_i = -\frac{\mu_i}{\lambda_i}, \qquad i = \overline{1, N}. \tag{2}$$

Однако нахождение собственных чисел $\{\lambda_i\}$ и собственных векторов $\{x^i\}$ матрицы A – самостоятельная и непростая задача. Можно обойти эту проблему следующим образом: пусть матрица Q преобразования переменных состоит из n векторов-столбцов $\{q^i\}$, $\quad q^i \neq 0$, удовлетворяющих условиям: $q^i \cdot A q^j = \tau_i \delta_{ij}$, $\quad i, j = \overline{1, n}$, где δ_{ij} – символ Кронекера: $\delta_{ij} = 0$ при $i \neq j$ и $\delta_{ij} = 1$.

Определение. Векторы, обладающие указанными выше свойствами, называются A-ортогональными или сопряженными по матрице A. ♣

Сделав замену $x = Qy$, получим:

$$f(x) = f(Qy) = \frac{1}{2} \sum_{i,j=1}^{n} (q^i \cdot A q^j) y_i y_j + \sum_{i=1}^{n} (b^T q^i) y_i = \sum_{i=1}^{n} \Big(\frac{1}{2} \tau_i y_i^2 + \sigma_i y_i \Big), \tag{3}$$

где $\tau_i = q^i \cdot A q^i$, $\sigma_i = b^T q^i$. Легко видеть, что векторы q^i, $i = \overline{1, n}$ линейно независимы, поэтому они образуют базис пространства R^n и вследствие этого

$$\min_{x \in R^n} f(x) = \min_{y \in R^n} f(Qy).$$

Используя условия минимума для (3) легко находим $\bar{y}$ а вместе с ним и $\bar{x}$:

$$\bar{y}_i = -\frac{\sigma_i}{\tau_i}, \quad i = \overline{1, n}; \quad \bar{x} = Q\bar{y}.$$

Остаётся указать способ построения A-ортогональных векторов.

8. Метод A-ортогонализации базиса

Пусть $e^1, e^2, \ldots, e^n$ – базис пространства R^n. Положим $q^1 = e^1,\ q^2 = e^2 + \nu_{21} q^1$, причём ν_{21} найдём из условия A-ортогональности q^1 и q^2 :

$$0 = q^1 \cdot Aq^2 = q^1 \cdot Ae^2 + \nu_{21} q^1 \cdot Aq^1, \quad \nu_{21} = -\frac{q^1 \cdot Ae^2}{q^1 \cdot Aq^1},$$

поскольку $q^1 \cdot Aq^1 > 0$. Аналогично на шаге i будем строить q^i :

$$q^i = e^i + \sum_{j=1}^{i-1} \nu_{ij} q^j,$$

выбирая $\{\nu_{ij}\}_{j=1}^{i}$ из условия A-ортогональности векторов q^i и $q^1, q^2, \ldots, q^{i-1}$:

$$0 = q^s \cdot Aq^i = q^s \cdot Ae^i + \sum_{j=1}^{i-1} \nu_{ij} q^s \cdot Aq^j, \quad s = \overline{1, i-1}, \quad \nu_{is} = -\frac{q^s \cdot Ae^i}{q^s \cdot Aq^s}.$$

Поскольку квадратичная форма $q^T Aq > 0$ для всех $q \in R^n$, то все знаменатели выписанных выше формул отличны от нуля и указанный алгоритм действительно приводит к построению A-ортогонального набора из n векторов $\{q^i\}$.

Метод сопряженных направлений имеет тот недостаток, что точка минимума функции $f(x)$ может быть найдена лишь после окончания всех построений, т.е. здесь отсутствует процесс приближения к искомому минимуму по шагам.

9. Метод сопряженных градиентов

Идея метода состоит в параллельном приведении матрицы A квадратичной формы к диагональному виду и получении приближений к точке минимума.

Пусть построены векторы $x^i,\ q^i = Ax^i + b$, и набор A-ортогональных векторов $p^i,\ i = \overline{0, k}$. Продолжим построения по формулам:

$$x^{k+1} = x^k + \bar{\mu}_k p^k, \tag{1}$$

причём $\bar{\mu}_k$ находится из условия:

$$f(x^{k+1}) = \min_{\mu} f(x^k + \mu p^k),\ \text{т.е. } \bar{\mu}_k = -\frac{(Ax^k + b)^T p^k}{p^k \cdot Ap^k} = -\frac{q^k \cdot p^k}{\tau_k}. \tag{2}$$

$$q^{k+1} = Ax^{k+1} + b = Ax^k + \bar{\mu}_k Ap^k + b = q^k + \bar{\mu}_k Ap^k. \tag{3}$$

Заметим, что вновь построенный вектор градиента ортогонален вектору p^k :

$$q^{k+1} \cdot p^k = q^k \cdot p^k + \bar{\mu}_k p^k \cdot Ap^k = q^k \cdot p^k - \frac{q^k \cdot p^k}{\tau_k} \tau_k = 0. \tag{4}$$

Пусть

$$p^{k+1} = q^{k+1} + \sigma_k p_k,\ \text{причём } p^{k+1} \cdot Ap^k = 0,\ \text{откуда } \sigma_k = -\frac{q^{k+1} \cdot Ap^k}{\tau_k}. \tag{5}$$

Шаг метода закончен. В качестве начальных векторов возьмём x^0 – произвольный вектор из R^n, $\quad q^0 = Ax^0 + b = p^0$.

Пусть уже построены k членов последовательностей $\{q^j\},\ \{p^j\},\ j = \overline{1,k}$ и $\|q^{k-1}\| \neq 0$.

Лемма. *Построенные последовательности обладают свойствами:*

1. $p^j \cdot q^k = 0, \quad j < k;$
2. $q^j \cdot q^k = 0, \quad j < k;$
3. если $\|q^k\| \neq 0$, то $\bar{\mu}_j \neq 0, \quad j < k;$
4. $f(x^0) > f(x^1) > \cdots > f(x^k);$
5. $p^j \cdot Ap^k = 0, \quad j < k.$ ♣

Доказательство леммы проводится методом математической индукции.

Теорема. *Последовательность* $\{x^k\}$, *построенная по формулам (1), обладает свойствами:*

1. $\exists\, \nu \leq n$ *такое, что* $x^\nu = \bar{x}$;
2. $f(x^{k+1}) = \min\limits_{\{c_i\}} f(x^0 + \sum\limits_{i=0}^{k} c_i p^i) = \min\limits_{\{d_i\}} f(x^0 + \sum\limits_{i=0}^{k} d_i q^i).$ ♣

Доказательство.
1. Если $q^k = 0$, то по (2) имеем: $\bar{\mu}_k = 0$ и из (3) получаем: $q^{k+1} = 0$. Следовательно, если $x^k \neq \bar{x}$, то $q^i \neq 0,\ i = \overline{1,k}$. Но тогда из пункта 2. Леммы следует, что векторы $q^i\ j = \overline{1,k}$ линейно независимы как набор попарно ортогональных векторов. Поскольку таких векторов в пространстве R^n не может быть более n, то при некотором $\nu \leq n$ будет выполнено условие $q^\nu = 0$, означающее одновременно и $x^\nu = \bar{x}$.
2. Покажем, что

$$f(x^{k+1}) = \min_{\{c_i\}} f(x^0 + \sum_{i=0}^{k} c_i p^i). \tag{6}$$

Выпишем необходимое условие в задаче минимума:

$$A(x^0 + \sum_{i=0}^{k} \bar{c}_i p^i) + b = 0,$$

умножив обе части скалярно на p^j, получим:

$$q^0 \cdot p^j + \sum_{i=0}^{k} \bar{c}_i p^j \cdot Ap^i = 0, \text{ откуда } \bar{c}_j = -\frac{q^0 \cdot p^j}{\tau_j}. \tag{7}$$

Из (3) имеем:

$$q^j = q^{j-1} + \bar{\mu}_{j-1} Ap^{j-1} = \cdots = q^0 + \sum_{s=1}^{j-1} \bar{\mu}_s Ap^s$$

и с учетом этого и пункта 1. Леммы выражение (2) для $\bar{\mu}_j$ даёт:

$$\bar{\mu}_j = -\frac{q^j \cdot p^j}{\tau_j} = -\frac{q^0 \cdot p^j}{\tau_j} = \bar{c}_j.$$

Следовательно, точка минимума в задаче (6) имеет вид:

$$\hat{x} \equiv x^0 + \sum_{i=0}^{k} \bar{c}_i p^i = x^0 + \sum_{i=0}^{k} \bar{\mu}_i p^i = ((x^0 + \mu_0 p^0) + \mu_1 p^1) + \cdots + \mu_k p^k = x^{k+1}$$

и первая часть второго утверждения теоремы доказана.

Для доказательства второй части достаточно убедиться, что линейные подпространства, в которых ищется минимум в пункте два теоремы, в обоих случаях совпадают. Обозначим

$$L_s = x^0 + \sum_{i=0}^{s} c_i p^i; \quad \hat{L}_s = x^0 + \sum_{i=0}^{s} c_i q^i.$$

Очевидно, что $L_1 = \hat{L}_1$. Допустим $L_s = \hat{L}_s$. Тогда

$$L_{s+1} = L_s + c_{s+1} p^{s+1} = L_s + c_{s+1}(q^{s+1} + \sigma_s p^s) =$$

$$= (L_s + c_{s+1}\sigma_s p^s) + c_{s+1} q^{s+1} = \hat{L}_{s+1}. \quad \clubsuit$$

Замечание 1. Метод сопряженных градиентов называют конечным, поскольку для квадратичной функции он за конечное число шагов приводит в точку минимума. На практике, однако, из-за погрешностей вычислений число шагов может оказаться бесконечным. $\clubsuit$

Замечание 2. (О практической сходимости итерационных методов.) Пусть итерационный метод таков, что имеет место оценка: $\|x^{k+1} - \bar{x}\| \le \gamma \|x^k - \bar{x}\|, \quad \gamma < 1$ и вычисления ведутся с погрешностью $\Delta x^k = x^k - \tilde{x}^k$, причём $\|\Delta x^k\| \le \sigma_k \le \sigma$.
Тогда при прежних предположениях относительно $f(\cdot)$ имеет место сходимость

$$\tilde{x}^k \to S_r(\bar{x}) = \{x : \ \|x - \bar{x}\| \le r\}, \qquad \text{где } r = \frac{\sigma}{1-\gamma}. \quad \clubsuit$$

Типовые задачи по теме "Методы поиска экстремума". Задание для самостоятельного выполнения: решить задачи 1-6 для функции

$$f(x_1, x_2, x_3) = x_1^2 + x_1 x_3 + x_2^2 + x_3^2 - 2x_1 + x_2 - x_3.$$

Ниже приводятся примеры выполнения задач.

Пример 1. Записать квадратичную функцию в матричном виде (1) и проверить выполнение условия (2).
а) $f(x_1, x_2) = x_1^2 + x_1 x_2 + x_2^2 - 2x_1 + x_2$.
б) $f(x_1, x_2) = x_1^2 + x_1 x_2 - x_1 + 2x_2$.
Решение. Для случая а):

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} -2 \\ 1 \end{pmatrix}.$$

Найдем собственные числа матрицы A, решив уравнение :

$$\det(A - \lambda E) = (2 - \lambda)^2 - 1 = 0$$

где E-единичная матрица, λ – скаляр. Собственные числа $\lambda_1 = 1, \ \lambda_2 = 3$, т.е. условие (2) выполнено и функция

$$f(x_1, x_2) = \frac{1}{2} x^T A x + x^T b$$

удовлетворяет условию соответствующей теоремы и данная функция имеет единственную точку минимума. В данном случае условие (2) для $f(\cdot)$ запишется так:

$$\|x\|^2 \le (x, Ax) \le 3\|x\|^2.$$

Для случая б):

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

Решив уравнение $\det(A - \lambda E) = 0$, получим $\lambda_{1,2} = \pm\sqrt{5}$, следовательно, функция:

$$f(x_1, x_2) = \frac{1}{2}x^T Ax + x^T b$$

не удовлетворяет условию (2).

Пример 2. Для случая а): $f(x_1, x_2) = x_1^2 + x_1x_2 + x_2^2 - 2x_1 + x_2$. По теореме 1 раздела 1.2 найти точку минимума $\bar{x}$ функции $f(x_1, x_2)$ и значение функции $f(\bar{x})$.
Решение. В соответствии с теоремой 1 решаем систему линейных алгебраических уравнений:

$$Ax + b = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} -2 \\ 1 \end{pmatrix} = 0$$

или же

$$\begin{cases} 2x_1 + x_2 &= 2 \\ x_1 + 2x_2 &= -1 \end{cases}$$

Отсюда: $x_1 = \frac{5}{3}, x_2 = -\frac{4}{3}$,

$$\bar{x} = \begin{pmatrix} \frac{5}{3} \\ -\frac{4}{3} \end{pmatrix}, \qquad f(\bar{x}) = -\frac{7}{3}.$$

Пример 3. Для функции $f(x) = x_1^2 + x_1x_2 + x_2^2 - 2x_1 + x_2$ найти первые три члена последовательности метода наискорейшего градиентного спуска.
Решение. Пусть

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} -2 \\ 1 \end{pmatrix}.$$

Первый член последовательности зададим в виде

$$x^1 = e^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad f(x^1) = -1.$$

Второй и третий члены последовательности получаем по методу наискорейшего градиентного спуска

$$q^1 = Ae^1 + b = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} + \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

$$Aq^1 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \qquad \mu_1 = -\frac{q^1 \cdot q^1}{q^1 \cdot Aq^1} = -\frac{1}{2},$$

$$x^2 = x^1 + \mu_1 q^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad f(x^2) = -2;$$

$$q^2 = Ax^2 + b = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} -2 \\ 1 \end{pmatrix} == \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

$$Aq^2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}, \qquad \mu_2 = -\frac{q^2 \cdot q^2}{q^2 \cdot Aq^2} = -\frac{1}{2},$$

$$x^3 = x^2 + \mu_2 q^2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ -1 \end{pmatrix}, \qquad f(x^3) = -\frac{9}{4}.$$

Пример 4 (метод Ричардсона). Пусть рассматривается стационарный двухшаговый метод спуска с матрицей

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

Её собственные числа суть $m = 1$, $M = 3$.

Полином

$$\Phi(\lambda) = 1 + c_1\lambda + c_2\lambda^2, \quad s = 2,$$

строится в соответствии с формулами (2) и (4) на стр. 38, в которых под a и b следует понимать m и M соответственно.

$$T_2(x) = 2x^2 - 1, \quad \bar{T}_2(x) = x^2 - \frac{1}{2}.$$

Пусть теперь выполнена следующая подстановка:

$$z = a\tau + b, \quad a = \frac{2}{M - m}, \quad b = -\frac{M + m}{M - m}.$$

В условиях нашей задачи

$$z = \tau - 2, \quad a = 1, \quad b = -2,$$

поэтому получаем такое представление для искомого полинома:

$$\bar{\Phi}(\tau) = \frac{T_2(\tau - 2)}{T_2(b)} = \frac{1}{7}(2\tau^2 - 8\tau + 7),$$

причем

$$\bar{\gamma}_2 = \frac{1}{|T_2(-2)|} < 1, \quad \bar{\gamma}_2 = \frac{1}{7}.$$

Подставляя в $\bar{\Phi}(\tau)$ вместо τ в качестве аргумента матрицу A, мы записываем итерационный процесс в виде

$$x^{k+1} = x^k - \frac{6}{7}q^k.$$

Первый член x^1 искомой последовательности и градиент q^1 мы зададим как и в задаче 3:

$$x^1 = e^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad q^1 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \qquad f(x^1) = -1.$$

Далее находим x^2:

$$x^2 = x^1 - \frac{6}{7}q^1 = \begin{pmatrix} 1 \\ -\frac{12}{7} \end{pmatrix}, \qquad f(x^2) = -1,49.$$

Нетрудно проверить, что

$$\|x^2 - \bar{x}\| < \gamma_2 \|x^1 - \bar{x}\|.$$

Пример 5. Для функции $f(x) = x_1^2 + x_1x_2 + x_2^2 - 2x_1 + x_2$, используя метод A-ортогонализации базиса, решить задачу минимизации функции $f(x)$.

Решение. Пусть

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad b = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$$

Пусть задан ортогональный базис

$$e^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad e^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Построим A-ортогональный базис

$$q^1 = e^1, \qquad q^2 = e^2 + \nu_{21}q^1,$$

где ν_{21} находится из условия

$$0 = q^1 \cdot Aq^2 = q^1 \cdot Ae^2 + \nu_{21}q^1 \cdot Aq^1.$$

Следовательно

$$\nu_{21} = -\frac{q^1 \cdot Ae^2}{q^1 \cdot Aq^1} = -\frac{1}{2},$$

$$q^2 = e^2 - \frac{1}{2}e^1 = \begin{pmatrix} -\frac{1}{2} \\ 1 \end{pmatrix}, \qquad Aq^2 = \begin{pmatrix} 0 \\ \frac{3}{2} \end{pmatrix},$$

а также

$$\tau_1 = q^1 \cdot Aq^1 = 2, \qquad \tau_2 = q^2Aq^2 = \frac{3}{2}.$$

Пусть теперь матрица Q составлена из столбцов q^1 , q^2:

$$Q = \begin{pmatrix} 1 & -\frac{1}{2} \\ 0 & 1 \end{pmatrix}.$$

Сделав замену переменной $x = Qy$, получим:

$$f(x) = f(Qy) = \sum_{i=1}^{2}(\frac{1}{2}\tau_i^2 y_i^2 + \sigma_i y_i),$$

где

$$\sigma_1 = b^T q^1 = -2, \qquad \sigma_2 = b^T q^2 = 2.$$

Используя условия минимума функции, как функции аргумента y, находим $\bar{y}$:

$$\bar{y}_1 = -\frac{\sigma_1}{\tau_1} = 1, \qquad \bar{y}_2 = -\frac{\sigma_2}{\tau_2} = -\frac{4}{3}.$$

Точка минимума $\bar{x}$ функции $f(x)$ находится как:

$$\bar{x} = Q\bar{y} = \begin{pmatrix} 1 & -\frac{1}{2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -\frac{4}{3} \end{pmatrix} = \begin{pmatrix} \frac{5}{3} \\ -\frac{4}{3} \end{pmatrix}.$$

Пример 6. Для функции $f(x) = x_1^2 + x_1x_2 + x_2^2 - 2x_1 + x_2$, используя метод сопряженных градиентов, решить задачу минимизации функции $f(x)$.

Решение. Идея метода состоит в параллельном проведении матрицы A квадратичной формы к диагональному виду и приближений к точке минимума. Одновременно строятся три последовательности векторов : последовательность x^k, вдоль которой функция $f(x^k)$ монотонно убывает, последовательность градиентов $q^k = Ax^k + b$, набор A-ортогональных векторов p^k.

Начальная точка x^1 может быть задана произвольно. Значения градиента: $q^1 = Ax^1 + b = p^1$. В качестве x^1 мы берем $x^1 = 0$,т.е. $q^1 = p^1 = b$. Следующая точка $x^2 = x^1 + \bar{\mu}_1 p^1$ определяется из условия минимального значения функции вдоль направления p^1:

$$\bar{\mu}_1 = -\frac{q^1 \cdot p^1}{p^1 \cdot Ap^1} = -\frac{5}{6}.$$

Таким образом

$$x^2 = \bar{\mu}_1 p_1 = \begin{pmatrix} \frac{5}{3} \\ -\frac{5}{6} \end{pmatrix}, \qquad q^2 = q^1 + \bar{\mu}_1 Ap^1 = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}.$$

Нетрудно видеть, что:

$$q^2 \cdot p^1 = 0.$$

Вектор p^2 находится с учетом условия $p^2 \cdot Ap^1 = 0$:

$$p^2 = q^2 + \sigma_1 p^1,$$

откуда

$$\sigma_1 = -\frac{q^2 \cdot Ap^1}{p^1 \cdot Ap^1} = \frac{1}{4},$$

и

$$p^2 = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix} + \frac{1}{4} \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{5}{4} \end{pmatrix}.$$

Далее находим

$$x^3 = x^2 + \bar{\mu}_2 p^2,$$

где

$$\bar{\mu}_2 = -\frac{q^2 \cdot p^2}{p^2 \cdot Ap^2} = -\frac{2}{5}.$$

Таким образом

$$x^3 = \begin{pmatrix} \frac{5}{3} \\ -\frac{5}{6} \end{pmatrix} - \frac{2}{5} \begin{pmatrix} 0 \\ -\frac{5}{4} \end{pmatrix} = \begin{pmatrix} \frac{5}{3} \\ -\frac{4}{3} \end{pmatrix}.$$

Искомая точка минимума $x^3 = \bar{x}$, так как согласно теореме число шагов, приводящих к результату, не превосходит размерности пространства плюс единица.

Глава 4. Интерполирование функций

1. Общая задача интерполирования

В вычислительной практике наиболее часто приходится иметь дело с функциями, заданными таблично: в $n+1$ узле $x_0, x_1, \ldots, x_n$ известны значения функции $f(x_0), f(x_1), \ldots, f(x_n)$. Если в плоскости взять декартову систему координат, то геометрически это будет означать задание в плоскости $n+1$ точки M_i с координатами $\{x_i, f(x_i)\}$ $(i = \overline{0,n})$. Будем считать, что функция $f(\cdot)$ рассматривается на некотором отрезке $[a,b]$. Если $f(\cdot)$ есть непрерывная функция, то такое множество будет некоторой линией над $[a,b]$. Она проходит через точки M_i, а в остальном эта линия произвольна и её ордината $f(\hat{x})$, $\hat{x} \neq x_i$, $i = \overline{0,n}$ может быть любой.

И тем не менее в этом случае строят функцию $\varphi(\cdot)$, совпадающую с $f(\cdot)$ в узлах $x_0, x_1, \ldots, x_n$ и простую в вычислительном плане. Задача построения такой функции называется *задачей интерполирования.*

Погрешность интерполирования $\varepsilon = f(x) - \varphi(x)$ зависит

- от исходных данных, в частности от числа узлов $n+1$;
- от расположения узлов x_i;
- от избранного правила интерполирования (выбора класса функций $\varphi(\cdot)$).

Всеми этими факторами – *параметрами интерполирования* – мы, как правило, может распоряжаться.

