Е. Е. Тыртышников

МЕТОДЫ ЧИСЛЕННОГО АНАЛИЗА

Москва 2006

Оглавление

Предисловие

Эта книга — существенно дополненная, переработанная и, уверен, улучшенная версия опубликованных мною в 1994 году лекций под названием “Краткий курс численного анализа”. Когда я готовил первую версию, то точно знал, почему это делаю. Несмотря на то, что есть много хороших и очень хороших книг по методам вычислений и смежным вопросам, я помню, что не понимал этого, будучи студентом. Поэтому я старался написать такие лекции, которые хотел бы иметь в то время, когда был студентом.

Таким образом, получается, что я писал книгу прежде всего для себя. Методы численного анализа, как я их вижу — это поразительное синергетическое сочетание красивых и глубоких идей и теорий из разных разделов математики: анализа, теории функций, теории операторов, теории приближений, линейной алгебры и матричного анализа. Предмет является синтетическим по сути. Поэтому он особенно труден для освоения и изложения, которое чаще всего сводится к пространному описанию вычислительных рецептов или разбору многочисленных примеров и приложений. Все это важно и нужно. Но очень хочется иметь руководство другого типа — с акцентом именно на идеях, достаточно краткое для того, чтобы не затуманить красоту идей и показать богатство связей с замечательными математическими теориями (детальное изучение которых требует, конечно, отдельных курсов, немалых усилий и времени), и в существенной части замкнутое — с полными доказательствами или указаниями на то, что нужно делать для их завершения.

Из сказанного ясно, что книга адресована математикам и тем, кто специализируется в области прикладной математики. Но думаю, что она будет полезной инженерам — и, возможно, даже в большей степени, так как дает шанс соприкоснуться, пусть и конспективно, с прекрасными разделами математики, которые как фундамент поддерживают разросшееся здание численного анализа.

Наверное, выбранные акценты отражают вкусы автора. Но они коррелируют и с современными тенденциями развития предмета. В книге немало

мест, которые пока еще нельзя найти в учебниках, а иногда и в монографиях. Поэтому надеюсь, что книга будет интересна как тем, кто учится, так и тем, кто учит.

Мне посчастливилось иметь хороших наставников, без них все было бы не так, в том числе и эта книга была бы другой. Прежде всего хочу поблагодарить Валентина Васильевича Воеводина — за постоянное внимание, всегда ценные, своевременные советы и поддержку во всем. Мне приятно выразить признательность Хакиму Дододжановичу Икрамову и Алексею Георгиевичу Свешникову. Не могу не отметить, что всегда ощущал заботу и поддержку на факультете вычислительной математики и кибернетики Московского университета им. М. В. Ломоносова — как раньше, во времена студенчества, аспирантуры, работы ассистентом, так и теперь, в качестве профессора.

Особая благодарность Гурию Ивановичу Марчуку и Валентину Павловичу Дымникову — за счастье работать в лучшем институте Российской академии наук — Институте вычислительной математики. А также и за их труд по воспитанию студентов, давший жизнь замечательным кафедрам в Московском физико-техническом институте и в Московском университете.

По-прежнему благодарен первым читателям “Краткого курса...” Сергею Анатольевичу Горейнову и Николаю Леонидовичу Замарашкину. Как и раньше, очень вдохновляет, что у них находится много поводов для того, чтобы поделиться замечаниями и предложениями — не только по поводу данной книги, но и по самым разным вопросам из излюбленной нами области матричных методов и численного анализа.

Глава 1

1.1 Метрическое пространство

Мы хотим научиться вычислять различные математические объекты. Но наши алгоритмы будут давать некоторые другие объекты, которые, как мы надеемся, будут "близки" к искомым. Таким образом, нам нужно уметь определять "близость" для различных объектов.

В общем виде понятие "близости" можно ввести с помощью *метрики*, или *расстояния*. Пусть M — непустое множество и $\rho(x, y)$ — неотрицательная функция, определенная для всех $x, y \in M$ и обладающая следующими свойствами:

(1) $\rho(x, y) \geq 0 \quad \forall\, x, y \in M; \qquad \rho(x, y) = 0 \quad \Leftrightarrow \quad x = y;$
(2) $\rho(x, y) = \rho(y, x) \quad \forall\, x, y \in M \quad$ (симметричность);
(3) $\rho(x, y) \leq \rho(x, y) + \rho(y, z) \quad \forall\, x, y, z \in M \quad$ (неравенство треугольника).

Такая функция $\rho(x, y)$ называется *метрикой*, или *расстоянием* (между x и y), а M при этом называется *метрическим пространством*.

Хорошо знакомый пример метрического пространства: M — все вещественные числа и $\rho(x, y) \equiv |x - y|$.

Другой поучительный пример: M — произвольное непустое множество, $\rho(x, y) = 0$ при $x = y$ и 1 при $x \neq y$.

1.2 Полезные определения

Последовательность $x_n \in M$ называется *сходящейся*, если $\exists\ x \in M : \lim\limits_{n\to\infty} \rho(x_n, x) = 0$. Легко доказать, что такая точка x может быть только одна; x называется *пределом* для x_n. Обозначение: $x = \lim\limits_{n\to\infty} x_n$.

Последовательность $x_n \in M$ называется *последовательностью Коши* (*фундаментальной*, *сходящейся в себе*), если

$$\forall\, \varepsilon > 0 \quad \exists\, N : \quad n, m \geq N \Rightarrow \rho(x_n, x_m) \leq \varepsilon.$$

Метрическое пространство M называется *полным*, если в нем любая последовательность Коши является сходящейся.

Множество $C \subset M$ называется *замкнутым*, если для любой сходящейся последовательности $x_n \in C$ ее предел принадлежит C.

Точка $x \in M$ называется *предельной точкой* множества $S \subset M$, если $\exists\ x_n \in S, x_n \neq x : x_n \to x$. *Замыканием* множества называется его объединение со всеми его предельными точками.

Замкнутое множество $C \subset M$ называется *компактным*, если в нем из любой последовательности можно выделить сходящуюся подпоследовательность.

Множество $B(a;r) \equiv \{x \in M : \rho(x,a) < r\}$ называется *открытым шаром* с центром в точке a и радиусом r. Множество $\overline{B}(a;r) \equiv \{x \in M : \rho(x,a) \leq r\}$ называется *замкнутым шаром*.

Множество $O \subset M$ называется *открытым*, если в нем любая точка x содержится вместе с некоторым открытым шаром $B(x,r)$.

Множество $S \subset M$ называется *ограниченным*, если S целиком содержится в некотором шаре.

Числовая функция $f(x), x \in M$, называется *непрерывной* в точке x_0, если для любой последовательности $x_n \neq x_0$, сходящейся к x_0, имеет место равенство $f(x_0) = \lim\limits_{n\to\infty} f(x_n)$.

1.3 Вложенные шары

Теорема 1.3.1 *Пусть в полном метрическом пространстве M имеются замкнутые шары $\overline{B}(a_n, r_n)$ такие, что:*

(1) $\overline{B}(a_1;r_1) \supset \overline{B}(a_2;r_2) \supset \ldots$;

(2) $\lim\limits_{n\to\infty} r_n = 0$.

Тогда пересечение этих шаров $P = \bigcap\limits_{n=1}^{\infty} \overline{B}(a_n, r_n)$ не пусто и содержит ровно одну точку.

Условие (2) существенно. Чтобы это показать, построим "экзотическое" метрическое пространство:

$$M = \{1, 2, \ldots\}, \quad \rho(m,n) = \begin{cases} 0, & m = n; \\ 1 + \max\left(\frac{1}{2^m}, \frac{1}{2^n}\right), & m \neq n. \end{cases}$$

Нетрудно проверить, что ρ является метрикой (то есть выполнены свойства (1)–(3) из п. 1.1). В пространстве M любая последовательность Коши

состоит из одинаковых членов начиная с некоторого номера и поэтому является сходящейся $\Rightarrow$ M — полное метрическое пространство. Заметим, что замкнутые вложенные шары

$$\overline{B}\left(1,\, 1+\tfrac{1}{2}\right) \supset \overline{B}\left(2,\, 1+\tfrac{1}{2^2}\right) \supset \overline{B}\left(3,\, 1+\tfrac{1}{2^3}\right) \supset \ldots$$

имеют пустое пересечение.

1.4 Нормированное пространство

Пусть V — вещественное или комплексное векторное (линейное) пространство, на котором определена неотрицательная функция $f(x)$ такая, что:

(1) $f(x) \geq 0 \quad \forall\, x \in V; \quad f(x) = 0 \Leftrightarrow x = 0;$

2) $f(\alpha x) = |\alpha| f(x), \quad \alpha$ — число, $x \in V;$

(2) $f(x+y) \leq f(x) + f(y)$ (неравенство треугольника).

Такая функция $f(x)$ называется *нормой* вектора x, а пространство V называется при этом *нормированным*. Обозначение: $\|x\| \equiv f(x)$.

Для любого нормированного пространства метрика вводится так:

$$\rho(x, y) \equiv \|x - y\|.$$

Сходимость и другие понятия, рассмотренные в п. 1.2, определяются именно для этой метрики. Полное нормированное пространство называется *банаховым*.

1.5 Популярные векторные нормы

Пусть $V = \mathbb{C}^n$ (или $\mathbb{R}^n$). Если $p \geq 1$ и $x = [x_1, \ldots, x_n]^T$, то положим $\|x\|_p \equiv \left(\sum\limits_{i=1}^{n} |x_i|^p \right)^{1/p}$ (p-норма вектора x).

Теорема 1.5.1 *$\|x\|_p$ является нормой.*

Свойства (1) и (2) нормы очевидны. Свойство (3) представляет собой неравенство Минковского, которое мы докажем ниже.

Лемма 1.5.1 *Пусть числа p, q образуют гельдеровскую пару в том смысле, что*

$$p,\, q > 1, \qquad \frac{1}{p} + \frac{1}{q} = 1.$$

Тогда для любых a, $b \geq 0$

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Доказательство легко получить, используя *вогнутость* логарифмической функции. Вогнутость (свойство, противоположное *выпуклости*) означает, что $\forall u, v > 0$

$$\alpha \log u + \beta \log v \leq \log (\alpha u + \beta v),$$

$$\forall \alpha, \beta \geq 0, \quad \alpha + \beta = 1.$$

Теорема 1.5.2 (Неравенство Гельдера) *Для любой гельдеровской пары* p, q *и любых векторов* $x = [x_1, \ldots, x_n]^T$, $y = [y_1, \ldots, y_n]^T$

$$\left| \sum_{i=1}^{n} x_i y_i \right| \leq \|x\|_p \|y\|_q.$$

Доказательство. Если $x = 0$ или $y = 0$, то неравенство очевидно. Поэтому рассмотрим ненулевые векторы x, y и положим

$$\tilde{x} = x/\|x\|_p, \qquad \tilde{y} = y/\|y\|_q.$$

Тогда $\|\tilde{x}\|_p = \|\tilde{y}\|_q = 1$. Согласно лемме 1.5.1,

$$|\tilde{x}_i| \, |\tilde{y}_i| \leq \frac{|\tilde{x}_i|^p}{p} + \frac{|\tilde{y}_i|^q}{q}, \quad i = 1, \ldots, n.$$

Складывая эти неравенства, получаем

$$\sum_{i=1}^{n} |\tilde{x}_i| \, |\tilde{y}_i| \leq \frac{\|\tilde{x}\|_p^p}{p} + \frac{\|\tilde{y}\|_q^q}{q} = 1. \quad \square$$

Теорема 1.5.3 (Неравенство Минковского)

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

Доказательство.

$$\|x + y\|_p^p = \sum_{i=1}^{n} |x_i + y_i|^p \leq \sum_{i=1}^{n} |x_i + y_i|^{p-1} \left(|x_i| + |y_i| \right)$$

(далее в силу неравенства Гельдера)

$$\leq \left(\sum_{i=1}^{n} \left(|x_i + y_i|^{p-1} \right)^q \right)^{\frac{1}{q}} \left(\|x\|_p + \|y\|_p \right).$$

Остается учесть, что $(p-1)q = p$. $\square$

К p-нормам относят также следующие нормы:

$$\|x\|_\infty \equiv \max_{1\le i\le n} |x_i|, \quad \|x\|_1 \equiv \sum_{i=1}^{n} |x_i|.$$

Легко доказать, что это действительно нормы и для них остаются в силе неравенства Гельдера (при $p = 1$ и $q = \infty$) и Минковского. Кроме того,

$$\|x\|_\infty = \lim_{p\to\infty} \|x\|_p.$$

Среди p-норм наиболее часто приходится иметь дело со значениями $p = 1, 2$ и ∞. Вот как выглядят единичные сферы для этих норм при $n = 2$:

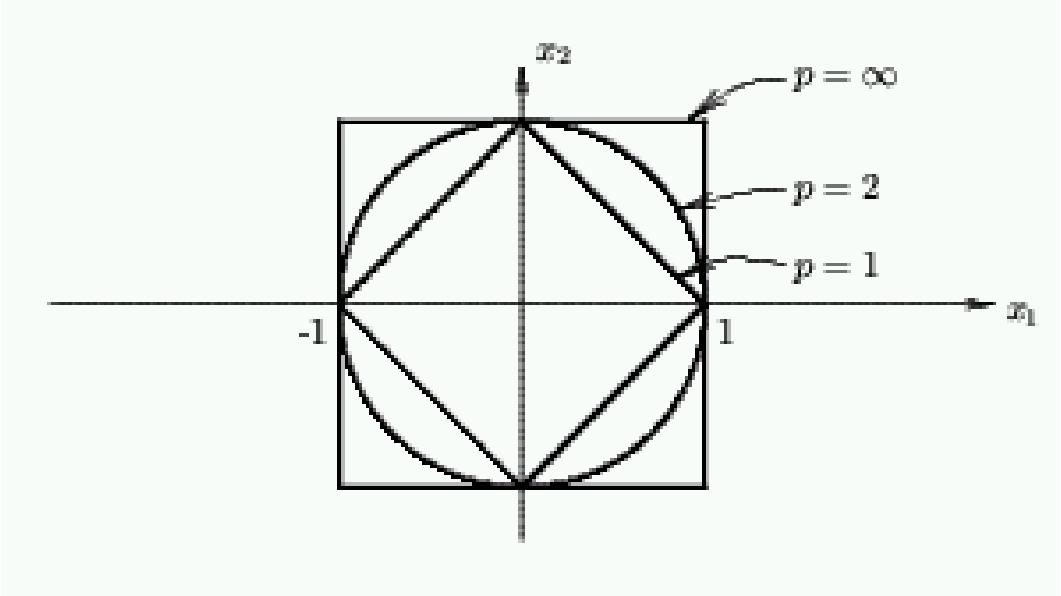


Рисунок 1.1. Единичные сферы при $p = 1, 2, 3$.

Почему 1-норму иногда называют октаэдрической, а ∞-норму — кубической?

1.6 Матричные нормы

Матрицы одинаковых размеров образуют конечномерное векторное пространство. Поэтому норма для матриц, в принципе, может порождаться любой векторной нормой. Однако под матричной нормой понимается нечто бóльшее.

Пусть каждой матрице A поставлено в соответствие число $\|A\|$. Тогда $\|A\|$ называется *матричной нормой*, если:

(1) на векторном пространстве матриц одинаковых размеров $\|A\|$ является векторной нормой;

(2) для любых матриц A и B, допускающих умножение,

$$\|AB\| \le \|A\|\|B\| \quad \text{(свойство мультипликативности)}.$$

Один из важнейших примеров матричной нормы — это *норма Фробениуса*:

$$\|A\|_F \equiv \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2}, \qquad A \in \mathbb{C}^{m\times n}.$$

В отечественной литературе $\|A\|_F$ называют *евклидовой нормой* матрицы и обозначают $\|A\|_E$.

Доказательство мультипликативности нормы Фробениуса. Пусть

$$A = [a_1, \ldots, a_n] \in \mathbb{C}^{m\times n} \qquad \text{и} \qquad B = \begin{bmatrix} b_1^T \\ \ldots \\ b_n^T \end{bmatrix} \in \mathbb{C}^{n\times k}.$$

Тогда

$$AB = a_1 b_1^T + \cdots + a_n b_n^T.$$

В силу неравенства треугольника

$$\begin{aligned} \|AB\|_F &\le \|a_1 b_1^T\|_F + \cdots + \|a_n b_n^T\|_F \\ &= \|a_1\|_2 \|b_1\|_2 + \cdots + \|a_n\|_2 \|b_n\|_2 \\ &\le \left(\sum_{i=1}^{n} \|a_i\|_2^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} \|b_i\|_2^2 \right)^{\frac{1}{2}} = \|A\|_F \|B\|_F. \ \square \end{aligned}$$

1.7 Эквивалентные нормы

Нормы $\|\cdot\|_*$ и $\|\cdot\|_{**}$ на одном векторном пространстве V называются *эквивалентными*, если существуют константы c_1, $c_2 > 0$ такие, что

$$c_1 \|x\|_* \le \|x\|_{**} \le c_2 \|x\|_* \qquad \forall\, x \in V.$$

Понятно, что эквивалентные нормы равноценны с точки зрения сходимости. Следующая теорема — фундаментальный факт, справедливый для конечномерных пространств.

Теорема 1.7.1 *Любые две нормы на конечномерном пространстве эквивалентны.*

Доказательство. В основе доказательства лежат три важных факта:

(1) компактность единичной сферы $S_n = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ относительно 2-нормы;

(2) непрерывность любой нормы $\|\cdot\|_*$ относительно 2-нормы;

(3) теорема Вейерштрасса о том, что функция, непрерывная на компактном множестве, является ограниченной.

Компактность единичной сферы. Рассмотрим последовательность $x^{(k)} = [x_1^{(k)}, \ldots, x_n^{(k)}]^T \in S_n$. Последовательность первых координат $x_1^{(k)}$ принадлежит отрезку $[-1, 1]$, и, значит, имеет сходящуюся подпоследовательность: $x_1^{(k_1)} \to x_1$, где k_1 стремится к бесконечности, пробегая какую-то подпоследовательность номеров 1, 2, ... Рассмотрим подпоследовательность $x^{(k_1)}$ и вторые координаты $x_2^{(k_1)}$. Пусть $x_2^{(k_2)} \to x_2$, где k_2 стремится к бесконечности, пробегая подпоследовательность первой подпоследовательности. Далее рассмотрим подпоследовательность $x^{(k_2)}$, третьи координаты $x_3^{(k_2)}$, и т.д. В итоге получим подпоследовательность векторов $x^{(k_n)}$ такую, что все координатные последовательности являются сходящимися:

$$x_i^{(k_n)} \to x_i, \quad 1 \le i \le n.$$

Пусть $x \equiv [x_1, \ldots, x_n]^T$. Тогда

$$\left\| x^{(k_n)} - x \right\|_2 = \left(\sum_{i=1}^{n} \left| x_i^{(k_n)} - x_i \right|^2 \right)^{\frac{1}{2}} \to 0.$$

Непрерывность нормы. Пусть $x^{(k)} \to x$ (значит, $\left\| x^{(k)} - x \right\|_2 \to 0$). Мы хотим доказать, что $\|x^{(k)}\|_* \to \|x\|_*$. Обозначим через $e_1, \ldots, e_n$ столбцы единичной матрицы. Тогда

$$\begin{aligned} \left| \|x^{(k)}\|_* - \|x\|_* \right| &\le \left\| x^{(k)} - x \right\|_* \\ &= \left\| \sum_{i=1}^{n} \left(x_i^{(k)} - x_i \right) e_i \right\|_* \le \sum_{i=1}^{n} \left| x_i^{(k)} - x_i \right| \|e_i\|_* \\ &\le \left\| x^{(k)} - x \right\|_2 \left(\sum_{i=1}^{n} \|e_i\|_*^2 \right)^{\frac{1}{2}} \to 0. \end{aligned}$$

Теорема Вейерштрасса. Пусть M — компактное множество и числовая функция $f(x)$ непрерывна в любой его точке. Предположим, что функция $f(x)$ не ограничена. Тогда существует последовательность $x^{(k)}$ такая, что

$\left|f(x^{(k)})\right| \geq k$. В силу компактности существует сходящаяся подпоследовательность: $x^{k'} \to x$. По непрерывности $f\left(x^{(k')}\right) \to f(x)$. Однако, этого не может быть, так как

$$k \leq \left|f\left(x^{(k')}\right)\right| \leq |f(x)| + \left|f\left(x^{(k')}\right) - f(x)\right|$$

для любого k. Следовательно, функция $f(x)$ ограничена.

Итак, функция $\|x\|_*$ непрерывна на компактном множестве S_n относительно 2-нормы и поэтому является ограниченной $\Rightarrow$ для некоторого $c_2 > 0$ имеем $\|x\|_* \leq c_2$. Функция $1 \,/\, \|x\|_*$ также является непрерывной на S_n $\Rightarrow$ для некоторого $c_1 > 0$ имеем $1 \,/\, \|x\|_* \leq c_1^{-1}$. Следовательно,

$$c_1 \leq \|x\|_* \leq c_2 \quad \forall\, x \in S_n.$$

Если $x \notin S_n$, $x \neq 0$, то $x \,/\, \|x\|_2 \in S_n$. Таким образом,

$$c_1\|x\|_2 \leq \|x\|_* \leq c_2\|x\|_2 \quad \forall\, x.$$

Мы доказали, что норма $\|\cdot\|_*$ эквивалентна норме $\|\cdot\|_2$. Отсюда ясно, что норма $\|\cdot\|_*$ эквивалентна и любой другой норме. $\square$

1.8 Операторные нормы

Пусть на $\mathbb{C}^m$ определена норма $\|\cdot\|_*$, а на $\mathbb{C}^n$ — норма $\|\cdot\|_{**}$. Тогда для $A \in \mathbb{C}^{m\times n}$ положим

$$\|A\|_{***} = \max_{x\neq 0} \frac{\|Ax\|_*}{\|x\|_{**}}.$$

Максимум существует в силу теоремы Вейерштрасса: если c — точная верхняя грань вещественной непрерывной функции $f(x)$ на компактном множестве $S \subset \mathbb{C}^n$, то из последовательности $x_k \in S$ со свойством $f(x_k) \to c$ можно выбрать сходящуюся подпоследовательность $x_{k_1} \to x \in S$ $\Rightarrow$ в силу непрерывности $f(x) = c$.

Несложно проверить, что $\|A\|_{***}$ является векторной нормой на $\mathbb{C}^{m\times n}$. Она называется *операторной нормой*, порожденной векторными нормами $\|\cdot\|_*$ и $\|\cdot\|_{**}$. Для операторной нормы имеет место следующее *свойство согласованности*:

$$\|Ax\|_* \leq \|A\|_{***}\|x\|_{**},$$

очевидно вытекающее из определения $\|A\|_{***}$.

Используя p-нормы векторов, получаем операторную норму

$$\|A\|_p \equiv \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}.$$

Это — матричная норма. В самом деле, если $AB \neq 0$, то

$$\begin{aligned} \|AB\|_p &= \max_{x\neq 0, Bx\neq 0} \frac{\|ABx\|_p}{\|Bx\|_p} \frac{\|Bx\|_p}{\|x\|_p} \\ &\leq \max_{x\neq 0, Bx\neq 0} \frac{\|ABx\|_p}{\|Bx\|_p} \cdot \max_{x\neq 0} \frac{\|Bx\|_p}{\|x\|_p} \leq \|A\|_p \|B\|_p. \ \square \end{aligned}$$

Отметим полезные формулы (докажите их!): если $A = [a_{ij}] \in \mathbb{C}^{m\times n}$, то

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|, \qquad \|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Мы скоро увидим, что $\|A\|_2^2$ есть максимальное собственное значение матриц A^*A и AA^* — поэтому операторную норму $\|A\|_2$ часто называют *спектральной нормой*.

Задачи

1. Докажите, что в банаховом пространстве последовательность вложенных замкнутых шаров, радиусы которых стремятся к нулю, имеет непустое пересечение, состоящее из единственной точки.

2. Покажите, что последовательность открытых вложенных шаров, радиусы которых стремятся к нулю, может иметь пустое пересечение.

3. Всегда ли замыкание открытого шара совпадает с замкнутым шаром с тем же центром и радиусом?

4. Норма называется *абсолютной*, если $||x|| = ||\,|x|\,||$, где $|x|$ обозначает вектор, составленный из абсолютных величин компонент вектора x. Приведите пример нормы, не являющейся абсолютной.

5. Для векторов $x, y \in \mathbb{R}^n$ выполнено равенство $||x+y||_2 = ||x||_2 + ||y||_2$. Докажите, что x и y линейно зависимы. Верно ли это, если $||x+y||_p = ||x||_p + ||y||_p$ при $p \neq 2$?

6. Докажите, что операторная норма является нормой.

7. Докажите формулы для $\|A\|_1$ и $\|A\|_\infty$ из п. 1.6.

8. Пусть норма $||\cdot||$ задана на $\mathbb{C}^n$. Операторная норма

$$||x||_* = \max_{y\neq 0} \frac{|x^T y|}{||y||}, \quad x \in \mathbb{C}^n,$$

называется *дуальной* к норме $||\cdot||$. Докажите, что для p-нормы дуальной является q-норма, где p и q образуют гельдеровскую пару.

9. Пусть матрица $A \in \mathbb{C}^{n\times n}$ сохраняет p-норму:

$$||Ax||_p = ||x||_p \quad \forall\, x \in \mathbb{C}^{n\times n}.$$

Докажите, что в этом и только в этом случае матрица A^T сохраняет q-норму:

$$||A^T x||_q = ||x||_q \quad \forall\, x \in \mathbb{C}^{n\times n}$$

($p \geq 1$ и $q \geq 1$ образуют гельдеровскую пару).

10. Докажите, что норма Фробениуса не является операторной нормой.

11. П. Гроен построил пример матричной нормы, принимающей значение 1 на единичной матрице, но не являющейся операторной нормой ни для каких векторных норм, сохраняющих одно и то же значение при любых перестановках координат векторов:

$$\|A\| \equiv \max_{1\leq i\leq n} \left(|a_{ii}| \ + \ c \sum_{j\neq i} |a_{ij}| \right), \quad c \ > \ 1, \quad A = [a_{ij}] \in \mathbb{C}^{n\times n}.$$

Докажите все это в случае $n = 2$.

12. Приведите пример неэквивалентных норм.

13. Докажите, что замкнутый шар $B = \overline{B}(0;1)$ для любой нормы в $\mathbb{R}^n$ обладает следующими свойствами:

(1) B — компактное множество относительно 2-нормы;
(2) если $x, y \in B$ и $0 \leq \alpha \leq 1$, то $\alpha x + (1-\alpha)y \in B$ (выпуклость);
(3) если $x \in B$ и $|\alpha| \leq 1$, то $\alpha x \in B$ (уравновешенность);
(4) $\exists\, r > 0 : \quad \{y \,:\, \|y\|_2 < r\} \subset B$.

Докажите, что если в $\mathbb{R}^n$ взять произвольное множество B, обладающее свойствами (1)–(4), то существует норма, для которой

$$B \ = \ \overline{B}(0,1).$$

14. Может ли норма подматрицы быть больше нормы самой матрицы?

15. Пусть A — подматрица матрицы B. Докажите, что $\|A\|_p \leq \|B\|_p$.

16. Элементы матриц A и B неотрицательны и $a_{ij} \leq b_{ij}$ для всех i, j. Верно ли, что $\|A\|_p \leq \|B\|_p$?