На практике целесообразно строить интерполяционные формулы, которые могут дать хорошие по точности результаты в некоторых достаточно широких классах функций $f(\cdot)$, содержащих в себе все практически важные функции с некоторыми общими для них свойствами.

Рассматривается класс F функций $f(\cdot)$, заданных на отрезке $[a,b]$. Для интерполирования $f(\cdot)$ выбирают семейство Φ функций $\varphi(\cdot)$:

- более простых, чем $f(\cdot)$;
- достаточно легко вычисляемых.

Затем среди функций $\{\varphi(\cdot)\}$ выбирают ту, которая имеет такие же исходные данные интерполирования, что и $f(\cdot)$:

$$\varphi(x_i) = f(x_i), \qquad i = \overline{0,n}.$$

После этого приближённо полагают

$$\varphi(x) \approx f(x), \qquad \forall x \in [a,b].$$

Рассмотрим линейное нормированное пространство $\Phi[a,b]$ и последовательность функций из этого пространства ω_i, $i = \overline{0,n}$, определенных на отрезке $[a,b]$ и линейно независимых там.

Определение 1. Функции бесконечной последовательности называются *линейно независимыми*, если взятые в любом конечном числе они оказываются линейно независимыми. ♣

Возьмём функции $\omega_i,\ i=\overline{0,n}$ и образуем линейную комбинацию (*обобщенный полином)*

$$s_n = a_0\omega_0 + a_1\omega_1 + \ldots + a_n\omega_n. \tag{1}$$

Здесь $\{a_i\}$ – произвольные постоянные . Они выбираются так, чтобы выполнялись условия $s_n(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$. Это значит, что параметры $\{a_i\}$ должны удовлетворять системе линейных алгебраических уравнений (СЛАУ):

$$a_0\omega_0(x_i) + a_1\omega_1(x_i) + \ldots + a_n\omega_n(x_i) = f(x_i), \qquad i=\overline{0,n}. \tag{2}$$

Решая её относительно $\{a_i\}$ и подставляя найденные значения в (2), получим *обобщенный полином* s_n, интерполирующий функцию $f(\cdot)$ по исходным данным.

Из сказанного выше вытекают естественные требования к функциям $\{\omega_i(\cdot)\}$, которые мы и рассмотрим.

Система (2) имеет единственное решение, если её определитель

$$D_{n+1} = D_{n+1}(x_0, x_1, \ldots, x_n) = \begin{vmatrix} \omega_0(x_0) & \omega_1(x_0) & \ldots & \omega_n(x_0) \\ \omega_0(x_1) & \omega_1(x_1) & \ldots & \omega_n(x_1) \\ \ldots & \ldots & \ldots & \ldots \\ \omega_0(x_n) & \omega_1(x_n) & \ldots & \omega_n(x_n) \end{vmatrix} \tag{3}$$

отличен от нуля:

$$D_{n+1}(x_0, x_1, \ldots, x_n) \neq 0. \tag{4}$$

Левая часть неравенства (4) зависит от $x_i,\ i=\overline{0,n}$ и функции $\omega_i\cdot$ принято выбирать так, чтобы оно выполнялось при любых $x_0, x_1, \ldots, x_n$, лежащих на $[a,b]$ и различных между собой. В таком случае *интерполирование функции* $f(\cdot)$ при помощи линейной комбинации s_n возможно и единственно при любом выборе узлов $x_i,\ i=\overline{0,n}$ на $[a,b]$, лишь бы они были не совпадающими.

Определение 2. Систему функций $\omega_i(\cdot)$, $i=\overline{0,n}$, для которой выполняется условие (4) при любых различных $x_i \in [a,b]$, называют *системой Чебышева* на интервале $[a,b]$. ♣

Дадим ещё одно, эквивалентное приведенному выше, определение чебышевской системы.

Определение 2^***.** Систему функций $\omega_i(\cdot)$, $i=\overline{0,n}$ называют *системой Чебышева* на $[a,b]$, если обобщенный полином (1) имеет на $[a,b]$ *не более* n корней при любом наборе коэффициентов $\{a_i\}_{i=0}^n$, $a_0^2 + a_1^2 + \ldots + a_n^2 \neq 0$. ♣

Пример 1. Положим $\omega_0(x) = 1,\ \omega_1(x) = x, \ldots,\ \omega_n(x) = x^{n-1}, \ldots$. Эта система функций является Чебышевской, так как определитель системы (4)

$$D_{n+1} = D_{n+1}(x_0, x_1, \ldots, x_n) = \begin{vmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & \ldots & x_n^n \end{vmatrix} \tag{5}$$

есть не что иное, как определитель Вандермонда, который отличен от нуля при всех различных между собой значениях $\{x_i\},\ i=\overline{0,n}$. Обобщенный многочлен s_n в данном случае

есть алгебраический многочлен степени n. Число исходных данных интерполирования при этом может быть произвольным. ♣

Пример 2. Пусть $\varphi_0(x) = 1, \varphi_1(x) = \cos x$, $[a, b] = [-1, 1]$. Рассмотрим обобщенный полином $s_1(x) = a_0 + a_1 \cos x$. Уравнение $s_2(x) = 0$ при $a_0 = \sqrt{3}$, $a_1 = -2$ имеет на интервале $[-1, 1]$ два корня $x_{1,2} = \pm\dfrac{\pi}{6}$ и в соответствии с определением 2^* не является чебышевской на нём. Если же взять интервал $[0, 1]$, то на нём та же система функций будет таковой, поскольку уравнение $s_2(x) = 0$ имеет на $[0, 1]$ не более одного решения при любом выборе коэффициентов $a_0,\ a_1 :\ a_0^2 + a_1^2 \neq 0$. ♣

Сформулируем теперь первое требование, накладываемое на систему функций $\omega_i(\cdot)$, $i = \overline{0, n}$: *при всех значениях* $n = 0, 1, \ldots$ *функции* $\omega_i(\cdot)$ $i = \overline{0, n}$ *должны составлять систему Чебышева на отрезке* $[a, b]$.

Второе условие связано с понятием полноты.

Определение 3. Семейство линейных комбинаций $\{s_k\}$ вида (1) называется *полным* в классе F функций $f(\cdot)$, если для любой функции $f(\cdot) \in F$ и любого $\varepsilon > 0$ существует такое N и такие коэффициенты $a_0, a_1, \ldots, a_N$, что для всех $x \in [a, b]$ выполняется неравенство $|f(x) - s_N(x)| < \varepsilon$. ♣

Если семейство $\{s_k\}$ не обладает этим свойством, то невозможно рассчитывать на то, чтобы с помощью комбинаций $\{s_k\}$ можно было выполнить сколь угодно точную аппроксимацию любой функции $f(\cdot) \in F$.

Поэтому второе условие сформулируем так: *система функций* $\omega_i,\ i = 1, 2, \ldots$ *должна быть такой, чтобы соответствующее ей семейство линейных комбинаций* $\{s_k\}$ *вида (1) было полным в классе* F *функций* $f(\cdot)$, *подлежащих интерполированию.*

Замечание. Необходимо отметить, что полнота еще не гарантирует возможность сколь угодно точной аппроксимации $f(\cdot)$. Это связано с тем, что полнота семейства $\{s_k\}$ в F даёт возможность сколь угодно точного приближения $f(\cdot)$ посредством s_n при некотором n в случае, когда на выбор s_n не налагается никаких ограничений.В задаче же интерполирования выбор s_n (при произвольном, но фиксированном n) определяется (ограничивается):

- выбором узлов x_i, $i = \overline{0, n}$
- необходимостью выполнения равенств $s_n(x_i) = f(x_i)$, $i = \overline{0, n}$.

Именно поэтому вопрос о возможности сколь угодно точного приближения $f(\cdot)$ при этих ограничениях в построении s_n подлежит исследованию для каждого конкретного случая. ♣

2. Алгебраическое интерполирование. Полином Лагранжа

Пусть на отрезке $[a, b] \subset R$ рассматривается функция $f(\cdot)$ и пусть известны её значения в $(n + 1)$ различных узлах $x_0, x_1, \ldots, x_n$, принадлежащих $[a, b]$.
Возьмём многочлен степени n

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n. \tag{1}$$

Напомним, что требования к системе функций $\omega_i = x^i$ в этом случае выполнены, так как :

- во-первых, она является Чебышевской;

- во-вторых, семейство алгебраических многочленов является полным в классе непрерывных на этом отрезке функций, ибо *по теореме Вейерштрасса если функция* $f(\cdot)$ *непрерывна на конечном замкнутом отрезке* $[a,b]$, *то для всякого* $\varepsilon > 0$ *существует многочлен* P_k *некоторой степени* k, *для которого при всех* $x \in [a,b]$ *выполняется* $|f(x) - P_k(x)| < \varepsilon$.

Коэффициенты a_i выбираем так, чтобы совпадали значения $f(\cdot)$ и $P_n(\cdot)$ в узлах интерполирования x_i:

$$P_n(x_i) = f(x_i), \qquad i = 0, 1, 2, \ldots, n. \tag{2}$$

Эти равенства дают квадратную систему линейных алгебраических уравнений относительно неизвестных $\{a_i\}$, откуда следует, что коэффициенты искомого обобщенного полинома $\{a_i\}$ линейно зависят от значений $f(x_k)$, $k = 0, 1, 2, \ldots, n$. Поэтому и многочлен $P_n(\cdot)$ линейно зависит от величин $f(x_k)$. Иначе говоря, он может быть представлен в виде:

$$P_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i). \tag{3}$$

Коэффициенты $l_i(x)$ можно найти, используя простые алгебраические соображения. Согласно с постановкой задачи в узлах интерполирования должны быть выполнены равенства (2):

$$f(x_k) = P_n(x_k) = \sum_{i=0}^{n} l_i(x_k) f(x_i), \quad k = 0, 1, 2, \ldots, n.$$

А это возможно, если

$$l_i(x_k) = \begin{cases} 1, & i = k, \\ 0 & i \neq k. \end{cases}$$

Таким образом, $l_k(x)$ – многочлен степени n, для которого все узлы x_i $(i = 0, 1, 2, \ldots, n; i \neq k)$ являются корнями. Количество корней – n, степень многочлена – n, а значит корни – простые. Это позволяет выписать разложение многочлена $l_k(x)$:

$$l_k(x) = A_k(x - x_0)\ldots(x - x_{k-1})(x - x_{k+1})\ldots(x - x_n), \tag{4}$$

Постоянный множитель A_k определяется из условия $l_k(x_k) = 1$, что даёт равенство

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

$$= \frac{\omega_{n+1}(x)}{(x - x_k)\omega'_{n+1}(x_k)}, \qquad \text{где} \qquad \omega_{n+1}(x) = (x - x_0)(x - x_1)\ldots(x - x_n). \tag{5}$$

Многочлены $l_k(x)$ называют *множителями Лагранжа*, а формулу

$$P_n(x) = L_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i) = \sum_{i=0}^{n} \frac{\omega_{n+1}(x)}{(x - x_i)\omega'_{n+1}(x_i)} f(x_i), \tag{6}$$

формулой Лагранжа для интерполирующего многочлена $P_n(x)$. Часто интерполяционный полином $L_n(\cdot)$ называют просто *полиномом Лагранжа*.

2.1. Погрешность метода. Остаточный член формулы Лагранжа

Оценку для $r_n(x) = f(x) - L_n(x)$ будем искать на классе функций $f(\cdot) \in C^{n+1}[a,b]$ в точке $x = x^* \in [a,b]$. Для этого рассмотрим вспомогательную функцию

$$\Delta(x) = r_n(x) - K \cdot \omega_{n+1}(x) \tag{1}$$

и потребуем, чтобы $\Delta(x^*) = 0$.

Это возможно, если

$$K = \frac{r_n(x^*)}{\omega_{n+1}(x^*)}$$

Функция

$$\Delta(x) = r_n(x) - \frac{r_n(x^*)}{\omega_{n+1}(x^*)}\omega_{n+1}(x) \tag{2}$$

имеет корни $x_0, x_1, \ldots, x_n, x^*$. Их общее количество равно $n+2$.

По теореме Ролля производная $\Delta'(x)$ функции $\Delta(x)$ обращается в нуль по крайней мере в $(n+1)$-ой точке.

Применяя теорему Ролля к производной $\Delta'(x)$, получаем, что $\Delta''(x)$, обращается в нуль по крайней мере в n точках.

Рассуждая таким образом, придём к тому, что $\Delta^{(n+1)}(x)$ имеет на $[a,b]$ по крайней мере один корень. Пусть это будет точка $\xi \in [a,b]$.

$$\Delta^{(n+1)}(\xi) = f^{(n+1)}(\xi) - L_n^{(n+1)}(\xi) - K \cdot (n+1)! = 0. \tag{3}$$

Тогда справедливо равенство

$$f^{(n+1)}(\xi) - \frac{r_n(x^*)}{\omega_{n+1}(x^*)} \cdot (n+1)! = 0, \tag{4}$$

что позволяет представить остаточный член в виде:

$$r_n(x^*) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot \omega_{n+1}(x^*). \tag{5}$$

Таким образом справедливо равенство:

$$r_n(x) = f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot \omega_{n+1}(x), \quad \xi \in (a,b). \tag{6}$$

В дальнейшем будем его называть формулой Лагранжа для остатка интерполирования или *методической погрешностью интерполирования.*

2.2. Выбор узлов интерполирования

Рассмотрим множество Φ_n всевозможных функций $f(\cdot)$, которые $(n+1)$ раз непрерывно дифференцируемы на $[a,b]$ и производная которых порядка $(n+1)$ ограничена по модулю числом M_{n+1}: $|f^{(n+1)(x)}| \le M_{n+1}$ $(x \in [a,b])$. В этом классе функций остаток интерполирования (методическая погрешность интерполирования) имеет оценку:

$$|r_n(x)| \le \frac{M_{n+1}}{(n+1)!}|x - x_0||x - x_1| \ldots |x - x_n|. \tag{1}$$

Она является *точной* и достигается в том случае, когда $f(\cdot)$ есть многочлен степени $n+1$ вида

$$f(x) = \frac{M_{n+1}}{(n+1)!}x^{n+1} + a_1x^n + a_2x^{n-1} + \dots .$$

Задача 1. *Функция $f(\cdot)$ задана таблично, $\hat{x}$ - не табличное значение аргумента, в котором необходимо приблизить функцию $f(\cdot)$ при помощи интерполяционного многочлена степени n, взяв за узлы любые табличные значения аргумента $x_{i_0}, x_{i_1}, \dots, x_{i_n}$.*

Множитель $\dfrac{M}{(n+1)!}$ не зависит от выбора узлов. Поэтому при фиксированном значении $\hat{x}$ необходимо выбрать $\{x_{i_k}\}$ так, чтобы произведение

$$|\hat{x} - x_{i_0}||\hat{x} - x_{i_1}| \dots |\hat{x} - x_{i_n}|$$

имело *наименьшее значение.*

Очевидно, что для этого набор $\{x_{i_k}\}_{k=0}^n$ следует выбирать из условий:

$$x_{i_0} = \min_i |\hat{x} - x_i|; \quad x_{i_1} = \min_{i \neq i_0} |\hat{x} - x_i|; \dots, \quad x_{i_n} = \min_{i \neq i_0, i_1, \dots, i_{n-1}} |\hat{x} - x_i|. \quad \clubsuit$$

Задача 2. *Выбрать узлы $x_0, x_1, \dots, x_n$ на $[a,b]$ так, чтобы правая часть оценки (1) принимала минимальное значение.*

Если считать $[a,b] = [-1,1]$, то учитывая свойства полиномов Чебышева можно утверждать,что $\omega_{n+1}(x) = \bar{T}_{n+1}(x) = \dfrac{1}{2^n}T_{n+1}(x)$. В общем же случае $[a,b] \neq [-1,1]$ достаточно сделать преобразование $[a,b] \to [-1,1]$ и получим искомые узлы $\{x_i\}$ в виде:

$$x_i = \frac{1}{2}\big[(b-a)\cos\frac{(2i+1)}{2(n+1)}\pi + (b+a)\big], \quad i = \overline{0,n}.$$

При таком выборе узлов оценка (1) для методической погрешности принимает вид:

$$\big|f(x) - L_n(x)\big| \leq \frac{M_{n+1}(b-a)^{n+1}}{2^{2n+1}(n+1)!}. \quad \clubsuit$$

2.3. О сходимости интерполяционного процесса

Решив задачу интерполирования по заданному числу узлов на интервале $[a,b]$, мы должны ответить на вопрос: как ведёт себя последовательность интерполяционных полиномов при неограниченном возрастании числа узлов на $[a,b]$? Будет ли (и если будет, то при каких условиях) иметь место свойство:

$$r_n(x_*, f) = f(x_*) - L_n(x_*, f) \to 0 \quad \text{при} \quad n \to \infty \ ?$$

Поскольку ответ на поставленный вопрос зависит в том числе и от свойств функции $f(\cdot)$, обозначаем здесь полином Лагранжа через $L_n(x,f)$, чтобы подчеркнуть указанное обстоятельство.

Уточним задачу. Рассмотрим бесконечную треугольную таблицу узлов, расположенных на $[a,b]$:

$$X = \left\{ \begin{array}{cccc} x_1^1 & & & \\ x_1^2 & x_2^2 & & \\ \dots & \dots & \dots & \dots \\ x_1^n & x_2^n & \dots & x_n^n \\ \dots & \dots & \dots & \dots \end{array} \right\}. \tag{1}$$

Определение 1. Построение интерполяционных полиномов $L_n(x,f)$ по таблице (1) для функции $f(\cdot)$ будем называть *интерполяционным процессом* (для экономии места – ИП). ♣

Определение 2. Если для $\hat{x} \in [a,b]$ имеет место: $r_n(\hat{x},f) \to 0$ при $n \to \infty$, то говорят, что ИП для $f(\cdot)$ по таблице (1) сходится в точке $\hat{x}$. ♣

Определение 3. Если сходимость ИП имеет место для всех $x \in [a,b]$, то говорят, что ИП сходится для $f(\cdot)$ на $[a,b]$. ♣

Определение 4. Если $r_n(x,f) \to 0$ при $n \to \infty$ *равномерно* на $[a,b]$, то говорят о *равномерной* сходимости ИП к $f(\cdot)$ на $[a,b]$. ♣

Некоторые факты.

- Для любой непрерывной функции $f(\cdot)$ можно выбрать узлы (1) так, что ИП будет равномерно на $[a,b]$ сходиться к этой функции. В основе утверждения – теорема Вейерштрасса о приближении функций полиномами.
- (т. Фабера) Для любой таблицы (1) существует непрерывная функция, для которой ИП не является равномерно сходящимся.
- Для целой функции ИП по любой таблице (1) равномерно на $[a,b]$ сходится к ней.

Пример Бернштейна. Для $f(x) = |x|, \quad x \in [-1,1]$ ИП по равноотстоящим узлам не сходится ни в одной точке, кроме точек $-1, 0, 1$ (при этом -1 и $+1$ считаются узлами интерполирования). ♣

Рассмотрим несколько примеров построения интерполяционного полинома и оценки методической погрешности.

Пример 1. С какой методической погрешностью можно найти $\sin\dfrac{\pi}{4}$, имея таблицу:

x	0	$\pi/6$	$\pi/2$
f(x)	0	1/2	1

Установим связь между параметрами задачи и параметрами, использованными при рассмотрении вопроса о построении интерполяционного полинома Лагранжа:

$$f(x) = \sin x,\ n = 2,\ a = 0,\ b = \pi/2,\ M_3 = 1.$$

$$\begin{aligned} \left|r_n\Big(\frac{\pi}{4}\Big)\right| &= \left|\sin\frac{\pi}{4} - L_2\Big(\frac{\pi}{4}\Big)\right| \le \frac{1}{3!}\left|\Big(\frac{\pi}{4}-0\Big)\Big(\frac{\pi}{4}-\frac{\pi}{6}\Big)\Big(\frac{\pi}{4}-\frac{\pi}{2}\Big)\right| = \\ &= \frac{1}{6}\Big(\frac{\pi}{4}\Big)^2 \cdot \frac{\pi}{12} \approx \frac{1}{6}\cdot\frac{10}{16}\cdot\frac{1}{4} \approx \frac{1}{37}, \end{aligned}$$

т.е. относительная погрешность вычисления указанного значения с помощью полинома Лагранжа составляет $\approx 4\%$. ♣

Пример 2. Построить полином Лагранжа по таблице:

x	0	2	3
f(x)	1	3	2

$$L_2(x) = \frac{(x-2)(x-3)}{(-2)(-3)} \cdot 1 + \frac{x(x-3)}{2(-1)} \cdot 3 + \frac{x(x-2)}{3 \cdot 1} \cdot 2 = -\frac{2}{3}x^2 + \frac{7}{3}x + 1. \quad \clubsuit$$

Пример 3. Построить полином Лагранжа по таблице:

x	0	2	3	5
f(x)	1	3	2	5

$$L_3(x) = \frac{(x-2)(x-3)(x-5)}{(-2)(-3)(-5)} \cdot 1 + \frac{x(x-3)(x-5)}{2(-1)(-3)} \cdot 3 + \frac{x(x-2)(x-5)}{3 \cdot 1 \cdot (-2)} \cdot 2 + \frac{x(x-2)(x-3)}{5 \cdot 3 \cdot 2} \cdot 5 = \frac{3}{10}x^3 - \frac{13}{6}x^2 + \frac{62}{15}x + 1. \quad \clubsuit$$

Примеры 2 и 3 показывают, что при добавлении узла в таблицу приходится заново проделывать всю работу. Этот недостаток может быть устранен при записи интерполяционного полинома в иной *форме*.

2.4. Разностные отношения (разделённые разности)

Разностные отношения (разделенные разности – РР) применяются при изучении функций, заданных на неравномерной сетке.

Полагаем, что для любых, но различных между собой значениях $x_0, x_1, \ldots$, даны значения функции $f(x_0), f(x_1), \ldots$.

Определение.

- Разностными отношениями *первого порядка* называются величины

$$f(x_0, x_1) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}, \quad f(x_1, x_2) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}, \ldots .$$

- Разностными отношениями *второго порядка* называются величины

$$f(x_0, x_1, x_2) = \frac{f(x_1, x_2) - f(x_0, x_1)}{x_2 - x_0}, \quad f(x_1, x_2, x_3) = \frac{f(x_2, x_3) - f(x_1, x_2)}{x_3 - x_1}, \ldots .$$

- Разностные отношения *любого порядка* $i+1, \quad i = 1, 2, \ldots$ определяются при помощи разностных отношений порядка i по формуле

$$f(x_0, x_1, \ldots, x_i, x_{i+1}) = \frac{f(x_1, x_2, \ldots, x_{i+1}) - f(x_0, x_1, \ldots, x_i)}{x_{i+1} - x_0}. \quad \clubsuit$$

Можно выписать простые выражения разностных отношений всех порядков через значения функции.