17. Дана матрица $A \in \mathbb{R}^{m\times n}$. Доказать замкнутость множества

$$\{y = Ax, \ \ x = [x_1, \ldots, x_n]^\top, \ \ x_1, \ldots, x_n \geq 0\}.$$

Глава 2

2.1 Скалярное произведение

Пусть V — вещественное или комплексное векторное пространство, на котором для каждой пары векторов x и y определено число (x, y) таким образом, что:

(1) $(x, x) \geq 0 \quad \forall x; \qquad (x, x) = 0 \Leftrightarrow x = 0;$

(2) $(x, y) = \overline{(y, x)};$

(3) $(\alpha x, y) = \alpha(x, y), \quad \alpha$ — число;

(4) $(x + y, z) = (x, z) + (y, z).$

Тогда (x, y) называется *скалярным произведением* векторов x и y.

Вещественное пространство со скалярным произведением называется *евклидовым*. Комплексное пространство со скалярным произведением называется *унитарным*.

Если $(x, y) = 0$, то векторы называются *ортогональными*. Пусть x_i и y_i — координаты векторов x и y в разложении их по некоторому базису в n-мерном V. Если $(x, y) = x_1\bar{y}_1 + \ldots + x_n\bar{y}_n$, то базис называется *ортонормированным*.

2.2 Длина вектора

Скалярное произведение позволяет естественным образом определить *длину* вектора x как $(x, x)^{1/2}$. То, что длина вектора является нормой, вытекает из *неравенства Коши–Буняковского–Шварца* (докажите):

$$|(x, y)| \leq (x, x)^{1/2}(y, y)^{1/2}.$$

Равенство достигается в том и только том случае, когда x и y линейно зависимы (докажите).

Если $\|x\| \equiv (x, x)^{1/2}$, то имеет место *тождество параллелограмма*:

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

Отсюда легко вывести, что существуют нормы, не порождаемые никаким скалярным произведением (например, $\|x\|_1$ — докажите).

Теорема 2.2.1 *Для того чтобы норма на векторном пространстве порождалась скалярным произведением, необходимо и достаточно, чтобы она подчинялась тождеству параллелограмма.*

Доказательство. Для простоты рассмотрим вещественное пространство; положим

$$(x,y) \equiv \tfrac{1}{2}\left(\|x+y\|^2 - \|x\|^2 - \|y\|^2\right)$$

и будем доказывать, что это — скалярное произведение. Свойства (1) и (2) очевидны. Свойство (4) равносильно тождеству

$$\begin{aligned} &\|x+y+z\|^2 - \|x+y\|^2 - \|z\|^2 \\ &= \left(\|x+z\|^2 - \|x\|^2 - \|z\|^2\right) + \left(\|y+z\|^2 - \|y\|^2 - \|z\|^2\right). \end{aligned} \qquad (*)$$

Чтобы его получить, будем опираться на тождество параллелограмма:

$$\|x+y+2z\|^2 = \|(x+y+z)+z\|^2 = 2\|x+y+z\|^2 + 2\|z\|^2 - \|x+y\|^2;$$
$$\|x+y+2z\|^2 = \|(x+z)+(y+z)\|^2 = 2\|x+z\|^2 + 2\|y+z\|^2 - \|x-y\|^2.$$

Из этих двух равенств находим

$$\|x+y+z\|^2 = \tfrac{1}{2}\|x+y+2z\|^2 - \|z\|^2 + \tfrac{1}{2}\|x+y\|^2;$$
$$\|x+z\|^2 + \|y+z\|^2 = \tfrac{1}{2}\|x+y+2z\|^2 + \tfrac{1}{2}\|x-y\|^2.$$

Первое подставляем в левую, второе — в правую часть $(*)$. Еще раз вспомнив о тождестве параллелограмма, видим, что обе части одинаковы.

Свойство (3) для рациональных α вытекает из свойства (4). В силу непрерывности по α оно выполняется для всех вещественных α. $\square$

2.3 Изометричные матрицы

Матрица $Q \in \mathbb{C}^{n\times n}$ называется *сохраняющей норму* $||\cdot||$ на $\mathbb{C}^n$, или *изометричной* относительно нормы $||\cdot||$ на $\mathbb{C}^n$, если

$$||Qx|| = ||x|| \quad \forall\, x \in \mathbb{C}^n.$$

Что можно сказать о матрицах, сохраняющих p-нормы? Очевидный пример таких матриц — любые матрицы, получающиеся из единичной матрицы с помощью перестановки строк (или столбцов), то есть *матрицы перестановки*. За исключением $p = 2$, изометричные матрицы не сильно отличаются от матриц перестановки.

Выбирая в качестве x столбцы единичной матрицы, заключаем, что p-норма каждого столбца матрицы Q равна 1. Кроме того, если $p \geq 1$ и $q \geq 1$ образуют гельдеровскую пару, то

$$\begin{aligned} ||Q^\top y||_q &= \max_{x \neq 0} \frac{|y^T Q x|}{||x||_p} = \max_{x \neq 0} \frac{|y^T Q x|}{||Qx||_p} \\ &= \max_{z \neq 0} \frac{|y^T z|}{||z||_p} = ||y||_q. \end{aligned}$$

Отсюда получаем, что q-норма каждой строки матрицы Q равна 1.

Пусть $p < 2$. Тогда 2-норма каждого столбца Q не меньше 1, а 2-норма каждой строки Q не больше 1. При этом сумма квадратов длин всех строк совпадает с суммой квадратов длин всех столбцов $\Rightarrow$ каждая строка и каждый столбец имеют 2-норму, равную 1. Если в каком-то столбце два или более ненулевых элемента, то каждый из них по модулю строго меньше 1 $\Rightarrow$ p-норма такого столбца должна быть строго меньше 1, что противоречит ранее установленному. Значит, в каждом столбце имеется ровно один ненулевой элемент. То же верно относительно строк. Случай $p > 2$ рассматривается аналогично.

Таким образом, при $p \neq 2$ матрица Q, сохраняющая p-норму, имеет вид

$$Q = P \operatorname{diag}(d_1, \ldots, d_n),$$

где P—матрица перестановки и $|d_i| = 1, i = 1, \ldots, n$.

2.4 Сохранение длин и унитарные матрицы

При $p = 2$ множество изометричных матриц существенно шире. В данном случае сохранение 2-нормы влечет за собой сохранение скалярного произведения (докажите!) $\Rightarrow$ столбцы $q_1, \ldots, q_n$ матрицы Q образуют ортонормированную систему:

$$q_i^* q_j = \delta_{ij} \Leftrightarrow Q^* Q = I$$

(δ—символ Кронекера: $\delta_{ij} = 1$ при $i = j$ и 0 при $i \neq j$).

Матрица $Q \in \mathbb{C}^{n \times n}$ такая, что $Q^* = Q^{-1}$, называется *унитарной*.

Унитарные матрицы интересны тем, что они и только они сохраняют длину вектора (2-норму) и скалярное произведение векторов.

Важное свойство унитарных матриц: их произведения и обратные к ним матрицы остаются унитарными (докажите).

2.5 Теорема Шура

Теорема 2.5.1 (Теорема Шура). *Для любой матрицы $A \in \mathbb{C}^{n\times n}$ с собственными значениями $\lambda_1, \ldots, \lambda_n$ существует унитарная матрица Q такая, что:*

(1) *матрица Q^*AQ верхняя треугольная;*

(2) $\mathrm{diag}(Q^*AQ) = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

Доказательство. Пусть $Av_1 = \lambda_1 v_1, ||v_1||_2 = 1$. Выберем $v_2, \ldots, v_n$ таким образом, чтобы матрица $V_1 = [v_1, v_2, \ldots, v_n]$ была унитарной. Тогда

$$V_1^* A V_1 = \begin{bmatrix} \lambda_1 & * & \ldots & * \\ 0 & & & \\ \ldots & & A_1 & \\ 0 & & & \end{bmatrix}.$$

Далее по индукции. $\square$

2.6 Нормальные матрицы

Матрица A называется *нормальной*, если $A^*A = AA^*$.

Важнейшие классы нормальных матриц:

(1) *эрмитовы* матрицы: $H^* = H$;

(2) *унитарные* матрицы: $U^* = U^{-1}$.

В вещественном случае эрмитова матрица называется *симметричной*, а унитарная — *ортогональной*.

Теорема 2.6.1 *Матрица $A \in \mathbb{C}^{n\times n}$ является нормальной тогда и только тогда, когда в $\mathbb{C}^n$ существует ортонормированный базис из ее собственных векторов.*

Доказательство. Для любой $A \in \mathbb{C}^{n\times n}$ существует унитарная матрица U такая, что $T = U^*AU$ есть верхняя треугольная матрица (теорема Шура). Далее, $A^*A = AA^*$ равносильно $T^*T = TT^*$, и легко убедиться, что верхняя треугольная матрица с таким свойством обязана быть диагональной. Таким образом, столбцы матрицы U образуют базис из собственных векторов матрицы A. $\square$

Теорема 2.6.2 *Нормальная матрица A является эрмитовой тогда и только тогда, когда все ее собственные значения вещественны.*

Теорема 2.6.3 *Нормальная матрица является унитарной тогда и только тогда, когда все ее собственные значения по модулю равны 1.*

Докажите эти теоремы.

Полезно иметь в виду, что любая матрица $A \in \mathbb{C}^{n\times n}$ представима и притом однозначно в виде

$$A = H + \mathrm{i}\, K, \quad H^* = H, \quad K^* = K, \quad \mathrm{i}^2 = 1.$$

Это так называемое *эрмитово разложение* матрицы A.

Тривиально доказывается, что нормальность матрицы A равносильна тому, что H и K коммутируют.

2.7 Знакоопределенные матрицы

Среди эрмитовых матриц выделяются *знакоопределенные* матрицы— для них скалярное произведение $(Ax, x) = x^*Ax$ имеет один и тот же знак для всех x.

Если (Ax, x) для всех $x \in \mathbb{C}^{n\times n}$, то матрица A называется *положительно полуопределенной*, или *неотрицательно определенной*. Обозначение: $A \geq 0$.

Если $(Ax, x) > 0$ для всех $x \neq 0$ из $\mathbb{C}^{n\times n}$, то матрица A называется *положительно определенной*. Обозначение: $A > 0$.

Эрмитовость вытекает из знакоопределенности. Для $A \geq 0$ рассмотрим ее эрмитово разложение $A = H + \mathrm{i}\, K$. Имеем

$$(Ax, x) = (Hx, x) + \mathrm{i}\, (Kx, x) \in \mathbb{R} \quad \forall\, x \in \mathbb{C}^n.$$

Следовательно, $(Kx, x) = 0 \quad \forall\, x \in \mathbb{C}^n \Rightarrow$ все собственные значение эрмитовой матрицы K равны $0 \quad \Rightarrow \quad K = 0$. $\square$

Заметим, что если $A \in \mathbb{R}^{n\times n}$ и $(Ax, x) \geq 0 \quad \forall\, x \in \mathbb{R}^n$, то матрица A не обязана быть симметричной.

Для неотрицательной (положительной) определенности матрицы $A \in \mathbb{C}^{n\times n}$ необходима и достаточна неотрицательность (положительность) ее собственных значений (докажите).

Неотрицательная (положительная) определенность матрицы влечет за собой неотрицательную (положительную) определенность всех ее ведущих подматриц. (Подматрица называется *ведущей*, если она занимает левый

верхний угол матрицы.) Для доказательства достаточно заметить, что

$$[y^* \quad 0] \begin{bmatrix} B & * \\ * & * \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix} = y^* B y \quad \forall\, y.$$

Нам скоро понадобится следующий простой факт: $A^*A \geq 0$ для любой матрицы A (докажите).

2.8 Сингулярное разложение матрицы

Теорема 2.8.1 *Пусть $A \in \mathbb{C}^{m\times n}$, $r = \mathrm{rank}A$. Тогда существуют положительные числа $\sigma_1 \geq \ldots \geq \sigma_r > 0$ и унитарные матрицы $U \in \mathbb{C}^{n\times n}$, $V \in \mathbb{C}^{m\times m}$ такие, что*

$$A = V\Sigma U^*, \tag{2.8.1}$$

где Σ — диагональная прямоугольная $m \times n$-матрица вида

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \\ & & \end{bmatrix}. \tag{2.8.2}$$

Доказательство. $A^*A \geq 0 \quad \Rightarrow \quad$ существует унитарная матрица $U = [u_1, \ldots, u_n] \in \mathbb{C}^{n\times n}$ такая, что

$$U^*A^*AU = \mathrm{diag}\,(\sigma_1^2, \ldots, \sigma_n^2), \quad \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n.$$

Предположим, что $\sigma_r > 0$ и $\sigma_i = 0$ при $i > r$. Пусть $U_r = [u_1, \ldots, u_r]$ и $\Sigma_r = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_r)$. Тогда

$$U_r^*\, A^*A\, U_r = \Sigma_r^2 \quad \Rightarrow \quad (\Sigma_r^{-1}U_r^*A^*)(AU_r\Sigma_r^{-1}) = I.$$

Следовательно, матрица $V_r = AU_r\Sigma_r^{-1}$ такова, что $V_r^*V_r = I \;\Rightarrow\; V_r$ имеет ортонормированные столбцы.

Если $1 \leq i \leq r$, то $Au_i = \sigma_i v_i$. При $r+1 \leq i \leq n$ находим $u_i^*A^*Au_i = ||Au_i||_2^2 = 0 \;\Rightarrow\; Au_i = 0$. Достраивая произвольным образом V_r до унитарной матрицы $V \in \mathbb{C}^{m\times m}$, получаем

$$AU = V \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \quad \Rightarrow \quad V^*AU = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}.$$

Поскольку умножение на невырожденные матрицы не меняет ранг,

$$r = \mathrm{rank}A. \qquad \square$$

Разложение (2.8.1) называется *сингулярным разложением* матрицы A. Числа $\sigma_1 \geq \ldots \geq \sigma_r > 0$ называются сингулярными числами, векторы u_i — правыми, v_i — левыми сингулярными векторами матрицы A. Обычно говорят, что помимо r ненулевых сингулярных чисел, матрица A имеет $\min(m,n) - r$ нулевых сингулярных чисел.

Следствие 2.8.1 *Сингулярные числа матрицы определяются однозначно.*

Следствие 2.8.2 *Если $\sigma_1 > \ldots > \sigma_r > 0$, то сингулярные векторы $u_1, \ldots, u_r$ и $v_1, \ldots, v_r$ определяются однозначно с точностью до множителя, равного по модулю 1.*

Следствие 2.8.3

$$Au_i = \begin{cases} \sigma_i v_i, & 1 \leq i \leq r, \\ 0, & r+1 \leq i \leq n. \end{cases} \tag{2.8.3}$$

$$A^* v_i = \begin{cases} \sigma_i u_i, & 1 \leq i \leq r, \\ 0, & r+1 \leq i \leq m. \end{cases} \tag{2.8.4}$$

Следствие 2.8.4 $A = \sum\limits_{i=1}^{r} \sigma_i v_i u_i^*$.

Следствие 2.8.5

$$\begin{aligned} \ker A &= \operatorname{span}\{u_{r+1}, \ldots, u_n\}, \\ \operatorname{im} A &= \operatorname{span}\{v_1, \ldots, v_r\}, \\ \ker A^* &= \operatorname{span}\{v_{r+1}, \ldots, v_m\}, \\ \operatorname{im} A^* &= \operatorname{span}\{u_1, \ldots, u_r\} \end{aligned}$$

2.9 Унитарно инвариантные нормы

Если $\|A\| = \|QAZ\|$ для любых унитарных Q, Z и любой матрицы A (при естественном согласовании размеров), то такая матричная норма называется *унитарно инвариантной*.

Важнейшие унитарно инвариантные нормы: $\|A\|_2$ и $\|A\|_F$. Вот доказательство унитарной инвариантности:

$$\begin{aligned} \|QAZ\|_2 &= \sup_{x \neq 0} \frac{\|QAZx\|_2}{\|x\|_2} = \sup_{x \neq 0} \frac{\|(QAZ)Z^*x\|_2}{\|Z^*x\|_2} = \\ &= \sup_{x \neq 0} \frac{\|QAx\|_2}{\|x\|_2} = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \|A\|_2; \end{aligned}$$

$$\|QAZ\|_F^2 = tr(QAZ)^*(QAZ) = trZ^*\left(A^*A\right)Z = trA^*A = \|A\|_F^2. \ \square$$

Согласно (2.8.1), для любой унитарно инвариантной нормы $\|A\| = \|\Sigma\|$ $\Rightarrow$ унитарно инвариантная норма определяется по сингулярным числам. Для спектральной и фробениусовой норм находим:

$$\|A\|_2 = \sigma_1, \tag{2.9.5}$$

$$\|A\|_F = \left(\sigma_1^2 + \ \ldots\ + \sigma_r^2\right)^{1/2}. \tag{2.9.6}$$

2.10 Короткий путь к сингулярному разложению

Норма $\|A\|_2$ определяется как операторная норма. В силу компактности единичной сферы существуют нормированные векторы x и y такие, что $Ax = \sigma y, \quad \sigma = \|A\|_2$. Возьмем унитарные матрицы $U = [x \ \ U_1], \quad V = [y \ \ V_1]$. Тогда

$$V^*AU = \begin{bmatrix} \sigma & w^* \\ 0 & B \end{bmatrix},$$

$$\|V^*AU \begin{bmatrix} \sigma \\ w \end{bmatrix}\|_2^2 \geq \left(\sigma^2 + w^*w\right)^2 \Rightarrow \|V^*AU\|_2^2 \geq \sigma^2 + w^*w.$$

Поскольку $\|V^*AU\|_2 = \|A\|_2$ (унитарная инвариантность спектральной нормы), $w = 0$. Далее по индукции.

2.11 Аппроксимации меньшего ранга

Теорема 2.11.1 *Пусть* $k < \mathrm{rank}A$, $A_k \equiv \sum\limits_{i=1}^{k} \sigma_i v_i u_i^*$. *Тогда*

$$\min_{\mathrm{rank}B=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}.$$

Доказательство. Поскольку $\mathrm{rank}B = k$, находим $\dim\ker B = n-k$. Тогда существует ненулевой вектор $z \in \ker B \cap \mathrm{span}\{u_1, \ldots, u_{k+1}\}$. (Почему?) Будем считать, что $\|z\|_2 = 1$. Тогда

$$\|A - B\|_2^2 \geq \|(A-B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i \left(u_i^*z\right)^2 \geq \sigma_{k+1}^2. \quad \square$$

В частности, младшее сингулярное число невырожденной матрицы — это расстояние (в спектральной норме) до ближайшей вырожденной матрицы.

Задачи

1. Найдите все $p \geq 1$, при которых норма $\|x\|_p$ порождается каким-либо скалярным произведением.

2. Докажите, что для любой матрицы $A \in \mathbb{C}^{m\times n}$ подпространства $\ker A$ и $\operatorname{im} A^*$ ортогональны и дают в прямой сумме $\mathbb{C}^n$.

3. Известно, что матрица A^{2006} нормальная. Будет ли нормальной A?

4. Докажите, что для любой эрмитовой матрицы H матрица

$$Q = (I - \mathrm{i}\, H)^{-1}(I + \mathrm{i}\, H)$$

будет унитарной. Верно ли, что любую унитарную матрицу можно представить таким образом?

5. Пусть $A = I + \alpha uu^*, \|u\|_2 = 1$. Найдите все комплексные α, при которых матрица A будет унитарной.

6. Отображение $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$ сохраняет скалярные произведения. Докажите, что оно является обратимым.

7. Докажите, что произведение эрмитовой матрицы и положительно определенной матрицы имеет вещественные собственные значения.

8. Найдите сингулярное разложение $n \times n$-матрицы

$$A = \begin{bmatrix} 1 \\ 2 \\ \dots \\ n \end{bmatrix} [1\ 1 \dots\ 1].$$

9. Докажите, что $\sigma_1(A) = \max\limits_{||u||_2=||v||_2=1} |u^*Av|$.

10. Чему равно расстояние от вырожденной матрицы A до ближайшей невырожденной?

11. Можно ли утверждать, что $B = A^*$, если $(Ax, x) = (x, Bx)$ для любого: (а) $x \in \mathbb{R}^n$; (б) $x \in \mathbb{C}^n$?

12. Докажите, что для квадратной матрицы A выполняется неравенство $\lambda_{min}(A+A^*) \leq 2\sigma_{min}(A)$, где $\lambda_{min}(\cdot)$ и $\sigma_{min}(\cdot)$ обозначают минимальное собственное значение и минимальное сингулярное число. Можно ли слева заменить λ_{min} на σ_{min}?

13. Докажите, что $\|AB\|_F \leq \|A\|_2 \|B\|_F$.

14. Докажите, что $\|A\|_F \leq \sqrt{\operatorname{rank}(\mathrm{A})}\, \|A\|_2$.

15. Матрица $A = [A_{ij}]$ составлена из блоков A_{ij}, а матрица $B = [b_{ij}]$ такова, что $b_{ij} = ||A_{ij}||_2$, $1 \leq i \leq m$, $1 \leq j \leq n$. Докажите, что

$$||A||_2 \leq ||B||_2.$$

16. Пусть L — нижняя треугольная матрица с нижней треугольной частью, взятой из матрицы $A \in \mathbb{C}^{n \times n}$. Докажите, что

$$||L||_2 \leq \log_2 2n \, ||A||_2.$$

17. Пусть $A \in \mathbb{C}^{n \times n}$. Докажите, что $\operatorname{tr} A = 0$ в том и только том случае, когда $||I + zA||_F \geq \sqrt{n}$ для всех $z \in \mathbb{C}$.

18. Нормальная матрица имеет блочно-треугольный вид

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$$

с квадратными блоками A_{11} и A_{22}. Докажите, что матрицы A_{11} и A_{22} нормальные и, кроме того, $A_{12} = 0$.

19. Пусть зафиксировано подпространство $L \in \mathbb{C}^n$ и рассматриваются матрицы P такие, что $P^2 = P$ и $\operatorname{im} P = L$. Докажите, что среди всех таких матриц P наименьшую 2-норму имеет эрмитова матрица.

Глава 3

3.1 Теория возмущений

Пусть по x вычисляется $f(x)$. Иногда алгоритмы дают большую погрешность. Думая об этом, мы, конечно, должны понимать, что "плохим" может быть не только алгоритм, но и сама задача. Важный вопрос: как сильно может измениться $f(x)$ при малых возмущениях x?

В простейшем случае $f(x+\delta) \simeq f(x) + f'\delta$ и, следовательно, мерой чувствительности задачи может служить $||f'||$. Если $f(x) \neq 0$ и $x \neq 0$, то

$$\frac{f(x+\delta) - f(x)}{||f(x)||} \simeq \left(\frac{f'(x)}{||f(x)||}||x||\right)\frac{\delta}{||x||}.$$

Поэтому относительной мерой чувствительности задачи (другими словами, ее *числом обусловленности*) может служить

$$\mathrm{cond}\,(f(x)) \equiv \frac{||f'(x)||}{||f(x)||}||x||.$$

3.2 Число обусловленности матрицы

Пусть A — невырожденная матрица и $f(A) = A^{-1}$. Тогда (проверьте!)

$$(A+\Delta)^{-1} - A^{-1} = -A^{-1}\Delta\,(A+\Delta)^{-1} \simeq -A^{-1}\Delta\,A^{-1}$$

$$\Rightarrow \quad \frac{||(A+\Delta)^{-1} - A^{-1}||}{||A^{-1}||} \quad \lesssim \quad (||A^{-1}||\;||A||)\frac{||\Delta||}{||A||}.$$

Величина

$$\mathrm{cond}(A) \equiv ||A^{-1}||\;||A||$$

называется *числом обусловленности матрицы* A. Оно зависит от нормы. Для p-нормы пишут cond_p. Обычно cond_2 называют спектральным числом обусловленности.

Для вырожденных матриц $\mathrm{cond} = \infty$. Обычно число обусловленности задачи обратно пропорционально ее расстоянию до множества вырожденных (в соответствующем смысле) задач. При обращении матриц множество вырожденных задач — это все вырожденные матрицы. Мы знаем, что

$$\min_{\det S=0} ||A - S||_2 = \sigma_{\min} \quad \text{(минимальное сингулярное число)}$$

и одновременно (докажите) $||A^{-1}||_2 = 1 \,/\, \sigma_{\min}$. Поэтому

$$\mathrm{cond}\ (A) = \frac{||A||_2}{\min\limits_{\det S=0} ||A - S||_2}.$$

Почему $(A + \Delta)^{-1} \approx A^{-1}$ при малых Δ ? Это вытекает из стандартных соображений непрерывности. Однако, в матричном анализе есть простая и полезная техника для таких случаев.

3.3 Сходящиеся матрицы и ряды

Ряд $\sum_{k=0}^{\infty} A_k$, где $A_k \in \mathbb{C}^{n\times n}$, называется *сходящимся*, если сходится последовательность его частичных сумм $S_N \equiv \sum_{k=0}^{N} A_k$. Для этого достаточно, чтобы сходился числовой ряд $\sum_{k=0}^{\infty} \|A_k\|$ (докажите!).

Ряд $\sum_{k=0}^{\infty} F^k$ называется *рядом Неймана*. Очевидно, он будет сходиться, если $\|F\| < 1$ (докажите). Менее очевидно, что он будет сходиться, если все собственные значения F по модулю меньше 1.

Максимальное по модулю собственное значение матрицы F называется ее *спектральным радиусом*. Обозначение: $\rho(F)$. Если $\rho(F) < 1$, то матрица F называется *сходящейся*.

Лемма 3.3.1 *Ряд Неймана с матрицей $F \in \mathbb{C}^{n\times n}$ сходится тогда и только тогда, когда матрица F является сходящейся.*

Достаточность. По теореме Шура, для некоторой унитарной P получаем верхнюю треугольную матрицу $T = [t_{ij}] = P^{-1}FP$. Докажем, что ряд Неймана сходится для некоторой матрицы, подобной F (это эквивалентно его сходимости для F — почему?).

Положим $D_\varepsilon = \mathrm{diag}\,(1, \varepsilon, \ldots, \varepsilon^{n-1})$. Тогда $\{D_\varepsilon^{-1} T D_\varepsilon\}_{ij} = \varepsilon^{j-i} t_{ij}$ при $i \le j$. Диагональные элементы этой матрицы по модулю меньше 1 $\Rightarrow$ при досточно малом ε имеем $||D_\varepsilon^{-1} T D_\varepsilon||_1 < 1$ $\Rightarrow$ ряд Неймана с матрицей $D_\varepsilon^{-1} T D_\varepsilon$ сходится.

Необходимость. Допустим, что $Fx = \lambda x$, $x \neq 0$, и $|\lambda| \ge 1$. Тогда $1 \le |\lambda|^k \le ||F^k||_2$ (почему?) $\Rightarrow$ $||F^k||_2 \not\to 0$ $\Rightarrow$ ряд Неймана с F расходится. $\square$

3.4 Простейший итерационный метод

Чтобы решить линейную систему $Ax = b$ с невырожденной матрицей коэффициентов, запишем ее в виде $x = Fx + \alpha\, b$, где $F = I - \alpha\, A,\ \alpha \neq 0$, и рассмотрим следующий итерационный метод:

$$\begin{aligned} x_0 &- \text{произвольный начальный вектор;} \\ x_k &= Fx_{k-1} + \alpha\, b \quad \text{при} \quad k = 1, 2, \ldots . \end{aligned}$$

Это так называемый *метод простой итерации* (иногда его называют методом Ричардсона).

Легко вывести (см. лемму 3.3.1), что $x_k \to x$ для любого начального вектора x_0 тогда и только тогда, когда матрица F сходящаяся.