Так, для разностного отношения первого порядка справедливо равенство:

$$f(x_0, x_1) = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0},$$

а для разностного отношения второго порядка:

$$\begin{aligned} f(x_0,x_1,x_2) &= \frac{1}{x_2-x_0}\{f(x_1,x_2)-f(x_0,x_1)\} = \\ &= \frac{f(x_0)}{(x_0-x_1)(x_0-x_2)}+\frac{f(x_1)}{(x_1-x_0)(x_1-x_2)}+\frac{f(x_2)}{(x_2-x_0)(x_2-x_1)}. \end{aligned}$$

Выполнив индукцию, можно показать, что при всяком n:

$$\begin{aligned} f(x_0,x_1,\ldots,x_n) &= \sum_{i=0}^{n}\frac{f(x_i)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)} = \\ &= \sum_{i=0}^{n}\frac{f(x_i)}{\omega'_{n+1}(x_i)}, \qquad \omega_{n+1}(x)=\prod_{i=0}^{n}(x-x_i). \end{aligned} \tag{1}$$

Полученное соотношение назовём *основным свойством РР*. Можно показать, что при всяком $n=1,2,\ldots$ справедливо равенство:

$$\begin{aligned} f(x_n) &= f(x_0)+(x_n-x_0)f(x_0,x_1)+(x_n-x_0)(x_n-x_1)f(x_0,x_1,x_2)+\ldots \\ &+ (x_n-x_0)\ldots(x_n-x_{n-1})f(x_0,x_1,\ldots,x_n), \end{aligned}$$

устанавливающее связь между значениями функции $f(x_n), f(x_0)$ и разностными отношениями.

2.5. Свойства разделённых разностей

- линейность по $f(\cdot)$, т.е. если $f(\cdot)=\alpha\,\varphi(\cdot)+\beta\psi(\cdot)$, то

$$f(x_0,x_1,\ldots,x_n)=\alpha\,\varphi(x_0,x_1,\ldots,x_n)+\beta\psi(x_0,x_1,\ldots,x_n);$$

- симметричность относительно узлов $\{x_i\}$, т.е.

$$f(x_{i_0},x_{i_1},\ldots,x_{i_k})=f(x_{j_0},x_{j_1},\ldots,x_{j_k})$$

при условии, что набор узлов в левой и правой частях формулы один и тот же. ♣

Оба свойства очевидны, если обратиться к представлению разделённых разностей по формуле (1). Учитывая второе свойство РР, можно всегда перечислять список аргументов РР в порядке их нумерации. Кроме того, примем соглашение (удобное при записи формул): отождествлять РР *нулевого порядка* с самой функцией, т.е. поскольку узлы считаются *различными* и упорядоченными в порядке нумерации, то $f(x_i,x_{i+1},\ldots,x_{i+k})=f(x_i)$ при $k=0$.

2.6. Интерполяционный полином Ньютона

Получим форму записи интерполяционного полинома, удобную при увеличении числа узлов.

Пусть $x_0,x_1,\ldots,x_n$ – узлы, расположенные на интервале $[a,b]$. Пусть по узлам $x_0,x_1,\ldots,x_{n-1}$ (и значениям $\{y_i\}$ функции $f(\cdot)$ в этих точках) построен полином Лагранжа $L_{n-1}(\cdot)$, а по узлам $x_0,x_1,\ldots,x_n$ – полином Лагранжа $L_n(\cdot)$. Рассмотрим их разность:

$$\Delta L_n(x)=L_n(x)-L_{n-1}(x)=A_n(x-x_0)(x-x_1)\ldots(x-x_{n-1}), \tag{1}$$

поскольку узлы $x_0, x_1, \ldots, x_{n-1}$ являются *корнями* полинома $\Delta L_n(x)$.
Но A_n, как это следует из (1), есть старший коэффициент полинома $L_n(x)$, т.е.

$$A_n = \sum_{i=0}^{n} \frac{y_i}{\omega'_{n+1}(x_i)} = f(x_0, x_1, \ldots, x_n). \tag{2}$$

Здесь использована формула (1) предыдущего раздела.

Теперь полином Лагранжа можно представить в виде:

$$\begin{aligned} L_n(x) &= L_0(x) + \big(L_1(x) - L_0(x)\big) + \big(L_2(x) - L_1(x)\big) + \ldots \big(L_n(x) - L_{n-1}(x)\big) = \\ &= L_0(x) + \sum_{i=1}^{n} \Delta L_i(x) = \sum_{i=0}^{n} f(x_0, x_1, \ldots, x_i)\omega_i(x) \equiv P_n(x). \end{aligned} \tag{3}$$

Здесь использованы соглашения о РР предыдущего раздела, дополненные следующими относительно ω_i:

$$\omega_0(x) = 1, \quad \omega_1(x) = x - x_0, \quad \omega_i = \prod_{k=0}^{i} (x - x_k).$$

Полученное представление интерполяционного полинома (3) носит название *полином Ньютона.*

Замечание. Вычисление РР удобно производить в таблице:

порядок РР:	0	1	2	3
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	$f(x_2)$		$f(x_1, x_2, x_3)$	
		$f(x_2, x_3)$		
x_3	$f(x_3)$			

Пример 4. Построим интерполяционный полином по исходным данным примера 3 (см. стр.54), однако используем результат примера 2, (см. стр.53), т.е. используя те же обозначения, запишем, что $L_3(x) = L_2(x) + f(x_0, x_1, x_2, x_3)\omega_3(x)$, где $\omega_3(x) = (x - x_0)(x - x_1)(x - x_2)$. Остаётся вычислить РР заданной таблично в примере 3 функции по узлам $0, 2, 3, 5$:

порядок РР:	0	1	2	3
$x_0 = 0$	1			
		1		
$x_1 = 2$	3		-2/3	
		-1		3/10
$x_2 = 3$	2		5/6	
		3/2		
$x_3 = 5$	5			

Теперь имеем:

$$L_3(x) = L_2(x) + \frac{3}{10} x(x-2)(x-3)$$

Раскрывая скобки, подставляя $L_2(x)$ из примера 2, получим тот же результат:

$$L_3(x) = -\frac{2}{3}x^2 + \frac{7}{3}x + 1 + \frac{3}{10}x(x^2 - 5x + 6) = \frac{3}{10}x^3 - \frac{13}{6}x^2 + \frac{62}{15}x + 1. \quad \clubsuit$$

2.7. Методическая погрешность полинома Ньютона

Получим выражение методической погрешности для полинома Ньютона.
По определению методической погрешности рассмотрим $r_n(x_*) = f(x_*) - P_n(x_*)$. Добавим к имеющимся узлам $x_0, x_1, \ldots, x_n$, по которым построен полином Ньютона, узел $x_* = x_{n+1}$ и решим задачу интерполирования по узлам $x_0, x_1, \ldots, x_n, x_{n+1}$, обозначив полученный полином $P_{n+1}(x)$. Для него имеем:

$$P_{n+1}(x_*) = f(x_*) = P_n(x_*) + f(x_0, x_1, \ldots, x_n, x_{n+1})\omega_{n+1}(x_*). \tag{1}$$

Откуда получаем:

$$r_n(x_*) = f(x_*) - P_{n+1}(x_*) = f(x_0, x_1, \ldots, x_n, x_{n+1})\omega_{n+1}(x_*). \tag{2}$$

Формула (2) и есть искомое представление.
Отметим, что РР $f(x_0, x_1, \ldots, x_n, x_{n+1})$ *не может* быть вычислена, т.к. для её вычисления необходимо иметь $f(x_*)$.

Побочным результатом формулы (2) является

Теорема. *Если узлы $x_0, x_1, \ldots, x_n$ принадлежат отрезку $[a, b]$ и $f(\cdot)$ имеет там непрерывную производную порядка n, то на (a, b) существует такая точка ξ, что для неё верно равенство*:

$$f(x_0, x_1, \ldots, x_n) = \frac{1}{n!} f^{(n)}(\xi). \quad \clubsuit \tag{3}$$

Утверждение теоремы непосредственно следует из сравнения полученной формулы (2) и выражения для методической погрешности полинома Лагранжа – *поскольку полиномы Лагранжа и Ньютона представляют собой лишь различную по форме запись интерполяционного полинома, то и методические погрешности для них совпадают.*

Следствие. Если $f(x)$ является полиномом степени k, то РР для такой функции порядка k по любой системе узлов есть константа, зависящая лишь от функции, а РР более высоких порядков равны нулю. $\clubsuit$

2.8. Конечные разности. Интерполирование по равноотстоящим узлам

Конечные разности (КР) являются рабочим аппаратом при изучении функций, заданных таблицей значений в равноотстоящих узлах. Пусть в равноотстоящих узлах $x_k = x_0 + kh$, $(k = 0, 1, 2, \ldots)$ известны соответствующие им значения функции: $y_k = f(x_0 + kh)$.

Определение.

- Конечными разностями *первого порядка* называются величины

$$\Delta y_k = y_{k+1} - y_k, \quad k = 0, 1, 2, \ldots$$

- Конечными разностями *второго порядка* называются величины

$$\Delta^2 y_k = \Delta y_{k+1} - \Delta y_k, \quad k = 0, 1, 2, \ldots$$

- Конечными разностями $(i+1)$-*го порядка* называются величины

$$\Delta^{i+1} y_k = \Delta^i y_{k+1} - \Delta^i y_k, \quad k = 0, 1, 2, \ldots$$

Нетрудно выразить конечные разности любого порядка через значения функции:

$$\Delta y_0 = y_1 - y_0,$$

$$\Delta^2 y_0 = \Delta y_1 - \Delta y_0 = (y_2 - y_1) - (y_1 - y_0) = y_2 - 2y_1 + y_0.$$

Продолжая вычисления, получим

$$\Delta^n y_0 = y_n - \frac{n}{1!} y_{n-1} + \frac{n(n-1)}{2!} y_{n-2} + \ldots + (-1)^n y_0. \tag{1}$$

Существенно упростить представление конечных разностей помогает введение оператора E, при действии которого на $f(x)$ её аргумент увеличивается на h:

$$Ef(x) = f(x+h).$$

При таких обозначениях разность первого порядка примет вид:

$$\Delta y_0 = Ey_0 - y_0 = (E - I)y_0.$$

Здесь I – тождественный оператор: $If(x) = f(x)$. Из определения оператора E следует, что $E^m f(x) = f(x + mh)$, а значит конечная разность второго порядка предстанет в виде:

$$\Delta^2 y_0 = E^2 y_0 - 2Ey_0 + y_0 = (E - I)^2 y_0.$$

И равенство (1) перепишется в краткой условной форме:

$$\Delta^n y_0 = (E - I)^n y_0. \tag{2}$$

Столь же просто получить выражения для значения функции y_n через начальное значение y_0 и значения конечных разностей $\Delta^k y_0$, $(k = 0, 1, \ldots, n)$, относящихся к начальной точке x_0.

Действительно:

$$\Delta y_0 = y_1 - y_0, \quad \text{откуда } y_1 = y_0 + \Delta y_0,$$

$$y_2 = y_1 + \Delta y_1 = (y_0 + \Delta y_0) + (\Delta y_0 + \Delta^2 y_0) = y_0 + 2\Delta y_0 + \Delta^2 y_0.$$

Продолжая вычисления, получим:

$$y_n = y_0 + \frac{n}{1!}\Delta y_0 + \frac{n(n-1)}{2!}\Delta^2 y_0 + \ldots + \Delta^n y_0 = (I + \Delta)^n y_0 = E^n y_0. \tag{3}$$

Для $n = 1$ очевидно, а для $n > 1$ легко доказывается по индукции соотношение между КР и РР:

$$f(x_0, x_1, \ldots, x_n) = \frac{\Delta^n f(x_0)}{n! h^n}, \quad h = x_{i+1} - x_i = const \tag{4}$$

Пусть узлы интерполирования $x_0, x_1, \ldots, x_n$ являются равнооотстоящими, т.е. $x_i = x_0 + ih$, $i = \overline{0,n}$ и требуется оценить значение $f(x)$, $x \in (x_0, x_1)$. Введём безразмерную переменную $t = \dfrac{x - x_0}{h}$ и подставим в (4):

$$\begin{aligned} P_n(x) &= f(x_0) + \sum_{i=1}^{n} (x - x_0)(x - x_0 - h) \ldots \big(x - x_0 - (i-1)h\big) \frac{\Delta^i f(x_0)}{i\,!\,h^i} = \\ &= f(x_0) + \sum_{i=1}^{n} \frac{t(t-1)\ldots(t-i+1)}{i\,!} \Delta^i f(x_0), \quad t \in (0,1). \end{aligned} \tag{5}$$

Напомним, что комбинаторная формула для *сочетаний* имеет вид:

$$C_n^k = \frac{n\,!}{k\,!\,(n-k)\,!} = \frac{n(n-1)\ldots(n-k+1)}{k\,!}.$$

Используя эту формулу и для нецелых n, формулу (5) перепишем в компактном виде:

$$P_n(x) = \sum_{i=0}^{n} C_t^i \Delta^i f(x_0), \quad t \in (0,1). \tag{6}$$

В частности, если $t = k$, то с учетом формулы (3) получаем $P(x_k) = y_k$.
Оценим коэффициенты в формуле (6) при $t \in (0,1)$:

$$\begin{aligned} &|C_t^1| < 1, \quad |C_t^2| < \frac{|t(t-1)|}{2} < \frac{1}{8}, \\ &|C_t^i| < \frac{\big|1 \cdot 1 \cdot 2 \cdot \ldots |i-1|\big|}{i\,!} < \frac{1}{i}, \quad i > 2. \end{aligned} \tag{7}$$

Из полученной оценки следует, что вклад слагаемых в сумме (6) в значение полинома уменьшается вместе с ростом номера слагаемого. К тому же сами КР, как правило, быстро убывают с ростом порядка КР. Полученное представление (6) носит название *интерполяционного полинома Ньютона для интерполяции вперед*. Аналогично могут быть получены полиномы Ньютона для интерполирования назад, если перенумеровать узлы в обратном порядке.

Рассмотрим полином Лагранжа для случая равноотстоящих узлов.
Сделаем замену переменной:

$$x = x_0 + th, \quad t = \frac{x - x_0}{h}, \quad t \in [0, n].$$

Тогда

$$L_n(x) = L_n(x_0 + th) \equiv \bar{L}_n(t) = \sum_{i=0}^{n} \frac{\bar{\omega}_{n+1}(t)}{(t-i)\bar{\omega}'_{n+1}(t_i)} y_i = \sum_{i=0}^{n} \bar{l}_i(t) y_i,$$

где

$$\bar{\omega}_{n+1}(t) = t(t-1)\ldots(t-n), \quad \bar{l}_i(t) = \frac{\bar{\omega}_{n+1}(t)}{(t-i)\bar{\omega}'_{n+1}(t_i)}$$

Отметим, что полиномы $\bar{l}_i(t) \quad i = 1, 2, \ldots, n$ не зависят от узлов, интервала и значений функции, а потому могут быть составлены независимо от параметров интерполирования.

2.9. Кратное интерполирование

Пусть в узлах $x_0, x_1, \ldots, x_n$, расположенных на $[a, b]$ заданы значения функции $f(\cdot)$ вместе с производными до некоторого порядка:

$$\left.\begin{array}{cccc} f(x_0), & f'(x_0), & \ldots & f^{(m_0-1)}(x_0) \\ f(x_1), & f'(x_1), & \ldots & f^{(m_1-1)}(x_1) \\ \ldots & \ldots & \ldots & \ldots \\ f(x_n), & f'(x_n), & \ldots & f^{(m_n-1)}(x_n) \end{array}\right\} \tag{1}$$

Рассмотрим задачу построения алгебраического полинома $P_m(\cdot)$, обладающего свойствами:

$$\begin{cases} P_m^{(k)}(x_j) = f^{(k)}(x_j) \\ j = \overline{0,n}; \quad k = \overline{0, m_{j-1}}. \end{cases} \tag{2}$$

Поставленная задача называется*задачей кратного интерполирования*. Ответим сначала на вопрос: какова должна быть степень искомого полинома $P_m(\cdot)$ для того, чтобы можно было надеяться на её разрешимость?

Поскольку $P_m(\cdot)$ имеет $(m+1)$ коэффициент, а СЛАУ (2) содержит $\sum\limits_{j=0}^{n} m_j$ уравнений относительно этих коэффициентов, то будем считать, что

$$m + 1 = \sum_{j=0}^{n} m_j. \tag{3}$$

Единственность интерполяционного полинома

Пусть равенствам (2) удовлетворяют *два* полинома $\bar{P}_m(\cdot)$ и $\tilde{P}_m(\cdot)$, *степень которых определена условием (3).*
Рассмотрим их разность:

$$\Delta P(\cdot) = \bar{P}_m(\cdot) - \tilde{P}_m(\cdot).$$

Обозначив $deg\, P(\cdot)$ степень полинома P, можем записать:

$$deg\, \Delta P(\cdot) \le m = \sum_{j=0}^{n} m_j \ - 1 \tag{4}$$

Но узел x_i для $\Delta P(\cdot)$ является корнем кратности не ниже m_j. Следовательно

$$\begin{cases} \Delta P(x) = q_s(x) \prod\limits_{i=0}^{n} (x - x_i)^{m_i} \\ \text{где } q_s - \text{полином: } s \ge 0, \quad s = deg\, q_s(x). \end{cases} \tag{5}$$

Но тогда получаем:

$$deg\, \Delta P(\cdot) = s + \sum_{j=0}^{n} m_j \ge \sum_{j=0}^{n} m_j,$$

что противоречит соотношению (4). Единственность искомого полинома *при условии его существования* доказана. ♣

Существование интерполяционного полинома

Для определения коэффициентов полинома $P_m(\cdot)$ мы имеем СЛАУ (2).
Рассмотрим её однородный аналог, т.е. положим в (2) $y_j^k = f^{(k)}(x_j) = 0$ для всех k, j. Тогда, очевидно, СЛАУ (2) *имеет решение* вида $a_s = 0, \quad s = \overline{0,m}$, а тогда по доказанному в п. а) это решение единственно. Поскольку определитель системы (2) не зависит от заданных значений функции y_j^k, то СЛАУ (2) имеет единственное решение для любой правой части.
Интерполяционный полином, решающий задачу кратного интерполирования, называется *полиномом Эрмита*. Обозначим его $H_m(\cdot)$. ♣

Методическая погрешность полинома Эрмита

Пусть $R_m(x) = f(x) - H_m(x)$. Будем считать, что $f(\cdot) \in C^{m+1}[a,b]$, узлы $\{x_i\}$ расположены на $[a,b]$. Пусть $\omega_{m+1}(x) = \prod\limits_{j=0}^{n}(x - x_j)^{m_j}$ – узловой полином. Оценку для $R_m(x)$ произведём в точке $x_* \in [a,b],\ x_* \neq x_j$.
Введём вспомогательную функцию $\varphi(x) = R_m(x) - k\omega_{m+1}(x)$. Постоянную k определим из условия: $\varphi(x_*) = 0$, т.е. $k = \dfrac{R_m(x_*)}{\omega_{m+1}(x_*)}$.
Таким образом, функция $\varphi(\cdot)$ имеет на $[a,b]$ следующие корни:

$$\begin{array}{l} x_k \text{ – корень кратности не ниже } m_k, \quad k = 0,1,2,\dots\ n; \\ x_* \text{ – корень кратности не ниже } 1. \end{array}$$

Итого на $[a,b]$ с учетом кратности $\varphi(\cdot)$ имеет $\sum\limits_{j=0}^{n} m_j \ + 1 = m + 2$ корня. Тогда

- $\varphi'(\cdot)$ имеет по теореме Ролля $(n+1)$ корень (хотя бы по одному между любой парой соседних корней функции $\varphi(\cdot)$);
- корень функции $\varphi(\cdot)$ кратности $m_j > 1$ остаётся корнем функции $\varphi'(\cdot)$ кратности $(m_j - 1)$.

Итого с учетом кратностей корней функция $\varphi'(\cdot)$ имеет на $[a,b]$ корней в количестве

$$(n+1) + \sum_{j=0}^{n}(m_j - 1) = \sum_{j=0}^{n} m_j = m + 1,$$

т.е. у производной общее количество корней на единицу меньше, чем у функции.
Продолжая рассуждения, придём к выводу о существовании на $[a,b]$ точки ξ такой, что $\varphi^{(m+1)}(\xi) = 0$. Отсюда получаем представление для методической погрешности интерполирования с помощью полинома Эрмита:

$$R_m(x_*) = \frac{f^{(m+1)}(\xi)}{(m+1)!}\omega_{m+1}(x_*). \quad ♣$$

2.10. Один численный метод построения полинома Эрмита

Пусть построен интерполяционный полином $P_{n,0}(\cdot)$ (Ньютона или Лагранжа) по узлам $x_i, \quad i=\overline{0,n}$, удовлетворяющий условиям:

$$P_{n,0}(x_i)=y_i^0=f(x_i), \quad i=\overline{0,n}.$$

Искомый полином $H_m(\cdot)$ представим в виде

$$H_m(x)=P_{n,0}(x)+\omega_{n+1,0}(x)P_{m-n-1}(x), \quad \omega_{n+1,0}(x)=\prod_{j=0}^{n}(x-x_j). \tag{1}$$

Очевидно, что полином (1) удовлетворяет условиям $H_m(x_i)=y_i^0$, поэтому остаётся выбрать полином $P_{m-n-1}(x)$ так, чтобы были выполнены остальные соотношения (2):

$$\begin{cases} P_m^{(k)}(x_j)=f^{(k)}(x_j) \\ j=\overline{0,n}; \quad k=\overline{1,m_{j-1}}. \end{cases} \tag{2}$$

Подставив (1) в (2), получим:

$$\begin{cases} y_j^k=P_{n,0}^{(k)}(x_j)+\sum\limits_{s=0}^{k}C_k^s\omega_{n+1,0}^{(s)}(x_j)P_{m-n-1}^{(k-s)}(x_j) \\ j=\overline{0,n}; \quad k=\overline{1,m_{j-1}}. \end{cases} \tag{3}$$

Соотношения (3) являются СЛАУ относительно коэффициентов полинома $P_{m-n-1}(x)$. Рассмотрим эти соотношения, отвечающие случаю $k=0,\ j=\overline{0,n}$:

$$y_j^1=P'_{n,0}(x_j)+\omega'_{n+1,0}(x_j)P_{m-n-1}(x_j)+\omega_{n+1,0}(x_j)P'_{m-n-1}(x_j). \tag{4}$$

Поскольку $\omega_{n+1,0}(x_j)=0,\ j=\overline{0,n}$, то из (4) находим $P_{m-n-1}(x_j),\ j=\overline{0,n}$:

$$P_{m-n-1}(x_j)=\frac{y_j^1-P'_{n,0}(x_j)}{\omega'_{n+1,0}(x_j)} \quad j=\overline{0,n}\,.$$

Далее, из (3) при $k=2$ найдём:

$$\begin{aligned} y_j^2=P''_{n,0}(x_j) \quad &+ \quad \omega_{n+1,0}(x_j)P''_{m-n-1}(x_j)+2\omega'_{n+1,0}(x_j)P'_{m-n-1}(x_j)+ \\ &+ \quad \omega''_{n+1,0}(x_j)P_{m-n-1}(x_j),\ j=\overline{0,n}, \end{aligned} \tag{5}$$

откуда определяются $\{P'_{m-n-1}(x_j),\ j=\overline{0,n}\}$ для тех узлов, в которых заданы значения $\{y_j^2\}$:

$$P'_{m-n-1}(x_j)=\frac{y_j^2-P''_{n,0}(x_j)-\omega''_{n+1,0}(x_j)P_{m-n-1}(x_j)}{2\omega'_{n+1,0}(x_j)}$$

.

Понятно, что через $\nu=\max\{m_j\}$ шагов получим задачу кратного интерполирования для $P_{m-n-1}(x)$, в которой условий на $(n+1)$ меньше, чем в исходной задаче.

Применив указанную процедуру к полиному $P_{m-n-1}(x)$, редуцируем полученную задачу к новой того же класса, но содержащую ещё меньше условий в таблице.

Наконец, через конечное число повторений описанной процедуры придём к простой задаче интерполирования, т.е. такой, которую решает полином Лагранжа. Обратным ходом найдём искомый полином Эрмита. ♣

Пример. Построить полином Эрмита по таблице:

x	-1	0	1
y	0	1	2
y'	5	0	
y''	-20		

Представим искомый полином $H_5(x)$ в виде: $H_5(x) = (x+1) + x(x^2-1)P_2(x)$; далее в соответствии с приведенными выше формулами получаем: $P_2(-1) = \dfrac{5-1}{2} = 2$; $P_2(0) = \dfrac{0-1}{-1} = 1$; $P_2'(-1) = \dfrac{-20+12}{4} = -2$.

Таким образом, для $P_2(x)$ получаем такую задачу кратного интерполирования:

x	-1	0
y	2	1
y'	-2	

Вновь $P_2(x)$ запишем в виде: $P_2(x) = (1-x) + x(x+1)P_0(x)$; далее как и выше получаем: $P_0(-1) = \dfrac{-2+1}{-1} = 1$. Остаётся подставить найденные полиномы в представление для $H_5(x)$:

$H_5(x) = (x+1) + x(x^2-1)\{(1-x) + x(x+1)\} = 1 + x^5$. ♣

3. Численное дифференцирование

3.1. Постановка задачи

Пусть функция $f(\cdot) \in C^m[a,b]$ представлена таблицей значений $\{y_i\}$ по узлам $\{x_i\}$, $i = \overline{0,n}$. Требуется оценить значение производной $f^{(k)}(\hat{x}) = \hat{y}^k$, где $\hat{x} \in [a,b]$, $k \le m$.

Основная идея численного дифференцирования заключается в замене функции $f(\cdot)$ её интерполяционным полиномом $Q_n(\cdot)$ (*в частности*, алгебраическим) и вычислении его производной требуемого порядка в нужной точке, полагая

$$f^{(k)}(\hat{x}) \approx Q_n^{(k)}(\hat{x}).$$

Здесь естественно считать, что $k \le n$, поскольку в противном случае, очевидно, $Q_n^{(k)}(x) \equiv 0$ и приближение производной указанным приемом для более высоких порядков теряет всякий смысл.

Если $r_n(x) = f(x) - Q_n(x)$, то $r_n^{(k)}(x) = f^{(k)}(x) - Q_n^{(k)}(x)$, т.е. методическая погрешность при таком подходе суть $r_n^{(k)}(x)$. При замене функции $f(\cdot)$ её интерполяционным полиномом $Q_n(\cdot)$ предполагают, что методическая погрешность $r_n(x)$ такой замены мала, из чего вовсе не следует малость $r_n^{(k)}(x)$. Практика показывает, что при таком способе вычисления производных получается сравнительно бóльшая погрешность, особенно при вычислении производных высоких порядков.

3.2. Формулы численного дифференцирования

Рассмотрим применение данного подхода при использовании в качестве $Q_n(\cdot)$ полиномов Лагранжа и Ньютона.

- При использовании в качестве интерполяционного полиномома $Q_n(\cdot)$ интерполяционного полиномома Ньютона

$$P_n(x) = \sum_{i=0}^{n} f(x_0, x_1, \ldots, x_i)\omega_i(x) \tag{1}$$

получаем следующие формулы для производных:

$$f^{(k)}(\hat{x}) \approx \sum_{i=0}^{n} f(x_0, x_1, \ldots, x_i)\omega_i^{(k)}(\hat{x}) \tag{2}$$

- При использовании полинома Лагранжа имеем:

$$f^{(k)}(\hat{x}) \approx \sum_{i=0}^{n} \frac{y_i}{\omega'_{n+1}(x_i)} \frac{d^k}{dx^k} \left[\frac{\omega_{n+1}(x)}{x - x_i} \right]_{x=\hat{x}} \tag{3}$$

Первая формула обладает достоинствами полинома Ньютона, т.е. легко преобразуется при увеличении числа узлов, вторая же формула требует выполнения всей работы заново.

Пример. Построим формулы для приближенного вычисления производных первого и второго порядков. Заметим, что $f^{(n)}(x) = n! f(x_0, x_1, \ldots, x_n)$. Поэтому

$$f'(x) \approx P'_1(x) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

$$f''(x) \approx P'_2(x) = 2\Big[\frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)} \Big].$$

В случае равноотстоящих узлов $x_{i+1} - x_i = h = const > 0$ последнее выражение упрощается:

$$f''(x) \approx P'_2(x) = \frac{\Delta^2 f(x_0)}{h^2} = \frac{1}{h^2} \big(f(x_0 + 2h) - 2f(x_0 + h) + f(x_0) \big). \qquad \clubsuit$$

Пусть состав узлов фиксирован и требуется оценивать производные разных порядков в *одной и той же* точке $\hat{x}$. Тогда в формуле (3) при y_i стоят некоторые коэффициенты c_i, $i = \overline{0, n}$, *не зависящие* от $f(\cdot)$, т.е.

$$f^{(k)}(\hat{x}) \approx \sum_{i=0}^{n} c_i f(x_i), \quad c_i = c_i(x_0, x_1, \ldots, x_n, \hat{x}). \tag{4}$$

Рассмотрим задачу о вычислении этих коэффициентов без обращения к их представлению по формулам (3).

3.3. Метод неопределенных коэффициентов

Идея метода состоит в том, чтобы выбрать $\{c_i\}$ в формуле (3) так, чтобы эта формула была *точной* на некотором множестве функций. Таким классом может служить класс полиномов $\mathcal{P}_n$, степени не выше n. В самом деле, если $f(\cdot) = Q_m(\cdot) \in \mathcal{P}_n$, то $L_n^{(k)}(\cdot) = Q_n^{(k)}(\cdot)$ *(по теореме единственности!)*.

Поэтому подставляя в формулы (4) $f(x) = x^s$, $s = \overline{0, n}$ и учитывая, что в этом случае имеет место *точное равенство*, получаем:

$$\begin{aligned}
s = 0: &\quad \sum_{i=0}^{n} c_i = 0 \\
s = 1: &\quad \sum_{i=0}^{n} c_i x_i = 0 \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
s = k: &\quad \sum_{i=0}^{n} c_i x_i^k = k! \\
&\cdots\cdots\cdots\cdots\cdots\cdots \\
s = n: &\quad \sum_{i=0}^{n} c_i x_i^n = \frac{n!}{(n-k)!} \hat{x}^{n-k}
\end{aligned} \tag{1}$$

Определитель системы есть определитель Вандермонда, а потому СЛАУ (1) имеет единственное решение.

3.4. Методическая погрешность формул численного дифференцирования

Пусть $\{x_0, x_1, \ldots, x_n\} \in [a,b]$ – узлы интерполирования, причем $x_0 < x_1 < \ldots < x_n$, $f(\cdot) \in C^{n+1}[a,b]$ – интерполируемая функция, $P_n(\cdot)$ – интерполяционный полином и $R_n(x) = f(x) - P_n(x)$ – погрешность интерполирования. Тогда $R_n^{(k)}(x) = f^{(k)}(x) - P_n^{(k)}(x)$ – погрешность численного дифференцирования. Получим представление для последней величины.

Поскольку $R_n(\cdot) \in C^{n+1}[a,b]$ и $R_n(x_i) = 0, \quad i = \overline{0,n}$, то можно k раз применить теорему Ролля, $k \le n$. При этом расположение нулей производных погрешности $R_n(x)$ можно охарактеризовать с помощью следующей таблицы:

$R_n(x)$	$R_n^{(1)}$	$R_n^{(2)}$	$\ldots$	$R_n^{(k)}$
x_0				
x_1	(x_0, x_1)			
x_2	(x_1, x_2)	(x_0, x_2)		
$\vdots$	$\vdots$	$\vdots$		
x_k	(x_{k-1}, x_k)	(x_{k-2}, x_k)	$\ldots$	(x_0, x_k)
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
x_n	(x_{n-1}, x_n)	(x_{n-2}, x_n)	$\ldots$	(x_{n-k}, x_n)

В k-том столбце таблицы указаны открытые интервалы (x_i, x_{i+k}), в каждом из которых в силу теоремы Ролля должен лежать по крайней мере один из корней ξ_i производной $R_n^{(k)}(x)$. Ясно, что точки ξ_i зависят от функции $f(\cdot)$, узлов интерполирования $\{x_i\}$, но не зависят от x.

Введем в рассмотрение функцию

$$F(x) = R_n^{(k)}(z) - \alpha \prod_{i=0}^{n-k} (z - \xi_i)$$

и заметим, что $F(\xi_i) = 0$ для $i = \overline{0, n-k}$. Для произвольного фиксированного $x \neq \xi_i$, $i = \overline{0, n-k}$ выберем $\alpha = \alpha(x)$ так, чтобы $F(x) = 0$. Тогда $F(\cdot)$ имеет $(n-k+2)$ различных корней и можно опять воспользоваться теоремой Ролля $(n-k+1)$ раз, поскольку $F(\cdot)$ имеет $(n-k+1)$-ю непрерывную производную. Это приводит к заключению, что $F^{(n-k+1)(z)}$ имеет нуль $\eta(x)$ на отрезке, содержащем точки $x, \xi_0, \ldots \xi_{n-k}$. Вычисляя указанную производную, получим:

$$0 = F^{(n-k+1)}(\eta) = R^{(k+1)}(\eta) - \alpha(n-k+1)! = f^{(n+1)}(\eta) - \alpha(n-k+1)!,$$

откуда $\alpha = \dfrac{f^{(n+1)}(\eta)}{(n-k+1)!}$. Подставляя найденное значение α в равенство $F(x) = 0$, приходим к формуле

$$R^{(k)}(x) = \frac{f^{(n+1)}(\eta(x))}{(n-k+1)!} \prod_{i=0}^{n-k} (x - \xi_i). \tag{1}$$

Нетрудно видеть, что эта формула остается верной и для $x = \xi_i$, причем η можно считать любой точкой из $[a,b]$.

3.5. Анализ полной погрешности формул численного дифференцирования

Рассмотрим простейший случай – вычисление производной первого порядка таблично заданной функции и полную погрешность (без учёта погрешностей округления) подхода, изложенного в предыдущем пункте. Будем считать, кроме того, что таблица содержит только два узла и вычисляется первая производная функции. Эти предположения значительно упрощают рассмотрение вопроса, позволяя сделать, тем не менее, принципиальные выводы.

Введём следующие обозначения:

- $r^F_{nk}(\hat{x})$ – полная погрешность вычисления производной k-го порядка по n узлам в точке $\hat{x}$;
- $r^M_{nk}(\hat{x})$ – методическая и
- $r^N_{nk}(\hat{x})$ – неустранимая погрешности.

Как известно, имеет место оценка:

$$|r^F_{nk}(\hat{x})| \le |r^M_{nk}(\hat{x})| + |r^N_{nk}(\hat{x})|. \tag{1}$$

Итак, пусть

$$n = 1,\ k = 1,\ x_1 = x_0 + h,\ y_0 = f(x_0),\ y_1 = f(x_1).$$

Тогда в этом частном случае

$$\omega_2(x) = (x - x_0)(x - x_0 - h), \quad P_1(x) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0),$$

и в соответствии с интерполяционным подходом, имеем:

$$f'(\hat{x}) \approx \frac{y_1 - y_0}{x_1 - x_0}.$$

Выпишем методическую погрешность по формуле (1), которая в данном случае $n = 1,\ k = 1$ имеет вид:

$$r^M_{11}(\hat{x}) = f''(\xi)(x - \xi_0) \tag{2}$$

Учтём неустранимую погрешность: пусть табличные значения содержат погрешности ε_0 и ε_1: $\tilde{y}_i = y_i + \varepsilon_i,\ i = 1, 2$. Тогда

$$r^N_{11}(x_0) = \frac{\tilde{y}_1 - \tilde{y}_0}{h} - \frac{y_1 - y_0}{h} = \frac{\varepsilon_1 - \varepsilon_0}{h}. \tag{3}$$

Подставляя полученные представления методической и неустранимой погрешностей в (1), получаем:

$$|r^F_{11}(x_0)| \le M_2 h + \frac{2\varepsilon}{h}, \quad M_2 = \sup_{[a,b]} |f''(x)|, \quad \varepsilon = \max\{|\varepsilon_0|, |\varepsilon_1|\}. \tag{4}$$

Обозначим через $\sigma(h)$ правую часть оценки (4):

$$\sigma(h) = M_2 h + \frac{2\varepsilon}{h}. \tag{5}$$

Очевидно, что $\sigma(h) \to \infty$ при $h \to 0$ и $\sigma(h) \to 0$ при $h \to \infty$. Следовательно, на интервале $(0, \infty)$ существует минимум функции $\sigma(h)$ и достигается он, как легко видеть, в точке $\bar{h} = \sqrt{2\varepsilon M_2^{-1}}$. Отсюда следует вывод: *для минимизации полной погрешности формул численного дифференцирования необходимо согласовать шаг сетки с уровнем неустранимой погрешности и свойствами функции* (выраженными в данном случае в константе M_2).

4. Обратное интерполирование

Пусть для таблично заданной на $[a,b]$ функции $f(\cdot)$ поставлена задача определения точки $\hat{x}$ такой, что в этой точке $f(\hat{x}) = \hat{y}$, где $\hat{y}$ – заданное значение. Будем считать, что $f(\cdot) \in C^{n+1}[a,b]$ и $\hat{y} \in [c,d]$, где $c = \min\{y_i\}$, $d = \max\{y_i\}$, $i = \overline{0,n}$. При сделанных предположениях относительно $f(\cdot)$ можно гарантировать существование искомой точки $\hat{x}$.

Рассмотрим два возможных случая:

- пусть $f(\cdot)$ монотонна на $[a,b]$. В этом случае для $f(\cdot)$ существует обратная функция $F(\cdot):\ F\big(f(x)\big) \equiv x,\ x \in [a,b]$. В этом случае числа $\{y_i\},\ i = \overline{0,n}$ могут быть приняты за узлы интерполирования для функции $F(\cdot)$, а $\{x_i\},\ i = \overline{0,n}$ – в качестве её значений в указанных узлах. Пусть $L_n(\cdot)$ – полином Лагранжа для $F(\cdot)$; тогда в качестве приближения к точке $\hat{x}$ примем $\tilde{x} = L_n(\hat{y})$, а методическая погрешность может быть представлена в виде:

$$|\hat{x} - \tilde{x}| = |F(\hat{y}) - L_n(\hat{y})| = \frac{F^{(n+1)}(\eta)}{(n+1)!}\omega_{n+1}(\hat{y}).$$

- Если же $f(\cdot)$ не является монотонной на $[a,b]$, то можно, заменив $f(\cdot)$ на её интерполяционный полином $L_n(\cdot)$, решить алгебраическое уравнение

$$L_n(x) = \hat{y}.$$

Пусть $\tilde{x}$ – вещественное решение этого уравнения (*которое существует ввиду сделанного предположения* $\hat{y} \in [c,d]$): $L_n(\tilde{x}) = f(\hat{x}) = \hat{y}$. Оценим методическую погрешность такого решения задачи, считая для определённости, что $\tilde{x} \le \hat{x}$.

С одной стороны по теореме Лагранжа

$$f(\tilde{x}) - f(\hat{x}) = (\tilde{x} - \hat{x})f'(\eta),\ \eta \in [\tilde{x},\hat{x}]. \tag{1}$$

С другой стороны имеем :

$$f(\tilde{x}) - f(\hat{x}) = f(\tilde{x}) - L_n(\tilde{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(\tilde{x}), \quad \xi \in [a,b]. \tag{2}$$

Сравнивая (1) и (2) заключаем, что

$$\tilde{x} - \hat{x} = \frac{f^{(n+1)}(\xi)}{f'(\eta)(n+1)!}\omega_{n+1}(\tilde{x}) \quad \text{при } f'(\eta) \neq 0. \tag{3}$$

Пусть известны оценки: $|f'(\eta)| \ge m_1 \neq 0$ на $[\tilde{x},\hat{x}]$ и $f^{(n+1)}(\xi) \le M_{n+1}$ на $[a,b]$. Тогда из (3) получаем:

$$|\tilde{x} - \hat{x}| \le \frac{M_{n+1}}{m_1(n+1)!}|\omega_{n+1}(\tilde{x})|. \tag{4}$$

5. Интерполирование функций многих переменных

5.1. Постановка и особенности задачи

Пусть $f(\cdot)$ – скалярная функция векторного аргумента, определённая в некоторой области $D \subset R^k$ и пусть $\{x^i\},\ i = \overline{0,n}$ набор векторов из D, которые мы будем называть *узлами*.

Задача. В узлах известны значения функции. Требуется построить *алгебраический* полином k переменных минимальной степени, который в заданных узлах совпадает с заданными значениями функции. ♣

Дальнейшие рассмотрения с целью снижения громоздкости записей проведём в предположении $k = 2$.

Итак, пусть $(x_i, y_i),\ i = \overline{0,n}$ узлы, в которых заданы значения функции $z_i = f(x_i, y_i),\ i = \overline{0,n}$. Искомый полином запишем в виде:

$$P_m(x,y) = \sum_{i+j=0}^{m} a_{ij} x^i y^j. \tag{1}$$

Для определения коэффициентов полинома $P_m(x,y)$ имеем СЛАУ:

$$P_m(x_i, y_i) = z_i. \tag{2}$$

Рассмотрим особенности этой задачи при $n = 2$. СЛАУ (2) примет вид:

$$\begin{aligned} a_{00} + a_{10}x_0 + a_{01}y_0 &= z_0 \\ a_{00} + a_{10}x_1 + a_{01}y_1 &= z_1 \\ a_{00} + a_{10}x_2 + a_{01}y_2 &= z_2. \end{aligned}$$

Для разрешимости этой системы при *любой* правой части необходимо потребовать, чтобы её определитель был отличен от нуля:

$$\begin{vmatrix} 1 & x_0 & y_0 \\ 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \end{vmatrix} \neq 0.$$

Последнее же означает, что указанные узлы $(x_i, y_i),\ i = 1,2,3$ *не должны лежать на одной прямой*. Таким образом, возникло первое принципиальное различие задач интерполирования функции многих переменных и функции одной переменной: *узлы в задаче интерполирования функции многих переменных не могут быть расположены произвольно.*
Так при $n = 6$ приходится рассматривать вопрос о принадлежности этих узлов одной кривой второго порядка, при $n = 10$ – третьего порядка и т.д., что довольно трудоемко. Поэтому обычно данную задачу интерполирования решают при специальном расположении узлов, относительно которых заранее известно, что СЛАУ (2) разрешима относительно коэффициентов искомого полинома.

Далее, СЛАУ (2) в общем случае является системой с

$$1 + 2 + \cdots + (m+1) = \frac{(m+1)(m+2)}{2}$$

неизвестными, а уравнений в системе $(n+1)$, поэтому не для любого числа узлов n можно обеспечить равенство

$$\frac{(m+1)(m+2)}{2} = n + 1,$$

а это означает, что *даже если решение задачи существует, оно не является, вообще говоря, единственным.*
Третья особенность рассматриваемой задачи состоит в том, что *нет аналога теоремы Ролля, что затрудняет оценку методической погрешности интерполирования.*

5.2. Линейная интерполяция

Пусть в пространстве R^n заданы узлы $\{x^i\}_{i=0}^n$ и в них известны значения функции $f(x^i) = f_i$. Рассмотрим задачу построения полинома первой степени n переменных, совпадающего в этих узлах с заданными значениями функции, т.е.

$$P_1(x) = \alpha_0 + a^T x, \quad P_1(x^i) = f_i, \tag{1}$$

т.е. имеем СЛАУ

$$\alpha_0 + a^T x = f_i, \quad i = \overline{0,n}. \tag{2}$$

Вычитая из последних n уравнений первое, рассмотрим СЛАУ

$$a^T(x^i - x^0) = f_i - f_0, \quad i = \overline{1,n}. \tag{3}$$

Обозначив $q^i = x^i - x^0$, сформулируем теорему:

Теорема. *Для того чтобы задача линейного интерполирования была однозначно разрешима для любой функции $f(\cdot)$, необходимо и достаточно, чтобы* $\det\{q^1, q^2 \ldots, q^n\} \neq 0$ ♣.