Часто пытаются найти расщепление $A = M - N$, для которого матрица $M^{-1}N$ будет сходящейся, а M *легко обращается* (это означает, что мы умеем эффективно решать системы с матрицей M). Тогда

$$x_k = M^{-1}(Nx_{k-1} + b) \ \to \ x.$$

3.5 Обратные матрицы и ряды

Лемма 3.5.1 *Если $\|F\| < 1$, то матрица $A = I - F$ обратима и при этом:*

$$\text{(a)} \quad (I - F)^{-1} = \sum_{k=0}^{\infty} F^k; \qquad \text{(b)} \quad \|(I - F)^{-1}\| \leq \frac{\|I\|}{1 - \|F\|}.$$

Доказательство. Легко проверить, что

$$(I - F)\left(\sum_{k=0}^{N} F^k\right) = I - F^{N+1} \ \longrightarrow I.$$

Чтобы получить (b), запишем

$$\|\sum_{k=0}^{N} F^k\| \leq \|I\| \ \sum_{k=0}^{N} \|F\|^k \ \leq \ \frac{\|I\|}{1 - \|F\|}. \quad \square$$

Следствие 3.5.1 *Если A — невырожденная матрица и E — ее возмущение, такое что $\|A^{-1}E\| < 1$, то:*

(a) *Матрица $A+E$ невырожденная и при этом*

$$(A+E)^{-1} = \sum_{k=0}^{\infty} \left(-A^{-1}E\right)^k A^{-1} \;=\; A^{-1} \sum_{k=0}^{\infty} \left(-EA^{-1}\right)^k;$$

(b) $$\frac{\|(A+E)^{-1} - A^{-1}\|}{\|A^{-1}\|} \;\le\; \frac{\|A^{-1}\|\|E\|}{1 - \|A^{-1}E\|}.$$

3.6 Обусловленность линейной системы

Рассмотрим систему $Ax = f,\ f \neq 0$, с невырожденной матрицей A и возмущенную систему $(A+\Delta A)\tilde{x} = f + \Delta f$. Как сильно $\tilde{x}$ может отличаться от x? Пусть $\left\|A^{-1}\Delta A\right\| < 1$. Тогда

$$\begin{aligned}
\tilde{x} - x &= (A+\Delta A)^{-1}(f+\Delta f) - A^{-1}f \\
&= \left[(A+\Delta A)^{-1} - A^{-1}\right] f + (A+\Delta A)^{-1}\Delta f \\
&= \left[\sum_{k=1}^{\infty}(-A^{-1}\Delta A)^k\right]\left(A^{-1}f\right) + \left[\sum_{k=0}^{\infty}\left(-A^{-1}\Delta A\right)^k\right] A^{-1}\Delta f \\
&\Rightarrow \\
&\frac{\|\tilde{x}-x\|}{\|x\|} \le \frac{\|A^{-1}\|\|A\|}{1 - \|A^{-1}\Delta A\|}\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta f\|}{\|f\|}\right). \qquad (3.6.1)
\end{aligned}$$

Число $\operatorname{cond} A = \|A^{-1}\|\|A\|$ (число обусловленности матрицы A) характеризует чувствительность решения x к малым возмущениям матрицы и правой части. О матрицах с "очень большим" и "не очень большим" числом обусловленности говорят как о плохо и хорошо обусловленных матрицах.

3.7 Согласованность матрицы и правой части

Оценка (3.6.1) не улучшаема на всем множестве матриц и возмущений. Однако плохая обусловленность матрицы не всегда адекватна "плохой обусловленности линейной системы".

Используя сингулярное разложение, запишем

$$A = \sum_{k=1}^{n} \sigma_k v_k u_k^* \quad \Rightarrow \quad A^{-1} = \sum_{k=1}^{n} \sigma_k^{-1} u_k v_k^* \quad \Rightarrow \quad x = \sum_{k=1}^{n} \frac{v_k^* f}{\sigma_k} u_k.$$

Пусть A фиксирована и $\Delta f = \xi_1 v_1 + \cdots + \xi_n v_n,\ \Delta x = \eta_1 u_1 + \cdots + \eta_k u_k$. Ясно, что

$$\sigma_1^2\eta_1^2 + \cdots + \sigma_n^2\eta_n^2 = \xi_1^2 + \cdots + \xi_n^2.$$

Значит, если Δf принадлежит шару радиуса ε в координатах $\{\xi_i\}$, то Δx принадлежит эллипсоиду в координатах $\{\eta_i\}$:

$$\xi_1^2 + \ldots + \xi_n^2 \le \varepsilon^2 \quad \Leftrightarrow \quad \frac{\eta_1^2}{1/\sigma_1^2} + \ldots + \frac{\eta_n^2}{1/\sigma_n^2} \le \varepsilon^2.$$

Отсюда видно, что $\|\Delta x\|$ существенно зависит от направления возмущения.

Если правая часть f имеет нулевые компоненты в направлении младших сингулярных векторов $v_{r+1}, \ldots, v_n$ (в этом случае говорят, что матрица и правая часть являются *согласованными*) и возмущения в этих направлениях тоже отсутствуют, то оценка (3.6.1), очевидно, улучшается:

$$\frac{\|\Delta x\|_2}{\|x\|_2} \le \frac{\sigma_1}{\sigma_r} \frac{\|\Delta f\|_2}{\|f\|_2}.$$

3.8 Возмущение собственных значений

Пусть $\lambda(A)$ обозначает спектр матрицы A.

Теорема Бауэра–Файка. *Если $\mu \in \lambda(A+F)$, но $\mu \notin \lambda(A)$, то*

$$\frac{1}{||(A-\mu I)^{-1}||_2} \le ||F||_2.$$

Доказательство. Матрица $(A+F) - \mu I = (A - \mu I) + F$ вырожденная $\Rightarrow$ матрица $I + (A-\mu I)^{-1}F$ вырожденная $\Rightarrow$ $||(A-\mu I)^{-1}F||_2 \ge 1$. $\square$

Теорема 3.8.1 *Пусть A диагонализуема:*

$$P^{-1}AP = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \equiv \Lambda. \tag{3.8.2}$$

Тогда если $\mu \in \lambda(A+F)$, то

$$\min_{1 \le i \le n} |\mu - \lambda_i| \le \|P^{-1}\|_2 \, \|P\|_2 \, \|F\|_2. \tag{3.8.3}$$

Доказательство. Неравенство очевидно, если $\mu \in \lambda(A)$. Если же $\mu \notin \lambda(A)$, то $\mu \notin \lambda(\Lambda)$ и $\mu \in \lambda(\Lambda + P^{-1}FP)$ — остается применить теорему Бауэра—Файка. $\square$

Итак, чувствительность спектра к малым возмущениям характеризуется числом обусловленности матрицы собственных векторов (это столбцы P).

Теорема 3.8.2 *Пусть $P^{-1}AP = J$ есть жорданова форма матрицы A и $\mu \in \lambda(A+F)$. Тогда существует $\lambda \in \lambda(A)$ такое, что*

$$\frac{|\mu - \lambda|^m}{1 + |\mu - \lambda| + \cdots + |\mu - \lambda|^{m-1}} \le ||P^{-1}||_2 ||P||_2 ||F||_2,$$

где m — максимальный порядок жордановых клеток, отвечающих λ.

Доказательство. Опять используем теорему Бауэра—Файка: если $\mu \notin \lambda(A)$, то

$$\frac{1}{||(J-\mu\, I)^{-1}||_2} \;\le\; ||P^{-1}FP||_2.$$

Пусть J состоит из жордановых клеток $J_1, \ldots, J_k$. Тогда

$$\frac{1}{||(J-\mu\, I)^{-1}||_2} \;\ge\; \frac{1}{\max\limits_{1\le i\le k}\; ||(J_i-\mu\, I)^{-1}||_2}.$$

Запишем $J_i = \lambda\, I + N_i$ и предположим, что эта жорданова клетка имеет порядок m. Тогда $N_i^m = 0$ и, кроме того, $||N_i||_2 = 1 \quad \Rightarrow$

$$||(J_i-\mu\, I)^{-1}||_2 = ||((\lambda-\mu)\, I + N_i)^{-1}||_2 \;\le\; ||(I+(\lambda-\mu)^{-1}N_i)^{-1}||_2 |(\lambda-\mu)^{-1}|$$

$$\le\; \left(1 + |(\lambda-\mu)^{-1}| + \cdots + |(\lambda-\mu)^{-1}|^{m-1}\right) |(\lambda-\mu)^{-1}|. \quad \square$$

Итак, если матрица с максимальным порядком жордановых клеток m возмущается величинами порядка ε, то любое собственное значение возмущенной матрицы может отличаться от некоторого собственного значения исходной матрицы на величину порядка $|\varepsilon|^{\frac{1}{m}}$.

Верно ли, что при малом возмущении собственные значения матриц A и $A+F$ можно разбить на пары близких значений? Чтобы ответить на этот вопрос, примем во внимание непрерывную зависимость корней полинома от его коэффициентов. Этот важный факт заслуживает специального обсуждения.

3.9 Непрерывность корней полинома

Теорема 3.9.1 *Рассмотрим параметризованное семейство полиномов*

$$p(x,t) = x^n + a_1(t)x^{n-1} + \cdots + a_n(t),$$

где $a_1(t), \ldots, a_n(t) \in C\,[\alpha,\beta]$. *Тогда существуют непрерывные функции*

$$x_1(t), \ldots, x_n(t) \in C\,[\alpha,\beta]$$

такие, что

$$p\,(x_i(t),t) = 0 \quad \textit{при} \quad \alpha \le t \le \beta, \qquad i = 1, \ldots, n.$$

Начиная доказательство, заметим, что достаточно установить существование *одной* непрерывной функции $x_n(t)$, такой что $p(x_n(t),t) = 0$ при $\alpha \le t \le \beta$. Если это сделано, запишем

$$p(x,t) = (x - x_n(t))\, q(x,t),$$

где $q(x,t) = x^{n-1} + b_1(t)x^{n-2} + \cdots + b_{n-1}(t)$. В силу известного алгоритма деления полиномов $b_1(t), \ldots, b_{n-1}(t) \in C[\alpha, \beta]$. Таким образом, можно использовать индукцию.

Итак, будем доказывать существование одного непрерывного корня. Сделаем это по аналогии с тем, как доказывается существование решения дифференциального уравнения $\frac{dy}{dt} = f(t,y)$ для непрерывной f с помощью ломаных Эйлера и теоремы Арцела.

Последовательность функций $y_m(t)$ называется *равностепенно непрерывной* при $t \in [\alpha, \beta]$, если $\forall\, \varepsilon > 0\, \exists\, \delta > 0 \ : \ |t_1 - t_2| \ \leq \ \delta \ \Rightarrow |y_m(t_1) - y_m(t_2)| \leq \varepsilon\, \forall\, m$. Последовательность функций $y_m(t)$ называется *равномерно ограниченной* при $t \in [\alpha, \beta]$, если $\exists\, c > 0 \ : \ |y_m(t)| \leq c\, \forall\, m, \ \forall\, t \in [\alpha, \beta]$.

Теорема 3.9.2 (Арцела) *Для любой последовательности равномерно ограниченных и равностепенно непрерывных функций на $[\alpha, \beta]$ существует подпоследовательность, равномерно сходящаяся на $[\alpha, \beta]$.*

Доказательство. Пронумеруем каким-либо способом все рациональные точки на $[\alpha, \beta] \ : \ t_1, t_2, \ldots$ Из исходной последовательности $y_m(t)$ выберем подпоследовательность $y_{1,m}(t)$, сходящуюся в точке t_1; из этой подпоследовательности выберем подпоследовательность $y_{2,m}(t)$, сходящуюся в точке t_2, и т.д. В итоге мы будем иметь "вложенные" подпоследовательности $y_{1,m}(t)$, $\ldots$, $y_{k,m}(t)$, $\ldots$, такие что $y_{k,m}(t)$ сходится при $t = t_1, \ldots, t_k$ (при этом $y_{k+1,m}(t)$ является подпоследовательностью для $y_{k,m}(t)$). Рассмотрим "диагональную" последовательность $y_{m,m}(t)$. Возьмем $\varepsilon > 0$ и $\delta > 0$, определенное условием равностепенной непрерывности. Для произвольной точки $t \in [\alpha, \beta]$ существует t_i такое, что $|t - t_i| \leq \delta$. При достаточно больших m, k получаем

$$\begin{aligned} |y_{mm}(t) - y_{kk}(t)| \ &\leq \ |y_{mm}(t) - y_{mm}(t_i)| + |y_{mm}(t_i) - y_{kk}(t_i)| + \\ &\quad + |y_{kk}(t_i) - y_{kk}(t)| \\ &\leq \ 2\varepsilon + |y_{mm}(t_i) - y_{kk}(t_i)| \leq 3\varepsilon. \end{aligned}$$

Значит, $y_{m,m}(t)$ — это последовательность Коши. $\square$

Построим на $[\alpha, \beta]$ последовательность равномерных сеток:

$$\alpha = t_{0m} < t_{1m} < \ldots < t_{mm} = \beta; \quad t_{i+1,m} - t_{im} = \frac{\beta - \alpha}{m}.$$

Пусть $y_m(t)$ есть ломаная с изломами в узлах $t_{0m}, t_{1m}, \ldots, t_{mm}$. Ее значения в этих узлах будут определяться следующим образом.

Фиксируем корень z_0 полинома $p(x, \alpha)$ и положим для всех m

$$y_m\left(t_{0m}\right) = z_{0m} \equiv z_0.$$

Далее, пусть z_{1m} — любой корень полинома $p\,(x, t_{1m})$, ближайший к z_{0m}, и, по индукции, пусть $z_{i+1,m}$ — корень полинома $p\,(x, t_{i+1,m})$, ближайший к z_{im}. Положим

$$y_m\,(t_{im}) = z_{im}, \qquad i = 1, \ldots, m.$$

Равномерная ограниченность ломаных $y_m(t)$ очевидна. Равностепенная непрерывность вытекает из следующего неравенства:

$$\begin{aligned} |z_{i+1,m} - z_{im}| \;&\leq\; |p\,(z_{im}, t_{i+1,m})|^{\frac{1}{n}} = |p\,(z_{im}, t_{i+1,m}) - p\,(z_{im}, t_{im})|^{\frac{1}{n}} \\ &\leq\; R\left(\max_{t_1,t_2} \sum_{j=1}^{n} |a_j\,(t_1) - a_j\,(t_2)|\right)^{\frac{1}{n}}, \end{aligned}$$

где максимум берется при условии

$$\alpha \leq t_1, t_2 \leq \beta, \quad |t_1 - t_2| \leq \frac{\beta - \alpha}{m},$$

а $R \geq 1$ есть радиус какого-либо круга, содержащего все корни всех полиномов $p(x, t)$ при $\alpha \leq t \leq \beta$.

Используя теорему Арцела, находим равномерно сходящуюся подпоследовательность. Примем во внимание, что предел равномерно сходящейся последовательности непрерывных функций на $[\alpha, \beta]$ является непрерывной функцией. Остается заметить, что предельная функция $y(t)$ удовлетворяет уравнению $p(y(t), t) = 0$ при $\alpha \leq t \leq \beta$. Тем самым теорема 3.9.1 доказана.

Задачи

1. Непрерывная зависимость корней полинома от коэффициентов не противоречит тому, что они могут очень сильно изменяться при малых возмущениях коэффициентов. Вот пример Уилкинсона:

$$p(x;\, \varepsilon) = (x-1)(x-2)\ldots(x-20) + \varepsilon\, x^{19} = \prod_{i=1}^{20}(x - x_i(\varepsilon)),$$

   где $x_i(\varepsilon)$ — непрерывные функции такие, что $x_i(0) = i$. При малых ε было замечено, что $x_1(\varepsilon) \approx 1$, но $x_{20}(\varepsilon)$ сильно отличается от 20. Чтобы объяснить это различие, сравните значения производных функций $x_1(\varepsilon)$ и $x_{20}(\varepsilon)$ при $\varepsilon = 0$.

2. Означает ли $|\det A| = 1$ хорошую обусловленность матрицы A? Означает ли $|\det A| << 1$ плохую обусловленность?

3. Найдите $\mathrm{cond}_\infty(A) = ||A^{-1}||_\infty \, ||A||_\infty$ для двухдиагональной матрицы

$$A = \begin{bmatrix} 1 & 2 & & & 0 \\ & 1 & 2 & & \\ & & \ddots & \ddots & \\ & & & 1 & 2 \\ 0 & & & & 1 \end{bmatrix}_{n \times n}.$$

4. Пусть $\rho(A)$ — спектральный радиус матрицы A. Докажите, что для произвольной операторной нормы $||\cdot||$

$$\rho(A) = \lim_{n \to \infty} ||A^n||^{\frac{1}{n}}.$$

5. Пусть $\lambda_{\min}(\cdot)$ — минимальное по модулю собственное значение, а $\sigma_{\min}(\cdot)$ — минимальное сингулярное число матрицы. Докажите, что для любой квадратной матрицы A

$$|\lambda_{\min}(A)| = \lim_{n \to \infty} (\sigma_{\min}(A^n))^{\frac{1}{n}}.$$

6. Для элементов квадратных матриц A и B имеют место неравенства

$$0 \le a_{ij} \le b_{ij}, \quad 1 \le i, j \le n.$$

Докажите, что $\rho(A) \le \rho(B)$, где $\rho(\cdot)$ — спектральный радиус матрицы.

7. Пусть A — произвольная невырожденная матрица. Всегда ли можно выбрать α так, чтобы матрица $I - \alpha A$ была сходящейся?

8. Пусть A — эрмитова положительно определенная матрица. Докажите, что при всех достаточно малых α матрица $I - \alpha A$ будет сходящейся.

9. Пусть $A = \alpha I - N$, где N — квадратная матрица с неотрицательными элементами и $\alpha > \rho(N)$, где $\rho(\cdot)$ — спектральный радиус матрицы. Докажите, что матрица A невырожденная и все элементы матрицы A^{-1} неотрицательны.

10. Для любой ли матрицы A существует расщепление $A = M - N$ со сходящейся матрицей $M^{-1}N$?

11. Пусть матрица M невырожденная и матрица $M^*M - N^*N$ неотрицательно определенная. Докажите, что $\rho(M^{-1}N) \le 1$ ($\rho(\cdot)$ — спектральный радиус матрицы).

12. Матрицы A и M вещественные, при этом $A = A^\top$ и выполняется *условие Самарского*

$$(Mx, x) > \frac{1}{2}(Ax, x) > 0 \quad \forall \ \ x \neq 0.$$

Докажите, что $\rho(I - M^{-1}A) < 1$.

13. Пусть $A = A^\top > 0$ и $A = L + D + L^\top$, где $D = \mathrm{diag}(A)$ и L — нижняя треугольная матрица. Докажите, что $\rho(I-(L+D)^{-1}A) < 1$. Получите отсюда сходимость *метода Гаусса–Зейделя*:

$$x_{i+1} = x_i + (L + D)^{-1}(b - Ax_i).$$

14. Докажите, что для любого $\varepsilon > 0$ существует нижняя треугольная матрица L такая, что $||L||_2 \leq \varepsilon$ и при этом $I+L$ имеет общую нижнюю треугольную часть с некоторой матрицей ранга 1.

15. Возьмем $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ и рассмотрим следующие матрицы Фробениуса:

$$\Phi = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & 0 & \dots & 0 & -a_1 \\ 0 & 1 & 0 & \dots & 0 & -a_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & -a_{n-2} \\ 0 & 0 & 0 & \dots & 1 & -a_{n-1} \end{bmatrix}, \quad \Psi = \begin{bmatrix} -a_{n-1} & 1 & 0 & \dots & 0 & 0 \\ -a_{n-2} & 0 & 1 & \dots & 0 & 0 \\ -a_{n-3} & 0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -a_1 & 0 & 0 & \dots & 0 & 1 \\ -a_0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

Докажите, что $p(x)$ есть характеристический полином для Φ и для Ψ. Можно ли трансформировать теорему 3.8.1 в теорему о возмущении корней полинома?

16. Найдите собственные значения возмущенной жордановой клетки

$$J(\varepsilon) = \begin{bmatrix} \lambda & 1 & & & 0 \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ \varepsilon & & & & \lambda \end{bmatrix}_{n \times n}.$$

17. Пусть собственные значения вещественной симметричной матрицы A попарно различны. Докажите, что при всех достаточно малых по норме вещественных возмущениях F собственные значения возмущенной матрицы $A + F$ будут вещественными.

Глава 4

4.1 Диагональное преобладание

Матрица $A \in \mathbb{C}^{n\times n}$ имеет *строчное диагональное преобладание*, если

$$|a_{ii}| \;>\; r_i \equiv \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \quad i = 1,\ldots,n, \tag{4.1.1}$$

и *столбцовое диагональное преобладание*, если

$$|a_{jj}| \;>\; c_j \equiv \sum_{\substack{i=1\\ i\neq j}}^{n} |a_{ij}|, \quad j = 1,\ldots,n. \tag{4.1.2}$$

Теорема 4.1.1 (Леви — Деспланк) *Матрица, имеющая строчное или столбцовое диагональное преобладание, является невырожденной.*

Доказательство. Примем обозначения

$$\operatorname{diag}(A) \equiv \operatorname{diag}(a_{11},\ldots,a_{nn}), \qquad \operatorname{off}(A) \equiv A - \operatorname{diag}(A).$$

Тогда неравенства (4.1.1) означают, что $\|[\operatorname{diag}(A)]^{-1}\operatorname{off}(A)\|_\infty < 1$, и поэтому матрица $A = \operatorname{diag}(A) + \operatorname{off}(A)$ невырожденная. Случай (4.1.2) сводится к (4.1.1) переходом к A^T. $\square$

4.2 Круги Гершгорина

Теорема 4.2.1 (Гершгорин) *Для $A \in \mathbb{C}^{n\times n}$ рассмотрим круги*

$$R_i \equiv \{z \in \mathbb{C} : |a_{ii} - z| \le r_i\}, \qquad C_i \equiv \{z \in \mathbb{C} : |a_{ii} - z| \le c_i\},$$

где r_i и c_i имеют вид (4.1.1) и (4.1.2). Тогда если $\lambda \in \lambda(A)$, то

$$\lambda \in \bigcup_{i=1}^{n} R_i \quad и \quad \lambda \in \bigcup_{i=1}^{n} C_i.$$

Доказательство. Если $\lambda \notin \bigcup_i R_i$, то матрица $A-\lambda I$ имеет строчное диагональное преобладание и в силу теоремы Леви — Деспланка невырожденная $\Rightarrow \quad \lambda \notin \lambda(A)$. $\square$

Теорема 4.2.2 *Если m кругов Гершгорина образуют область G, изолированную от других кругов, то в G находятся ровно m собственных значений.*

Доказательство. Положим $A(t) = \operatorname{diag}(A) + t \operatorname{off}(A), 0 \le t \le 1$. Обозначим через $G(t)$ объединение кругов Гершгорина с теми же центрами, что и круги в G, и через $G'(t)$ — объединение остальных кругов. Очевидно, $G(t) \subset G(1) = G$ и $G'(t) \subset G'(1) \equiv C'$.

Поскольку $G \cap G' = \emptyset$, имеем $G(t) \cap G'(t) = \emptyset$ при всех t. Согласно теореме о непрерывной зависимости корней полинома от его коэффициентов, существуют непрерывные функции $\lambda_1(t), \ldots, \lambda_m(t)$ такие, что

$$\begin{array}{ll} \text{(a)} & \{\lambda_1(0), \ldots, \lambda_m(0)\} = G(0); \\ \text{(b)} & \lambda_1(t), \ldots, \lambda_m(t) \in \lambda(A(t)). \end{array}$$

Пусть $t_i \equiv \sup t \ : \ \lambda_i(t) \in G(t)$. Если $t_i < 1$, то при всех $t > t_i$ получаем $\lambda_i(t) \in G'(t)$ (следствие теоремы 4.2.1). Отсюда $G(t_i) \cap G'(t_i) \neq \emptyset$, что невозможно. Поэтому $t_i = 1$ при $i = 1, \ldots, m$. $\square$

Следствие 4.2.1 *Если круги Гершгорина попарно не пересекаются, то каждый из них содержит ровно одно собственное значение.*

4.3 Малые возмущения собственных значений и векторов

Предположим что матрица A имеет лишь простые (то есть попарно различные) собственные значения, а возмущенная матрица имеет вид

$$A(\varepsilon) = A + A_1\varepsilon + \mathcal{O}\left(\varepsilon^2\right).$$

Пусть P — матрица собственных векторов для A; тогда $\Lambda \equiv P^{-1}AP$ — диагональная матрица собственных значений для A. Положим

$$\Omega(\varepsilon) \equiv P^{-1}A(\varepsilon)P = \Lambda + \Omega_1\varepsilon + \mathcal{O}\left(\varepsilon^2\right), \qquad \Omega_1 = P^{-1}A_1P.$$

Матрица $\Omega(\varepsilon)$ имеет те же собственные значения, что и $A(\varepsilon)$. По теоремам Гершгорина, при всех малых ε она имеет простые собственные значения (докажите).

Диагональную матрицу собственных значений для $\Omega(\varepsilon)$ и соответствующую матрицу собственных векторов запишем в виде

$$\Lambda(\varepsilon) = \Lambda + \Lambda_1\varepsilon + \hat{\Lambda}(\varepsilon), \qquad \Lambda(0) = \Lambda;$$

$$Z(\varepsilon) = I + Z_1\varepsilon + \hat{Z}(\varepsilon), \qquad Z(0) = I.$$

Матрица $Z(\varepsilon)$ определяется с точностью до нормировки столбцов. Тем не менее, при любой нормировке при всех достаточно малых $\varepsilon > 0$ имеем $\mathrm{diag}Z(\varepsilon) \neq 0$ (докажите). Поэтому мы потребуем, чтобы

$$\mathrm{diag}Z(\varepsilon) = I, \quad \mathrm{diag}Z_1 = 0 \quad \Rightarrow \quad \mathrm{diag}\hat{Z}(\varepsilon) = 0.$$

Рассмотрим равенство

$$\left(\Lambda + \Omega_1\varepsilon + \mathcal{O}\left(\varepsilon^2\right)\right)\left(I + Z_1\varepsilon + \hat{Z}\right) = \left(I + Z_1\varepsilon + \hat{Z}\right)\left(\Lambda + \Lambda_1\varepsilon + \hat{\Lambda}\right). \quad (*)$$

Перепишем его в виде

$$(\Lambda Z_1 - Z_1\Lambda + \Omega_1 - \Lambda_1)\varepsilon + \ldots = 0.$$

и определим Λ_1 и Z_1 из уравнения

$$\Lambda Z_1 - Z_1\Lambda = \Lambda_1 - \Omega_1 \quad \Rightarrow$$

$$\Lambda_1 = \mathrm{diag}\Omega_1, \quad \Lambda Z_1 - Z_1\Lambda = -\mathrm{off}\Omega_1. \qquad (**)$$

Итак, пусть Λ_1 и Z_1 удовлетворяют $(**)$. Будем рассматривать $(*)$ как уравнение относительно $\hat{\Lambda}$ и $\hat{Z}$. Сначала докажем, что $\hat{\Lambda} = \mathcal{O}\left(\varepsilon^2\right)$. Для этого заметим, что

$$\left(\Lambda + \Omega_1\varepsilon + \mathcal{O}\left(\varepsilon^2\right)\right)(I + Z_1\varepsilon) - (I + Z_1\varepsilon)(\Lambda + \Lambda_1\varepsilon) = \mathcal{O}\left(\varepsilon^2\right).$$

При малых ε имеем $\left\|(I + Z_1\varepsilon)^{-1}\right\| = \mathcal{O}(1)$. Следовательно, в силу теорем Гершгорина собственные значения матрицы $\Lambda + \Omega_1\varepsilon + \mathcal{O}\left(\varepsilon^2\right)$ с точностью $\mathcal{O}\left(\varepsilon^2\right)$ суть диагональные элементы $\Lambda_0 + \Lambda_1\varepsilon \;\Rightarrow\; \hat{\Lambda} = \mathcal{O}\left(\varepsilon^2\right)$.