5.3. Квадратичное интерполирование

Рассмотрим задачу построения квадратичной функции

$$P_2(x) = \alpha + 2a^T x + x^T A x, \quad A^T = A = \{a_{ij}\}_{i,j=1}^n, \tag{1}$$

которая совпадала бы со значениями заданной функции $f(\cdot)$ в

$$1 + n + \frac{(n+1)n}{2} = \frac{(n+1)(n+2)}{2}$$

узлах (именно столько параметров имеет функция (1)).

В качестве узлов возьмём векторы

$$\begin{cases} x^i = x^0 + h_1 e^i \\ \\ x^{n+i} = x^0 + h_2 e^i, \quad i = \overline{1,n}, \quad h_1 \neq h_2. \end{cases} \tag{2}$$

Здесь $e^i = (\underbrace{0, \ldots, 0, 1}_{i}, \ldots, 0)^T$. Записав функцию (1) в виде

$$P_2(x) = P_2(x^0) + 2(Ax^0 + a)^T(x - x^0) + (x - x^0)^T A(x - x^0),$$

подставим сюда узлы (2) и учтём, что

$$P_2(x^k) = f(x^k) = f_k, \quad k = \overline{1,2n}$$

Получим следующие соотношения:

$$\left\{ \begin{array}{lcl} f_i & = & f_0 + 2h_1 w_i + h_1^2 a_{ii}, \\ \\ f_{n+i} & = & f_0 + 2h_2 w_i + h_2^2 a_{ii}, \quad i = \overline{1,n}. \end{array} \right. \tag{3}$$

Здесь

$$w_i = (Ax^0 + a)^T e^i. \tag{4}$$

Определитель системы (3), очевидно, отличен от нуля при $h_1 \neq h_2$, т.к.

$$\Delta = \left| \begin{array}{cc} 2h_1 & h_1^2 \\ 2h_2 & h_2^2 \end{array} \right| = 2h_1 h_2 (h_2 - h_1).$$

Следовательно, из (3) однозначно определяются величины $\{a_{ii}, w_i\},\ i = \overline{1,n}$. Далее рассмотрим систему узлов-векторов

$$x^{ij} = x^0 + h_3 e^i + h_4 e^j, \quad i \neq j,\ h_1 h_2 \neq 0,$$

в которых *заданы* значения функции f_{ij}. Условие совпадения значений заданной и квадратичной функций в узлах даёт соотношения:

$$f_{ij} = f_0 + 2h_3 w_i + 2h_4 w_j + h_3^2 a_{ii} + h_4^2 a_{jj} + 2h_3 h_4 a_{ij},$$

откуда находятся внедиагональные элементы $\{a_{ij}\},\ i \neq j$ матрицы A, а из соотношений (4) находим компоненты вектора $a = \{a_i\},\ i = \overline{1,n}$:

$$a_i = w_i - Ax^0 \cdot e^i, \quad i = \overline{1,n}.$$

Остаётся используя, например, значение f_0 выразить α:

$$\alpha = f_0 - 2a^T x^0 + x^0 \cdot Ax^0.$$

Пример 1. (квадратичное интерполирование).

Построим ИП $P_2(x) = \alpha + 2a^T x + x^T Ax$, $a = (a_1,\ a_2)^T,\ x = (u, v)^T$, $A = \{a_{i,j}\},\ i, j = 1, 2$ для функции $f(x) = u^2 v$. Поскольку искомый полином имеет 6 параметров, то для их определения следует взять 6 узлов.

В качестве узлов интерполирования берем

$$\left\{ \begin{array}{ll} x^i = x^0 + h_1 e^i, & x^0 = (0,0)^T,\ h_1 = 1,\ e^1 = (1,0)^T,\ e^2 = (0,1)^T, \\ x^{n+i} = x^0 + h_2 e^i, & n = 2,\ h_2 = 2. \end{array} \right.$$

В данном случае $x^1 = (1,0), \quad x^2 = (0,1)^T, \quad x^3 = (2,0)^T, \quad x^4 = (0,2)^T$. Нетрудно убедиться, что $f_i = f(x^i) = 0$. Система уравнений для вспомогательных параметров $\{w_i\}$ и диагональных элементов матрицы A имеет вид:

$$\left\{ \begin{array}{l} 0 = 2w_i + a_{ii}, \\ 0 = 4w_i + 4a_{ii}, \quad i = 1, 2, \end{array} \right.$$

откуда получаем $w_i = 0, \quad a_{ii} = 0,\ i = 1, 2$.

В качестве последнего шестого узла возьмем $x^{12} = x^0 + h_3 e^1 + h_4 e^2$ при $h_1 = 1,\ h_2 = 2$, т.е. $x^{12} = (1, 2)^T$. Вычислив $f_{12} = f(x^{12} = 2)$, находим $a_{12} = a_{21}$:

$$f_{12} = 2 = 2 \cdot 2 \cdot a_{12}, \quad a_{12} = \frac{1}{2}.$$

Компоненты вектора $a = (a_1, a_2)^T$ находим из $w_i = (Ax^0 + a)^T e^i$: $a_1 = 0,\ a_2 = 0$. Для α получаем: $\alpha = f_0 - 2a^T - x^0 \cdot Ax^0 = 0$. Тем самым искомый ИП имеет вид: $P_2(x) = uv$. ♣

5.4. Интерполирование функции двух переменных по прямоугольной таблице

Пусть задана прямоугольная таблица значений функции $z = f(x, y)$:

x \ y	y_0	y_1	$\dots$	y_m
x_0	z_{00}	z_{01}	$\dots$	z_{0m}
x_1	z_{10}	z_{11}	$\dots$	z_{1m}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_n	z_{n0}	z_{n1}	$\dots$	z_{nm}

Зафиксируем столбец, соответствующий столбцу с ординатой y_s, и построим интерполяционный полином по узлам $x_0, x_1, \dots, x_n$ и указанному столбцу:

$$P_n(x, y_s) = \sum_{i=0}^{n} a_i^s x^i, \quad s = \overline{0, m}. \tag{1}$$

Пусть $a_0(y), a_1(y), \dots, a_n(y)$ – полиномы, принимающие значения

$$\{a_0^s\}_{s=0}^m, \quad \{a_1^s\}_{s=0}^m, \quad \dots, \quad \{a_n^s\}_{s=0}^m,$$

в узлах $y_0, y_1, \dots, y_m$ соответственно. Представим их в виде

$$a_i(y) = \sum_{j=0}^{m} b_{ij} y^j.$$

Очевидно, что полином $P_{n+m}(\cdot, \cdot)$ степени $(n + m)$

$$P_{n+m}(x, y) = \sum_{i=0}^{n} a_i(y) x^i = \sum_{i=0}^{n} \sum_{j=0}^{m} b_{ij} x^i y^j$$

удовлетворяет условиям задачи.

Замечание. Построенный полином не является, вообще говоря, полиномом минимальной степени, решающим рассматриваемую задачу, и единственным полиномом степени $(n+m)$. ♣

Изучим вопрос о представлении методической погрешности в рассмотренной задаче. Введём обозначения:

$$\begin{aligned} f(\overline{x_0, x_i}; \overline{y_0, y_j}) &= f(x_0, x_1, \dots, x_i; y_0, y_1, \dots, y_j), \\ f(\overline{x_0, x}; \overline{y_0, y}) &= f(x_0, x_1, \dots, x_n, x; y_0, y_1, \dots, y_m, y). \end{aligned}$$

Записывая $f(\cdot, \cdot)$ через полином Ньютона по переменной x с остаточным членом, имеем:

$$f(x, y) = \sum_{i=0}^{n} f(\overline{x_0, x_i}; y)\, \omega_i(x) + f(\overline{x_0, x}; y)\, \omega_{n+1}(x). \tag{2}$$

Со слагаемым, стоящим под знаком суммы, поступим аналогично:

$$f(\overline{x_0, x_i}; y) = \sum_{j=0}^{m} f(\overline{x_0, x_i}; \overline{y_0, y_j})\, \omega_j(y) + f(\overline{x_0, x_i}; \overline{y_0, y})\, \omega_{m+1}(y). \tag{3}$$

Подставив (3) в (2), получим:

$$\begin{aligned} f(x,y) &= \sum_{i=0}^{n}\sum_{j=0}^{m} f(\overline{x_0, x_i}; \overline{y_0, y_j})\, \omega_i(x)\, \omega_j(y) + \\ &+ \left[\sum_{i=0}^{n} f(\overline{x_0, x_i}; \overline{y_0, y})\, \omega_i(x)\right] \omega_{m+1}(y) + f(\overline{x_0, x}; y)\, \omega_{n+1}(x). \end{aligned} \tag{4}$$

Сумма, стоящая в скобках, участвует в формуле

$$f(x; \overline{y_0, y}) = \sum_{i=0}^{n} f(\overline{x_0, x_i}; \overline{y_0, y})\, \omega_i(x) + f(\overline{x_0, x}; \overline{y_0, y})\, \omega_{n+1}(x). \tag{5}$$

Выражая её отсюда и подставляя в предыдущую формулу, получаем:

$$\begin{aligned} f(x,y) &= P_{m+n}(x,y) + \{f(x; \overline{y_0, y})\, \omega_{m+1}(y) + \\ &+ f(\overline{x_0, x}; y)\, \omega_{n+1}(x) - f(\overline{x_0, x}; \overline{y_0, y})\, \omega_{n+1}(x)\, \omega_{m+1}(y)\}. \end{aligned} \tag{6}$$

Здесь $P_{m+n}(x,y)$ – интерполяционный полином, построенный по прямоугольной таблице, а выражение в фигурных скобках как раз и даёт представление методической погрешности в рассматриваемой задаче. Используя связь разделенной разности с производной, окончательно имеем:

$$\begin{aligned} r_{nm}(x,y) &= \frac{\omega_{n+1}(x)}{(n+1)!} \cdot \frac{\partial^{n+1} f(\xi, y)}{\partial x^{n+1}} + \frac{\omega_{m+1}(y)}{(m+1)!} \cdot \frac{\partial^{m+1} f(x, \eta)}{\partial y^{m+1}} - \\ &- \frac{\omega_{n+1}(x)\, \omega_{m+1}(y)}{(n+1)!\,(m+1)!} \cdot \frac{\partial^{m+n+2} f(\mu, \nu)}{\partial x^{n+1} \partial y^{m+1}}. \end{aligned}$$

Здесь ξ, η, μ, ν – из интервалов $[x_0, x_n]$ и $[y_0, y_m]$.

Пример 2. Пусть задана прямоугольная таблица значений функции $z = f(x,y)$:

y \ x	-1	0	1
-1	4	3	6
0	2	0	2
1	4	3	6

Интерполяционные полиномы для столбцов, как легко видеть, имеют вид:

$$l_1(y) = 2y^2 + 2; \quad l_2(y) = 3y^2; \quad l_3(y) = 4y^2 + 2.$$

Теперь построим ИП $q_i(x),\ i = 0,1,2$ по таблицам, в которых собраны коэффициенты полиномов $l_i(y),\ i = 0,1,2$ при $y^2,\ y,\ y^0$ соответственно:

x	-1	0	1
q_0	2	0	2

, | x | -1 | 0 | 1 |
|---|---|---|---|
| q_1 | 0 | 0 | 0 |

, | x | -1 | 0 | 1 |
|---|---|---|---|
| q_2 | 2 | 3 | 4 |

.

Решения этих задач легко выписываются: $q_0(x) = 2x^2,\ q_1(x) = 0,\ q_2(x) = x + 3$, после чего строится искомый интерполяционный полином $P(x,y)$:

$$P(x,y) = q_2(x)y^2 + q_1(x)y + q_0 = 2x^2 + 3y^2 + xy^2. \quad \clubsuit$$

6. Интерполирование с помощью сплайнов

6.1. Понятие сплайна

Пусть на $[a,b] \in R$ задана сетка $\Delta_N = \{x_0, x_1, \ldots, x_N\}$, причём $x_i < x_{i+1}$ и $x_0 = a$, $x_N = b$. Обозначим через $\mathcal{P}_n$ множество полиномов с действительными коэффициентами степени не выше n. Примем, что $C^0[a,b]$ есть множество непрерывных на $[a,b]$ функций, $C^{-1}[a,b]$ – множество кусочно непрерывных на $[a,b]$ функций.

Определение. Функция $s_{n,k}(x)$, определённая на $[a,b]$, называется сплайном степени n гладкости k на сетке Δ_N, если она удовлетворяет условиям:

1) $s_{n,k}(x) \in \mathcal{P}_n$, при $x \in [x_i, x_{i+1}]$, $i = \overline{0, N-1}$

2) $s_{n,k}(x) \in C^k[a,b], \quad -1 \le k \le n$ ♣.

Непосредственно из определения получаем

Следствие 1. При фиксированных n, k, Δ_N множество сплайнов образует линейное пространство $\mathcal{S}_{n,k}(\Delta_N)$. При фиксированной сетке Δ_N имеет место соотношение:

$$\mathcal{S}_{m,p} \subset \mathcal{S}_{n,q} \text{ при } \begin{cases} m \le n \\ p \ge q \end{cases} . \quad ♣$$

Следствие 2. При $k = n$ имеет место $\mathcal{S}_{n,n} = \mathcal{P}_n$ ♣.

6.2. Базис пространства сплайнов

Введём в рассмотрение срезки степенных функций вида:

$$(x - x_i)_+^j = \begin{cases} (x - x_i)^j & \text{при} \quad x \ge x_i \\ 0 & \text{при} \quad x < x_i \end{cases} . \tag{1}$$

Очевидно, что $(x - x_i)_+^\nu \in C^{\nu-1}[a,b]$.

Теорема. *$\mathcal{S}_{n,k}(\Delta_N)$ имеет размерность $(n+1) + (N-1)(n-k)$. В качестве базиса этого пространства можно взять набор функций*

$$\begin{cases} x^j, & j = \overline{0,n} \\ (x - x_i)_+^\nu, & k+1 \le \nu \le n, \ i = \overline{1, N-1} \end{cases} . \quad ♣ \tag{2}$$

Доказательство. Покажем сначала, что любой сплайн $s_{n,k}(x) \in \mathcal{S}_{n,k}$ может быть представлен в виде линейной комбинации функций (2):

$$s_{n,k}(x) = \sum_{\nu=0}^{n} a_\nu x^\nu + \sum_{i=1}^{N-1} \sum_{\nu=k+1}^{n} a_{i\nu} (x - x_i)_+^\nu . \tag{3}$$

Докажем этом методом математической индукции по интервалам $[x_0, x_i]$, $i = \overline{1, N}$.

1) Пусть $x \in [x_0, x_1]$. На этом интервале сплайн суть полином степени не выше n, а поэтому он записывается в виде (3) при должном выборе коэффициентов a_ν, поскольку все слагаемые под знаком двойной суммы равны нулю по определению функций (3). Тем самым создана база для индукции.

2) Пусть формула (3) верна для $x \in [x_0, x_j]$, т.е.

$$s_{n,k}(x) = \sum_{\nu=0}^{n} a_\nu x^\nu + \sum_{i=1}^{j-1} \sum_{\nu=k+1}^{n} a_{i\nu}(x - x_i)_+^\nu. \tag{4}$$

Покажем, что она верна и для $x \in [x_0, x_{j+1}]$. Для этого рассмотрим интервал $[x_j, x_{j+1}]$. На нём сплайн является по определению полиномом $s_{n,k}(\cdot) \in \mathcal{P}_n$. Пусть

$$s_{n,k}(x) = Q_n^{j+1}(x).$$

На интервале $[x_{j-1}, x_j]$ он также является полиномом $s_{n,k}(x) = Q_n^j(x)$, причём согласно индуктивному предположению на этом интервале для него верно представление (4). На $[x_j, x_{j+1}]$ можем записать:

$$Q_n^{j+1}(x) - Q_n^j(x) = \sum_{\nu=0}^{n} a_{j\nu}(x - x_j)^\nu = \sum_{\nu=0}^{n} a_{j\nu}(x - x_j)_+^\nu, \tag{5}$$

причём из условий согласованности в узле x_j получаем:

$$\frac{d^l}{dx^l} Q_n^{j+1}(x)\big|_{x=x_j} = \frac{d^l}{dx^l} Q_n^j(x)\big|_{x=x_j}, \quad l = \overline{0,k}, \tag{6}$$

откуда следует $a_{j0} = 0,\ a_{j1} = 0, \ldots, a_{jk} = 0$. С учетом (5) получаем:

$$Q_n^{j+1}(x) = Q_n^j(x) + \sum_{\nu=k+1}^{n} a_{j\nu}(x - x_j)_+^\nu,$$

что доказывает справедливость формулы (3).

Остаётся доказать линейную независимость функций (2), что делается вполне аналогично индукцией по интервалам $[x_0, x_i],\ i = \overline{1,N}$. Действительно, пусть $s_{n,k}(x) \equiv 0$ на $[a, b]$. При $x \in [x_0, x_1]$ имеем

$$s_{n,k}(x) = \sum_{\nu=0}^{n} a_\nu x^\nu \equiv 0 \quad \Longleftrightarrow \quad a_\nu = 0,\ \nu = \overline{0,n}.$$

При $x \in [x_0, x_2]$

$$s_{n,k}(x) = \sum_{\nu=k+1}^{n} a_{1\nu}(x - x_1)^\nu \equiv 0 \quad \Longleftrightarrow \quad a_{1\nu} = 0,\ \nu = \overline{k+1,n}.$$

Дальнейшее очевидно.

Общее число функций в системе (2) и даёт размерность пространства $\mathcal{S}_{n,k}$, указанную в теореме. ♣

6.3. Интерполирование сплайнами $s_{1,0}(\cdot)$

Пусть на $[a,b]$ задана сетка $\delta_M = \{\xi_0, \xi_1, \ldots, \xi_M\}$, в узлах которой известны значения функции $f(\cdot)$: $f_i = f(\xi_i),\ i = \overline{0,M}$.

На том же интервале задана сетка $\Delta_N = \{x_0, x_1, \ldots, x_N\}$. Зададим на ней пространство сплайнов $\mathcal{S}_{1,0}$ и поставим задачу: построить сплайн $s_{1,0}(\cdot)$ такой,что

$$s_{1,0}(\xi_i) = f_i, \quad i = \overline{0,M}.$$

Рассмотрим эту задачу в частном случае $M = N,\ \delta_M = \Delta_N$. Тогда, очевидно, существует единственный *интерполяционный* сплайн данного класса, удовлетворяющий заданным условиям. График этого сплайна представляет собой ломанную, соединяющую точки плоскости с координатами $(x_i, f_i),\ i = \overline{0,N}$.

Пусть $h_i = x_{i+1} - x_i$. Тогда на каждом интервале $[a,b]$ искомый сплайн можно представить в виде:

$$s_{1,0}(x) = \frac{x_{i+1} - x}{h_i} f_i + \frac{x - x_i}{h_i} f_{i+1}. \tag{1}$$

Получим оценку методической погрешности $r_1(x) = f(x) - s_{1,0}(x)$ интерполирования сплайнами из $\mathcal{S}_{1,0}$: поскольку на каждом интервале $[x_i, x_{i+1}]$ сплайн есть интерполяционный полином (Лагранжа), то

$$r_1(x) = \frac{f''(\xi)}{2}(x - x_i)(x - x_{i+1}), \quad x, \xi \in [x_i, x_{i+1}]. \tag{2}$$

Легко видеть, что

$$\max_{x \in [x_i, x_{i+1}]} \left|(x - x_i)(x - x_{i+1})\right| = \frac{h_i^2}{4} \quad \text{при } x = \frac{x_i + x_{i+1}}{2},$$

поскольку оцениваемая функция есть квадратный трехчлен. Если известна оценка $|f''(x)| \le M_2$ для $x \in [a,b]$, и $h = \max\{h_i\}$, то

$$|r_1(x)| \le \frac{M_2 h^2}{8} \to 0 \text{ при } h \to 0,$$

т.е. интерполяционный процесс, построенный с использованием сплайнов из $\mathcal{S}_{1,0}$, *равномерно на* $[a,b]$ *сходится к интерполируемой функции при условии* $f(\cdot) \in C^2[a,b]$. Это условие гладкости необходимо для возможности использования оценки (2).

Замечание. Используя теорему Кантора (непрерывная на замкнутом интервале функция равномерно непрерывна на нём) нетрудно показать, что для сходимости интерполяционного процесса на основе сплайнов из $\mathcal{S}_{1,0}$ достаточно потребовать, чтобы $f(\cdot) \in C[a,b]$. ♣

6.4. Интерполирование сплайнами $s_{2,0}(\cdot)$

Пусть $\delta_{2m} = \{x_0, x_1, \ldots, x_{2m}\}$ – узлы на $[a,b]$, $x_0 = a,\ x_{2m} = b$, в которых заданы значения функции $f(\cdot)$. В качестве сетки для введения пространства сплайнов $\mathcal{S}_{2,0}$ возьмём $\Delta_m = \{x_0, x_2, \ldots, x_{2m}\}$. Поставим задачу: найти $s_{2,0}(\cdot)$ такой, чтобы

$$s_{2,0}(x_i) = f_i, \quad f_i = f(x_i), \quad i = \overline{0,2m} \tag{1}$$

Существование и единственность поставленной задачи для *любой* функции $f(\cdot)$ очевидны, ибо на каждом интервале $[x_{2i}, x_{2(i+1)}]$ сплайн $s_{2,0}(\cdot)$ совпадает с полиномом, решающим задачу интерполирования по узлам $x_{2i}, x_{2i+1}, x_{2(i+1)},\ i = \overline{0, m-1}$.

Введём в рассмотрение сплайны $s^i(\cdot),\ i = \overline{0, 2m}$, определённые условиями:

$$s^i(\cdot) \in \mathcal{S}_{2,0}, \quad s^i(x_j) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{2}$$

С их помощью интерполяционный сплайн может быть записан в виде, вполне аналогичном записи полинома Лагранжа:

$$s_{2,0}(x) = \sum_{i=0}^{2m} s^i(x) f_i. \tag{3}$$

Отметим важную особенность сплайна $s^i(\cdot)$: его носитель (т.е. множество, где он отличен от нуля) сосредоточен вокруг точки x_i, поэтому добавление узлов интерполирования добавляет несколько слагаемых в представлении (3) интерполяционного сплайна.

Оценим методическую погрешность $r_2(x) = f(x) - s_{2,0}(x)$, предполагая, что $f(\cdot) \in C^3[a, b]$. На интервале $[x_{2i}, x_{2(i+1)}]$ сплайн совпадает с полиномом Лагранжа, поэтому на этом интервале

$$r_2(x) = \frac{f'''(\xi)}{3!}(x - x_{2i})(x - x_{2i+1})(x - x_{2i+2}). \tag{4}$$

Обозначим $h = max_i\{h_i\},\ h_i = x_{i+1} - x_i$. Тогда, при условии, что $|f'''(x)| \leq M_3$ на $[a, b]$, получим

$$|r_2(x)| \leq \frac{M_3 h^3}{3!}. \tag{5}$$

Из этой оценки следует, что *интерполяционный процесс, построенный на основе сплайнов из $\mathcal{S}_{2,0}$, равномерно на $[a, b]$ сходится к интерполируемой функции.*

Более того, производная сплайна $s_{2,0}(\cdot)$ (вне узлов, где производная не существует), хорошо приближает производную от $f(\cdot)$. Чтобы убедиться в этом, оценим величину $r_2'(x) = f'(x) - s_{2,0}'(x)$. Обозначим для краткости $\alpha = x_{2i}$, $\beta = x_{2i+1}$, $\gamma = x_{2(i+1)}$.
Поскольку $r_2(\alpha) = r_2(\beta) = r_2(\gamma) = 0$, то существуют $\xi_1, \xi_2 \in [\alpha, \gamma]$, в которых $r_2'(\xi_1) = r_2'(\xi_2) = 0$. Из последнего следует опять же существование точки $\eta \in [\alpha, \gamma]$, в которой $r_2''(\eta) = 0$. Поскольку $r_2'''(x) = f'''(x)$, то

$$|r_2''(x)| = \Big|\int_\eta^x r_2'''(t)dt\Big| \leq \left\{ \begin{array}{l} \displaystyle\int_\eta^x |f'''(t)|dt \text{ при } x \geq \eta, \\ \\ \displaystyle\int_x^\eta |f'''(t)|dt \text{ при } x \leq \eta \end{array} \right\} \leq \int_\alpha^\gamma |f'''(t)|dt \leq 2hM_3.$$

Аналогично получаем

$$|r_2'(x)| = \Big|\int_{\xi_1}^x r_2''(t)dt\Big| \leq \int_\alpha^\gamma |r_2''(t)|dt \leq 4h^2 M_3,$$

что и доказывает высказанное утверждение.

6.5. Интерполирование кубическими сплайнами Эрмитовы кубические сплайны

Пусть $a = x_0 < x_1 < \cdots < x_n = b$ – сетка на $[a, b]$, в узлах которой заданы значения функции $f(\cdot)$ и её первой производной. На этой же сетке зададим пространство сплайнов $\mathcal{S}_{3,1}$ и поставим задачу построения сплайна из этого пространства такого, что

$$s_{3,1}^{(r)}(x_i) = f_i^{(r)}, \quad i = \overline{0, N},\ r = 0, 1. \tag{1}$$

На каждом из интервалов $[x_i, x_{i+1}]$ запишем сплайн в виде:

$$s_{3,1}(x) = \sum_{s=0}^{3} a_{is}(x - x_i)^s \tag{2}$$

и рассмотрим задачу (1):

$$s_{3,1}^{(r)}(x_j) = f_j^{(r)}, \quad j = i, i+1;\ r = 0, 1. \tag{3}$$

Это есть задача кратного интерполирования и потому её решение существует и единственно. Введём переменную $t = \dfrac{x - x_i}{h_i}, \quad h_i = x_{i+1} - x_i,\ t \in [0, 1]$. Тогда решение задачи (3) может быть записано в виде:

$$\sigma(t) = s_{3,1}(x_i + h_i t) = \alpha\,(t) f_i + \beta(t) f_{i+1} + \gamma(t) h_i f_i' + \delta(t) h_i f_{i+1}', \tag{4}$$

где $\alpha(t),\ \beta(t),\ \gamma(t),\ \delta(t) \in \mathcal{P}_3$ и *не зависят от номера i интервала* при условии, что эти функции являются решениями следующих задач кратного интерполирования:

t	0	1
α	1	0
α'	0	0

,

t	0	1
β	0	1
β'	0	0

,

t	0	1
γ	0	0
γ'	1	0

,

t	0	1
δ	0	0
δ'	0	1

,

Нетрудно получить и явные выражения этих функций:

$$\begin{aligned} \alpha\,(t) &= (1-t)^2(1+2t) \geq 0, \quad & \beta(t) &= t^2(3-2t) \geq 0, \\ \gamma(t) &= t\,(1-t)^2 \geq 0, & \delta(t) &= -t^2(1-t) \leq 0. \end{aligned} \tag{5}$$

Лемма 1. *Если $f(\cdot) \in C[a, b]$ и $p \cdot q > 0$, то существует $\xi \in [a, b]$ такая, что*

$$pf(a) + qf(b) = (p+q)f(\xi) \quad \clubsuit \tag{6}$$

Доказательство. Рассмотрим вспомогательную функцию

$$\psi(x) = pf(a) + qf(b) - (p+q)f(x).$$

Имеем для неё:

$$\left.\begin{aligned} \psi(a) &= q\big(f(a) - f(b)\big) \\ \psi(b) &= p\big(f(b) - f(a)\big) \end{aligned}\right\} \quad \Longrightarrow \psi(a)\psi(b) < 0.$$

Следовательно, существует точка $\xi \in [a, b]$ такая, что $\psi(\xi) = 0$, что равносильно утверждению леммы. $\clubsuit$

Лемма 2. *Функции* $\alpha(\cdot)$ *и* $\beta(\cdot)$ обладают свойством:

$$\alpha(\cdot)+\beta(\cdot)=1. \quad \clubsuit \tag{7}$$

Доказательство. Утверждение леммы доказывается непосредственной проверкой с учетом формул (5). $\clubsuit$

Оценим методическую погрешность $R(x)=f(x)-s_{3,1}(x)$.
Поскольку $x_i=x-th_i$ и $x_{i+1}=x+(1-t)h_i$, то применяя формулу конечных приращений Лагранжа, имеем:

$$f_i=f(x)-th_if'(\xi), \quad f_{i+1}=f(x)+(1-t)h_if'(\eta). \tag{8}$$

С учетом леммы 1 получаем для $R(x)$:

$$\begin{aligned} |R(x)| &= |f(x)-\sigma(t)| = \big|f(x)-\big(\alpha(t)\big(f(x)-th_if'(\xi)\big)+ \\ &+ \beta(t)\big(f(x)+(1-t)h_if'(\eta)\big)+\gamma(t)h_if_i'+\delta(t)h_if_{i+1}'\big)\big| = \\ &= \big|-\alpha(t)th_if'(\xi)+\beta(t)(1-t)h_if'(\eta)+\gamma(t)h_if_i'+\delta(t)h_if_{i+1}'\big)\big|. \end{aligned} \tag{9}$$

Сгруппировав первое с четвертым и третье со вторым слагаемым, преобразуем их, используя для каждой пары лемму 2:

1. $\big(-\alpha(t)t\big)h_if'(\xi)+\delta(t)h_if_{i+1}'=h_i\big(-\alpha(t)t+\delta(t)\big)f(\hat{\xi})$;
2. $\beta(t)(1-t)h_if'(\eta)+\gamma(t)h_if_i'=h_i\big(\beta(t)(1-t)+\gamma(t)\big)f'(\hat{\eta})$.

Здесь $\hat{\xi},\hat{\eta}\in[a,b]$. Заметив, что

$$-\big(-\alpha(t)t+\delta(t)\big)=\beta(t)(1-t)+\gamma(t)=t(1-t)(1+2t-2t^2)\equiv\varphi(t), \tag{10}$$

получаем такую оценку для $R(x)$:

$$|R(x)|\le 2h_i|\varphi(t)|M_1, \tag{11}$$

где $M_1\ge|f'(x)|,\ x\in[a,b]$. Нетрудно убедиться, что

$$\psi(z)=\varphi(z+1/2)=(1/4-z^2)(3/2-2z^2), \quad z\in[-1/2,1/2],$$

откуда следует, что $\psi(z)\ge 0$ и $\min\limits_{[-1/2,1/2]}\psi(z)=3/8$. Поэтому оценка (11) принимает вид

$$|R(x)|\le\frac{3}{4}hM_1, \qquad \text{где } h=\max_i h_i,\ x\in[a,b].$$

6.6. Интерполирование сплайнами $s_{3,2}(\cdot)$

Для краткости обозначим $s_{3,2}(\cdot)=s(\cdot)$ и рассмотрим следующую задачу: в узлах сетки $\Delta_N=\{x_0,x_1,\dots,x_N\}$, $x_0=a$, $x_N=b$ заданы значения функции $f(x_i)=f_i$. Построить сплайн $s(\cdot)\in\mathcal{S}_{3,2}$, удовлетворяющий условиям:

$$\begin{aligned} &s^{(k)}(x_i-0)=s^{(k)}(x_i+0), \quad i=\overline{1,N-1},\ k=0,1,2; \\ &s(x_i)=f_i, \quad i=\overline{0,N}. \end{aligned} \tag{1}$$

Всего имеем $(4N-2)$ уравнения относительно параметров сплайна $s(\cdot)$, число которых $4N$. Поэтому обычно задают ещё два дополнительных (*краевых*) условия, наиболее распространенные из которых имеют вид:

1. $s'(a) = f'(a), \quad s'(b) = f'(b);$
2. $s''(a) = f''(a), \quad s''(b) = f''(b);$
3. $s^{(k)}(a) = s^{(k)}(b) \quad k = 1, 2;$
4. $s'''(x_i - 0) = s'''(x_i + 0), \quad i = 1, N-1.$

Обозначим $s'(x_i) = m_i$. Тогда $s(\cdot)$ можно рассматривать как эрмитов кубический сплайн и, следовательно, $s(\cdot) \in C^1[a, b]$:

$$\begin{aligned} \sigma(t) &= s(x_i + h_i t) = \alpha(t) f_i + \beta(t) f_{i+1} + \gamma(t) h_i m_i + \delta(t) h_i m_{i+1} = \\ &= (1-t)^2(1+2t) f_i + t^2(3-2t) f_{i+1} + \\ &+ t(1-t)^2 m_i h_i - t^2(1-t) h_i m_{i+1}. \end{aligned} \tag{2}$$

Выберем параметры $m_i,\ i = \overline{1, N-1}$ так, чтобы обеспечить $s(\cdot) \in C^2[a, b]$:

$$\begin{aligned} s''(x) &= \frac{6(1-2t)(f_{i+1} - f_i)}{h_i^2} + \frac{2\big((3t-2)m_i + (3t-1)m_{i+1}\big)}{h_i}, \\ s''(x_i + 0) &= s''(x)\Big|_{t=0} = \frac{6(f_{i+1} - f_i)}{h_i^2} + \frac{2(-2m_i - m_{i+1})}{h_i}. \end{aligned} \tag{3}$$

Чтобы получить $s''(x_i - 0)$ по формуле (3), следует заменить в ней i на $i-1$ и положить $t = 1$:

$$s''(x_i - 0) = \frac{6(f_{i-1} - f_i)}{h_{i-1}^2} + \frac{2(m_{i-1} + 2m_i)}{h_{i-1}}. \tag{4}$$

Условие непрерывности $s''(\cdot)$ в точке x_i даёт:

$$\lambda_i m_{i-1} + 2m_i + \mu_i m_{i+1} = 3\Big(\mu_i \frac{f_{i+1} - f_i}{h_i} + \lambda_i \frac{f_i - f_{i-1}}{h_{i-1}}\Big) \equiv c_i, \quad i = \overline{1, N-1} \tag{5}$$

Здесь обозначено $\mu_i = \dfrac{h_{i-1}}{h_{i-1} + h_i}$, $\lambda_i = 1 - \mu_i$.

К полученной системе следует добавить краевые условия. Так, для краевых условий первого типа получаем СЛАУ для определения $m_i,\ i = \overline{0, N}$:

$$\begin{aligned} &m_0 = f'_0, \\ &\lambda_i m_{i-1} + 2m_i + \mu_i m_{i+1} = c_i,\ i = \overline{1, N-1}, \\ &m_N = f'_N. \end{aligned} \tag{6}$$

Отметим, что полученная система есть СЛАУ с диагональным преобладанием, а поэтому решение её существует и единственно.

Приведём без доказательства оценку методической погрешности интерполирования с помощью рассмотренных сплайнов для $f(\cdot) \in C^1[a, b]$:

$$\max_{[a,b]} |f(x) - s(x)| \le \frac{7}{4} M_1 h,$$

где M_1 и h те же, что и в предыдущем пункте.

Задачи, предлагаемые для лучшего усвоения материала данной главы.

Пусть $P_4(x) = x^4 - 2x^2 + 3x - 1$.
1. Построить интерполяционный полином для $P_4(\cdot)$ по узлам:

- (-1,0,1,2);
- (-2,0,1,2);
- (-2,-1,0,1,2).

2. Провести квадратичное интерполирование для функции $F(x, y) = x^3 + y^3 + xy$.

3. Для функции $f(x, y) = x^2 + y$ решить задачу интерполирования по прямоугольной таблице, взяв в качестве узлов $x_i, y_i \in \{-1, 0, 1\},\ i = \overline{1, 3}$.

4. Решить задачу кратного интерполирования для указанного выше полинома $P_4(\cdot)$ по данным: $P_4(-1), P_4(0), P_4'(-1), P_4'(0)$.

5. Построить для $y(\cdot) = P_4(\cdot)$ интерполяционный сплайн $s_{2,0}$, заданный на сетке $(-2, 0, 2)$ с узлами интерполирования $(-2, -1, 0, 1, 2)$.

Глава 5. Аппроксимация функций в метрических пространствах

1. Линейная задача наименьших квадратов

Пусть, как и в общей задаче интерполирования, на интервале $[a,b]$ задана некоторая функция $f(\cdot)$, причём её значения $\{y_i = f(x_i)\}$ в узлах сетки $\{x_0, x_1, \ldots, x_n\} \subset [a,b]$ известны с погрешностями $\{\varepsilon_i\}$, т.е. вместо набора значений $\{y_i\}$ имеем набор $\{\tilde{y}_i = y_i + \varepsilon_i\}$. (Далее под $\{y_i\}$ будем понимать заданные значения функции, т.е. с погрешностями, вводя обозначение $\tilde{y}_i$ лишь в случае необходимости). Также на $[a,b]$ определены функции $\varphi_j(\cdot) \in \Phi,\ j = \overline{0,m}$. Пусть $P_m(x) = \sum\limits_{j=0}^{m} a_j \varphi_j(x)$ – обобщенный полином.

Введём вектор a коэффициентов полинома $P_m(\cdot)$: $a = (a_0, a_1, \ldots, a_m)^T$, вектор $y = (y_0, y_1, \ldots, y_n)^T$ и вектор-функцию $\varphi(x) = (\varphi_0(x), \varphi_1(x), \ldots, \varphi_m(x))^T$, а также функции

$$\sigma(a,y) = \sum_{i=0}^{n} (P_m(x_i) - y_i)^2, \quad \delta(a,y) = \sqrt{\frac{\sigma(a,y)}{n+1}}. \tag{1}$$

Функцию $\delta(a,y)$ назовём *среднеквадратичным уклонением* обобщенного полинома $P_m(\cdot)$ от функции $f(\cdot)$ на системе узлов $x_0, x_1, \ldots, x_n$. Отказываясь теперь от условия $P_m(x_i) = y_i$, поставим задачу так: *найти обобщенный полином* $\bar{P}_m(\cdot) = \bar{a}^T \varphi(\cdot)$, *для которого среднеквадратичное уклонение минимально*:

$$\delta(\bar{a}, y) = \min_{a} \delta(a,y). \tag{2}$$

Поставленную задачу называют *линейной задачей метода наименьших квадратов* или просто методом наименьших квадратов (МНК). Если искомый полином существует, то будем называть его *многочленом наилучшего среднеквадратичного приближения* (МНСП).

Отметим, что минимум функций $\sigma(\cdot)$ и $\delta(\cdot)$ достигается *на одном и том же векторе* $\bar{a}$, поэтому фактически будем вести минимизацию функции $\sigma(a,y)$.

Простейший подход к решению задачи – использование необходимых условий в задаче поиска экстремума для дифференцируемой функции:

$$\frac{\partial \sigma(a,y)}{\partial a_k}\Big|_{a=\bar{a}} = 0, \quad k = \overline{0,m}. \tag{3}$$

В дополнение к уже введенным обозначениям примем ещё следующее:

$$Q = \begin{pmatrix} \varphi_0(x_0), & \varphi_1(x_0), & \ldots & \varphi_m(x_0), \\ \varphi_0(x_1), & \varphi_1(x_1), & \ldots & \varphi_m(x_1), \\ \ldots & \ldots & \ldots & \ldots \\ \varphi_0(x_n), & \varphi_1(x_n), & \ldots & \varphi_m(x_n) \end{pmatrix}.$$

Кроме того будем считать, что под скалярным произведением двух векторов $u, v \in R^k$ понимается число

$$(u, v) = u^T v = u \cdot v = \sum_{i=1}^{k} u_i v_i,$$

а под нормой векторов будем понимать евклидову норму: $\|y\|^2 = (y, y)$. Тогда в новых обозначениях

$$\begin{aligned} \sigma(a, y) &= \|Qa - y\|^2 = (Qa - y, Qa - y) = (Qa, Qa) - 2(Qa, y) + (y, y) = \\ &= (a, Q^T Qa) - 2(a, Q^T y) + \|y\|^2. \end{aligned} \tag{4}$$

Для определения параметров искомого полинома в соответствии с формулой (3) имеем СЛАУ:

$$Ha = b, \quad \text{где } H = Q^T Q,\ b = Q^T y. \tag{5}$$

Лемма 1. Пусть $\bar{a}$ – решение системы (5). Тогда для введенной выше функции $\sigma(a, y)$ верно равенство: $\sigma(\bar{a} + \Delta a, y) = \sigma(\bar{a}, y) + \|Q\Delta a\|^2$. ♣

Доказательство. Утверждение леммы следует из следующей цепочки равенств:

$$\begin{aligned} \sigma(\bar{a} + \Delta a, y) &= \|Q(\bar{a} + \Delta a) - y\|^2 = (Q\bar{a} - y + Q\Delta a, Q\bar{a} - y + Q\Delta a) = \\ &= \|Q\bar{a} - y\|^2 + 2(Q\bar{a} - y, Q\Delta a) + \|Q\Delta a\|^2 = \\ &= \sigma(\bar{a}, y) + (Q^T Q\bar{a} - Q^T y, \Delta a) + \|Q\Delta a\|^2 = \sigma(\bar{a}, y) + \|Q\Delta a\|^2, \end{aligned}$$

поскольку $(Q^T Q\bar{a} - Q^T y, \Delta a) = (H\bar{a} - b, \Delta a) = 0$ в силу (5). ♣

Введём векторы $\bar{\varphi}_j = (\varphi_j(x_0), \varphi_j(x_1), \ldots, \varphi_j(x_n))^T,\ \bar{\varphi}_j \in R^{n+1},\ j = \overline{0, m}$.

Определение. Будем говорить, что система функций $\varphi_0(\cdot), \varphi_1(\cdot), \ldots, \varphi_m(\cdot)$ *линейно зависима* в точках $x_0, x_1, \ldots, x_n$, если один из векторов $\bar{\varphi}_j$ системы $\bar{\varphi}_0, \bar{\varphi}_1, \ldots, \bar{\varphi}_m$ может быть представлен в виде линейной комбинации остальных, т.е. $\bar{\varphi}_j = \sum_{k \neq j} \alpha_k \bar{\varphi}_k$. В противном случае указанную систему функций будем называть *линейно независимой* в точках $x_0, x_1, \ldots, x_n$. ♣

Лемма 2. Если система функций $\varphi_0(\cdot), \varphi_1(\cdot), \ldots, \varphi_m(\cdot)$ линейно независима в точках $x_0, x_1, \ldots, x_n$, то $\det H \neq 0$. ♣

Доказательство. По условию

$$\psi(c) = \|\sum_{i=0}^{m} c_i \bar{\varphi}_i\|^2 > 0 \text{ при } \sum_{i=0}^{m} c_i^2 > 0.$$

Но функция $\psi(c)$ может быть записана и в виде

$$\psi(c) = \sum_{i,j=0}^{m} c_i c_j (\bar{\varphi}_i, \bar{\varphi}_j) = c^T H c > 0, \text{ при } c \neq 0,$$

т.е. матрица H является положительно определённой. ♣

Теорема. *Если функции $\{\varphi_i(x)\}$ линейно независимы в точках $x_0, x_1, \ldots, x_n$, то существует единственный многочлен наилучшего среднеквадратичного приближения .* ♣

Доказательство. Во-первых, в силу леммы 2 решение системы (5) существует и единственно, а во-вторых, на этом решении в соответствии с леммой 1 достигается минимум функции $\sigma(a, y)$. ♣

Замечание 1. При $m = n$ и выполнении условий теоремы МНСП совпадает с интерполяционным многочленом (при условии его существования), т.к. для него $\sigma(a, y) = 0$. ♣

Замечание 2. Часто принимают $m \ll n$. В этом случае метод обладает сглаживающими свойствами. ♣
Часто для приближений по МНК используют алгебраические многочлены степени $m \leq n$.

Лемма 3. При $m \leq n$ система функций $1, x, \ldots, x^m$ линейно независима в точках $x_0, x_1, \ldots, x_n$. ♣
Доказательство. Если это не так, то по определению линейной зависимости функций в узлах $x_0, x_1, \ldots, x_n$ при некотором j имеет место равенство:

$$x_i^j = \sum_{\substack{k=0 \\ k \neq j}}^{m} \alpha_k x_i^k, \quad i = \overline{0, n}.$$

Рассмотрим полином

$$P_m(x) = x^j - \sum_{\substack{k=0 \\ k \neq j}}^{m} \alpha_k x^k. \tag{6}$$

Для него $P_m(x_i) = 0$, $i = \overline{0, n}$, а т.к. $m \leq n$, то $P_m(x) \equiv 0$, что противоречит (6), поскольку в этой записи $\alpha_j = 1$. ♣

2. Наилучшие приближения в линейных нормированных пространствах

2.1. Постановка задачи

Пусть U – линейное нормированное пространство функций, $\{\varphi_i(\cdot)\} \subset U$, $i = \overline{0, n}$ – линейно независимые функции, $f(\cdot) \in U$. Введём подпространство $\bar{U}$ пространства U: $\bar{U} = \{\varphi(\cdot) \in U : \varphi(\cdot) = \sum_{i=0}^{n} c_i \varphi_i(\cdot),\ \varphi_i(\cdot) \in U\}$, и расстояние $\rho(f, \varphi) = \|f(\cdot) - \varphi(\cdot)\| \geq 0$, т.е. $\inf \rho(f, \varphi) \geq 0$.
Рассмотрим вопрос: существует ли $\bar{\varphi}(\cdot) \in \bar{U}$ такой, что $\rho(f, \bar{\varphi}) = \inf_{\varphi \in \bar{U}} \rho(f, \varphi)$?