Теперь докажем, что $\hat{Z} = \mathcal{O}\left(\varepsilon^2\right)$. Для этого заметим, что элементы матрицы $\hat{Z}$ являются единственным решением системы линейных уравнений

$$\left(\Lambda + \Omega_1\varepsilon + \mathcal{O}\left(\varepsilon^2\right)\right)\hat{Z} - \hat{Z}\left(\Lambda + \Lambda_1\varepsilon + \mathcal{O}\left(\varepsilon^2\right)\right) = \mathcal{O}\left(\varepsilon^2\right), \quad \mathrm{diag}\hat{Z} = 0,$$

а матрица коэффициентов этой системы есть ε-возмущение невырожденной матрицы коэффициентов аналогичной системы при $\varepsilon = 0$.

Таким образом, получена

Теорема 4.3.1 *Пусть $P^{-1}AP = \Lambda$ — диагональная матрица с попарно различными собственными значениями для A. Тогда при малых ε матрица $A(\varepsilon) = A + A_1\varepsilon + \mathcal{O}\left(\varepsilon^2\right)$ диагонализуема:*

$$P^{-1}(\varepsilon)A(\varepsilon)P(\varepsilon) = \Lambda(\varepsilon),$$

и при этом

$$\Lambda(\varepsilon) = \Lambda + \Lambda_1\varepsilon + \mathcal{O}\left(\varepsilon^2\right), \qquad P(\varepsilon) = P\left(I + Z_1\varepsilon + \mathcal{O}\left(\varepsilon^2\right)\right),$$

где

$$\Lambda_1 = \operatorname{diag}\left(P^{-1}A_1P\right),$$

а Z_1 однозначно определяется из уравнений

$$\operatorname{diag}Z_1 = 0, \quad \Lambda Z_1 - Z_1\Lambda = -\operatorname{off}\left(P^{-1}A_1P\right).$$

Следствие 4.3.1 *Собственные значения $\lambda_i(\varepsilon)$ для $A(\varepsilon)$ имеют вид*

$$\lambda_i(\varepsilon) = \lambda_i + q_i^TA_1p_i\varepsilon + \mathcal{O}\left(\varepsilon^2\right),$$

где q_i^T суть строки матрицы P^{-1}.

4.4 Обусловленность простого собственного значения

Будем считать, что $\|A_1\|_2 = 1$. Тогда, с учетом равенства $q_i^Tp_i = 1$,

$$\|q_i^TA_1p_i\|_2 \;\le\; \frac{\|q_i^T\|_2\|p_i^T\|_2}{|q_i^Tp_i|}.$$

Число

$$s(\lambda_i) \equiv \frac{\|q_i^T\|_2\|p_i^T\|_2}{|q_i^Tp_i|}$$

называется *обусловленностью* (или *коэффициентом перекоса*) собственного значения λ_i. Равенства $Ap_i = \lambda_ip_i$ и $q_i^TA = \lambda_iq_i^T$ означают, что векторы p_i и q_i — это соответственно левый и правый собственные векторы матрицы A. Ясно, что обусловленность не зависит от нормировки векторов p_i и q_i.

Обусловленность корректно определяется для простого собственного значения и в случае недиагонализуемой матрицы. Важно, что следствие 4.3.1 остается в силе для любого простого собственного значения (при условии нормировки $q_i^Tp_i = 1$). (Это можно доказать с помощью перехода к слабо возмущенной, но уже диагонализуемой матрице, имеющей те же векторы p_i и q_i для простого собственного значения λ_i).

Обусловленность собственного значения характеризует расстояние до матриц, для которых данное собственное значение является кратным.

Теорема 4.4.1 (Уилкинсон) *Пусть A имеет простое собственное значение λ_i с обусловленностью $s(\lambda_i)$. Тогда существует матрица $A+E$, для которой λ_i — кратное собственное значение и при этом*

$$\|E\|_2 \le \frac{\|A\|_2}{\sqrt{s^2(\lambda_i)-1}}. \tag{4.4.3}$$

Доказательство. Не ограничивая общности, можно считать, что A имеет форму Шура

$$A = \begin{bmatrix} \lambda_i & z^\top \\ 0 & B \end{bmatrix}.$$

Правый и левый собственные векторы для λ_i имеют вид

$$p = [1\ 0\ \ldots\ 0]^\top \quad \text{и} \quad q^\top = \begin{bmatrix} 1\ v^\top \end{bmatrix} \quad \Rightarrow \quad s(\lambda_i) = \left(\|v\|_2^2+1\right)^{1/2}.$$

Очевидно, $v^\top B + z^\top = \lambda_i v^\top \ \Rightarrow\ v^T(B-\lambda_i I) = z^\top \ \Rightarrow$ матрица $\tilde{B} \equiv B + \frac{vz^\top}{\|v\|_2^2}$ имеет λ_i своим собственным значением, то есть можно взять

$$E = \begin{bmatrix} 0 & 0 \\ 0 & \frac{vz^T}{\|v\|_2^2} \end{bmatrix} \quad \Rightarrow \quad \|E\|_2 \ \le\ \frac{\|z^T\|_2}{\|v\|_2} \ \le\ \frac{\|A\|_2}{\|v\|_2} \ =\ \frac{\|A\|_2}{\sqrt{s^2(\lambda_i)-1}}.$$

□

4.5 Аналитические возмущения

Пусть ряд $A(\varepsilon) = \sum_{k=0}^{\infty} A_k \varepsilon^k$ сходится при всех $|\varepsilon| < \varepsilon_0$. Тогда если A_0 имеет лишь простые собственные значения, то при всех досаточно малых ε матрица $A(\varepsilon)$ диагонализуется при помощи $P(\varepsilon)$:

$$P^{-1}(\varepsilon)A(\varepsilon)P(\varepsilon) = \Lambda(\varepsilon), \tag{4.5.4}$$

где

$$\Lambda(\varepsilon) = \sum_{k=0}^{\infty} \Lambda_k \varepsilon^k, \qquad P(\varepsilon) = \sum_{k=0}^{\infty} P_k \varepsilon^k. \tag{4.5.5}$$

Существование и сходимость при всех малых ε ряда $\Lambda(\varepsilon)$ вытекает из "аналитического" варианта теоремы о неявной функции.

Матрицы Λ_k, P_k легко определяются. Положим $Z_k \equiv P_0^{-1}P_k$, $\Omega_k \equiv P_0^{-1}A_kP_0$. Тогда

$$(\Lambda_0 + \Omega_1\varepsilon + \ldots)(I + Z_1\varepsilon + \ldots) = (I + Z_1\varepsilon + \ldots)(\Lambda_0 + \Lambda_1\varepsilon + \ldots).$$

Приравнивания коэффициенты при ε^k, находим

$$\Lambda_0 Z_k - Z_k\Lambda_0 = \Lambda_k - \Phi_k, \tag{4.5.6}$$

где

$$\Phi_k = \sum_{i=1}^{k-1} (\Omega_i Z_{k-i} - Z_{k-i}\Omega_i) + \Omega_k. \tag{4.5.7}$$

Отсюда

$$\Lambda_k = \mathrm{diag}\Phi_k, \qquad \Lambda_0 Z_k - Z_k \Lambda_0 = -\mathrm{off}\Phi_k. \tag{4.5.8}$$

Зная Λ_i, Z_i при $i \le k-1$, мы можем получить Λ_k, Z_k из 4.5.8.

Рассмотрим оператор $\mathcal{A}:\ Z \mapsto \Lambda_0 Z - Z\Lambda_0$ на пространстве матриц Z таких, что $\mathrm{diag}Z = 0$. В силу простоты собственных значений оператор $\mathcal{A}$ обратимый. Тогда при малых ε будет обратимым и оператор $\mathcal{A} + \varepsilon\mathcal{B}$, где

$$\mathcal{B}:\ Z \mapsto \Omega_1(\varepsilon)Z - Z\Lambda_1(\varepsilon),$$
$$\Omega_1(\varepsilon) = \sum_{k=0}^{\infty} \Omega_{k+1}\varepsilon^k, \quad \Lambda_1(\varepsilon) = \sum_{k=0}^{\infty} \Lambda_{k+1}\varepsilon^k.$$

Положим

$$Z_1(\varepsilon) = \sum_{k=0}^{\infty} Z_{k+1}\varepsilon^k.$$

Тогда

$$[\mathcal{A} + \varepsilon\mathcal{B}]\, Z_1(\varepsilon) = \Lambda_1(\varepsilon) - \Omega_1(\varepsilon).$$

Отсюда видно, что $Z_1(\varepsilon)$ представляется сходящимся степенным рядом (если известно, что $\Lambda_1(\varepsilon)$ представляется таковым рядом).

Если A_0 имеет кратные собственные значения, то собственные значения и векторы представляются рядами Пюизье (то есть рядами по дробным степеням ε).

Задачи

1. Пусть $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, где λ_i — вещественные попарно различные числа. Пусть возмущенная матрица $A(\varepsilon) = A + A_1\varepsilon$ является эрмитовой и $\mathrm{diag}(A_1) = 0$. Докажите, что собственные значения $A(\varepsilon)$ имеют вид
$$\lambda_i(\varepsilon) = \lambda_i + \mathcal{O}\left(\varepsilon^2\right).$$

2. Пусть λ — простое собственное значение матрицы A; p и q — соответствующие ему левый и правый собственные векторы такие, что $q^T p = 1$. Докажите, что возмущенная матрица $A(\varepsilon) = A + A_1\varepsilon$ при всех достаточно малых ε имеет простое собственное значение
$$\lambda(\varepsilon) = \lambda + q^T A_1 p\varepsilon + \mathcal{O}\left(\varepsilon^2\right).$$

3. Если есть большой коэффициент перекоса, то хотя бы еще один будет большим. Объясните, почему.

4. Пусть A имеет простые собственные значения с коэффициентами перекоса $s_1, \ldots, s_n$. Докажите, что если P — матрица из собственных векторов, то
$$\operatorname{cond}_2 P \geq \max_{1 \leq i \leq n} s_i.$$

5. Пусть A имеет простые собственные значения с коэффициентами перекоса $s_1, \ldots, s_n$. Докажите, что $s_1 = \ldots = s_n = 1$ тогда и только тогда, когда матрица A нормальная.

6. Объясните, почему почти треугольная матрица
$$A = \begin{bmatrix} n & n-1 & n-2 & \ldots & 3 & 2 & 1 \\ n-1 & n-1 & n-2 & \ldots & 3 & 2 & 1 \\ & n-2 & n-2 & \ldots & 3 & 2 & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ & & & & 2 & 2 & 1 \\ & & & & & 1 & 1 \end{bmatrix}_{n \times n}$$
имеет плохо обусловленные собственные значения при больших n.

7. Пусть A диагонализуется матрицей P:
$$P^{-1} A P = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$
и пусть $A + F$ — возмущенная матрица. Пусть среди кругов
$$B_i = \left\{ z : \ |z - \lambda_i| \leq \|P^{-1} F P\|_2 \right\}, \quad i = 1, \ldots, n.$$
объединение каких-то m из них образует область M, не пересекающуюся с остальными кругами. Докажите, что в M содержится ровно m собственных значений матрицы $A + F$.

8. Пусть все элементы матрицы A отличны от нуля. Тогда любое собственное значение $\lambda \in \lambda(A)$ либо является внутренней точкой объединения кругов Гершгорина $R_1 \cup \ldots \cup R_n$, либо является общей граничной точкой для всех кругов Гершгорина $R_1, \ldots, R_n$. Докажите.

9. Пусть $\sigma_1 \geq \ldots \geq \sigma_n$ — сингулярные числа $n \times n$-матрицы
$$A = \begin{bmatrix} 1 & 2 & & & \\ & 1 & 2 & & \\ & & \ddots & \ddots & \\ & & & 1 & 2 \\ & & & & 1 \end{bmatrix}.$$
Докажите, что $1 \leq \sigma_{n-1} \leq \ldots \leq \sigma_1 \leq 3$ и, кроме того, $0 < \sigma_n < 2^{-n+1}$.

10. Докажите, что все собственные значения матрицы $A = [a_{ij}] \in \mathbb{C}^{n\times n}$ принадлежат объединению следующих множеств (известных как *овалы Кассини*):

$$M_{ij} \;=\; \{z \in \mathbb{C} : |a_{ii} - z||a_{jj} - z| \le r_i r_j\}, \quad 1 \le i, j \le n, \;\; i \ne j,$$

где

$$r_i = \sum_{l \ne i} |a_{il}|.$$

11. Приведите пример матрицы $A = [a_{ij}]$, для которой не все собственные значения принадлежат объединению множеств

$$M_{ijk} = \{z \in \mathbb{C} : |a_{ii} - z||a_{jj} - z||a_{kk} - z| \le r_i r_j r_k\},$$

$$1 \le i, j, k \le n, \;\; i \ne j, \;\; j \ne k, \;\; i \ne k.$$

12. Дана матрица A порядка n с положительными элементами. Докажите, что минимальное значение функционала $f(D) = ||D^{-1}AD||_\infty$ на множестве всех диагональных матриц с положительными диагональными элементами достигается на некоторой матрице $\widetilde{D}$ такой, что для $\widetilde{D}^{-1}A\widetilde{D}$ все строчные суммы одинаковы. Докажите, что вектор $[\widetilde{d}_{11}, ..., \widetilde{d}_{nn}]^\top$ является собственным вектором матрицы A, а соответствующее собственное значение равно спектральному радиусу матрицы A (*теорема Перрона–Фробениуса*).

13. Пусть $A = [a_{ij}] \in \mathbb{C}^{n\times n}$ и $|a_{ij}| \le b_{ij}$, где $b_{ij} > 0$ для всех i, j. Используя предыдущую задачу, докажите, что все собственные значения A принадлежат объединению кругов

$$\Phi_i = \{z \in \mathbb{C} : |a_{ii} - z| \le \rho(B) - b_{ii}, \quad 1 \le i \le n,$$

где $\rho(\cdot)$ — спектральный радиус (*теорема Фань–Цзы*).

Глава 5

5.1 Спектральные расстояния

Формулировки следствий теоремы Бауэра–Файка "несимметричны" относительно A и $A+F$: матрица $A+F$, в отличие от A, может быть недиагонализуемой или иметь другие порядки жордановых клеток. Теперь мы займемся "симметричными" теоремами — в них будет оцениваться некоторое расстояние между спектрами матриц.

Хаусдорфово спектральное расстояние между A и B с собственными значенями $\{\lambda_i\}$ и $\{\mu_j\}$ определяется так:

$$\mathrm{hd}(A,\ B) \equiv \max\{\max_i \min_j |\lambda_i - \mu_j|, \max_j \min_i |\lambda_i - \mu_j|\}.$$

Спектральное p-расстояние определяется так:

$$\mathrm{d_p}\,(A,\ B) \equiv \min_P ||\lambda(A) - P\lambda(B)||_p,$$

где минимум берется по всем матрицам перестановки P и подразумевается, что

$$\lambda(A) = [\lambda_1, \ldots, \lambda_n]^\top, \quad \lambda(B) = [\mu_1, \ldots, \mu_n]^\top.$$

5.2 "Симметричные" теоремы

Теорема 5.2.1 (Элснер).

$$\mathrm{hd}(A,\ B) \ \leq\ (||A||_2 + ||B||_2)^{1-\frac{1}{n}} ||A-B||_2^{\frac{1}{n}}.$$

Доказательство. Пусть векторы $x_1, \ldots, x_n$ образуют столбцы унитарной матрицы X и $Bx_1 = \mu x_1$. Если $\lambda_1, ..., \lambda_n$ — собственные значения A, то

$$\prod_{i=1}^{n} |\lambda_i - \mu| = |\det\left((A - \mu\, I)\, X\right)|$$

(вспомним *неравенство Адамара*: модуль определителя не превосходит произведения 2-норм столбцов $\Leftrightarrow$ объем n-мерного кубоида не превосходит произведения его длин)

$$\begin{aligned} &\leq \prod_{i=1}^{n} ||(A-\mu I)\,x_i||_2 \;\leq\; ||(A-B)\,x_1||_2 \prod_{i=2}^{n} ||(A-\mu I)\,x_i||_2 \\ &\leq ||A-B||_2\,(||A||_2+||B||_2)^{n-1}. \quad \square \end{aligned}$$

Теорема 5.2.2 (Островский—Элснер).

$$\mathrm{d}_\infty(A,\ B) \;\leq\; (2n-1)\ \mathrm{hd}(A,\ B).$$

Доказательство. Рассмотрим круги

$$D_i = \{z:\ |z-\lambda_i| \;\leq\; \mathrm{hd}\,(A,\ B)\} \quad \text{и} \quad D_i(\tau) = \{z:\ |z-\lambda_i| \;\leq\; \varepsilon(\tau)\},$$

где

$$\varepsilon(\tau) \equiv \left(2 \max_{0\leq t\leq 1} ||A+t\,(B-A)||_2\right)^{1-\frac{1}{n}}\ ||\tau\,(B-A)||^{\frac{1}{n}}, \quad 0\leq\tau\leq 1.$$

Согласно теореме 5.2.1, все собственные значения матрицы $A+\tau\,(B-A)$ принадлежат объединению кругов $D_i(\tau)$. По аналогии с кругами Гершгорина, если m кругов $D_i(\tau)$ изолированы от остальных кругов, то их объединение содержит в точности m собственных значений матрицы $A+\tau\,(B-A)$. $\Rightarrow$ Если m кругов $D_i\,(1)$ изолированы от остальных кругов, то объединение соответствующих кругов D_i содержит в точности m собственных значений μ_i (почему?). Пусть нумерация кругов и собственных значений μ_i для B таковы, что

$$\mu_i \in \bigcup_{1\leq k\leq m} D_k, \quad i=1,\ldots,m.$$

Тогда $|\mu_i-\lambda_j| \leq (2m-1)\,\mathrm{hd}(A,\ B),\ 1\leq i,j\leq m.$ $\square$

5.3 Теорема Виландта—Хоффмана

Теорема 5.3.1 (Виландт—Хоффман). *Для нормальных матриц A и B*

$$\mathrm{d}_2(A,\ B) \;\leq\; ||A-B||_F.$$

Доказательство. Поскольку матрицы A и B нормальные, с помощью унитарных матриц P и Q мы можем получить диагональные матрицы

$$D_A = \mathrm{diag}(\lambda_i) = P^*AP \quad \text{и} \quad D_B = \mathrm{diag}(\mu_i) = Q^*BQ.$$

Положим $Z \equiv P^*Q$. Учитывая унитарную инвариантность нормы Фробениуса, находим

$$\begin{aligned} ||A-B||_F^2 &= ||D_A - ZD_BZ^*||_F^2 \\ &= \mathrm{tr}(D_A - ZD_BZ^*)\,(D_A - ZD_BZ^*)^* \\ &= ||D_A||_F^2 + ||D_B||_F^2 - 2\,\mathrm{Re}\ \mathrm{tr}(ZD_BZ^*D_A^*). \end{aligned}$$

Чтобы получить оценку снизу, запишем

$$\gamma \equiv 2\,\mathrm{Re}\ \mathrm{tr}(ZD_BZ^*D_A^*) = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\alpha_{ij}, \quad s_{ij} = |z_{ij}|^2, \quad \alpha_{ij} = 2\,\mathrm{Re}\ \lambda_i^*\mu_j.$$

При фиксированных α_{ij} функционал $\gamma = \gamma\,(S)$ является линейным на множестве матриц $S = [s_{ij}]$ и нас интересует его максимум на множестве матриц S с неотрицательными элементами и всеми строчными и столбцовыми суммами, равными 1. Такие матрицы S называются *двоякостохастическими*. Для них имеет место следующая

Теорема Биркгоффа. *Любая двоякостохастическая матрица S представима в виде выпуклой комбинации конечного числа матриц перестановки P_k (сами матрицы перестановки и их число зависят от S)*:

$$S = \sum_{k=1}^{m} \nu_k\, P_k, \quad \nu_1 + ... + \nu_m = 1, \quad \nu_1, ..., \mu_m \ \geq\ 0.$$

В силу теоремы Биркгоффа

$$\gamma\,(S) \leq \max_{1\leq k\leq m} \gamma\,(P_k),$$

то есть максимальное значение функции γ достигается на некоторой матрице перестановки. Обозначим ее через Π. Тогда

$$\begin{aligned} ||A-B||_F^2 &= ||D_A||_F^2 + ||D_B||_F^2 - 2\,\mathrm{Re}\ \mathrm{tr}(ZD_BZ^*D_A^*) \\ &\geq ||D_A||_F^2 + ||\Pi D_B\Pi^*||_F^2 - 2\,\mathrm{Re}\ \mathrm{tr}(\Pi D_B\Pi^*D_A^*) \\ &= ||D_A - \Pi D_B\Pi^*||_F^2 \ \geq\ \mathrm{d}_2\,(A,\,B). \quad \square \end{aligned}$$

Следущий раздел — для тех, кто хотел бы узнать, как доказывается теорема Биркгоффа.

5.4 Перестановочные диагонали

Под перестановочной диагональю $n \times n$-матрицы A, отвечающей матрице перестановки P, понимается вектор, составленный из компонент главной

диагонали матрицы $\mathrm{diag}\,(P^T A)$. Позиции извлекаемых из A элементов совпадают с позициями единиц в P.

Теорема Биркгоффа — почти очевидный факт для тех, кому кажется очевидным, что в любой двоякостохастической матрице можно выделить ненулевую перестановочную диагональ. Действительно, если ν_1 — минимальная компонента ненулевой перестановочной диагонали, отвечающей матрице перестановки P_1, то $S - \nu_1 P_1 = \phi_1 S_1$, где $0 \leq \nu_1, \phi_1 \leq 1$, $\nu_1 + \phi_1 = 1$ и S_1 есть двоякостохастическая матрица, в которой, по меньшей мере, одним нулевым элементом больше, чем в исходной матрице S. Далее по индукции.

Но попробуйте все же *доказать* существование ненулевой перестановочной диагонали — это не очень просто! Мы сделаем это, вооружившись следующей (нетривиальной) теоремой.

Теорема Холла. *Для того чтобы любая перестановочная диагональ в $n \times n$-матрице A содержала нуль, необходимо и достаточно, чтобы в A существовала нулевая $p \times q$-подматрица такая, что $p + q > n$.*

Необходимость. Пусть любая перестановочная диагональ матрицы A порядка $n > 1$ содержит нуль. Если $A = 0$, то все доказано. Если $A \neq 0$, то будем считать, что $a_{1n} \neq 0$, и запишем

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1\,n-1} & a_{1n} \\ & & & a_{2n} \\ & B & & \cdots \\ & & & a_{nn} \end{bmatrix}.$$

Тогда матрица B порядка $n-1$ обязана иметь нуль в любой своей перестановочной диагонали. Предположим (индукция по n), что для B существует нулевая $p \times q$-подматрица такая, что $p + q > n - 1$. Если $p + q > n$, то эта подматрица является искомой.

Предположим, что $p + q = n$ и A имеет вид

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0_{p\times q} & A_{22} \end{bmatrix},$$

где блоки A_{11} и A_{22} квадратные порядка q и p, соответственно. Если A_{22} имеет перестаночную диагональ без нулей, то в любую перестановочную диагональ для A_{11} обязан входить нуль. По индуктивному предположению, в A_{11} есть нулевая $r \times s$-подматрица с размерами что $r + s > q$. Не ограничивая общности, предположим, что эта подматрица находится в левом нижнем углу матрицы A_{11}. Тогда в левом нижнем углу матрицы A располагается нулевая подматрица размеров $(r + p) \times s$ и при этом $(r + p) + s = (r + s) + p > q + p = n$.

Если любая перестановочная диагональ A_{22} содержит нуль, то индуктивное предположение можно применить непосредственно к A_{22}.

Достаточность. Не ограничивая общности, можно считать, что $p \geq q$ и нулевая подматрица размеров $p \times q$ занимает нижний левый угол. Тогда A имеет вид

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0_{(n-q)\times q} & A_{22} \end{bmatrix}, \qquad (*)$$

причем последняя строка в A_{11} нулевая, так как $p > n - q$. Поэтому любая перестановочная диагональ матрицы A_{11} обязана содержать элемент этой строки, то есть нуль. $\square$

Следствие. *В любой двоякостохастической матрице существует ненулевая перестановочная диагональ.*

Доказательство. Очевидно, если $n = 1$. Далее по индукции. Если это не так для двоякостохастической матрицы A порядка $n > 1$, то в силу теоремы Холла в A существует нулевая $p \times q$-подматрица такая, что $p + q > n$. Не ограничивая общности, можно полагать, что A имеет вид $(*)$. Сумма всех элементов матрицы A_{11} равна q $\Rightarrow$ сумма всех элементов матрицы A_{12} равна нулю (почему?) $\Rightarrow$ $A_{12} = 0$ $\Rightarrow$ каждая из матриц A_{11} и A_{22} является двоякостохастической. Перестановочная диагональ для A может быть составлена из ненулевых перестановочных диагоналей для A_{11} и A_{22}. $\square$

5.5 "Ненормальное" обобщение

В конце 1980-х годов получено обобщение теоремы Виландта – Хоффмана на случай произвольных диагонализуемых матриц (не обязательно нормальных).

Теорема 5.5.1 (Сан–Зхенг) *Пусть A и B диагонализуемы, а P и Q — соответствующие матрицы собственных векторов. Тогда*

$$\mathrm{d}_2(A,\, B) \;\leq\; \mathrm{cond}_2(P)\, \mathrm{cond}_2(Q)\, ||A - B||_F,$$

где

$$\mathrm{cond}_2(P) \equiv ||P^{-1}||_2 ||P||_2, \quad \mathrm{cond}_2(Q) \equiv ||Q^{-1}||_2 ||Q||_2.$$

Доказательство.

$$\begin{aligned} ||A - B||_F &= ||PD_AP^{-1} - QD_BQ^{-1}||_F \\ &= ||P\{D_A(P^{-1}Q) - (P^{-1}Q)D_B\}Q^{-1}||_F \\ &\geq \frac{1}{||P^{-1}||_2||Q||_2}||D_AZ - ZD_B||_F, \quad \text{где } Z \equiv P^{-1}Q. \end{aligned}$$

Рассмотрим сингулярное разложение $Z = V\Sigma U^*$. Тогда

$$||D_A Z - Z D_B||_F = ||V\{(V^* D_A V)\Sigma - \Sigma(U^* D_B U)\}U^*||_F = ||M\Sigma - \Sigma N||_F,$$

где матрицы $M = V^* D_A V$ и $N = U^* D_B U$, очевидно, нормальные. В силу теоремы Виландта—Хоффмана,

$$\mathrm{d}_2(A,\, B) = \mathrm{d}_2(D_A,\, D_B) \;\leq\; ||M - N||_F.$$

Поэтому достаточно убедиться в том, что

$$||M\Sigma - \Sigma N||_F \;\geq \sigma_{\min}||M - N||_F.$$

Положим $\Omega = \Sigma - \sigma_{\min} I$. Тогда

$$\begin{aligned} ||M\Sigma - \Sigma N||_F^2 &= ||(M\Omega - \Omega N) + \sigma_{\min}(M - N)||_F^2 \\ &= ||M\Omega - \Omega N||_F^2 + \sigma_{\min}^2||M - N||_F^2 \\ &+ \sigma_{\min}\mathrm{tr}\,(\,(M\Omega - \Omega N)(M - N)^* + (M - N)(M\Omega - \Omega N)^*\,)\,. \end{aligned}$$

С использованием нормальности M и N последнее слагаемое преобразуется к виду

$$\sigma_{\min}\,\mathrm{tr}\,\Omega\,(\,(M - N)(M - N)^* + (M - N)^*(M - N)\,) \geq 0.$$

□

5.6 Собственные значения эрмитовых матриц

Теорема 5.6.1 (Курант–Фишер) *Пусть* $\lambda_1 \geq \cdots \geq \lambda_n$ *— собственные значения эрмитовой матрицы* A *порядка* n*. Тогда*

$$\lambda_k \;=\; \max_{\dim L = k}\; \min_{\substack{x\in L\\ x\neq 0}} \frac{x^* A\, x}{x^* x} \tag{5.6.1}$$

и одновременно

$$\lambda_k \;=\; \min_{\dim L = n-k+1}\; \max_{\substack{x\in L\\ x\neq 0}} \frac{x^* A\, x}{x^* x}. \tag{5.6.2}$$

Доказательство. Пусть $u_1, \ldots, u_n$ — ортонормированнные собственные векторы, отвечающие соответственно $\lambda_1, \ldots, \lambda_n$. Если

$$x \;=\; \sum_{i=1}^{k} \xi_i\, u_i, \quad \text{то} \quad \frac{x^* A\, x}{x^* x} \geq \lambda_k.$$

Поэтому правая часть (5.6.1) не может быть меньше λ_k (почему?).