Определение. Всякий элемент $\bar{\varphi}(\cdot) \in \bar{U}$, для которого выполняется последнее равенство, называется элементом *наилучшего приближения* для $f(\cdot)$ в $\bar{U}$ (или *проекцией* $f(\cdot)$ на $\bar{U}$). ♣

2.2. Существование элемента наилучшего приближения

Теорема 1. *В подпространстве $\bar{U}$ для любой функции $f(\cdot) \in U$ существует элемент наилучшего приближения $\bar{\varphi}(\cdot)$.*♣

Доказательство. Используя свойства нормы имеем неравенства:

$$\|f_1(\cdot)\| = \|\big(f_1(\cdot) - f_2(\cdot)\big) + f_2(\cdot)\| \le \|f_1(\cdot) - f_2(\cdot)\| + \|f_2(\cdot)\|;$$
$$\|f_2(\cdot)\| = \|\big(f_2(\cdot) - f_1(\cdot)\big) + f_1(\cdot)\| \le \|f_2(\cdot) - f_1(\cdot)\| + \|f_1(\cdot)\|;$$

откуда следует:

$$\|f_1(\cdot)\| - \|f_2(\cdot)\| \le \|f_1(\cdot) - f_2(\cdot)\|,$$
$$\|f_1(\cdot)\| - \|f_2(\cdot)\| \ge -\|f_1(\cdot) - f_2(\cdot)\|,$$

и, следовательно, имеет место соотношение:

$$\big|\|f_1(\cdot)\| - \|f_2(\cdot)\|\big| \le \|f_2(\cdot) - f_1(\cdot)\|.$$

Используя это неравенство, получаем:

$$\big|\|f(\cdot) - \sum_{i=0}^{n} c_i^1\varphi_i(\cdot)\| - \|f(\cdot) - \sum_{i=0}^{n} c_i^2\varphi_i(\cdot)\|\big| \le \big\|\sum_{i=0}^{n}(c_i^2 - c_i^1)\varphi_i(\cdot)\big\| \le$$
$$\le \sum_{i=0}^{n} |c_i^2 - c_i^1|\|\varphi_i(\cdot)\|, \qquad (1)$$

откуда следует, что функция $F(f,c) = \|f(\cdot) - \sum_{i=0}^{n} c_i\varphi_i(\cdot)\|$ непрерывна по своему аргументу $c = (c_1, c_2, \ldots, c_n)^T$ для любой функции $f(\cdot) \in U$. Функция $F(0,c)$, будучи непрерывной и на сфере $\|c\| = 1$, достигает на ней своей точной нижней грани в некоторой точке c^0:

$$\mu = \inf_{\|c\|=1} F(0,c).$$

Очевидно, что $\mu > 0$ при условии линейной независимости функций $\{\varphi_i\}$, $i = \overline{0,n}$. Для любого вектора $c \in R^{n+1}$, $c \ne 0$ имеет место неравенство:

$$F(0,c) = \|c\|\, F\Big(0, \frac{c}{\|c\|}\Big) \ge \mu\|c\|. \qquad (2)$$

Обозначим $r = 2\dfrac{\|f(\cdot)\|}{\mu}$ и рассмотрим шар $S = \{c:\ \|c\| \le r\}$. Поскольку $F(f,c)$ неотрицательна и непрерывна в R^{n+1}, то в некоторой точке $\bar{c} \in S$ она достигает своей точной нижней грани:

$$F(f,\bar{c}) = \inf_{S} F(f,c) \equiv \nu.$$

Заметив, что $F(f,0) = \|f(\cdot)\|$, получаем неравенство (т.к. $0 \in S$):

$$\|f(\cdot)\| = F(f,0) \ge \nu \qquad (3)$$

С другой стороны вне шара имеем, учитывая неравенство (2):

$$\begin{aligned} F(f,c) &= \|f(\cdot) - \sum_{i=0}^{n} c_i\varphi_i(\cdot)\| \ge \big\|\sum_{i=0}^{n} c_i\varphi_i\big\| - \|f(\cdot)\| = \\ &= F(0,c) - \|f(\cdot)\| \ge \mu r - \|f(\cdot)\| = \|f(\cdot)\| \ge \nu. \end{aligned} \qquad (4)$$

Таким образом, верно неравенство

$$F(f,c) \ge \nu = F(f,\bar{c}) \quad \text{для всех } c \in R^{n+1},$$

что и доказывает утверждение теоремы. ♣

2.3. Единственность элемента наилучшего приближения

Определение. Линейное нормированное пространство U называется *строго нормированным*, если из условия

$$\|f_1(\cdot) + f_2(\cdot)\| = \|f_1(\cdot)\| + \|f_2(\cdot)\|,$$

где $f_1(\cdot) \neq 0$, $f_2(\cdot) \neq 0$ следует равенство:

$$f_2(\cdot) = \alpha f_1(\cdot),\ \alpha > 0. \quad \clubsuit$$

Теорема 2. *Если U – строго нормированное пространство, то в $\bar{U}$ существует единственный элемент наилучшего приближения для любой функции $f(\cdot) \in U$.* $\clubsuit$

Доказательство. Допустим существование двух *различных* элементов наилучшего приближения, т.е. $\bar{\varphi}(\cdot), \hat{\varphi}(\cdot) \in \bar{U}$ и

$$\|f(\cdot) - \bar{\varphi}(\cdot)\| = \|f(\cdot) - \hat{\varphi}(\cdot)\| = \nu = \inf_{\varphi(\cdot)\in\bar{U}} \|f(\cdot) - \varphi(\cdot)\|.$$

Отметим, что здесь $\nu > 0$, поскольку иначе $\bar{\varphi}(\cdot) = \hat{\varphi}(\cdot) = f(\cdot)$, что противоречит только что сделанному предположению $\bar{\varphi}(\cdot) \neq \hat{\varphi}(\cdot)$.
Кроме того, имеем оценку:

$$\begin{aligned} \left\|f(\cdot) - \frac{\bar{\varphi}(\cdot) + \hat{\varphi}(\cdot)}{2}\right\| &= \left\|\frac{f(\cdot) - \hat{\varphi}(\cdot)}{2} + \frac{f(\cdot) - \bar{\varphi}(\cdot)}{2}\right\| \leq \\ &\leq \frac{\|f(\cdot) - \hat{\varphi}(\cdot)\|}{2} + \frac{\|f(\cdot) - \bar{\varphi}(\cdot)\|}{2} = \nu \end{aligned} \tag{1}$$

Поскольку $\dfrac{\bar{\varphi}(\cdot) + \hat{\varphi}(\cdot)}{2} \in \bar{U}$, то $\left\|f(\cdot) - \dfrac{\bar{\varphi}(\cdot) + \hat{\varphi}(\cdot)}{2}\right\| \geq \nu$ и сравнивая с (1) заключаем, что

$$\left\|\frac{f(\cdot) - \hat{\varphi}(\cdot)}{2} + \frac{f(\cdot) - \bar{\varphi}(\cdot)}{2}\right\| = \frac{\|f(\cdot) - \hat{\varphi}(\cdot)\|}{2} + \frac{\|f(\cdot) - \bar{\varphi}(\cdot)\|}{2} = \nu.$$

Учитывая строгую нормированность пространства $\bar{U}$, имеем равенство:

$$\frac{f(\cdot) - \bar{\varphi}(\cdot)}{2} = \alpha \frac{f(\cdot) - \hat{\varphi}(\cdot)}{2}, \quad \alpha > 0.$$

Если здесь $\alpha \neq 1$, то $f(\cdot) = \dfrac{\bar{\varphi}(\cdot) - \alpha\hat{\varphi}(\cdot)}{1 - \alpha} \in \bar{U}$ и, следовательно, $\nu = 0$, что мы уже отвергли.
Если же $\alpha = 1$, то $\hat{\varphi}(\cdot) = \bar{\varphi}(\cdot)$, что противоречит предположению. $\clubsuit$

3. Наилучшие приближения непрерывных функций

3.1. Наилучшие приближения в пространстве C

Пусть $U = C[a, b]$,

$$\|f(\cdot)\| = \|f(\cdot)\|_C = \sup_{x\in[a,b]} |f(\cdot)|,$$

$\{\varphi_i(\cdot)\},\ i = \overline{0, n}$ – линейно независимые функции в $C[a, b]$,

$$\bar{U} = \bar{C} = \{\varphi(\cdot) : \quad \varphi(\cdot) = \sum_{i=0}^{n} c_i\varphi_i(\cdot)\}, \quad c_i \in R.$$

Определение. Элемент $\bar{\varphi}(\cdot) \in \bar{C}$ называется элементом *наилучшего равномерного приближения*, если

$$\|f(\cdot) - \bar{\varphi}(\cdot)\| = \min_{\varphi(\cdot)\in\bar{C}} \|f(\cdot) - \varphi(\cdot)\|. \quad \clubsuit$$

На основании предыдущего рассмотрения такой элемент существует, однако пространство $C[a,b]$ не является строго выпуклым, поскольку, например, для функций $f_1(x) = 1,\ f_2(x) = x$, рассматриваемых на $[-1,1]$, имеем: $\|f_1(\cdot)\| = 1,\ \|f_2(\cdot)\| = 1,\ \|f_1(\cdot) + f_2(\cdot)\| = 2 = \|f_1(\cdot)\| + \|f_2(\cdot)\|$, однако $f_1(\cdot) \neq \alpha f_2(\cdot)$.

Теорема (Хаара). *Для того, чтобы для любой $f(\cdot) \in C[a,b]$ существовал* **единственный** *обобщенный многочлен наилучшего равномерного приближения необходимо и достаточно, чтобы функции $\{\varphi_i(\cdot)\},\ i = \overline{0,n}$ образовывали систему Чебышева.* $\clubsuit$

3.2. Наилучшие приближения алгебраическими полиномами

Будем далее считать, что $\varphi_k(x) = x^k,\ k = \overline{0,n}$. Тогда обобщенный многочлен становится алгебраическим многочленом $Q_n(\cdot)$. Обозначим через $\bar{Q}_n(\cdot)$ *алгебраический полином наилучшего (равномерного) приближения* (АПНП) и $E_n = \|f(\cdot) - \bar{Q}_n(\cdot)\|$.

Теорема Чебышева (об альтернансе). *Чтобы алгебраический многочлен $Q_n(\cdot)$ был многочленом наилучшего равномерного приближения функции $f(\cdot) \in C[a,b]$ необходимо и достаточно существования на $[a,b]$ по крайней мере $(n+2)$ точек $x_0, x_1, \ldots, x_{n+1}$ таких, что*

$$f(x_i) - Q_n(x_i) = \alpha(-1)^i \|f(\cdot) - Q_n(\cdot)\|,\ i = \overline{0,n+1} \tag{1}$$

причём здесь $\alpha = 1$ или $\alpha = -1$ для всех узлов одновременно. $\clubsuit$

Замечание. Точки $x_0, x_1, \ldots, x_{n+1}$, удовлетворяющие условиям теоремы, называются *точками Чебышевского альтернанса.* $\clubsuit$

Замечание. Алгебраический многочлен наилучшего равномерного приближения единствен в силу теоремы Хаара (см. с.86). $\clubsuit$

Пример 1. Покажем, что для функции $f(x) = \sin 2x$ на интервале $[0, 2\pi]$ многочленом наилучшего приближения второй степени является полином $Q_2(x) \equiv 0$. Для этого найдём точки Чебышевского альтернанса: $L = \sup|\sin 2x - 0| = 1$; уравнение $|\sin 2x| = 1$ имеет на $[0, 2\pi]$ четыре корня $(n = 2, m = n + 2 = 4)$

$$x_k = \frac{\pi}{4} + \frac{\pi k}{2}, \quad k = 0, 1, 2, 3.$$

В этих точках выполнены соотношения:

$$\sin 2x_k - 0 = (-1)^k \cdot 1, \quad k = 0, 1, 2, 3,$$

что в соответствии с теоремой Чебышева и доказывает утверждение. $\clubsuit$

Пример 2. Покажем, что полином

$$Q_{n-1} = x^n - \bar{T}_n(x), \quad \bar{T}_n(x) = \frac{1}{2^{n-1}} \cos(n \arccos x)$$

является полиномом наилучшего приближения степени $(n-1)$ на интервале $[-1,1]$ для полинома вида $f(x)=x^n$.
Действительно:

$$L=\sup_{x\in[-1,1]}|f(x)-Q_{n-1}|=\sup_{x\in[-1,1]}|\bar{T}_n(x)|=\frac{1}{2^{n-1}};$$

точки экстремума полинома Чебышева $T_n(x)$ являются точками Чебышевского альтернанса для $Q_{n-1}(x)$, т.к.

$$x_k^n-Q_{n-1}(x_k)=(-1)^kL,\quad x_k=\cos\frac{\pi k}{n},\ k=\overline{0,n},\ m=n+1$$

что в соответствии с теоремой Чебышева $(m=(n-1)+2)$ и доказывает утверждение. ♣

Пример 3. Покажем, что АПНП нечётной функции есть функция нечетная.
Пусть $\bar{Q}_n(x)$ – АПНП для $f(x)$, $f(-x)=-f(x)$. Согласно сделанным предположениям

$$E_n(f)=\sup_x|f(x)-\bar{Q}_n(x)|=\sup_x|-f(-x)-(-\bar{Q}_n(-x))|.$$

Отсюда следует, что полином $-\bar{Q}_n(-x)$ является АПНП для функции $-f(-x)=f(x)$, а в силу единственности такого полинома $\bar{Q}_n(x)=-\bar{Q}_n(-x)$. ♣

Замечание. Для чётной функции АПНП – функция четная. ♣

Пример 4. Построить полином минимальной степени, равномерно на $[-1,1]$ приближающий функцию $f(x)=\sin x$ с точностью $\varepsilon=10^{-3}$.
Возьмём отрезок разложения данной функции в ряд Тейлора

$$P_{2n+1}(x)=\sum_{k=0}^{n}(-1)^k\frac{x^{2k+1}}{(2k+1)\,!},$$

погрешность замены которым функции оценивается величиной $\varepsilon_0=\dfrac{1}{(2n+3)!}$, поскольку данный ряд является Лейбницевым. При $n=2$ имеем: $degP_{2n+1}(x)=5$, $\varepsilon_0\approx 2\cdot10^{-4}<\varepsilon$. Понижение степени (пример 2) вносит дополнительную погрешность $\varepsilon_1=\dfrac{|a_5|}{2^4}\approx 5.2\cdot10^{-4}$. Поскольку $\varepsilon_0+\varepsilon_1<\varepsilon$, то понижение степени допустимо:

$$\begin{aligned}Q_3(x) &= P_5(x)-\frac{1}{2^4}\frac{1}{5\,!}T_5(x)=x-\frac{x^3}{6}+\frac{x^5}{120}-\frac{1}{1920}(16x^5-20x^3+5x) &(2)\\ &= \frac{383}{384}x-\frac{5}{32}x^3. &(3)\end{aligned}$$

Поскольку $\varepsilon_2\approx\dfrac{5}{32}\cdot\dfrac{1}{2^2}>\varepsilon=10^{-4}$, то дальнейшее понижение степени невозможно. ♣

3.3. Полиномы Бернштейна

Построение АПНП представляется непростой задачей, в то же время равномерное приближение непрерывной функции полиномом может оказаться полезным во многих вычислительных задачах. В частности, рассматриваемые ниже полиномы Бернштейна позволяют получить равномерное приближение непрерывной функции с любой наперёд заданной точностью.

Вспомогательные алгебраические соотношения

а) $A(x) = \sum\limits_{k=0}^{n} C_n^k x^k (1-x)^{n-k} = 1$, поскольку $A(x) = \big(x + (1-x)\big)^n = 1$.

б) Обозначим

$$S_p(x) = \sum_{k=0}^{n} k^p C_n^k x^k (1-x)^{n-k}, \quad p \geq 1. \tag{1}$$

Заметим, что $S_0(x) = A(x)$, если принять соглашение: $0^0 = 1$. Возьмём производную от $S_p(x)$ и проделав несложные преобразования, получим:

$$\begin{aligned} S_p'(x) &= \sum_{k=0}^{n} C_n^k \Big(k^{p+1} x^{k-1} (1-x)^{n-k} - k^p (n-k) x^k (1-x)^{n-k-1} \Big) = \\ &= \frac{S_{p+1}(x)}{x} - \frac{1}{1-x}\big(n S_p(x) - S_{p+1}(x)\big), \end{aligned}$$

откуда получаем рекуррентное соотношение:

$$S_{p+1}(x) = x(1-x) S_p'(x) + nx S_p(x).$$

В частности, имеют место соотношения:

$$\begin{aligned} S_1(x) &= x(1-x) S_0'(x) + nx S_0(x) = nx. \\ S_2(x) &= x(1-x) S_1'(x) + nx S_1(x) = nx(1-x) + (nx)^2. \end{aligned}$$

Используя полученные формулы, получим представление для следующей суммы:

$$\begin{aligned} \sigma_n(x) &= \sum_{k=0}^{n} (k - nx)^2 C_n^k x^k (1-x)^{n-k} = \\ &= S_2(x) - 2nx S_1(x) + (nx)^2 S_0(x) = nx(1-x). \end{aligned} \tag{2}$$

Рассмотрим функцию $\sigma_n(x) = nx(1-x)$. Легко видеть, что на интервале $[0,1]$ она удовлетворяет соотношениям:

$$\sigma_n(x) \geq 0, \quad \max_{x \in [0,1]} \sigma_n(x) = \sigma_n\Big(\frac{1}{2}\Big) = \frac{n}{4}. \tag{3}$$

Пусть функция $f(\cdot)$ определена на $[a,b]$.

Определение. Полином вида

$$B_n(x) = \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big) C_n^k x^k (1-x)^{n-k}$$

называется полиномом Бернштейна для функции $f(\cdot)$. ♣

Теорема 1. *Если функция $f(\cdot)$ удовлетворяет на $[0,1]$ условию Липшица с константой L, то*

$$|B_n(x) - f(x)| \le \frac{L}{2\sqrt{n}}. \quad \clubsuit \tag{4}$$

Доказательство. Введём векторы $u^1, u^2 \in R^{n+1}$:

$$u^1 = \left\{ \left| \frac{k}{n} - x \right| \sqrt{C_n^k x^k (1-x)^{n-k}} \right\}_{k=0}^{n}, \quad u^2 = \left\{ \sqrt{C_n^k x^k (1-x)^{n-k}} \right\}_{k=0}^{n}.$$

Оценим методическую погрешность

$$\begin{aligned} \Delta_n &= |B_n(x) - f(x)| \le \sum_{k=0}^{n} \left| f\left(\frac{k}{n}\right) - f(x) \right| C_n^k x^k (1-x)^{n-k} \le \\ &\le L \sum_{k=0}^{n} \left| \frac{k}{n} - x \right| C_n^k x^k (1-x)^{n-k} = L u^1 \cdot u^2 \le \\ &\le L \|u^1\| \cdot \|u^2\| \le L \sqrt{\frac{1}{n^2} \sigma_n(x)} \sqrt{A(x)} \le \frac{L}{2\sqrt{n}}. \quad \clubsuit \end{aligned}$$

Доказательство.

- *Достаточность.* Предполагая существование точек альтернанса докажем, что $Q_n(\cdot)$ является АПНП для $f(\cdot)$.
  Пусть $L = \|f(\cdot) - Q_n(\cdot)\| = \sup\limits_{[a,b]} |f(x) - Q_n(x)|$ и $\mu = \min\limits_{i=\overline{0,n+1}} |f(x_i) - Q_n(x_i)|$. Из условия (1) следует, что $L = \mu$. С другой стороны, $E_n = \|f(\cdot) - \bar{Q}_n(\cdot)\| \ge \mu$. При $\mu = 0$ это очевидно. При $\mu > 0$, если допустить противное, т.е. $E_n < \mu$, то имеем: $\Delta Q_n(x) = Q_n(x) - \bar{Q}_n(x) \ne 0$, поскольку допущение $E_n < \mu$ равносильно $E_n < L$. С другой стороны

  $$\Delta Q_n(x) = \big(Q_n(x) - f(x)\big) - \big(\bar{Q}_n(x) - f(x)\big). \tag{5}$$

  В силу сделанного предположения $E_n < \mu$ знак правой части в узлах $\{x_i\}$ определяется знаком первого слагаемого, т.е. полином $\Delta Q_n(\cdot)$ меняет знак на каждом из интервалов $[x_i, x_{i+1}]$, $i = \overline{0,n}$, и поэтому имеет на $[a,b]$ по крайней мере $(n+1)$ корень, что невозможно.

  Поскольку $\bar{Q}_n(\cdot)$ является многочленом наилучшего равномерного приближения для $f(\cdot)$, то из установленного неравенства $E_n(f) \ge \mu$ следует $E_n(f) = \mu = L$, т.е. с учетом теоремы Хаара $Q_n(\cdot) = \bar{Q}_n(\cdot)$.

- *Необходимость.* Пусть $Q_n(\cdot) = \bar{Q}_n(\cdot)$, $L = \|f(\cdot) - Q_n(\cdot)\|$. Покажем существование точек Чабышевского альтернанса.

  Обозначим $\Delta(\cdot) = f(\cdot) - Q_n(\cdot)$. Пусть

  $$y_1 = \inf_{x \in [a,b]} \{x : \quad |\Delta(x)| = L\}.$$

  Такая точка существует ввиду непрерывности функции $\Delta(\cdot)$ и замкнутости интервала $[a,b]$. Будем считать для определённости, что $\Delta(y_1) = +L$.