Теперь возьмем произвольное k-мерное подпространство M. Существует ненулевой вектор (почему?) $z \in M \cap K$, $K \equiv \mathrm{span}\{u_k, u_{k+1}, \ldots, u_n\}$, и для него

$$\frac{z^* A z}{z^* z} \le \max_{\substack{x \in K \\ x \neq 0}} \frac{x^* A x}{x^* x} \le \lambda_k.$$

Поэтому правая часть (5.6.1) не может быть больше λ_k. Соотношение (5.6.2) доказывается аналогично. $\square$

5.7 Соотношения разделения

Теорема 5.7.1 *Пусть A — эрмитова матрица порядка n, и B — ее ведущая подматрица порядка $n-1$. Тогда для собственных значений $\lambda_1 \ge \cdots \ge \lambda_n$ матрицы A и собственных значений $\mu_1 \ge \cdots \ge \mu_{n-1}$ матрицы B выполняются следующие соотношения разделения:*

$$\lambda_k \ge \mu_k \ge \lambda_{k+1}, \quad k = 1, \ldots, n-1. \tag{5.7.3}$$

Доказательство. Пусть M — подпространство векторов $x \in \mathbb{C}^n$ вида

$$x = \begin{bmatrix} \hat{x} \\ 0 \end{bmatrix}, \quad \hat{x} \in \mathbb{C}^{n-1} \quad \Rightarrow \quad \frac{x^* A x}{x^* x} = \frac{\hat{x}^* B \hat{x}}{\hat{x}^* \hat{x}}.$$

Согласно (5.6.1) и (5.6.2),

$$\mu_k = \max_{\substack{\dim L = k \\ L \subset M}} \min_{\substack{x \in L \\ x \neq 0}} \frac{x^* A x}{x^* x} \le \max_{\dim L = k} \min_{\substack{x \in L \\ x \neq 0}} \frac{x^* A x}{x^* x} = \lambda_k.$$

$$\mu_k = \min_{\substack{\dim L = (n-1)-k+1 \\ L \subset M}} \max_{\substack{x \in L \\ x \neq 0}} \frac{x^* A x}{x^* x} \ge \min_{\dim L = n-(k+1)+1} \max_{\substack{x \in L \\ x \neq 0}} \frac{x^* A x}{x^* x} = \lambda_{k+1}. \quad \square$$

Теорема 5.7.2 *Пусть A и B — эрмитовы матрицы порядка n с собственными значениями $\lambda_1 \ge \ldots \ge \lambda_n$ и $\mu_1 \ge \ldots \ge \mu_n$. Тогда если*

$$B = A + \varepsilon p p^*, \quad \varepsilon > 0, \quad p \in \mathbb{C}^n, \quad ||p||_2 = 1, \tag{5.7.4}$$

то выполняются соотношения разделения

$$\mu_1 \ge \lambda_1 \ge \mu_2 \ge \ldots \ge \lambda_{n-1} \ge \mu_n \ge \lambda_n \tag{5.7.5}$$

и при этом

$$\mu_k = \lambda_k + t_k \varepsilon, \quad t_k \ge 0, \quad k = 1, \ldots, n; \qquad t_1 + \ldots + t_n = 1.$$

Доказательство. По теореме 5.6.1,

$$\lambda_k = \max_{\dim L=k} \min_{\substack{x\in L\\ x\neq 0}} \frac{x^*Ax}{x^*x} \le \max_{\dim L=k} \min_{\substack{x\in L\\ x\neq 0}} \frac{x^*Bx}{x^*x} = \mu_k.$$

При $k \ge 2$ находим

$$\begin{aligned} \mu_k &\le \max_{\dim L=k} \min_{\substack{x\in L, x\perp p\\ x\neq 0}} \frac{x^*Bx}{x^*x} = \max_{\dim L=k} \min_{\substack{x\in L, x\perp p\\ x\neq 0}} \frac{x^*Ax}{x^*x} \\ &\le \max_{\dim L=k-1} \min_{\substack{x\in L\\ x\neq 0}} \frac{x^*Ax}{x^*x} = \lambda_{k-1}. \end{aligned}$$

Ясно, что $\mu_1 \le \lambda_1 + \varepsilon$ (почему?). Поэтому возможность записи

$$\mu_k = \lambda_k + t_k\varepsilon, \quad 0 \le t_k \le 1, \quad 1 \le k \le n,$$

очевидна. Сложив эти равенства при всех k, получим

$$\operatorname{tr} B = \operatorname{tr} A + \varepsilon(t_1 + ... + t_n).$$

В то же время из (5.7.4) вытекает, что

$$\operatorname{tr} B = \operatorname{tr} A + \varepsilon \quad \Rightarrow \quad t_1 + ... + t_n = 1. \qquad \square$$

5.8 Что такое кластеры?

Рассмотрим последовательность наборов (комплексных) чисел

$$z_{11}; \quad z_{12},\, z_{22}; \quad z_{13},\, z_{23},\, z_{33}; \quad \dots \ .$$

Кластер — это множество M на комплексной плоскости, "притягивающее" к себе почти все числа n-го набора при $n \to \infty$.

Чтобы дать строгое определение, обозначим через $\gamma_n(\varepsilon)$, $\varepsilon > 0$, число ε-*удаленных элементов* n-го набора — отстоящих от любого элемента M на расстояние больше, чем ε. Множество M называется (общим) *кластером* для z_{in}, если

$$\lim_{n\to\infty} \frac{\gamma_n(\varepsilon)}{n} = 0 \quad \forall\, \varepsilon > 0,$$

и *собственным* (или *сильным*) *кластером*, если существует функция $c(\varepsilon)$ такая, что

$$\gamma_n(\varepsilon) \le c(\varepsilon) \quad \forall\, n, \quad \forall\, \varepsilon > 0.$$

В приложениях часто возникают последовательности матриц $A_n \in \mathbb{C}^{n\times n}$, порожденные каким-то общим процессом (например, дискретизации

линейного оператора на последовательности все более мелких сеток). В этих случаях полезно знать, какие кластеры имеются у соответствующих им наборов сингулярных чисел A_n или наборов собственных значений A_n. Особый интерес представляют кластеры, состоящие из одной или нескольких точек.

5.9 Кластеры сингулярных чисел

Чтобы доказать существование кластера сингулярных чисел для какой-то последовательности матриц, достаточно найти "близкую" последовательность, для которой наличие кластера очевидно. В данном случае понятие "близости" может пониматься в весьма широком смысле. [1]

Теорема 5.9.1 *Пусть даны последовательности матриц A_n и B_n такие, что*

$$||A_n - B_n||_F^2 = o(n) \tag{5.9.6}$$

или

$$\operatorname{rank}(A_n - B_n) = o(n). \tag{5.9.7}$$

В любом из двух случаев всякий кластер сингулярных чисел A_n является кластером сингулярных чисел B_n.

Доказательство. Пусть $\sigma_1(A_n) \geq ... \geq \sigma_n(A_n)$ и $\sigma_1(B_n) \geq ... \geq \sigma(B_n)$ — сингулярные числа матриц A_n и B_n.

Сначала рассмотрим случай (21.8.9). Нетрудно проверить, что $\pm\sigma_i(A_n)$ и $\pm\sigma_i(B_n)$ будут собственными значениями эрмитовых матриц

$$\widetilde{A}_n = \begin{bmatrix} 0 & A_n \\ A_n^* & 0 \end{bmatrix}, \quad \widetilde{B}_n = \begin{bmatrix} 0 & B_n \\ B_n^* & 0 \end{bmatrix}.$$

Применяя теорему Виландта–Хоффмана к эрмитовым матрицам (в нашем случае к $\widetilde{A}_n$ и $\widetilde{B}_n$), можно использовать естественное упорядочение собственных значений (докажите!) $\Rightarrow$

$$\sum_{k=1}^{n} (\sigma_k(A_n) - \sigma_k(B_n))^2 \leq ||A_n - B_n||_F^2.$$

Возьмем произвольное $\delta > 0$ и обозначим через $\alpha_n(\delta)$ число номеров $k \in \{1, ..., n\}$, для которых $|\sigma_k(A_n) - \sigma_k(B_n)| \geq \delta$. Тогда, в силу (21.8.9),

$$\alpha_n(\delta)\delta^2 = o(n) \quad \Rightarrow \quad \alpha_n(\delta) = o(n).$$

[1] E. E. Tyrtyshnikov, A unifying approach to some old and new theorems on distribution and clustering, *Linear Algebra Appl.* 323 (1996), 1–43.

Пусть $M \subset \mathbb{C}$ — кластер сингулярных чисел A_n. Определим для него $\gamma_n^A(\varepsilon)$ и $\gamma_n^B(\varepsilon)$ — количества ε-удаленных элементов среди сингулярных чисел матриц A_n и B_n. Выберем $\delta = \varepsilon/2$, тогда

$$\gamma_n^B(\varepsilon) \le \gamma_n^A(\delta) + \alpha_n(\delta).$$

Правая часть есть $o(n) \quad \Rightarrow \quad \gamma_n^B(\varepsilon) = o(n)$.

Перейдем к случаю (21.8.10). Заметим, что $\sigma_i(A_n)$ и $\sigma_i(B_n)$ — это квадратные корни из собственных значений эрмитовых матриц $A_n^* A_n$ и $B_n^* B_n$. При этом $\operatorname{rank}(A_n^* A_n - B_n^* B_n) = o(n)$ (почему?). Без потери общности можно считать, что A_n и B_n эрмитовы. Кроме того, B_n получается из A_n путем прибавления и вычитания эрмитовых матриц ранга 1 в количестве $m = \operatorname{rank}(A_n - B_n)$. Применяя m раз соотношения разделения теоремы 5.7.2, получаем

$$\gamma_n^B(\varepsilon) \le \gamma_n^A(\varepsilon) + \operatorname{rank}(A_n - B_n). \qquad \square$$

5.10 Кластеры собственных значений

При изучении кластеров собственных значений свобода в определении "близких" последовательностей матриц A_n и B_n существенно уменьшается.

Теорема 5.10.1 *Предположим, что для любого n матрицы A_n и B_n порядка n диагонализуемы с помощью матриц собственных векторов P_n и Q_n, соответственно. Тогда если*

$$\operatorname{cond}_2^2 P_n \operatorname{cond}_2^2 Q_n \, ||A_n - B_n||_F^2 = o(n),$$

то всякий кластер собственных значений A_n является кластером собственных значений B_n.

Доказательство. Обратившись к "ненормальному" обобщению теоремы Виландта–Хоффмана, мы почти дословно можем повторить рассуждения теоремы 5.9.1. $\square$

Пусть $M = \{0\}$ — кластер состоит из одной точки 0. Если сингулярные числа кластеризуются в точке 0, то при весьма слабых предположениях то же верно и для собственных значений. [2]

Пусть A_n имеет сингулярные числа $\sigma_1(A_n) \ge \ldots \ge \sigma_n(A_n)$ и собственные значения $\lambda_i(A_n)$, упорядоченные по невозрастанию модуля: $|\lambda_1(A_n)| \ge$

[2] E. E. Tyrtyshnikov, N. L. Zamarashkin, On eigen and singular clusters, *Calcolo*, 33 (1997), 71–78.

$... \geq |\lambda_n(A_n)|$. Используя теорему Шура и соотношения разделения, можно получить следующие *неравенства Вейля*:

$$\prod_{i=1}^{l} |\lambda_i(A_n)| \leq \prod_{i=1}^{l} \sigma_i(A_n), \quad 1 \leq l \leq n.$$

Теперь предположим, что нуль является кластером сингулярных чисел A_n. Определим номера $k = k(\delta, n)$ и $m = m(\varepsilon, n)$ с помощью неравенств

$$\sigma_k(A_n) \geq \delta > \sigma_{k+1}(A_n) \quad \text{и} \quad |\lambda_m(A_n)| \geq \varepsilon > |\lambda_{m+1}(A_n)|.$$

Если кластер собственный (сильный), то $k(\delta) \leq c(\delta) \quad \forall\, n$. Фиксируем произвольное $\varepsilon > 0$. Допустим, что $m(\varepsilon, n) \to \infty$ при $n \to \infty$. Тогда для любого фиксированного $\delta > 0$ при всех лостаточно больших n имеем $m(\varepsilon, n) > k(\delta, n)$. Вследствие неравенств Вейля

$$\varepsilon^m \leq ||A_n||_2^k \, \delta^{m-k} \quad \Rightarrow \quad \left(\frac{\varepsilon}{\delta}\right)^m \leq \left(\frac{||A_n||_2}{\delta}\right)^k. \qquad (*)$$

Выбрав, например, $\delta = \varepsilon/2$, приходим к выводу о том, что $||A_n||_2 \to \infty$. Очевидная модификация нашего рассуждения доказывает, что неограниченность последовательности $m(\varepsilon, n)$ влечет за собой неограниченность нормы $||A_n||_2$. Таким образом, доказана

Теорема 5.10.2 *Если спектральные нормы A_n равномерно ограничены по n и сингулярные числа A_n имеют собственный кластер в нуле, то собственные значения A_n также имеют собственный кластер в нуле.*

Заметим также, что из $(*)$ следует, что

$$\left(\frac{\varepsilon}{\delta}\right)^{m/n} \leq \left(\frac{||A_n||_2}{\delta}\right)^{k/n}.$$

В случае общего кластера в нуле для $\sigma_i(A_n)$ имеем $k(\delta, n)/n \to 0$. Поэтому если $m(\varepsilon, n)/n \not\to 0$, то левую часть можно сделать сколь угодно большой за счет выбора достаточно малого $\delta > 0 \quad \Rightarrow \quad$ правая часть не может быть равномерно ограниченной при всех n и всех достаточно малых $\delta > 0$. Значит, справедлива

Теорема 5.10.3 *Предположим, что сингулярные числа A_n имеют кластер в нуле с числом δ-удаленных элементов $k(\delta, n)$ и для некоторого $c > 0$*

$$|\ln ||A_n||_2| \leq c \frac{n}{k(\delta, n)}$$

при всех n и достаточно малых $\delta > 0$. Тогда собственные значения A_n также имеют кластер в нуле.

Задачи

1. Известно сингулярное разложение матрицы $A = V\Sigma U^*$. Найдите базис из собственных векторов и собственные значения матрицы
$$\widetilde{A} = \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}.$$

2. Пусть эрмитовы $n \times n$-матрицы A и B имеют собственные значения $\lambda_1 \geq \cdots \geq \lambda_n$ и $\mu_1 \geq \cdots \geq \mu_n$. Докажите, что $\sum\limits_{i=1}^{n} |\lambda_i - \mu_i|^2 \leq \|A-B\|_F^2$.

3. Пусть $n \times n$-матрицы A и B имеют сингулярные числа $\lambda_1 \geq \cdots \geq \lambda_n$ и $\mu_1 \geq \cdots \geq \mu_n$. Докажите, что $\sum\limits_{i=1}^{n} |\lambda_i - \mu_i|^2 \leq \|A - B\|_F^2$.

4. Дана эрмитова матрица $A \in \mathbb{C}^{n\times n}$ и $B = P^*AP$, где матрица $P \in \mathbb{C}^{k\times n}$ имеет ортонормированные столбцы. Доказать, что
$$\lambda_{n-k+1}(A) + ... + \lambda_n(A) \leq \operatorname{tr} B \leq \lambda_1(A) + ... + \lambda_k(A).$$

5. Пусть A — произвольная матрица порядка n. Докажите неравенства Вейля
$$\prod_{i=1}^{l} |\lambda_i(A_n)| \leq \prod_{i=1}^{l} \sigma_i(A_n), \quad 1 \leq l \leq n.$$

6. Докажите, что для любой $n \times n$-матрицы
$$\sum_{i=1}^{l} |\lambda_i(A_n)|^2 \leq \sum_{i=1}^{l} \sigma_i(A_n)^2, \quad 1 \leq l \leq n.$$

7. Приведите пример последовательности матриц, для которых собственные значения имеют кластер в некоторой точке, но ни одна точка не является кластером для сингулярных чисел.

8. Приведите пример последовательности матриц, для которых сингулярные числа имеют кластер в нуле, а собственные значения не имеют.

9. Даны последовательности эрмитовых матриц порядка $n = 1, 2, ...$ вида
$$A_n = \begin{bmatrix} 2 & 1 & & & \\ 1 & 2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 2 & 1 \\ & & & 1 & 2 \end{bmatrix}, \quad B_n = \begin{bmatrix} 2 & c_n & & & c_n \\ c_n & 2 & c_n & & \\ & \ddots & \ddots & \ddots & \\ & & c_n & 2 & c_n \\ c_n & & & c_n & 2 \end{bmatrix}, \quad c_n = \frac{n-1}{n}.$$
Докажите, что собственные значения матриц $A_nB_n^{-1}$ имеют кластер в точке 1. Будет ли кластер собственным?

Глава 6

6.1 Машинные числа

Имеется лишь конечный набор чисел, представимых в компьютере. Это так называемые *машинные числа* вида

$$a = \pm \left(\frac{d_1}{p} + \frac{d_2}{p^2} + \ldots + \frac{d_t}{p^t} \right) \cdot p^{\alpha}.$$

Здесь $p, \alpha, d_1, \ldots, d_t$ — целые числа. Число $p > 0$ называется *основанием* арифметики. Число в скобках называется *мантиссой*, а число α — *порядком* машинного числа a. Числа $d_i \in \{0, 1, \ldots, p-1\}$ называются *разрядами*, а t — *длиной мантиссы*. Обычно считают, что $d_1 \neq 0$. Кроме того, имеются целые L и U, задающие границы для α: $L \leq \alpha \leq U$. Особым машинным числом является $a = 0$.

Таким образом, множество машинных чисел определяется параметрами p, t, L и U.

6.2 Аксиомы машинной арифметики

При вводе чисел в компьютер и при выполнении операций с машинными числами обычно происходит округление чисел. Округление — это некоторое отображение вещественных чисел в множество машинных чисел.

Если x — вещественное число и $fl(x)$ — результат отображения, то имеет место аксиома

$$fl(x) = x(1 + \varepsilon), \tag{6.2.1}$$

где в случае $fl(x) \neq 0 \quad |\varepsilon| \leq \eta$. Будем считать, что η есть точная верхняя грань для $|\varepsilon|$. При школьном правиле округления (докажите!)

$$\eta = \frac{1}{2} p^{1-t}. \tag{6.2.2}$$

Обычно результат операции $*$ с машинными числами a и b обозначается через $fl(a * b)$. При этом предполагается, что если $fl(a * b) \neq 0$, то

$$fl(a * b) = a * b(1 + \varepsilon), \qquad |\varepsilon| \leq \eta. \tag{6.2.3}$$

Это соотношение является основной аксиомой, позволяющей изучать влияние ошибок округления в алгоритмах.

Заметим, что относительная ошибка ε мала только в том случае, когда результат операции не является машинным нулем!

Иногда округление производится путем отбрасывания "лишних" разрядов. В этом случае равенство $x = a * b$, вообще говоря, не влечет за собой $fl(a*b) = fl(x)$. Например, пусть $p = 2$, $t = 2$. Пусть $a = 0.11$, $b = 0.0001$ и $x = a - b = 0.1011$. В данном случае $fl(x) = 0.10$, однако $fl(a-b) = 0.11$ (в силу того, что действие с числами в конечном случае сводится к операции с t-разрядными числами на "сумматоре").

При округлении отбрасыванием разрядов $\eta = p^{1-t}$.

6.3 Ошибки округления для скалярного произведения

Точные соотношения между реально вычисленными величинами должны содержать большое число различных $\varepsilon_1, \varepsilon_2, \ldots$. Чтобы не загромождать формулы, будем все эти $\varepsilon_1, \varepsilon_2, \ldots$ обозначать одной и той же буквой ε. Примем также обозначение

$$(1+\varepsilon)^n \equiv \prod_{i=1}^{n} (1+\varepsilon_{k_i}) .$$

Если таких степеней несколько, то ε_{k_i} в разложениях считаются разными. При получении неравенств это не вызывает проблем, так как любое ε удовлетворяет (6.2.3) (если не встретился машинный нуль).

Пусть $\widetilde{\alpha}$ есть полученное на машине скалярное произведение

$$\alpha = x^T y, \qquad x = [x_1, \ldots, x_n]^T, \qquad y = [y_1, \ldots, y_n]^T \in \mathbb{R}^n,$$

вычисленное по предписанию

$$\begin{aligned} \alpha = 0; \quad & \text{DO } i = 1,\ n \\ & \qquad \alpha = \alpha + x_i y_i \\ & \text{END DO} \end{aligned}$$

Тогда согласно (6.2.3) находим

$$\widetilde{\alpha} = \sum_{i=1}^{n} x_i y_i (1+\varepsilon)^{n+1-i}. \tag{6.3.4}$$

6.4 Прямой и обратный анализ

Формулу (6.3.4) можно интерпретировать по-разному. Действуя в духе *прямого анализа*, мы можем оценить близость точной и реально вычисленной величины:

$$|\widetilde{\alpha} - \alpha| \le n\eta |x^T||y| + \mathcal{O}(\eta^2). \tag{6.4.5}$$

Здесь и отныне мы полагаем, что если $A = [a_{ij}]$, то $|A| = [\ |a_{ij}|\ ]$.

Обратный анализ предлагает представить реально вычисленную величину как результат точного вычисления с возмущенными данными и дать оценку этого (называемого *эквивалентным*) возмущения:

$$\begin{aligned} &\widetilde{\alpha} = \widetilde{x}^T \widetilde{y}, \\ &|\tilde{x} - x| \le \tfrac{1}{2} n\eta |x| + \mathcal{O}(\eta^2), \quad |\tilde{y} - y| \le \tfrac{1}{2} n\eta |y| + \mathcal{O}(\eta^2). \end{aligned} \tag{6.4.6}$$

Очевидно, в данном примере возможны варианты при распределении возмущений между x и y.

6.5 Немного философии

При анализе ошибок округления типичным является невнимание к членам порядка $\mathcal{O}(\eta^2)$. Как подчеркивал классик этой области Дж. Х. Уилкинсон, главное в такого рода анализе — это не точные неравенства, а обнаружение (и устранение) возможных "узких мест" в алгоритме. Неправильно думать, что большая погрешность в решении задачи есть следствие большого числа операций (и значит, большого числа ошибок округления). Чаще всего все "портит" какая-то одна операция.

6.6 Пример "плохой" операции

Плохую репутацию имеет операция вычитания близких чисел одного знака. Эта операция сама по себе не хуже остальных, то есть ее собственная ошибка округления есть величина порядка η. Но эта операция может *катастрофически усилить* ранее имевшиеся погрешности. Именно, пусть $\widetilde{a} \approx a$, $\widetilde{b} \approx b$. Тогда

$$fl(\widetilde{a} - \widetilde{b}) = (\widetilde{a} - \widetilde{b})(1 + \varepsilon) = (a - b)(1 + \varepsilon + \delta),$$

где

$$\delta = \left(\frac{(\widetilde{a} - a) - (\widetilde{b} - b)}{a - b} \right) (1 + \varepsilon).$$

Очевидно, δ может оказаться очень большим.

6.7 Еще один пример

При вычислении собственных векторов блочно треугольной матрицы с блоками 2×2 по некоторой программе обнаружилось, что в точности один собственный вектор имеет невязку, превышающую на 3 порядка невязки для других векторов.

В данном случае "узким местом" оказалось решение однородной системы с двумя уравнениями и неизвестными:

$$a_1x_1 + a_2x_2 = 0, \quad b_1x_1 + b_2x_2 = 0, \qquad ||x||_\infty = 1.$$

Пусть векторы $a = [a_1,\ a_2]^T$ и $b = [b_1,\ b_2]^T$ ненулевые и приближенно коллинеарны; положим $|a_1b_2 - a_2b_1| = \delta$. Можно взять

$$x_1 = -a_2/\|a\|_\infty, \qquad x_2 = a_1/\|a\|_\infty. \tag{$*$}$$

Тогда получаем невязку

$$\begin{aligned} r_1 &\equiv |a_1\tilde{x}_1 + a_2\tilde{x}_2| \le 2\eta\|a\|_\infty, \\ r_2 &\equiv |b_1\tilde{x}_1 + b_2\tilde{x}_2| \le \frac{\delta}{\|a\|_\infty} + 2\eta\|b\|_\infty. \end{aligned} \tag{6.7.7}$$

Можно поступить иначе:

$$x_1 = -b_2/\|b\|_\infty, \qquad x_2 = b_1/\|b\|_\infty. \tag{$**$}$$

Тогда

$$\begin{aligned} r_1 &\le \frac{\delta}{\|b\|_\infty} + 2\eta\|a\|_\infty, \\ r_2 &\le 2\eta\|b\|_\infty. \end{aligned} \tag{6.7.8}$$

Оценки (6.7.7) и (6.7.8) отличаются вхождением члена $\delta/\|a\|_\infty$ либо $\delta/\|b\|_\infty$. Если величина $\|b\|_\infty$ превышает $\|a\|_\infty$ на 3 порядка, то невязка в первом способе, вообще говоря, может быть больше на 3 порядка.

Очевидным образом возникает рекомендация: если $\|a\|_\infty \ge \|b\|_\infty$, то использовать $(*)$; в противном случае использовать $(**)$.

6.8 Идеальные и машинные тесты

Обычно алгоритмы тестируют: сравнивают вычисленный ответ с заранее известным точным ответом. Незадачливые вычислители иногда не замечают разницы между идеальными и машинными тестами и из-за этого делают неверные выводы.

Например, программа, решающая линейные системы по методу Гаусса, опробывалась на тесте $Ax = b$ с матрицей Гильберта

$$A = \left[\frac{1}{i+j-1}\right]_{ij=1}^{n}$$

и решением $x = [1, \ldots, 1]^\top$. Правая часть для теста вычислялась умножением матрицы A на вектор x. Оказалось, что при $n = 10$ для вычисленного решения $\tilde{x}$ (в режиме REAL*8 на Фортране)

$$||\tilde{x} - x||_\infty \approx 0.9 \cdot 10^2 \ .$$

Но не спешите говорить, что алгоритм выдал "плохое" решение.

В действительности вместо точных A и b машина получила близкие, но другие $\hat{A}$ и $\hat{b}$, то есть машинный тест отличается от идеального теста. Матрица Гильберта плохо обусловлена; поэтому точное решение $\hat{x}$ для машинного теста будет сильно отличаться от x:

$$\hat{x} \approx \begin{bmatrix} 1.042595644 \\ 0.459944616 \\ 1.284169655 \\ 2.772926997 \\ 2.217756963 \\ 3.252956378 \\ -\ 5.410267887 \\ -46.120499977 \\ 93.504857996 \\ -43.063313904 \end{bmatrix} .$$

А это значит, что вычисленное решение $\tilde{x}$ нужно сравнивать с $\hat{x}$, а не с x! При этом относительная ошибка в компонентах решения не превышает 10^{-8}.

6.9 Вверх или вниз

Рассмотрим пример неустойчивого алгоритма, связанного с вычислением интегралов [1]

$$a_n = \int_0^1 x^n \, \mathrm{e}^{x-1} \, dx, \quad n = 0, 1, \ldots .$$

Интегрируя по частям, получаем простое рекуррентное соотношение. Реализуя его "снизу вверх", получаем следующий алгоритм:

$$a_0 \ = \ 1/\mathrm{e}, \qquad a_n \ = \ 1 - n\, a_{n-1}, \quad n = 1, 2, \ldots . \qquad (*)$$

[1] Пример из книги: Дж.Форсайт, М.Малькольм, К.Моулер, *Машинные методы математических вычислений*, Мир, 1980.

Пропустите этот алгоритм на своем компьютере. Несмотря на то, что должно быть (проверьте!)

$$0 \;<\; a_n \;<\; \frac{1}{n+1},$$

Вы будете получать очень большие и даже отрицательные a_n. Дело в том, что даже небольшая погрешность в a_0 при выполнении алгоритма $(*)$ умножится в a_n на $n!$. Отсюда вытекает, что если

$$a_0 \;=\; c, \qquad a_n \;=\; 1 - n\,a_{n-1}, \quad n = 1, 2, \ldots,$$

то (докажите!)

$$\lim_{n\to\infty} a_n \;=\; \begin{cases} 0 \;, & c = 1/\mathrm{e}, \\ \infty, & \text{иначе.} \end{cases}$$

Что же делать? Хорошая идея — взять достаточно большое N и выполнять то же рекуррентное соотношение "сверху вниз":

$$a_N = 1, \qquad a_n = (1 - a_{n+1})/n, \quad n = N-1, N-2, \ldots 0. \qquad (**)$$

Теперь погрешность в a_N будет быстро уменьшаться! В результате алгоритм $(**)$ можно использовать, например, для вычисления величины $\mathrm{e} = 2.718281828\ldots$ с машинной точностью.

6.10 Решение треугольных систем

Для решения системы $Lx = b$ с нижней треугольной матрицей $L = [\, l_{ij} \,]$ обычно используется очевидный *метод прямой подстановки*:

$$\begin{array}{l} \text{DO } i = 1,\ n \\ \qquad x_i = \left(b_i - \displaystyle\sum_{j=1}^{i-1} l_{ij} x_j \right) / l_{ii} \\ \text{END DO} \end{array}$$

Если $\tilde{x}_i$ есть реально вычисленная величина, то получаем

$$\tilde{x}_i = \left(b_i - \sum_{j=1}^{i-1} l_{ij} \tilde{x}_j (1+\varepsilon)^{i-j} \right) (1+\varepsilon)^2 / l_{ii}.$$

Положим

$$\tilde{l}_{ij} = \begin{cases} l_{ij}\,(1+\varepsilon)^{i-j}, & i > j, \\ l_{ii}\,/\,(1+\varepsilon)^2, & i = j, \\ 0, & i < j. \end{cases}$$

Тогда

$$\tilde{x}_i = \left(b_i - \sum_{j=1}^{i-1} \tilde{l}_{ij}\tilde{x}_j \right) / \tilde{l}_{ii},$$

то есть $\tilde{x}_i$ есть компонента точного решения для системы с той же правой частью, но с возмущенной матрицей $\tilde{L} = \left[\tilde{l}_{ij} \right]$. Близость $\tilde{L}$ и L легко оценивается:

$$\left| \tilde{l}_{ij} - l_{ij} \right| \le \begin{cases} |l_{ij}|\,(i-j)\eta + \mathcal{O}(\eta^2), & i > j, \\ |l_{ii}|\,2\eta + \mathcal{O}(\eta^2), & i = j, \\ 0, & i < j. \end{cases} \qquad (6.10.9)$$

Таким образом, справедлива следующая

Теорема 6.10.1 *Для алгоритма прямой подстановки реально вычисленное решение $\tilde{x}$ системы $Lx = b$ удовлетворяет возмущенной системе $\tilde{L}\tilde{x} = b$, где $\tilde{L}$ — нижняя треугольная матрица такая, что*

$$\left| \tilde{L} - L \right| \ \le \ n\eta|L| + \mathcal{O}(\eta^2). \qquad (6.10.10)$$

Неравенство 6.10.10 представляет собой компактную, но огрубленную запись неравенств 6.10.9. Полученный результат — это практически идеал того, что следует ожидать от алгоритма с точки зрения обратного анализа ошибок округления.

Аналогичный результат имеет место и для *метода обратной подстановки* — для решения систем с верхней треугольной матрицей.

Задачи

1. Верно ли, что всегда $fl(\frac{a+b}{2}) \in [a, b]$?

2. Чтобы найти e^x при $x = -13$, некто суммирует ряд

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

   и получает чудовищно большую погрешность. Почему? Что делать?

3. Известно, что обратная подстановка всегда дает малую невязку. Почему? Напишите программу для обратной подстановки и решите систему

$$\begin{bmatrix} 1 & 2 & & & \\ & 1 & 2 & & \\ & & \ddots & \ddots & \\ & & & 1 & 2 \\ & & & & 1 \end{bmatrix}_{n\times n} \begin{bmatrix} x_1 \\ \\ \vdots \\ \\ x_n \end{bmatrix} = \begin{bmatrix} 1 \\ \\ \vdots \\ 1 \\ 1/3 \end{bmatrix}$$

   при $n = 50$. Мала ли невязка? Какую точность имеет решение?

4. Придумайте алгоритм суммирования n чисел, для которого

$$fl(a_1 + \cdots + a_n) = \tilde{a}_1 + \cdots + \tilde{a}_n,$$

где

$$|\tilde{a}_i - a_i| \ \leq \ \eta \ \log_2 \ n|a_i| \ + \ \mathcal{O}(\eta^2).$$

Глава 7

7.1 Прямые методы для линейных систем

Мы строим алгоритмы, исходя из какого-то набора элементарных операций. Если для решения задачи требуется конечное число элементарных операций (в точной арифметике), то соответствующий метод называют *прямым*.

Не для каждой задачи можно придумать прямой метод. Например, с помощью конечного числа арифметических операций нельзя решить уравнение $x^2 = 2$.

Если извлечение корня считать элементарной операцией, то прямой метод уже существует. Но мы знаем (благодаря Галуа, Абелю и Руффини), что и в этом случае невозможно построить прямой метод нахождения корней произвольного полинома степени 5 и выше. Поэтому не стоит пытаться придумывать прямой метод вычисления собственных значений для произвольной матрицы (почему?).

Для линейных систем прямые методы существуют. В качестве элементарных операций рассматриваются арифметические операции; иногда к ним добавляется операция извлечения квадратного корня.

Классические прямые методы для *плотных* матриц (то есть матриц без какой-либо специфики) — это, прежде всего, метод Гаусса и методы, получающие нули с помощью преобразований вращения или отражения. Эти методы связаны с получением LU-разложения и QR-разложения для матрицы коэффициентов.

7.2 Теория LU-разложения

Матрица A называется *строго регулярной*, если все ее ведущие подматрицы (в том числе и A) невырожденные.

Под *LU-разложением* матрицы A понимается равенство $A = LU$, где матрица L — нижняя унитреугольная (треугольная с единичной главной диагональю), а U — невырожденная верхняя треугольная матрица.

Теорема 7.2.1 *Для того чтобы матрица A имела LU-разложение, необходимо и достаточно, чтобы она была строго регулярной.*

Доказательство. Необходимость очевидна (почему?). Достаточность докажем по индукции. Запишем

$$A = \begin{bmatrix} a & c^\top \\ b & D \end{bmatrix}. \tag{7.2.1}$$

Тогда

$$\begin{bmatrix} 1 & 0 \\ -z & I \end{bmatrix} \begin{bmatrix} a & c^\top \\ b & D \end{bmatrix} = \begin{bmatrix} a & c^\top \\ 0 & A_1 \end{bmatrix}, \qquad z = \frac{1}{a}b, \quad A_1 \equiv D - \frac{1}{a}bc^\top.$$

Легко видеть, что матрица A_1 также будет строго регулярной. По индуктивному предположению она имеет LU-разложение: $A_1 = L_1U_1$. Положим

$$L = \begin{bmatrix} 1 & 0 \\ z & L_1 \end{bmatrix}, \qquad U = \begin{bmatrix} a & c^\top \\ 0 & U_1 \end{bmatrix} \tag{7.2.2}$$

и вычислим

$$LU = \begin{bmatrix} a & c^\top \\ b & L_1U_1 + \frac{1}{a}bc^\top \end{bmatrix} = \begin{bmatrix} a & c^\top \\ b & D \end{bmatrix} = A. \quad \square$$

Следствие 7.2.1 *LU-разложение определяется однозначно.*

Равенство $L_1U_1 = L_2U_2$ влечет $L_2^{-1}L_1 = U_2U_1^{-1} \equiv D$. Произведение нижних треугольных матриц и обратная к такой матрице остаются нижними треугольными матрицами. То же верно для верхних треугольных матриц. Поэтому D одновременно является нижней и верхней треугольной матрицей $\Rightarrow$ она является диагональной. Поскольку матрица $L_2^{-1}L_1$ унитреугольная, получаем $D = I$. $\square$

Следствие 7.2.2 *Все ведущие миноры матрицы A положительны тогда и только тогда, когда U имеет положительные диагональные элементы.*

Следствие 7.2.3 *Если строго регулярная матрица $A \in \mathbb{C}^{n\times n}$ симметрична ($A = A^\top$), то для нее существует $LDL^\top$-разложение:*

$$A = LDL^\top,$$

где L — нижняя унитреугольная, D — невырожденная диагональная.

Положим $D = \operatorname{diag}(U)$. Тогда $A = LDD^{-1}U = A^\top = D^{-1}U^\top(DL^\top)$. Очевидно, матрица $D^{-1}U^\top$ — нижняя унитреугольная. В силу единственности LU-разложения $DL^\top = U$. $\square$

Следствие 7.2.4 *Если строго регулярная матрица $A \in \mathbb{C}^{n\times n}$ эрмитова ($A = A^*$), то для нее существует LDL^*-разложение, т.е. $A = LDL^*$, где L — нижняя унитреугольная, D — невырожденная диагональная.*

Докажите!

Пусть C — нижняя треугольная матрица с положительной главной диагональю. Разложение $A = CC^*$ называется *разложением Холецкого.*

Теорема 7.2.2 *Для того чтобы матрица $A \in \mathbb{C}^{n\times n}$ имела разложение Холецкого, необходимо и достаточно, чтобы она была эрмитовой с положительными ведущими минорами.*

Доказательство. Необходимость очевидна. Достаточность: возьмем разложение $A = LDL^T$ (или $A = LDL^*$) и положим $C = L\mathrm{diag}\left(\sqrt{d_1}, \ldots, \sqrt{d_n}\right)$, где $D = \mathrm{diag}\,(d_1, \ldots, d_n)$. □ □

7.3 Ошибки округления для LU-разложения

Доказательство теоремы об LU-разложении было конструктивным. По сути, в нем содержится рекурсивный алгоритм, совпадающий с тем, что обычно называется *алгоритмом Гаусса.*

Теорема 7.3.1 *Пусть A — представимая в машине вещественная строго регулярная матрица порядка n. Тогда для реально вычисленных по алгоритму Гаусса матриц $\widetilde{L}$, $\widetilde{U}$ выполняется неравенство (при условии, что в элементарных операциях не возникал машинный нуль)*

$$\left|\widetilde{L}\widetilde{U} - A\right| \le 3n\eta\left(|A| + |\widetilde{L}||\widetilde{U}|\right) + \mathcal{O}(\eta^2). \tag{7.3.3}$$

Доказательство. Пусть матрицы в точном равенстве $A = LU$ имеют вид (7.2.1), (7.2.2). Для реально вычисленных матриц находим

$$\begin{bmatrix} 1 & 0 \\ \widetilde{z} & \widetilde{L}_1 \end{bmatrix}\begin{bmatrix} a & c^\top \\ 0 & \widetilde{U}_1 \end{bmatrix} - \begin{bmatrix} a & c^\top \\ b & D \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ (\widetilde{z} - z)a & \widetilde{L}_1\widetilde{U}_1 - H \end{bmatrix}, \quad H = D - \widetilde{z}c^\top.$$

Положим $\widetilde{H} = fl(D - \widetilde{z}c^\top)$. Матрицы $\widetilde{L}_1$ и $\widetilde{U}_1$ образуют реально вычисленное LU-разложение матрицы $\widetilde{H}$. Предположим по индукции, что уже установлено, что

$$|\widetilde{L}_1\widetilde{U}_1 - \widetilde{H}| \le 3(n-1)\eta(|\widetilde{H}| + |\widetilde{L}_1||\widetilde{U}_1|) + \mathcal{O}(\eta^2).$$

В силу аксиомы машинной арифметики (6.2.3)

$$|\widetilde{H} - H| \le (|D| + 2|\widetilde{z}||c^\top|)\eta + \mathcal{O}(\eta^2).$$

Отсюда $|\widetilde{H}| \le |H| + O(\eta) \le |D| + |\widetilde{z}||c^\top| + O(\eta)$ и, окончательно,

$$\begin{aligned} |\widetilde{L}_1\widetilde{U}_1 - H| &\le (3n-2)\eta|D| + (3n-1)\eta|\widetilde{z}||c\top| + (3n-3_\eta|\widetilde{L}_1||\widetilde{U}_1| + \mathcal{O}(\eta^2) \\ &\le 3n(|D| + |\widetilde{z}||c^\top| + |\widetilde{L}_1||\widetilde{U}_1|) + \mathcal{O}(\eta^2). \end{aligned}$$

Кроме того, согласно (6.2.3), $|(\widetilde{z} - z)a| \le \eta|b|$. $\square$

7.4 Выбор ведущего элемента

Без члена $|\widetilde{L}||\widetilde{U}|$ оценка 7.3.3 была бы идеальной. Но по всей видимости член $|\widetilde{L}||\widetilde{U}|$ присутствует в ней по существу и указывает на "узкое место" алгоритма Гаусса — рост элементов в треугольных сомножителях.

Сколь опасен рост элементов, можно увидеть уже в случае $n = 2$. Пусть p — основание машинной арифметики и t — длина мантиссы. Возьмем

$$A \equiv \begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} p^{-t} & 1. \\ 1. & 1. \end{bmatrix}.$$

Тогда

$$\widetilde{L} = \begin{bmatrix} 1. & 0 \\ p^t & 1. \end{bmatrix}, \qquad \widetilde{U} = \begin{bmatrix} p^{-t} & 1. \\ 0 & -p^t \end{bmatrix}$$

и, следовательно,

$$\widetilde{L}\widetilde{U} - A = \begin{bmatrix} 0 & 0 \\ 0 & 1. \end{bmatrix}.$$

Причина чудовищно большой погрешности — малая величина ведущего элемента a.

Очевидно, необходим *выбор ведущего элемента*. Возможность выбора имеется. Например, с помощью перестановки строк ведущим можно сделать любой элемент очередного столбца. Конечно, это должен быть элемент, максимальный по модулю. Тогда выбор в столбце обеспечивает ограниченность элементов в L ($|l_{ij}| \le 1$). В то же время

$$\rho \equiv \frac{\max\limits_{i,j} |u_{ij}|}{\max\limits_{i,j} |a_{ij}|} \le 2^{n-1}. \tag{7.4.4}$$

Если ведущим делается максимальный по модулю элемент с минимальным строчным инлексом, то эта оценка для коэффициента роста ρ достигается

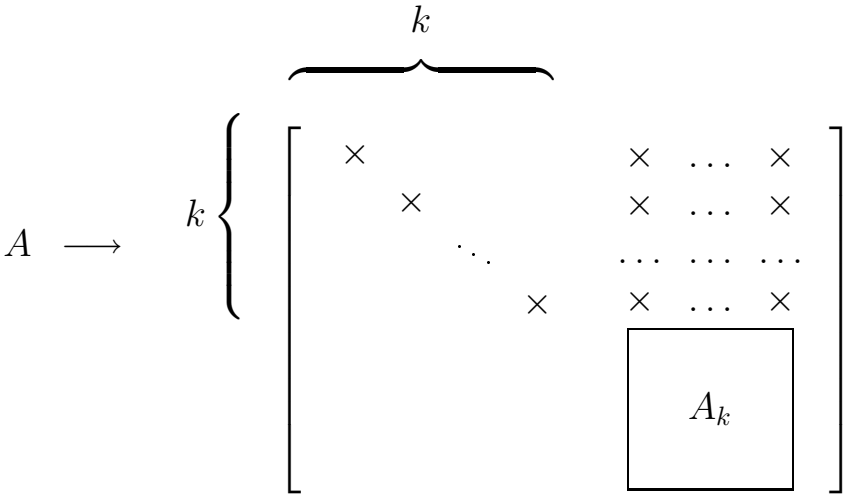
на матрице

$$A = \begin{bmatrix} 1 & & & & 1 \\ -1 & 1 & & 0 & 1 \\ -1 & -1 & \ddots & & 1 \\ \vdots & \vdots & & \ddots & \vdots \\ -1 & -1 & \dots & -1 & 1 \end{bmatrix}. \qquad (7.4.5)$$

В практических вычислениях такой рост при выборе в столбце наблюдается не часто (Кахан утверждает, что "с ростом элементов имеют дело только те вычислители, которые специально его ищут").

7.5 Полный выбор

После выполнения k шагов метода Гаусса, ведущим элементом можно сделать любой элемент *активной подматрицы* A_k:



Для этого потребуются перестановки строк и столбцов. Пусть в активной подматрице выбирается элемент, максимальный по модулю — это так называемый *полный выбор* (из полного набора возможностей) . Долгое время существовала гипотеза Уилкинсона о том, что для полного выбора $\rho \leq n$. В 1991 году эта гипотеза была опровергнута: Н.Гоулд [1] привел пример вещественной матрицы 13-го порядка, для которой $\rho > 13$.

Наиболее радикальное средство борьбы с ростом элементов — получать нули с помощью ортогональных преобразований.

7.6 Метод Холецкого

Пусть A — вещественная симметричная матрица с положительными ведущими минорами. Возьмем $n = 3$ и посмотрим, как можно удовлетворить

[1] N.Gould, On growth in Gaussian elimination with complete pivoting, SIAM J. Matrix Anal. Appl. 12 (2): 354–361 (1991).

уравнение

$$\begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} c_{11} & & 0 \\ c_{21} & c_{22} & \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} c_{11} & c_{21} & c_{31} \\ & c_{22} & c_{32} \\ 0 & & c_{33} \end{bmatrix} \quad \Rightarrow$$

$$\begin{aligned} &c_{11} = \sqrt{a_{11}}, && c_{21} = a_{21}/c_{11}, && c_{31} = a_{31}/c_{11}; \\ &c_{22} = \sqrt{a_{22} - c_{21}^2}, && c_{32} = (a_{32} - c_{31}c_{21})/c_{22}; \\ &c_{33} = \sqrt{a_{33} - c_{31}^2 - c_{32}^2}. \end{aligned}$$

Мы получили *алгоритм Холецкого*. Вот как он выглядит для произвольного n:

DO $k = 1,\ n$

$$c_{kk} = \left(a_{kk} - \sum_{j=1}^{k-1} c_{kj}^2 \right)^{1/2} \qquad (*)$$

DO $i = k + 1,\ n$

$$c_{ik} = \left(a_{ik} - \sum_{j=1}^{k-1} c_{ij} c_{kj} \right) / c_{kk} \qquad (**)$$

END DO

END DO

Предположим, что выражения $(*)$ и $(**)$ вычисляются таким образом, что для относительной ошибки ε любой арифметической операции и операции извлечения квадратного корня обеспечивается оценка $|\varepsilon| \le \eta$. Тогда для реально вычисленных величин $\tilde{c}_{ij}$ находим:

$$\begin{aligned} \tilde{c}_{kk}^2 \,/\, (1+\varepsilon)^3 &= a_{kk} - \sum_{j=1}^{k-1} \tilde{c}_{kj}^2 (1+\varepsilon)^{k-j} \qquad \text{для } (*); \\ \tilde{c}_{ik}\, \tilde{c}_{kk} \,/\, (1+\varepsilon)^2 &= a_{ik} - \sum_{j=1}^{k-1} \tilde{c}_{ij}\, \tilde{c}_{kj}\, (1+\varepsilon)^{k-j} \quad \text{для } (**). \end{aligned}$$

Первое соотношение дает

$$\sqrt{\sum_{j=1}^{k} |\tilde{c}_{kj}|^2} \;\le\; \sqrt{a_{kk} \;+\; \mathcal{O}(\eta)} \;\le\; \sqrt{a_{kk}} \;+\; \mathcal{O}(\eta).$$

Следовательно,

$$\begin{aligned}
\left|\sum_{j=1}^{k} \tilde{c}_{ij}\tilde{c}_{kj} - a_{ik}\right| &\leq \eta\,(k+1)\ \sum_{j=1}^{k} |\tilde{c}_{ij}|\,|\tilde{c}_{kj}| \ +\ \mathcal{O}(\eta^2) \\
&\leq \eta\,(k+1)\ \sqrt{\sum_{j=1}^{k} |\tilde{c}_{ij}|^2}\sqrt{\sum_{j=1}^{k} |\tilde{c}_{kj}|^2} \ +\ \mathcal{O}(\eta^2) \\
&\leq \eta\,(n+1)\ \sqrt{\sum_{j=1}^{k} |\tilde{c}_{ij}|^2}\sqrt{\sum_{j=1}^{i} |\tilde{c}_{kj}|^2} \ +\ \mathcal{O}(\eta^2) \\
&\leq \eta\,(n+1)\ \sqrt{a_{ii}}\ \sqrt{a_{kk}} \ +\ \mathcal{O}(\eta^2).
\end{aligned}$$

Таким образом,

$$|\tilde{C}\,\tilde{C}^T - A| \leq \eta\,(n+1)\ \begin{bmatrix} \sqrt{a_{11}} \\ \dots \\ \sqrt{a_{nn}} \end{bmatrix}\ [\sqrt{a_{11}}, \dots, \sqrt{a_{nn}}] \ +\ \mathcal{O}(\eta^2).$$

Отсюда видно, что метод Холецкого свободен от неприятностей, связанных с ростом элементов.

7.7 Треугольные разложения и решение систем

Процесс решения системы $Ax = b$ с невырожденной матрицей A может состоять из трех этапов:

1) $A = LU$ (вычисление LU-разложения);

2) $Ly = b$ (прямая подстановка);

3) $Ux = y$ (обратная подстановка).

Наиболее трудоемким является первый этап: при использовании метода Гаусса требуется $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ арифметических операций. В общем случае рекомендуется применять столбцовое пивотирование. Второй и третий этапы требуют $n^2 + \mathcal{O}(n)$ арифметических операций каждый.

Заметим, что на параллельном компьютере временно́е соотношение между этапами может измениться. Если вообразить абстрактный параллельный компьютер с произвольно большим числом процессоров и мгновенным обменом информацией между всеми участниками вычислений, то для каждого из трех этапов существуют алгоритмы с временем $\mathcal{O}(\log_2^2 n)$. Никто не знает, можно ли снизить это время хотя бы для треугольных систем.

7.8 Как уточнить решение

Предположим, что по методу Гаусса вычислено приближенное решение $\tilde{x}_0$. Рассмотрим следующий процесс его уточнения:

(1) вычисляем невязку $r_{i-1} = b - A\tilde{x}_{i-1}$;

(2) решаем систему относительно поправки c_{i-1}: $Ac_{i-1} = r_{i-1}$;

(3) $x_i = \tilde{x}_{i-1} + c_{i-1}$.

Если эти действия выполнить точно, то вектор x_i будет точным решением системы $Ax = b$. В условиях ошибок округления мы получим новое приближение $\tilde{x}_i$ к точному решению x. Будет ли новое приближение лучше прежнего?

Предположим, что реально вычисленная и точная поправки ($\tilde{c}_i$ и c_i) связаны неравенством

$$\|\tilde{c}_i - c_i\|_2 \le \tau\|c_i\|_2, \tag{7.8.6}$$

где $0 < \tau < 1/(1+\eta)$. Тогда для векторов $c_i = A^{-1}(b - A\tilde{x}_i) = x - \tilde{x}_i$ имеем

$$\|c_i\|_2 \le q\|c_{i-1}\|_2 + \eta\|x\|_2, \qquad q = \tau(1+\eta). \tag{7.8.7}$$

Отсюда легко вывести, что

$$\|c_i\|_2 \le q^i\|c_0\|_2 + \left(q^{i-1} + q^{i-2} + \ldots, +1\right)\eta\|x\|_2;$$

Следовательно,

$$\frac{\|c_i\|_2}{\|x\|_2} \le \frac{\eta}{1-q} + O(q^i). \tag{7.8.8}$$

Итак, если исходное приближение не слишком плохое (выполнено неравенство (7.8.6)), то процесс уточнения позволяет получить приближенное решение системы с погрешностью не выше $\eta/(1-q)$.

Вообще говоря, уже исходное приближение может иметь малую невязку. Поэтому рекомендуется:

(а) невязку r_{i-1} вычислять с повышенной точностью и лишь затем округлять до обычной точности;

(б) в системе уравнений для поправки сделать норму правой части величиной порядка 1 путем *масштабирования*:

$$b_i = r_{i-1}/s,$$

где s имеет смысл выбирать в виде некоторой степени основания машинной арифметики — чтобы избежать ошибок округления; решив систему

$$A\hat{c}_{i-1} = b_i,$$

присвоить

$$c_{i-1} = \hat{c}_{i-1}s.$$

Процесс уточнения рассматривается как относительно дешевый, поскольку при решении системы относительно поправки следует воспользоваться уже известным LU-разложением матрицы A.

Вообще говоря, процесс уточнения позволяет добиваться очень высокой точности даже для плохо обусловленных матриц. Это имеет смысл, если матрица и правая часть в компьютере заданы *точно*. В противном случае уточнение вряд ли можно рассматривать как средство "борьбы" с плохой обусловленностью задачи — по крайней мере, его применение требует специального обоснования.

Задачи

1. Покажите, что метод Гаусса с выбором ведущего элемента по столбцу эквивалентен (в точной и машинной арифметике) методу Гаусса без выбора ведущего элемента, примененного к той же матрице с переставленными строками.

2. Докажите, что в правой части неравенства (7.3.3) коэффициент n можно заменить на $n-1$.

3. Пусть система $Ax = b$ решается в три этапа согласно разделу 7.7. Предположим, что при реализации не возникало машинных нулей и были реально получены матрицы $\tilde{L}$, $\tilde{U}$ и решение $\tilde{x}$. Докажите, что $(A+E)\tilde{x} = b$, где

$$|E| \le n\eta \left(4|A| + 5|\tilde{L}||\tilde{U}|\right) + \mathcal{O}(\eta^2).$$

4. Пусть LU-разложение вычисляется по методу Гаусса без выбора ведущего элемента для матрицы $A \in \mathbb{R}^{n\times n}$, обладающей строчным диагональным преобладанием. Докажите, что в этом случае коэффициент роста элементов

$$\rho \equiv \frac{\max\limits_{i,j}|u_{ij}|}{\max\limits_{i,j}|a_{ij}|}$$

не превосходит 2.

5. Если LU-разложение вычисляется по методу Гаусса без выбора ведущего элемента для вещественной симметричной положительно определенной матрицы, то коэффициент роста элементов равен 1. Докажите.

6. Если A – эрмитова положительно определенная матрица, то спектральное число обусловленности активной подматрицы каждого шага метода Гаусса не превосходит спектрального числа обусловленности матрицы A. Докажите.

7. Пусть $\mathcal{M}$ — множество $n \times n$-матриц вида $A = \alpha I - N$, где матрица N имеет неотрицательные элементы и все ее собственные значения по модулю меньше α. Докажите, что если $A \in \mathcal{M}$, то для нее существует LU-разложение и при этом $L \in \mathcal{M}$ и $U \in \mathcal{M}$.