Аналогично определим точки $y_2, \ldots, y_k$:

$$\begin{aligned} y_2 &= \inf_{x\in(y_1,b]}\{x: \quad \Delta(x) = -L\},\\ y_3 &= \inf_{x\in(y_2,b]}\{x: \quad \Delta(x) = +L\},\\ &\cdots\cdots\cdots\cdots\cdots\cdots\\ y_k &= \inf_{x\in(y_{k-1},b]}\{x: \quad \Delta(x) = (-1)^{k-1}L\} \end{aligned}$$

Ввиду непрерывности $\Delta(\cdot)$ имеем равенства:

$$\Delta(y_k) = (-1)^{k-1}L. \tag{6}$$

Процесс построения указанных точек продолжим до значения $y_m = b$ либо такого, что $|\Delta(x)| < L$ при $x \in (y_m, b]$. Если $m \geq n+2$, то теорема доказана.

Пусть $m < n+2$. Ввиду непрерывности $\Delta(\cdot)$ для любого $k: \ 1 < k \leq m$ найдётся z_{k-1} такое, что при $x \in [z_{k-1}, y_k)$ имеет место неравенство $|\Delta(x) < L|$. Положим $z_0 = a, \ z_m = b$.

Согласно построению на каждом интервале $[z_{i-1}, z_i], \ i = \overline{1,m}$ существуют точки (например $\{y_i\}$), в которых $\Delta(x) = (-1)^{i-1}L$ – см. (6), и нет точек, в которых $\Delta(x) = (-1)^i L$. (*)

Положим

$$v(x) = \prod_{j=1}^{m-1}(z_j - x), \quad Q_n^\varepsilon(\cdot) = Q_n(\cdot) + \varepsilon v(\cdot), \ \varepsilon > 0, \tag{7}$$

и рассмотрим функцию

$$\delta^\varepsilon(\cdot) = f(\cdot) - Q_n^\varepsilon(\cdot) = f(\cdot) - Q_n(\cdot) - \varepsilon v(\cdot) = \Delta(\cdot) - \varepsilon v(\cdot), \tag{8}$$

на каждом из отрезков $[z_{j-1}, z_j], \ j = \overline{1,m}$.

Пусть сначала $j = 1$. На $[z_0, z_1]$ согласно (7) и (6) $v(x) > 0$, поэтому

$$\delta^\varepsilon(x) \leq L - \varepsilon v(x) < L. \tag{9}$$

В то же время на нём $\Delta(x) > -L$ (см. (*), $i = 1$). Отсюда следует, что существует достаточно малое ε_1 такое, что имеет место неравенство:

$$|\Delta(x)| < L \quad \text{при } \varepsilon < \varepsilon_1, \ x \in [z_0, z_1), \tag{10}$$

а в точке $z = z_1$ по (7) $v(z_1) = 0$, поэтому с учетом (8) имеем:

$$|\delta^{\varepsilon_1}(z_1)| = |\Delta(z_1)| < L. \tag{11}$$

Из (10) и (11) следует, что *на всём интервале* $[z_0, z_1]$ верно неравенство (10).

Аналогично для всякого интервала $[z_{i-1}, z_i]$ найдётся своё ε_i, при котором неравенство (10) верно на этом интервале при $\varepsilon < \varepsilon_i$. Выбрав $\bar{\varepsilon} = \max\{\varepsilon_i\}$ получим, что на всём интервале $[a, b]$ имеет место (10), что противоречит предположению о том, что $Q_n(\cdot)$ является полиномом наилучшего приближения. ♣

Теорема 2. *Если* $f(\cdot) \in C^k[0,1]$, *то* $B_n^{(k)}(x) \to f^{(k)}(x)$ *при* $k \to \infty$ *равномерно на интервале* $[0,1]$. ♣

4. Приближение функций в гильбертовых пространствах

Будем рассматривать в дальнейшем только вещественные пространства. Напомним, что линейное пространство называется гильбертовым, если

- в нём введено скалярное произведение $(\cdot,\cdot)$;
- оно сепарабельно, т.е. в нём существует счётное всюду плотное множество.

Норма в гильбертовом пространстве H вводится как $\|h\|^2 = (h,h)$.
Произвольная система функций $\{\varphi_i\}$ в гильбертовом пространстве называется ортонормированной, если $(\varphi_i,\varphi_j) = \delta_{ij}$, где δ_{ij} – символ Кронекера.
Ортонормированная система $\{\varphi_i\}$ называется полной, если из $(\varphi_i,\psi) = 0$ следует $\psi = 0$.
В гильбертовом пространстве любая ортонормированная система не более чем счётна.
Приведём некоторые теоремы.

Теорема 1. *В гильбертовом пространстве существует не более чем счётная полная система функций.* ♣

Теорема 2. *В гильбертовом пространстве ряд Фурье для любого элемента этого пространства по полной ортонормированной системе элементов сходится к этому элементу.* ♣

Рассмотрим гильбертово пространство U и его подпространство H. Пусть в U определена функция $f(\cdot)$. Поставим задачу: *найти элемент $h_0 \in H$ (элемент наилучшего приближения), для которого выполнено равенство*

$$\|f(\cdot) - h_0(\cdot)\| = \inf_{h\in H} \|f(\cdot) - h(\cdot)\|.$$

Замечание. Здесь не предполагается, что H есть линейная оболочка, натянутая на конечное число элементов из U. ♣

4.1. Основные теоремы теории приближения

Теорема 3. *Если в H существует элемент наилучшего приближения h_0, то $(f-h_0,h) = 0$ для любого $h \in H$.* ♣

Доказательство. Пусть $h_0 \in H$ существует, но одновременно существует и $h_1 \in H$: $(f - h_0, h_1) = \alpha \neq 0$. Рассмотрим $h_2 = h_0 + \alpha h_1$, считая $\|h_1\| = 1$. Имеем:

$$\|f - h_2\|^2 = \|f - h_0\|^2 - 2\alpha(f - h_0, h_1) + \alpha^2 = \|f - h_0\|^2 - \alpha^2,$$

что противоречит определению h_0. ♣

Теорема 4. *В подпространстве $H \subset U$ не может существовать двух элементов наилучшего приближения.* ♣

Доказательство. Пусть это не так, т.е. существуют *два* элемента наилучшего приближения $h_0, h_0',\ h_0 \neq h_0'$; тогда по теореме 3 имеем

$$\begin{cases} (f - h_0, h) &= 0, \\ (f - h_0', h) &= 0 \quad \text{для любого } h \in H, \text{ например } h = \Delta h = h_0 - h_0'. \end{cases} \tag{1}$$

Тогда

$$\|\Delta h\|^2 = \big((h_0 - f) + (f - h_0'), \Delta h\big) = (h_0 - f, \Delta h) + (f - h_0', \Delta h) = 0$$

в соответствии с (1), откуда $\|\Delta h\| = 0, \Rightarrow \Delta h = 0$, что противоречит предположению. ♣

Пусть теперь $H = H_n = \{h : \ h = \sum_{i=1}^{n} \alpha_i h_i\}$, $\{h_i\}$ $i = \overline{1,n}$ – *линейно независимы*, причём $\{h_i\}$ $i = \overline{1,\infty}$ – *полная* система [1] элементов в *полном*[2] гильбертовом пространстве U. На основании предыдущих рассмотрений элемент наилучшего приближения существует и единствен. Рассмотрим вопрос о построении элемента наилучшего приближения h_0.
Поскольку $h_0 \in H_n$, то на основании теоремы 3 должны быть выполнены равенства

$$(f - h_0, h_i) = 0, \quad h_i \in H_n, \ i = \overline{1,n}. \tag{2}$$

Это есть СЛАУ $Ax = b$ с матрицей Грама $A = \{(h_i, h_j)\}$, поэтому $\det A \neq 0$ и, следовательно, для любой функции $f(\cdot)$ существует единственный элемент наилучшего приближения $h_0 \in H_n$:

$$h_0 = \sum_{i=1}^{n} \alpha_i h_i.$$

Уклонение этого элемента от функции $f(\cdot)$ может быть представлено в виде:

$$\delta^2 = \|f - h_0\|^2 = (f - h_0, f - h_0) = (f, f) - \sum_{i=1}^{n} \alpha_i (h_i, f).$$

При условии *ортонормированности* элементов $\{h_i\}$ в соответствии с равенством Парсеваля имеет место оценка

$$\delta^2 = \|f - h_0\|^2 = \|f\|^2 - \sum_{i=1}^{n} \alpha_i^2.$$

4.2. Приближения алгебраическими многочленами

Пусть $U = L_2[a,b]$. Скалярное произведение и норма, как известно, задаются в этом пространстве формулами:

$$(f,g) = \int_a^b f(x)g(x)\,dx, \quad \|f\| = \int_a^b f^2(x)\,dx.$$

Пусть $p(x) \geq 0$, причём $p(x) = 0$ не более, чем на множестве меры нуль. Введём пространство $L_2(p)$: считаем, что $f \in L_2(p)$, если существует интеграл

$$\int_a^b p\,(x) f^2(x)\,dx.$$

Очевидно, что это пространство является линейным. Скалярное произведение в $L_2(p)$ зададим равенством:

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

Сходимость в пространстве L_2 – известная в анализе *сходимость в среднем*, а в $L_2(p)$ – *сходимость в среднем с весом* $p(x)$. В дальнейшем будем рассматривать пространство $L_2(p)$ как более общий случай.

[1] ортонормированная система элементов $\{\varphi_i\}$ пространства U называется полной, если не существует элемента $h \in U, \ h \neq 0$ такого, что $(\varphi_i, h) = 0 \ \forall i$

[2] пространство называется полным, если любая фундаментальная (сходящаяся в себе) последовательность элементов этого пространства сходится к некоторому элементу этого пространства

Функции $1, x, x^2, \ldots, x^n, \ldots$, взятые в любом числе, линейно независимы в $L_2(p)$. Множество $\mathcal{P}_n$ полиномов степени не выше n можно рассматривать как линейную оболочку, натянутую на функции $1, x, x^2, \ldots, x^n$, поэтому на основании предыдущих результатов можем утверждать, что в $\mathcal{P}_n$ существует единственный полином $\bar{P}_n(\cdot)$, дающий наилучшее приближение функции $f(\cdot) \in L_2(p)$ по метрике этого пространства, т.е.

$$\delta_n^2 = \|f - \bar{P}_n\|^2 = \int_a^b p\,(x)\big(f(x) - \bar{P}_n(x)\big)^2\,dx = \inf_{P_n \in \mathcal{P}_n} \int_a^b p\,(x)\big(f(x) - \bar{P}_n(x)\big)^2\,dx. \tag{1}$$

Если ввести обозначения

$$s_k = \int_a^b p\,(x)x^k\,dx, \quad m_k = \int_a^b p\,(x)f(x)x^k\,dx, \tag{2}$$

то коэффициенты $\{a_i\}$ полинома $\bar{P}_n(\cdot)$ могут быть найдены как решение СЛАУ

$$\big(f(x) - P_n(x), x^k\big) = 0, \quad k = \overline{0, n}, \tag{3}$$

или с использованием введенных обозначений

$$\left\{\begin{array}{lcl} a_0 s_0 + a_1 s_1 + \cdots + a_n s_n & = & m_0, \\ a_0 s_1 + a_1 s_0 + \cdots + a_n s_{n+1} & = & m_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots & & \\ a_0 s_n + a_1 s_{n+1} + \cdots + a_n s_{2n} & = & m_n. \end{array}\right.$$

Это есть система с определителем Грама, построенная по системе линейно независимых элементов $\{x^k\}$ и потому имеет единственное решение для любой функции f, однако с ростом n её обусловленность ухудшается, поэтому выгоднее выбирать ортонормированную систему полиномов. Метод ортогонализации позволяет построить такую систему:

$$\{Q_0(x), Q_1(x), \ldots, Q_n(x)\}: \quad \int_a^b p\,(x)Q_i(x)Q_j(x)\,dx = \delta_{ij}.$$

Имеет место **Теорема 1.** *Ортонормированная система полиномов в пространстве* $L_2(p)$ определяется единственным образом. ♣

Доказательство. Покажем сначала, что существует единственный полином $Q_n \in \mathcal{P}_n^1$ (т.е. полином вида $Q_n(x) = a_0 + a_1 x + \ldots + a_{n-1}x^{n-1} + x^n$), который ортогонален относительно введенного скалярного произведения в $L_2(p)$ всем многочленам из степени $(n-1)$. Запишем эти условия:

$$\left\{\begin{array}{l} a_0 s_0 + a_1 s_1 + \ldots + a_{n-1} s_{n-1} + s_n = 0 \\ a_0 s_1 + a_1 s_2 + \ldots + a_{n-1} s_n + s_{n+1} = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_0 s_{n-1} + a_1 s_n + \ldots + a_{n-1} s_{2n-2} + s_{2n-1} = 0 \end{array}\right.$$

Эта СЛАУ имеет единственное решение, поскольку матраца её является матрицей Грама, построенной по линейно независимым элементам. Наряду с $Q_n(\cdot)$ этим же свойством ортогональности к множеству $\mathcal{P}_{n-1}$ обладает, очевидно и полином $\alpha Q_n(\cdot)$. Покажем, что других ортогональных к $\mathcal{P}_{n-1}$ полиномов из $\mathcal{P}_n$ нет.

Действительно, если $\hat{Q}_n(x) = \alpha_n x^n + \ldots \perp \mathcal{P}_{n-1}$, то $q = \hat{Q}_n - \alpha_n Q_n$ имеет степень не выше $n-1$ и обладает свойством ортогональности ко всем полиномам из $\mathcal{P}_{n-1}$. Следовательно, $q \perp q$, т.е. $q = 0$ и потому $\hat{Q}_n = \alpha_n Q_n$ ♣.

Определение. Многочлен $\alpha_n Q_n$ $\alpha \neq 0$ будем называть ортогональным (относительно веса $p(x)$ и отрезка $[a,b]$).♣.

Теорема 2. Корни ортогонального многочлена вещественны, просты и расположены на интервале (a,b).♣

Доказательство. Поскольку

$$\int_a^b p(x)Q_n(x)dx = 0,$$

то $Q_n(x)$ имеет на (a,b) точки перемены знака, т.е. корни нечётной кратности, расположенные внутри (a,b). Пусть это $x_1, x_2, \ldots, x_m,\ m < n$ (если $m = n$ – утверждение верно). Образуем полином $q(x) = (x-x_1)(x-x_2)\ldots(x-x_m)$ и рассмотрим полином $q(x)Q_n(x)$. Это есть полином, у которого на $[a,b]$ все корни чётной кратности, а потому он сохраняет знак на $[a,b]$. Но тогда имеем

$$\int_a^b p(x)q(x)Q_n(x)dx \neq 0,$$

т.е. $Q_n(\cdot)$ не ортогонален полиному $q(\cdot) \in \mathcal{P}_{n-1}$, что противоречит определению ортогонального полинома.♣

Теорема 3. Пусть $Q_n(\cdot)$ – ортогональный многочлен на $[-a,a]$ $a > 0$ и при этом $p(\cdot)$ – чётная функция. Тогда $Q_n(\cdot)$ чётен или нечётен вместе со степенью n.♣

Доказательство. Представим $Q_n(\cdot)$ в виде: $Q_n(x) = r_n(x) + q_{n-1}(x)$, где r_n и q_n содержат члены одинаковой чётности с n и $n-1$ соответственно. Поскольку $q_{n-1} \in \mathcal{P}_{n-1}$, то $q_{n-1} \perp Q_n$ и поэтому

$$0 = \int_{-a}^a p(x)Q_n(x)q_{n-1}(x)dx = \int_{-a}^a p(x)r_n(x)q_{n-1}(x)dx + \int_{-a}^a p(x)q_{n-1}^2(x)dx.$$

Но

$$\int_{-a}^a p(x)r_n(x)q_{n-1}(x)dx = 0$$

как интеграл от нечётной функции, а следовательно и

$$\int_{-a}^a p(x)q_{n-1}^2(x)dx = 0,$$

откуда $q_{n-1}(x) = 0$, т.е. $Q_n(x) = r_n(x)$. ♣

Укажем некоторые частные случаи ортогональных систем многочленов.

- **Многочлены Якоби:** $p\,(x) = (1-x)^a(1+x)^b,\ a,b > -1; x \in [-1,1]$

$$P_n^{(a,b)}(x) = \frac{(-1)^n}{n!2^n}(1-x)^{-a}(1+x)^{-b} \cdot \frac{d^n}{dx^n}\big[(1-x)^{a+n}(1+x)^{b+n}\big].$$

- **Многочлены Лежандра:** $p\,(x) = 1,\ x \in [-1,1]$,

$$L_n(x) = \frac{1}{n!2^n} \cdot \frac{d^n}{dx^n}\big[(1-x^2)^n\big].$$

(Частный случай полиномов Якоби при $a = b = 0$)

- **Многочлены Чебышева первого рода:** $p\,(x) = (1-x^2)^{-\frac{1}{2}},\ x \in [-1,1]$,

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

- **Многочлены Чебышева второго рода:** $p\,(x) = (1-x^2)^{\frac{1}{2}},\ x \in [-1,1]$,

$$U_n(x) = \frac{\sin\big[(n+1)\arccos x\big]}{\sqrt{1-x^2}}.$$

- **Многочлены Лагерра:** $p\,(x) = x^a e^{-x},\ a > -1,\ x \in [0,+\infty)$

$$L_n^{(a)}(x) = (-1)^n x^{-a} e^x \frac{d^n}{dx^n}\big[x^{a+n}e^{-x}\big].$$

- **Многочлены Эрмита:** $p\,(x) = e^{-x^2},\ x \in (-\infty,+\infty)$,

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}\big[e^{-x^2}\big].$$

Для выбранной системы ортогональных многочленов $\{Q_0(x), Q_1(x), \ldots, Q_n(x)\}$ многочлен наилучшего приближения $\bar{Q}_n(x) \in \mathcal{P}_n$ запишется в виде

$$\bar{Q}_n(x) = c_0 Q_0(x) + c_1 Q_1(x) + \cdots + c_n Q_n(x),$$

причём коэффициенты $\{c_i\}$ на основании общей теории вычисляются по формулам

$$c_k = \frac{\displaystyle\int_a^b p\,(x) f(x) Q_k(x)\,dx}{\displaystyle\int_a^b p\,(x) Q_k^2(x)\,dx}.$$

Величина отклонения δ_n наилучшего приближения от аппроксимируемой функции определится по формуле

$$\delta_n^2 = \int_a^b p\,(x) f^2(x)\,dx - \sum_{k=0}^{n} c_k^2 \int_a^b p\,(x) Q_k^2(x)\,dx.$$

Задачи, предлагаемые для лучшего усвоения материала данной главы.

Для функции $y(x) = x^4 + x^3 + 2x^2 + 3x - 4$

1. Построить алгебраический полином не выше третьей степени $p^0(\cdot) \in \mathcal{P}_3$, являющийся наилучшим среднеквадратичным приближением для $y(\cdot)$, т.е.

$$\int_{-1}^{1} \big(p^0(t) - y(t)\big)^2 dt = \min_{p(\cdot)\in\mathcal{P}_3} \int_{-1}^{1} \big(p(t) - y(t)\big)^2 dt$$

и вычислить величину методической погрешности $r = \|y(\cdot) - p^0(\cdot)\|_{L_2[-1,1]}$.

Предлагается при выполнении задания использовать полиномы *Лежандра*: $P_n(x) = \dfrac{1}{2^n n!}\dfrac{d^n}{dx^n}(x^2-1)^n$, обладающие свойством ортогональности в пространстве $L_2[-1,1]$.

2. Для $y(x)$ построить полином $Q_3(x) \in \mathcal{P}_3$ путём замены полинома $q(x) = x^4$ полиномом его наилучшего равномерного приближения на интервале $[-1,1]$ и оценить погрешность $\|y(x) - Q_3(x)\|_{C[-1,1]}$.

Литература

1. Бахвалов Н.С., Жидков Н.П., Кобельков Г.М. Численные методы. – Физматлит, М.-СПб, 2000.

2. Березин И.С, Жидков Н.П. Методы вычислений, т.2 – М., 1962.

3. Воеводин В.В. Вычислительные основы линейной алгебры. – М., 1977.

4. Крылов В.И., Бобков В.В., Монастырный П.И. Вычислительные методы. – М., 1976.

5. Рябенький В.С. Введение в вычислительную математику. – М., 1994.

6. Фаддеев Д.К., Фаддеева В.Н. Вычислительные методы линейной алгебры. – М., 1963.

7. Амосов А.А., Дубинский Ю.А., Копчёнова Н.В. Вычислительные методы для инженеров. – М., 1994.

8. Мысовских И.П. Лекции по методам вычислений. – СПб., 1998.

Вопросы к экзамену по МЕТОДАМ ВЫЧИСЛЕНИЙ

ЧИСЛЕННОЕ РЕШЕНИЕ УРАВНЕНИЙ

1. Решение скалярных уравнений. Метод Чебышева.
2. Метод Чебышева решения скалярных уравнений. Теорема сходимости.
3. Решение скалярных уравнений. Метод итераций. Теорема сходимости.
4. Границы корней алгебраического уравнения.
5. Системы линейных алгебраических уравнений (СЛАУ). Число обусловленности, его свойства.
6. Точные методы решения СЛАУ. Методы Гаусса и Жордана. Применение к задаче построения обратной матрицы.
7. Метод простой итерации. Теорема сходимости.
8. Приведение СЛАУ с положительно определенной матрицей к виду, пригодному для применения метода простой итерации. Метод Зейделя.
9. Метод прогонки решения СЛАУ с трехдиагональной матрицей.
10. Метод Ньютона решения систем нелинейных уравнений.

МИНИМИЗАЦИЯ КВАДРАТИЧНОЙ ФУНКЦИИ

1. Постановка задачи минимизации квадратичной функции. Существование и единственность точки минимума.
2. Свойства минимизирующей последовательности.
3. Одношаговые градиентные методы. Метод наискорейшего градиентного спуска.
4. Одношаговые градиентные методы. Теорема о сходимости минимизирующей последовательности.
5. Многошаговые градиентные методы. Метод наискорейшего градиентного спуска.
6. Стационарный s-шаговый метод спуска.
7. Оптимальный многошаговый стационарный метод спуска.
8. Метод сопряженных направлений. Метод A-ортогонализации базиса.
9. Метод сопряжённых градиентов.