8. Придумайте параллельный алгоритм, решающий треугольную систему за время $\mathcal{O}(\log_2^2 n)$ (подсказка: эта задача не является очень трудной!)

9. Покажите, что разложение Холецкого можно найти за $O(n)$ параллельных шагов.

Глава 8

8.1 QR-разложение квадратной матрицы

Разложение $A = QR$, где Q — унитарная, R — верхняя треугольная матрица, называется *QR-разложением* (квадратной) матрицы A. Длины соответствующих столбцов R и A одинаковы $\Rightarrow$ $\max\limits_{i,j} |r_{ij}| \le \sqrt{n} \max\limits_{i,j} |a_{ij}|$. Поэтому роста элементов при получении QR-разложения опасаться не следует.

Теорема 8.1.1 *QR-разложение существует для любой квадратной матрицы.*

Доказательство. Пусть A невырожденная $\Rightarrow$ A^*A положительно определенная $\Rightarrow$ все ведущие подматрицы в A^*A положительно определенные $\Rightarrow$ все ведущие миноры в A^*A положительны $\Rightarrow$ существует разложение Холецкого $A^*A = R^*R$ (R — верхняя треугольная матрица) $\Rightarrow$ матрица $Q \equiv A\,R^{-1}$ унитарная:

$$Q^*Q = (A\,R^{-1})^*(A\,R^{-1}) = R^{-*}(A^*A)R^{-1} = (R^{-*}R^*)\,(R\,R^{-1}) = I.$$

Если матрица A вырожденная, то для всех достаточно больших n возмущенная матрица $A_n = A + \frac{1}{n}\,I$ будет невырожденной (почему?). Поэтому существует QR-разложение $A_n = Q_n R_n$. Множество унитарных матриц является компактным (почему?) $\Rightarrow$ существует сходящаяся подпоследовательность

$$Q_{n_k} \;\to\; Q \quad \Rightarrow \quad Q_{n_k}^* A \;\to\; Q^*A \equiv R.$$

Легко видеть, что матрица Q унитарная, а R — верхняя треугольная. $\square$

Следствие. *Для невырожденной A матрицы Q и R определяются однозначно, если требовать положительность главной диагонали для R.*

8.2 QR-разложение прямоугольной матрицы

Пусть $A \in \mathbb{C}^{m\times n}$ и $m \ge n$. Тогда существует разложение $A = QR$, где R — квадратная верхняя треугольная матрица порядка n, а матрица Q имеет ортонормированные столбцы.

Для доказательства достаточно вложить A в квадратную матрицу, заполнив недостающие позиции нулями.

8.3 Матрицы отражения

Матрица $H = H(u) = I - 2\,u\,u^*$, где $||u||_2 = 1$, называется *матрицей отражения*, или матрицей Хаусхолдера. Проверьте, что:

(a) H унитарная;

(b) H эрмитова;

(c) $Hu = -u$ и $Hv = v \quad \forall\ v \perp u$.

Лемма 8.3.1 *Для любых векторов a, $b \in \mathbb{C}^n$ одинаковой длины существуют число γ и матрица отражений H такие, что*

$$Ha = \gamma\, b, \quad |\gamma| = 1.$$

Доказательство. Если $H = H(u)$, то должно быть

$$a - 2\,(u^*a)\,u = \gamma\, b. \qquad (*)$$

Если a и b — ненулевые коллинеарные векторы, то можно взять $u = a/||a||_2$. В противном случае положим

$$u = \frac{a - \gamma\, b}{||a - \gamma\, b||_2}.$$

Отсюда, согласно $(*)$,

$$2\,(u^*a) = ||a - \gamma\, b||_2 \quad \Leftrightarrow$$

$$2\,(a^*a - \gamma^* b^* a) = ||a - \gamma\, b||_2^2 = ||a||_2^2 + ||b||_2^2 - 2\,\mathrm{Re}\,(\gamma^* b^* a).$$

Поскольку $||a||_2 = ||b||_2$, получаем $\gamma^* b^* a = \mathrm{Re}\,(\gamma^* b^* a) \quad \Leftrightarrow \quad$ число $\gamma^*\, b^* a$ вещественное. Если $b^* a = 0$, то можно взять любое γ, $|\gamma| = 1$. Иначе, есть две возможности:

$$\gamma = b^* a/|b^* a| \quad \text{или} \quad \gamma = -b^* a/|b^* a|. \quad \square$$

8.4 Исключение элементов с помощью отражений

Согласно лемме 8.3.1, для любого столбца $a \in \mathbb{C}^n$ существует матрица отражений H такая, что

$$H\,a = \gamma\ [\|a\|_2,\, 0, \ldots, 0]^T, \quad |\gamma| = 1.$$

В данном случае H определяется вектором $u = v/||v||_2$, где

$$v = [a_1 - \gamma\,\|a\|_2,\, a_2, \ldots,\, a_n]^T.$$

Если $a_1 \neq 0$, то рекомендуем взять

$$\gamma = -a_1/|a_1|.$$

(При этом не будут вычитаться числа одного знака!)

Для любой матрицы $A \in \mathbb{C}^{n\times n}$ существуют матрицы отражения $H_1, \ldots, H_{n-1}$ такие, что матрица

$$H_{n-1} \ldots H_1 A = R$$

является верхней треугольной.

Будем определять H_i с помощью вектора u_i, имеющего нули в первых $i-1$ компонентах и такого, что H_i аннулирует поддиагональные элементы i-го столбца матрицы $H_{i-1} \ldots H_1 A$.

Произведение унитарных матриц $Z \equiv H_{n-1} \ldots H_1$ есть унитарная матрица. Находим $A = Q\,R,\ Q = Z^*$, — мы получили еще одно (конструктивное) доказательство существования QR-разложения.

Чтобы получить QR-разложение с помощью матриц отражения, потребуется $\frac{4}{3}\,n^3 + \mathcal{O}\,(n^2)$ арифметических операций (докажите). Для сравнения: метод Гаусса находит LU-разложение за $\frac{2}{3}\,n^3 + \mathcal{O}\,(n^2)$ арифметических операций.

8.5 Матрицы вращения

Матрица $G_{kl} \in \mathbb{R}^{n\times n}$ называется *матрицей вращения*, или матрицей Гивенса, если она отличается от единичной матрицы лишь 2×2-подматрицей

$$M\,(\phi) = \begin{bmatrix} \cos\,\phi & -\sin\,\phi \\ \sin\,\phi & \cos\,\phi \end{bmatrix},$$

расположенной на строках и столбцах с номерами k и l.

Проверьте, что матрица G_{kl} является ортогональной.

8.6 Исключение элементов с помощью вращений

Если вектор $[a_1,\, a_2]^T \in \mathbb{R}^2$ ненулевой, то выбор

$$\cos\,\phi = \frac{a_1}{\sqrt{a_1^2 + a_2^2}}, \quad \sin\,\phi = \frac{-a_2}{\sqrt{a_1^2 + a_2^2}}$$

позволяет исключить его вторую компоненту:

$$M\,(\phi)\,\left[\begin{array}{c} a_1 \\ a_2 \end{array}\right] = \left[\begin{array}{c} \alpha \\ 0 \end{array}\right].$$

Очевидно, при умножении матрицы A слева на матрицу вращения G_{kl} можно получить нуль в любой позиции k-й или l-й строки. Следовательно, матрицу A можно привести к верхнему треугольному виду R с помощью умножений слева на последовательность матриц вращения:

$$G_{n-1\,n} \ldots G_{1n}\, \ldots\, G_{13}\, G_{12}\, A = R.$$

Чтобы получить QR-разложение с помощью вращений, потребуется $2\,n^3 + \mathcal{O}\,(n^2)$ арифметических операций (докажите).

8.7 Машинные реализации отражений и вращений

Реально вычисленные с помощью отражений или вращений $\tilde{Q}$ и $\tilde{R}$ таковы, что

$$||A - \tilde{Q}\,\tilde{R}|| \;\leq\; c_1\,(n)\,\eta\,||A|| \;+\; \mathcal{O}\,(\eta^2), \tag{8.7.1}$$

$$||\tilde{Q}^*\,\tilde{Q} - I|| \;\leq\; c_2\,(n)\,\eta \;+\; \mathcal{O}\,(\eta^2), \tag{8.7.2}$$

где $c_1\,(n)$ и $c_2\,(n)$ — некоторые функции от n (они зависят от используемых норм и особенностей реализации).

Алгоритмы отражений и вращений различаются с точки зрения параллельных вычислений. Используя нестандартный способ вычисления сумм (попарное суммирование), с помощью отражений мы можем найти QR-разложение за $\mathcal{O}\,(n\,\log\,n)$ параллельных шагов. В то же время, стандартный алгоритм вращений обладает "скрытым" параллелизмом: для его реализации достаточно $\mathcal{O}\,(n)$ параллельных шагов (проверьте!).

8.8 Метод ортогонализации

QR-разложение можно получить без помощи вращений или отражений. Пусть $n = 3$ и мы хотим удовлетворить равенство $(A = Q\,R)$

$$[a_1\;\, a_2\;\, a_3] = [q_1\;\, q_2\;\, q_3]\;\left[\begin{array}{ccc} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{array}\right].$$

Для удобства мы будем вводить ненормированные векторы p_1, p_2, p_3, коллинеарные соответственно q_1, q_2, q_3. Первые столбцы будут равными, если мы возьмем $p_1 = a_1,\ r_{11} = ||p_1||_2, \quad q_1 = p_1/r_{11}$. Приравняем вторые столбцы:

$$a_2 = q_1\, r_{12} \ + \ q_2\, r_{22}.$$

Вектор q_2 должен быть ортогонален к q_1. Умножая обе части на q_1^*, находим:

$$\begin{aligned} r_{12} &= q_1^* a_2, \qquad p_2 = a_2 - q_1\, r_{12}, \\ r_{22} &= ||p_2||_2, \quad q_2 = p_2/r_{22}. \end{aligned}$$

Приравняем третьи столбцы:

$$a_3 = q_1\, r_{13} \ + \ q_2\, r_{23} \ + \ q_3\, r_{33}.$$

Умножая обе части по очереди на q_1^* и q_2^*, находим:

$$\begin{aligned} & r_{13} = q_1^*\, a_3, \ \ r_{23} = q_2^*\, a_3, \\ & p_3 = a_3 - q_1\, r_{13} - q_2\, r_{23}, \quad r_{33} = ||p_3||_2,\ q_3 = p_3/r_{33}. \end{aligned}$$

То, что мы получили, — это классический *процесс ортогонализации Грама–Шмидта*. Он может применяться для получения ортонормированного базиса в линейной оболочке $\operatorname{span}\{a_1, \ldots, a_k\}$, натянутой на линейно независимые векторы $a_1, \ldots, a_k$. Вот как он выглядит в общем случае:

$$p_j = a_j - \sum_{i=1}^{j-1} q_i\, (q_i^* a_j), \quad q_j = p_j/||p_j||_2, \quad j = 1, \ldots, k. \tag{8.8.3}$$

При этом $r_{ij} = q_i^*\, a_j$.

Для того чтобы ортогонализовать n векторов размерности n, потребуется $2\, n^3 + \mathcal{O}\,(n^2)$ арифметических операций (проверьте).

8.9 Потеря ортогональности

В условиях машинной арифметики линейная оболочка для реально вычисленных векторов $\tilde{q}_1, \ldots, \tilde{q}_k$, полученных по формулам (8.8.3), совпадает с линейной оболочкой, натянутой на возмущенные векторы $a_1 + f_1, \ldots, a_k + f_k$. Можно гарантировать, что возмущения $f_1, \ldots, f_k$ малы (докажите!).

Однако, векторы $\tilde{q}_1, \ldots, \tilde{q}_k$ часто далеки от ортогональных. Попробуем понять, почему. Введем матрицы

$$\tilde{Q}_i \equiv [\tilde{q}_1, \ldots, \tilde{q}_i], \quad i = 1, \ldots, k.$$

Тогда естественной мерой ортогональности векторов $\tilde{q}_1, \ldots, \tilde{q}_i$ является величина

$$\delta_i \equiv ||\tilde{Q}_i^*\, \tilde{Q}_i - I||_2. \tag{8.9.4}$$

Предположим, что на $i+1$-ом шаге вычисления выполняются абсолютно точно. Находим

$$\begin{aligned}\beta_{i+1} \equiv \tilde{Q}_i^*\,\tilde{q}_{i+1} &= \frac{1}{\|\tilde{p}_{i+1}\|_2}\,\tilde{Q}_i^*\,\left(a_{i+1} - \tilde{Q}_i\,\tilde{Q}_i^*\,a_{i+1}\right)\\ &= \frac{1}{\|\tilde{p}_{i+1}\|_2}\,\left(I - \tilde{Q}_i^*\,\tilde{Q}_i\right)\,\tilde{Q}_i^*\,a_{i+1}.\end{aligned}$$

Следовательно,

$$\|\beta_{i+1}\|_2 \;\le\; \delta_i\,\sqrt{1+\delta_i}\,\frac{\|a_{i+1}\|_2}{\|\tilde{p}_{i+1}\|_2}. \tag{8.9.5}$$

Для наших целей достаточно очевидной оценки

$$\delta_{i+1} = \left\|\,\begin{bmatrix}\tilde{Q}_i^*\,\tilde{Q} - I & \beta_{i+1}\\ \beta_{i+1}^* & 0\end{bmatrix}\,\right\|_2 \;\le\; \delta_i \;+\; 2\,\|\beta_{i+1}\|_2. \tag{8.9.6}$$

Более аккуратная оценка имеет вид

$$\delta_{i+1} \;\le\; \frac{\delta_i \;+\; \sqrt{\delta_i^2 + 4\,\|\beta_{i+1}\|_2^2}}{2}. \tag{8.9.7}$$

Таким образом,

$$\delta_{i+1} \;\le\; \text{const}\,\frac{\|a_{i+1}\|_2}{\|\tilde{p}_{i+1}\|_2}\,\delta_i\,\sqrt{1+\delta_i}.$$

Даже если δ_i мало, δ_{i+1} может оказаться очень большим, если мала величина $\|\tilde{p}_{i+1}\|_2$.

8.10 Как бороться с потерей ортогональности

Пусть вычислены векторы $\tilde{q}_1, \ldots, \tilde{q}_i$, которые "не слишком далеки" от ортогональных: пусть $\delta_i < 1$. Чтобы поддержать ортогональность на $i+1$-ом шаге, рассмотрим следующий *процесс реортогонализации*:

$$\begin{aligned}p^{(0)} &= a_{i+1};\\ p^{(j)} &= (I - \tilde{Q}_i\,\tilde{Q}_i^*)\,p^{(j-1)}, \quad j = 1, 2, \ldots.\end{aligned}$$

Итерация при $j = 1$ отвечает обычному шагу процесса Грама–Шмидта. Остановившись на j-ой итерации, находим $q_{i+1} = p^{(j)}/\|p^{(j)}\|_2$.

Легко видеть, что

$$\beta^{(j)} \;\equiv\; \tilde{Q}_i^*\,p^{(j)} = (I - \tilde{Q}_i^*\,\tilde{Q}_i)^j\,\tilde{Q}_i^*\,p^{(0)}.$$

При $\delta_i < 1$ матрица $I - \tilde{Q}_i^*\,\tilde{Q}_i$ будет сходящейся $\Rightarrow$ $\beta^{(j)} \to 0$ при $j \to \infty$. Таким образом, вектор $\tilde{q}_i$ можно сделать с высокой степенью

точности ортогональным всем предыдущим векторам — даже в том случае, когда последние ортогональны с существенно меньшей степенью точности.

В.Хоффман рекомендует [1] проводить реортогонализацию до тех пор, пока

$$\frac{\|p^{(j)}\|_2}{\|p^{(j-1)}\|_2} \;\le\; \frac{1}{2}. \qquad (*)$$

8.11 Модифицированный алгоритм Грама–Шмидта

Рассмотрим две программы, реализующие формулы (8.8.3):

DO	$j = 1,\ k$	DO	$j = 1,\ k$
	$p_j = a_j$		$p_j = a_j$
	DO $i = 1,\ j-1$		DO $i = 1,\ j-1$
	$\quad p_j = p_j - q_i\ (q_i^*\ a_j)$		$\quad p_j = p_j - q_i\ (q_i^*\ p_j)$
	END DO		END DO
	$q_j = p_j / \|p_j\|_2$		$q_j = p_j / \|p_j\|_2$
END	DO	END	DO

Вторая реализует *модифицированный алгоритм Грама–Шмидта*. Она получается из первой заменой a_j на p_j во внутреннем цикле.

В точной арифметике модифицированный алгоритм эквивалентен исходному. Однако, в отличие от исходного варианта, модифицированный алгоритм уже дает некоторую гарантию ортогональности в условиях машинной арифметики. А.Бьорк показал, что

$$\delta_k \;\le\; c(n,\,k)\,\frac{\sigma_{\max}}{\sigma_{\min}}\,\eta,$$

где $\sigma_{\max}$ и $\sigma_{\min}$ — максимальное и минимальное сингулярные числа $n \times k$-матрицы $A = [a_1,\,\ldots\,,a_k],\ n \ge k$.

Недавно Ч. Шеффилд обнаружил, что в условиях машинной арифметики модифицированный алгоритм Грама–Шмидта совпадает с алгоритмом отражений, строящим QR-разложение расширенной нулями прямоугольной матрицы

$$\hat{A} = \begin{bmatrix} 0_{k\times k} \\ A \end{bmatrix}.$$

В 1992 году Бьорк и Пейдж [2] использовали это наблюдение для нового вывода оценки Бьорка.

---

[1] W.Hoffmann, Iterative algorithms for Gram–Schmidt orthogonalization, Computing 41: 335–348 (1989). (В этой статье приводится много числовых примеров в пользу (*), хотя строгой оценки для меры ортогональности там все же не получено.)

[2] A.Bjorck, C.C.Paige, Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm, SIAM J. Matrix Anal. Appl. 13 (1): 176–190 (1992).

Если ортогональность векторов в модифицированном алгоритме нас все-таки не устраивает, его нужно дополнить процессом реортогонализации.

8.12 Двухдиагонализация

Матрица $B = [b_{ij}]$ называется (верхней) *двухдиагональной*, или *бидиагональной*, если $b_{ij} = 0$, когда $i > j$ или $i + 1 < j$.

Любую $n \times n$-матрицу A можно привести к двухдиагональному виду

$$B = P\,A\,Q,$$

где P и Q — произведения конечного числа матриц отражения (или вращения).

Пусть используются отражения. Сначала мы умножаем A слева на матрицу отражения, аннулирующую все поддиагональные элементы первого столбца. Затем умножаем результат справа на матрицу отражения, аннулирующую элементы первой строки в позициях с 3-й по n-ю. Важно, что ранее полученные нули в первом столбце сохранятся!

Далее умножением слева аннулируем все поддиагональные элементы второго столбца, затем умножением справа получаем нули во второй строке в позициях с 4-й по n-ю, и т.д.. После каждого умножения на матрицу отражения все ранее полученные нули остаются (проверьте).

С помощью унитарной двухдиагонализации некоторые (не все!) задачи для матрицы A сводятся к задачам для двухдиагональной матрицы B. Вот важные примеры:

- *Сингулярное разложение* произвольной матрицы A можно получать, научившись находить сингулярное разложение двухдиагональной матрицы B.
- *Задача наименьших квадратов* для A, то есть задача минимизации $\|A\,x - b\|_2$ по x, с помощью замены переменных $x = Q\,u$, $b = P\,f$ сводится к задаче минимизации $\|B\,u - f\|_2$ по u для двухдиагональной матрицы B.

8.13 Приведение к почти треугольной форме

Матрица $H = [h_{ij}]$ называется (верхней) *почти треугольной*, или *хессенберговой*, если $h_{ij} = 0$ при $i > j + 1$.

Любую $n \times n$-матрицу A можно привести к *унитарно подобной* почти треугольной матрице

$$H = P\,A\,P^*,$$

где P - произведение конечного числа отражений (или вращений).

Пусть используются отражения. Выберем матрицу отражения

$$P_1 = I - 2uu^*$$

так, чтобы первая компонента вектора u была нулевой и при этом

$$P_1 \, [a_{11}, a_{21}, a_{31}, \ldots, a_{n1}]^T = [a_{11}, *, 0, \ldots, 0]^T .$$

Когда матрица умножается на P_1 слева, в ней не изменяются элементы первой строки. Поэтому когда матрица умножается на $P_1^* = P_1$ справа, в ней не изменяются элементы первого столбца. Поэтому в матрице $P_1 \, A \, P_1$ первый столбец имеет нули в позициях с 3-й по n-ю. Далее выбираем матрицы отражения $P_2, \ldots, P_{n-2}$ так, чтобы умножение слева на P_i давало нули в позициях i-го столбца с $i+2$-й по n-ю. В итоге $P = P_{n-2} \ldots P_1$.

Подобные матрицы имеют одинаковый спектр. Таким образом, если требуется найти собственные значения и собственные векторы для произвольной матрицы A, нам достаточно научиться решать эту задачу для почти треугольной матрицы H. Важное свойство: в силу унитарности матрицы P простые собственные значения для H будут иметь те же коэффициенты перекоса, что и для A (докажите).

Если матрица A эрмитова, то почти треугольная матрица H будет в действительности трехдиагональной (докажите!).

Задачи

1. Верно ли, что для любых векторов a, $b \in \mathbb{C}^n$ одинаковой длины существует матрица отражения H такая, что $H\,a = b$?

2. Верно ли, что любую матрицу вращения можно представить конечным произведением матриц отражения?

3. Верно ли, что любую матрицу отражения можно представить конечным произведением матриц вращения?

4. Пусть $H = H^*$. Докажите неравенство

$$\left\| \begin{bmatrix} H & \beta \\ \beta^* & 0 \end{bmatrix} \right\|_2 \leq \frac{\|H\|_2 + \sqrt{\|H\|_2^2 + 4\,\|\beta\|_2^2}}{2}.$$

   (Отсюда вытекает оценка (8.9.7)).

5. Докажите, что любая матрица порядка 3 унитарно подобна трехдиагональной матрице.

6. Матрица $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ составлена из квадратных блоков A_{ij} порядка n. Докажите, что если A является одновременно ортогональной и верхней хессенберговой (почти треугольной), то $\mathrm{rank} A_{12} \leq 1$.

7. Покажите, что в условиях машинной арифметики модифицированный алгоритм Грама–Шмидта для ортогонализации столбцов $n \times k$-матрицы A эквивалентен алгоритму отражений, получающему QR-разложение расширенной нулями прямоугольной матрицы

$$\hat{A} = \begin{bmatrix} 0_{k \times k} \\ A \end{bmatrix}.$$

8. Пусть $H = P^* A P$, где P — унитарная матрица. Докажите, что любое простое собственное значение для A имеет тот же коэффициент перекоса, что и для H.

9. Напишите программу, реализующую модифицированный алгоритм Грама-Шмидта и примените ее для ортогонализации столбцов матрицы $A = [1/(i + j - 1)]$ порядка $n = 13$. После завершения счета вычислите скалярные произведения "ортонормированных" столбцов и убедитесь в том, что некоторые из них далеки от машинного нуля. Посмотрите, что на этом примере дает реортогонализация.

10. Докажите, что метод QR-разложения, основанный на вращениях, реализуется за $O(n)$ параллельных шагов.

Глава 9

9.1 Проблема собственных значений

Под проблемой собственных значений понимается совокупность задач, связанных с вычислением собственных значений и векторов.

Пусть требуется найти все собственные значения матрицы $A \in \mathbb{C}^{n\times n}$. Как подступиться к этой задаче? Довольно старая идея — найти коэффициенты характеристического полинома и свести задачу к вычислению корней полинома.

Для получения коэффициентов характеристического полинома можно построить прямой метод, требующий $\mathcal{O}(n^3)$ арифметических операций. Однако, в компьютерную эру идея была отвергнута. Дело в том, что собственные значения могут быть слабо чувствительны к малым возмущениям элементов матрицы, но сильно чувствительны к малым возмущениям коэффициентов ее характеристического полинома: "хорошая" задача сводится к "плохой"!

Трудными нужно считать случаи, когда собственные значения сильно чувствительны к малым возмущениям элементов матрицы. Если эти собственные значения возникли на промежуточном этапе решения какой-то задачи, то, вероятно, имеет смысл подумать о том, нельзя ли ее решить без вычисления отдельных собственных значений.

Современная точка зрения [1] на решение спектральных задач связана с изучением так называемых *спектральных портретов*: для заданной матрицы A и параметра $\varepsilon > 0$ это множества вида

$$S(\varepsilon) = \{z \in \mathbb{C} : f(\lambda) \equiv \sigma_{\min}(A - zI) \leq \varepsilon\},$$

где $\sigma_{\min}(B)$ обозначает минимальное сингулярное число матрицы B.

Очевидно, спектр матрицы A содержится в $S(\varepsilon)$ (докажите). Возмущения порядка ε позволяют собственным значениям изменяться в пределах множества $S(\varepsilon)$. Поэтому ответ к задаче о вычислении собственных значе-

[1] С. К. Годунов, *Современные аспекты линейной алгебры*, Научная книга, Новосибирск, 1997.

ний полезно давать в виде *линий уровня* функции $f(\lambda)$, то есть кривых, определенных условием $f(\lambda) = \varepsilon$ при различных $\varepsilon > 0$.

Тем не менее, мы должны изучить прежде всего классические методы, дающие в качестве ответа отдельные собственные значения. Во многих случаях спектральные задачи решаются с их помощью очень успешно (в частности, для нормальных и "достаточно близких к ним" матриц).

9.2 Степенной метод

Пусть A имеет базис из собственных векторов: $A\,z_i = \lambda_i\,z_i$, и предположим, что

$$|\lambda_1| \;>\; |\lambda_2| \;>\; \ldots \;>\; |\lambda_n|.$$

Тогда λ_1 можно найти, применив *степенной метод*:

y_0 – ненулевой начальный вектор, $x_0 \;=\; y_0/\|y_0\|$;
$y_k \;=\; A\,x_{k-1}, \quad x_k \;=\; y_k/\|y_k\|, \quad k = 1, 2, \ldots.$
(Почему необходимы нормировки?)

Если $y_0 = \mu_1\,z_1 + \cdots + \mu_n\,z_n$ и $\mu_1 \neq 0$, то x_k коллинеарен вектору

$$\frac{A^k\,y_0}{\mu_1\,\lambda_1^k} = z_1 \;+\; \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \quad\Longrightarrow\quad x_k^*\,A\,x_k \;\longrightarrow\; \lambda_1.$$

Чтобы найти λ_i при $i \neq 1$, можно попытаться подобрать α так, чтобы $\lambda_i - \alpha$ стало старшим собственным значением для матрицы $A - \alpha\,I$. Всегда ли это возможно?

Другой подход — провести один шаг индуктивного доказательства теоремы Шура и получить приближение к матрице порядка $n-1$ с собственными значениями $\lambda_2, \ldots, \lambda_n$ (эту манипуляцию иногда называют *исчерпыванием*).

В чистом виде степенной метод используется не часто. Но многие современные алгоритмы явно или неявно эксплуатируют именно его идею.

9.3 Итерации подпространства

Естественным развитием идеи степенного метода можно считать *метод итерации подпространства*:

Y_0 – $n \times m$-матрица ранга m,
$Y_0 \;=\; X_0\,R_0$ (QR-разложение прямоугольной матрицы Y_0);
$Y_k \;=\; A\,X_{k-1}, \quad Y_k = X_k\,R_k$ (QR-разложение), $\quad k = 1, 2, \ldots.$

Метод порождает последовательность подпространств $L_k = \operatorname{im} X_k$. При этом роль QR-разложения сродни нормировкам в степенном методе (оно дает ортонормированные базисы в L_k).

Если подпространство L_k инвариантно относительно A (это означает, что $A\,L_k \subset L_k$), то $A_k = X_k^*\,A\,X_k$ есть диагональный блок блочно треугольной матрицы, подобной A (почему?). Поэтому

$$\lambda\,(A_k) \subset \lambda\,(A).$$

При определенных условиях с ростом k подпространство L_k будет приближаться к некоторому инвариантному подпространству матрицы A. Поэтому собственные значения $m \times m$-матрицы A_k (обычно m много меньше n) будут приближать собственные значения матрицы A.

Для строгого анализа, очевидно, нужно как-то определить расстояние между подпространствами.

9.4 Расстояние между подпространствами

Расстояние между вектором $x \in \mathbb{C}^n$ и подпространством $M \subset \mathbb{C}^n$ определяется так:

$$\rho\,(x,\,M) \equiv \min_{y\,\in\,M} \|x - y\|_2.$$

Поэтому, казалось бы, можно определить расстояние между подпространствами L и M как

$$\rho\,(L,\,M) \equiv \max_{x\,\in\,L,\;\|x\|_2\,=\,1} \rho\,(x,\,M).$$

Вопрос: удовлетворяет ли ρ аксиомам метрического пространства?

Мы скоро увидим, что это так, если рассматривать лишь подпространства одинаковой размерности. Если же размерности не одинаковы, то в общем случае $\rho\,(L,\,M) \neq \rho\,(M,\,L)$. Проверьте, что если L не ортогонально M и $\dim L > \dim M$, то $\rho\,(L,\,M) = 1$, но $\rho\,(M,\,L) < 1$.

Все аксиомы метрического пространства будут очевидным образом выполнены, если расстояние ввести следующим образом:

$$\operatorname{dist}\,(L,\,M) \equiv \max\{\rho\,(L,\,M),\,\rho\,(M,\,L)\}.$$

9.5 Подпространства и ортопроекторы

Матрица P называется *ортопроектором*, если $P^2 = P \quad$ и $\quad P^* = P$.

Утверждение 1. *Если $L = \operatorname{im} P$, то вектор $P\,x$ — это ортогональная проекция вектора x на подпространство L.*

Доказательство.

$$y \in L \;\Rightarrow\; y = P\,v \;\Rightarrow$$

$$y^*\,(x - P\,x) = v^*\,P^*\,(x - P\,x) = v^*\,(P\,x - P^2\,x) = 0. \quad \square$$

Понятно, что каждому подпространству L отвечает единственный ортопроектор P_L такой, что $\operatorname{im} P_L = L$ (докажите!).

Утверждение 2. *Если столбцы матрицы Q образуют ортонормированный базис в L, то $P_L = Q\,Q^*$.*

Доказательство.

$$(Q\,Q^*)^2 = Q\,(Q^*\,Q)\,Q^* = Q\,Q^*; \qquad (Q\,Q^*)^* = (Q^*)^*\,Q^* = Q\,Q^*. \quad \square$$

9.6 Расстояния и ортопроекторы

Лемма 9.6.1 *Пусть ортопроекторы P_L и P_M отвечают подпространствам L и M. Тогда*

$$\rho\,(L,\,M) = ||(I - P_M)\,P_L||_2.$$

Доказательство.

$$\begin{aligned} \rho\,(L,\,M) &= ||P_L\,x - P_M\,P_L\,x||_2 \quad (\text{для некоторого } x = P_L\,x,\ \|x\|_2 = 1) \\ &\le ||P_L - P_M\,P_L||_2 = ||(I - P_M)\,P_L||_2. \end{aligned}$$

Теперь покажем, что $||(I - P_M)\,P_L||_2 \le \rho\,(L,\,M)$. Пусть

$$||(I - P_M)\,P_L||_2 = ||(I - P_M)\,P_L\,y\|_2,\ \|y\|_2 = 1.$$

Если $P_L\,y = 0$, то это очевидно. Если $P_L\,y \ne 0$, то

$$||(I - P_M)\,P_L\,y||_2 \;\le\; \left\| \frac{P_L\,y}{\|P_L\,y\|_2} - P_M\,\frac{P_L\,y}{\|P_L\,y\|_2} \right\|_2 \|P_L\,y\|_2 \;\le\; \rho\,(L,\,M). \;\square$$

Теорема 9.6.1 *Пусть ортопроекторы P_L и P_M отвечают подпространствам L и M. Тогда*

$$\operatorname{dist}\,(L,\,M) = \|P_L - P_M\|_2.$$

Доказательство. Согласно лемме 9.6.1,

$$\rho\,(L,\,M) = ||(I - P_M)\,P_L||_2 = ||(P_L - P_M)\,P_L||_2 \le ||P_L - P_M||_2.$$

$$\Rightarrow \quad \mathrm{dist}\,(L,\,M) \;\le\; \|P_L - P_M\|_2.$$

Далее, $\exists\ x \in \mathbb{C}^n,\ \|x\|_2 = 1:\ \ ||P_L - P_M||_2 \;=\; ||(P_L - P_M)\,x||_2$. По теореме Пифагора,

$$\begin{aligned} ||(P_L \;-\; P_M)\,x||_2^2 & \\ &= ||P_L\,(P_L - P_M)\,x||_2^2 + ||(I - P_L)\,(P_L - P_M)\,x||_2^2 \\ &= ||P_L\,(I - P_M)\,x||_2^2 + ||(I - P_L)\,P_M\,x||_2^2 \\ &\le ||P_L\,(I - P_M)||_2^2\,||(I - P_M)\,x||_2^2 + ||(I - P_L)\,P_M||_2^2\,||P_M\,x||_2^2 \\ &\le \max\,\{\|(P_L\,(I - P_M)\|_2^2,\ \|(I - P_L)\,P_M\|_2^2\}. \end{aligned}$$

Остается заметить, что

$$||P_L\,(I - P_M)||_2 = ||(P_L\,(I - P_M))^*||_2 = ||(I - P_M)\,P_L||_2 \;\le\; \rho\,(L,\,M). \quad \square$$

Следствие 9.6.1 *Расстояние между подпространствами равно расстоянию между их ортогональными дополнениями.*

9.7 Подпространства одинаковой размерности

Лемма 9.7.1 *Если матрица*

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}, \quad Q_{11} \in \mathbb{C}^{m\times m},\ Q_{22} \in \mathbb{C}^{k\times k},$$

унитарная, то $\|Q_{12}\|_2 = \|Q_{21}\|_2$.

Доказательство. Рассмотрим сингулярное разложение для первого диагонального блока: $Q_{11} = U_1\,\Sigma_1\,V_1^*$, и перейдем к другой унитарной матрице

$$\hat{Q} \;=\; \begin{bmatrix} U_1^* & 0 \\ 0 & I \end{bmatrix}\,Q\,\begin{bmatrix} V_1 & 0 \\ 0 & I \end{bmatrix} \;=\; \begin{bmatrix} \Sigma_1 & \hat{Q}_{12} \\ \hat{Q}_{21} & \hat{Q}_{22} \end{bmatrix}.$$

Ясно, что $\|Q_{12}\|_2 = \|\hat{Q}_{12}\|_2, \quad \|Q_{21}\|_2 = \|\hat{Q}_{21}\|_2$. В то же время,

$$\begin{aligned} \hat{Q}\,\hat{Q}^* = I \quad &\Rightarrow \quad \hat{Q}_{12}\,\hat{Q}_{12}^* = I - \Sigma_1^2, \\ \hat{Q}^*\,\hat{Q} = I \quad &\Rightarrow \quad \hat{Q}_{21}^*\,\hat{Q}_{21} = I - \Sigma_1^2. \quad \square \end{aligned}$$

Теорема 9.7.1 *Пусть подпространства L и M имеют одинаковую размерность и унитарные матрицы $U = [U_1 \quad U_2]$ и $V = [V_1 \quad V_2]$ таковы, что*

$$\operatorname{im} U_1 = L, \quad \operatorname{im} V_1 = M.$$

Тогда

$$\operatorname{dist}(L,\, M) = \|U_1^*\, V_2\|_2 = \|U_2^*\, V_1\|_2.$$

Доказательство.

$$U^*\,(U_1\,U_1^* - V_1\,V_1^*)\,V = \begin{bmatrix} 0 & U_1^*\,V_2 \\ -U_2^*\,V_1 & 0 \end{bmatrix}. \quad \square$$

Следствие 9.7.1 *Если* $\dim L = \dim M$, *то* $\rho(L,\, M) = \rho(M,\, L)$.

Доказательство.

$$\begin{aligned} U^*\,(U_1\,U_1^*)\;(I - (V_1\,V_1^*))\;V &= \begin{bmatrix} 0 & U_1^*\,V_2 \\ 0 & 0 \end{bmatrix}, \\ V^*\,(V_1\,V_1^*)\;(I - (U_1\,U_1^*))\;U &= \begin{bmatrix} 0 & V_1^*\,U_2 \\ 0 & 0 \end{bmatrix}. \quad \square \end{aligned}$$

9.8 Углы между подпространствами и CS-разложение

Расстояние между подпространствами — это лишь одна из характеристик взаиморасположения двух подпространств. Более детальную информацию дают так называемые главные углы между подпространствами. Чтобы ввести их, требуется CS-разложение некоторой унитарной матрицы, связанной с двумя подпространствами.

Теорема 9.8.1 *Пусть Q — унитарная матрица порядка n. Тогда для любого $m \le n/2$ существуют унитарные матрицы U_1, V_1 порядка m и унитарные матрицы U_2, V_2 порядка $n - m$ такие, что*

$$\begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix} Q \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix} = \begin{bmatrix} C & S & 0 \\ -S & C & 0 \\ 0 & 0 & I_{n-2\,m} \end{bmatrix}, \qquad (*)$$

$$C = \operatorname{diag}(c_1, \ldots, c_m),\; S = \operatorname{diag}(s_1, \ldots, s_m),$$

$$c_1 \ge \cdots \ge c_m \ge 0,$$

$$0 \le s_1 \le \cdots \le s_m,$$

$$c_i^2 + s_i^2 = 1, \quad i = 1, \ldots, m.$$

Разложение $(*)$ называется CS-разложением унитарной матрицы Q.

Для доказательства представим Q в блочном виде

$$\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}, \quad Q_{11} \in \mathbb{C}^{m\times m},$$

и рассмотрим сингулярное разложение блока $Q_{11} = U\,C\,V^*$. Получаем

$$\begin{bmatrix} U^* & 0 \\ 0 & I_{n-m} \end{bmatrix} Q \begin{bmatrix} V & 0 \\ 0 & I_{n-m} \end{bmatrix} = \begin{bmatrix} C & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \equiv W.$$

Равенства $W\,W^* = W^*\,W = I$ влекут за собой

$$W_{12}\,W_{12}^* = W_{21}^*\,W_{21} = I_m - C^2 \equiv S^2.$$

Отсюда видно, что существуют унитарные матрицы X, Y порядка $n-m$ такие, что

$$W_{12}\,Y = [S,\ 0], \quad X\,W_{21} = [-S,\ 0]^T.$$

(Здесь есть некоторый произвол: например, можно было бы S и $-S$ поменять местами или умножить их на любые унитарные диагональные матрицы.) Доведите доказательство до конца!

Следствие 9.8.1 *Пусть подпространства L и M имеют одинаковую размерность $m \le \frac{n}{2}$. Тогда существуют ортонормированные базисы $u_1, \ldots, u_m \in L$ и $v_1, \ldots, v_m \in M$ такие, что*

$$u_i^*\,v_j = \begin{cases} c_i, & i = j, \\ 0, & i \ne j, \end{cases}$$

где числа $c_1 \ge \cdots \ge c_m \ge 0$ определяются однозначно.

Чтобы доказать следствие, достаточно рассмотреть CS-разложение унитарной матрицы $U^*\,V$, где $U = [U_1\ U_2]$, $V = [V_1\ V_2]$ — унитарные матрицы такие, что $L = \operatorname{im} U_1$, $M = \operatorname{im} V_1$.

Можно записать

$$c_i = \cos\,\phi_i, \quad 0 \le \phi \le \frac{\pi}{2}.$$

Углы ϕ_i называются *главными углами* между подпространствами L и M.

9.9 Сходимость для блочно диагональной матрицы

Пусть

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \quad A_1 \in \mathbb{C}^{m\times m}, \ \ A_2 \in \mathbb{C}^{r\times r}, \quad \exists\ A_1^{-1}. \qquad (*)$$

Введем m-мерное подпространство M, натянутое на первые m столбцов единичной матрицы. Очевидно, M инвариантно относительно A. Чтобы "заставить" метод итерации подпространства сходиться к M, потребуем, чтобы начальное m-мерное подпространство L не было "слишком далеким" от M (используем обозначения раздела 9.4):

$$\beta \equiv \rho\,(L,\, M) \;<\; 1. \qquad (**)$$

Лемма 9.9.1 *В условиях* $(*)$ *и* $(**)$

$$\rho\,(A\,L,\, M) \;\leq\; \frac{\beta}{\sqrt{1-\beta^2}}\,||A_2||_2\,||A_1^{-1}||_2. \qquad (9.9.1)$$

Доказательство. Пусть $y = A\,x,\; y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},\; x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},\; x_1,\; y_1 \in \mathbb{C}^m.$ Предположим, что $x \;\in\; L, \quad \rho\,(A\,L,\, M) \;=\; \rho\,(y,\, M), \quad ||y||_2 = 1.$ Тогда

$$\rho\,(\frac{x}{||x||_2},\, M) = \frac{||x_2||_2}{||x||_2} \leq \beta \quad \Rightarrow \quad ||x_2||_2 \leq \frac{\beta}{\sqrt{1-\beta^2}}\,||x_1||_2 \quad \Rightarrow$$

$$\begin{aligned} \rho\,(y,\, M) = ||y_2||_2 \quad &= \quad ||A_2\,x_2||_2 \;\leq\; ||A_2||_2\,||x_2|| \\ &\leq \quad ||A_2||_2\,\frac{\beta}{\sqrt{1-\beta^2}}\,||x_1||_2 \\ &\leq \quad ||A_2||_2\,\frac{\beta}{\sqrt{1-\beta^2}}\,||A_1^{-1}\,y_1||_2 \\ &\leq \quad ||A_2||_2\,\frac{\beta}{\sqrt{1-\beta^2}}\,||A_1^{-1}||_2. \quad \square \end{aligned}$$

Обозначим через λ_{m+1} — максимальное по модулю собственное значение блока A_2, через λ_m — минимальное по модулю собственное значение блока A_1, и положим

$$\gamma = \gamma\,(A_1,\, A_2) \equiv |\lambda_{m+1}| \;/\; |\lambda_m|. \qquad (9.9.2)$$

Следствие 9.9.1 *Если* $\gamma \;<\; 1$, *то* $\forall\, q \in (\gamma,\, 1)\;\exists\;\; c = c\,(q)$:

$$\rho\,(A^k\,L,\, M) \;\leq\; c\,q^k, \quad k = 1, 2, \ldots.$$

Доказательство. Применим лемму 9.9.1 к матрице A^k:

$$\rho\,(A^k\,L,\, M) \;\leq\; \frac{\beta}{\sqrt{1-\beta^2}}\,||A_2^k||_2\,||(A_1^{-1})^k||_2.$$

Мы знаем, что для любой матрицы A последовательность $\left(||A^k||_2\right)^{\frac{1}{k}}$ стремится при $k \to \infty$ к спектральному радиусу матрицы A (максимальному по модулю ее собственному значению). Следовательно, $\forall\ \delta > 0\ \exists\ k_\delta$: $k > k_\delta \Rightarrow$

$$||A_2^k||_2\,||(A_1^{-1})^k||_2 \ \le\ (|\lambda_{m+1}| + \delta)^k \ \left(\frac{1}{|\lambda_m|} + \delta\right)^k .$$

Для любого $q > \gamma$ при всех достаточно малых $\delta > 0$ правая часть неравенства будет меньше q^k. $\square$

9.10 Сходимость в общем случае

Пусть

$$A = Z\,\Lambda\,Z^{-1}, \quad \Lambda = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \tag{9.10.3}$$
$$A_1 \in \mathbb{C}^{m\times m}, \quad A_2 \in \mathbb{C}^{r\times r}, \quad \exists\, A_1^{-1}.$$

Лемма 9.10.1 *Для любой невырожденной матрицы Z и подпространств $\mathcal{L}$, $\mathcal{M}$ выполняется неравенство*

$$\rho\,(Z\,\mathcal{L},\, Z\,\mathcal{M}) \le \mathrm{cond}_2\, Z\ \rho\,(\mathcal{L},\, \mathcal{M}).$$

Доказательство. Пусть $x \in \mathcal{L},\ ||x||_2 = 1$ и при этом

$$\rho\,(Z\,\mathcal{L},\, Z\,\mathcal{M}) = \rho\,(\frac{Z\,x}{||Z\,x||_2},\, Z\,\mathcal{M}).$$

Находим:

$$\begin{aligned}
\rho\,(\frac{Z\,x}{||Z\,x||_2},\, Z\,\mathcal{M}) \ &=\ \rho\,(\frac{Z\,x}{||Z\,x||_2},\, \frac{Z}{||Z\,x||_2}\,\mathcal{M}) \\
&\le\ \frac{1}{||Z\,x||_2}\,||Z\,(x-z)||_2 \quad (\ \forall\, z \in \mathcal{M}\ ) \\
&\le\ \frac{||Z||_2}{||Z\,x||_2}\,||x-z||_2 \\
&\le\ ||Z||_2\,||Z^{-1}||_2\,\rho\,(\mathcal{L},\, \mathcal{M}). \quad \square
\end{aligned}$$

Теорема 9.10.1 *Пусть A имеет вид (9.10.3), подпространство M есть линейная оболочка, натянутая на первые m столбцов матрицы Z, а m-мерное подпространство L таково, что*

$$\beta \equiv \rho\,(Z^{-1}\,L,\, Z^{-1}\,M) \ <\ 1.$$

Пусть γ *имеет вид* (9.9.2). *Тогда если* $\gamma < 1$, *то* $\forall\, q \in (\gamma,\, 1)\ \exists\ c = c\,(q)$:

$$\rho\,(A^k\,L,\,M) \ \leq\ c\,q^k, \quad k = 1, 2, \ldots .$$

Доказательство. Общий случай сводится к случаю блочно диагональной матрицы Λ, имеющей инвариантное подпространство $Z^{-1}\,M$ (почему?). Мы уже знаем, что для Λ итерации начального подпространства $Z^{-1}\,L$ будут сходиться к $Z^{-1}\,M$: при $q \in (\gamma,\, 1)$

$$\rho\,(\Lambda^k\,(Z^{-1}\,L),\,(Z^{-1}\,M)) \ \leq\ c\,q^k.$$

Поскольку $A^k = Z\,\Lambda^k\,Z^{-1}$, с помощью леммы 9.10.1 находим

$$\rho\,(A^k\,L,\,M) = \rho\,(Z\,(\Lambda^k\,Z^{-1}\,L),\,Z\,(Z^{-1}\,M) \leq \mathrm{cond}_2\,Z\,c\,q^k. \quad \square$$

Если этот анализ показался Вам совершенно естественным и простым, примите все же во внимание, что добиться этой простоты стоило немалых усилий. [2]

Задачи

1. Пусть H — верхняя хессенбергова матрица порядка n с ненулевой поддиагональю, а $T(\lambda)$ — верхняя треугольная матрица вида

$$T\,(\lambda) = \begin{bmatrix} 1 & & & \\ 0 & & & \\ \ldots & & H - \lambda\,I & \\ 0 & & & \\ 0 & 0 & \ldots & 0 \quad 1 \end{bmatrix}.$$

   Докажите, что система уравнений

$$T\,(\lambda)\,[\phi_0\,(\lambda), \ldots, \phi_n\,(\lambda)]^T = [0, \ldots, 0,\ 1]^T$$

   относительно полиномов $\phi_0\,(\lambda), \ldots, \phi_n\,(\lambda)$ имеет единственное решение и при этом $\phi_0\,(\lambda)$ лишь ненулевым коэффициентом отличается от характеристического полинома матрицы H.

2. Придумайте алгоритм, получающий коэффициенты характеристического полинома произвольной хессенберговой матрицы порядка n за $\mathcal{O}\,(n^3)$ арифметических операций.

[2] D.S.Watkins, L.Elsner, Convergence of algorithms of decomposition type for the eigenvalue problem, Linear Algebra Appl. 143: 19–47 (1991).

3. Придумайте алгоритм, получающий коэффициенты характеристического полинома произвольной $n \times n$-матрицы за $\mathcal{O}(n^3)$ арифметических операций.

4. Дана матрица A порядка n. Придумайте алгоритм, вычисляющий следы матриц $A,\ A^2,\ \ldots,\ A^n$ за $O(\log^2 n)$ параллельных шагов. Зная эти следы, коэффициенты характеристического полинома можно найти за $O(\log^2 n)$ параллельных шагов — как?

5. Докажите, что $\operatorname{dist}(L,\, M) = 1$ тогда и только тогда, когда L содержит ненулевой вектор, ортогональный M.

6. Докажите, что расстояние между любыми двумя подпространствами в $\mathbb{C}^n$ равно расстоянию между их ортогональными дополнениями.

7. Дайте полное доказательство теоремы о CS-разложении унитарной матрицы.

8. Унитарная матрица представлена в блочном виде:

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}, \quad Q_{11},\, Q_{22} \in \mathbb{C}^{m \times m}.$$

   Докажите, что $|\det(Q_{12})| = |\det(Q_{21})|$.

9. Пусть A — невырожденная матрица порядка n; L и M – подпространства в $\mathbb{C}^n$. Докажите, что

$$\operatorname{dist}(A\,L,\, A\,M) \le \operatorname{cond}_2(A)\operatorname{dist}(L,\, M).$$

10. Если подпространства L и M имеют одинаковую размерность, то для любого ортонормированного базиса в L (столбцы матрицы U) существует такой ортонормированный базис в M (столбцы матрицы V), что

$$||U - V||_2 \le \sqrt{2}\operatorname{dist}(L,\, M).$$

    Докажите.

11. Верно ли, что если $\rho(L,\, M) < 1, \quad L,\, M \in \mathbb{C}^n$, то $\rho(Z\,L,\, Z\,M) < 1$ для любой невырожденной $n \times n$-матрицы Z ?

12. Докажите, что если для ортопроекторов P и Q имеет место неравенство $||P - Q||_2 < 1$, то $\operatorname{rank} P = \operatorname{rank} Q$.

13. Столбцы матрицы Q размеров $n \times (n-1)$ образуют ортонормированную систему. Докажите, что в Q можно найти невырожденную подматрицу M порядка $n-1$, для которой имеет место неравенство

$$||M^{-1}||_2 \leq \sqrt{n}.$$

Глава 10

10.1 QR-алгоритм

Если нас интересуют матрицы общего вида, порядок которых не больше тысячи (нескольких тысяч), то для вычисления всех собственных значений (и собственных векторов) можно рекомендовать QR-алгоритм. Его придумали в начале 60-х годов В. Н. Кублановская (Россия) и Дж. Фрэнсис (Англия).

Мы начнем с обсуждения QR-алгоритма в его изначальной (ортодоксальной) форме. Прежде всего попытаемся разобраться, почему и при каких условиях QR-алгоритм сходится. Важное замечание: в современном понимании QR-алгоритм представляет собой модификацию ортодоксальной схемы с некоторой совокупностью рецептов, делающих его действительно эффективным.

QR-алгоритм (в ортодоксальной форме):

$A_0 = A$ — исходная матрица;
$A_{k-1} = Q_k\, R_k$ (QR-разложение), $A_k = R_k\, Q_k, \quad k = 1, 2, \ldots$.

10.2 Основные соотношения

Вот основные соотношения, необходимые для анализа QR-алгоритма:

$$A_k = Z_k^*\, A\, Z_k, \quad Z_k = Q_1 \ldots Q_k; \tag{10.2.1}$$

$$A^k = Z_k\, U_k, \quad U_k = R_k \ldots R_1. \tag{10.2.2}$$

Доказательство. Чтобы получить (10.2.1), достаточно заметить, что матрица A_k унитарно подобна A_{k-1}:

$$A_k \;=\; Q_k^*\,(Q_k\, R_k)\, Q_k \;=\; Q_k^*\, A_{k-1}\, Q_k.$$

Чтобы получить (10.2.2), запишем

$$A^k \;=\; (Q_1\, R_1)^k \;=\; Q_1\,(R_1\, Q_1)^{k-1}\, R_1 \;=\; Q_1\, A_1^{k-1}\, R_1.$$

Если уже доказано, что $A_1^{k-1} = (Q_2 \dots Q_k)(R_k \dots R_2)$, то находим

$$A^k = Q_1 (Q_2 \dots Q_k)(R_k \dots R_2) R_1. \quad \square$$

Выполнив k итераций QR-алгоритма, мы сводим задачу для A к той же задаче для A_k. Попробуем понять, почему решать задачу для A_k будет проще.

10.3 Сходимость QR-алгоритма

Сделаем три предположения:

(1) $A = X \Lambda X^{-1}, \quad \Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}, \ \Lambda_1 \in \mathbb{C}^{m \times m}, \ \Lambda_2 \in \mathbb{C}^{r \times r};$

(2) $|\lambda_1| \geq \dots \geq |\lambda_m| > |\lambda_{m+1}| \geq \dots \geq |\lambda_{m+r}| > 0,$
$\{\lambda_1, \dots, \lambda_m\} = \lambda(\Lambda_1), \quad \{\lambda_{m+1}, \dots, \lambda_{m+r}\} = \lambda(\Lambda_2);$

(3) ведущая подматрица порядка m в X^{-1} невырожденная.

Теорема 10.3.1 *Пусть для матрицы A выполнены предположения* (1), (2), (3) *и QR-алгоритм порождает последовательность матриц*

$$A_k = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & A_{22}^{(k)} \end{bmatrix}.$$

Тогда

$$A_{21}^{(k)} \to 0 \quad \textit{при} \quad k \to 0.$$

Более того, если $\gamma \equiv |\lambda_{m+1}|/|\lambda_m|$, то

$$\forall \ q \in (\gamma, 1) \quad \exists \ c = c(q): \quad ||A_{21}^{(k)}||_2 \leq c\, q^k, \quad k = 1, 2, \dots.$$

Прежде чем доказывать эту теорему, обсудим ее.

Сходимость QR-алгоритма трактуется в ней как сходимость к нулю поддиагонального блока матрицы A_k. Малость поддиагонального блока означает, что матрица близка к некоторой верхней блочно треугольной матрице. Если мы решили, что блок достаточно мал, то заменяем его нулем и затем ищем (например, с помощью того же QR-алгоритма) собственные значения для диагональных блоков $A_{11}^{(k)}$ и $A_{22}^{(k)}$.

Предположим, что условия (1), (2), (3) выполнены для всех $1 \leq m \leq n-1$. Тогда сходятся к нулю *все* поддиагональные блоки. Другими словами:

все поддиагональные элементы матриц A_k стремятся к нулю $\Rightarrow$ диагональные элементы матриц A_k сходятся к искомым собственным значениям матрицы A.

Пессимистическое замечание: если рассмотренные выше предположения не выполнены, то, возможно, ни один из поддиагональных блоков не сходится к нулю. Это означает, что в общем случае QR-алгоритм не обязан сходиться (приведите пример).

Оптимистическое замечание: внеся в элементы матрицы сколь угодно малые возмущения, мы всегда можем удовлетворить предположения (1), (2), (3) (почему?).

10.4 Доказательство теоремы о сходимости

Докажем теорему 10.3.1.

Доказательство. Согласно условиям теоремы, матрица A невырожденная (почему?). Из соотношения (10.2.2) получаем $Z_k = A^k\,U_k^{-1}$ и $Z_k^{-1} = U_k\,A^{-k}$. Следовательно,

$$A_k = (U_k\,A^{-k})\,A\,(A^k\,U_k^{-1}).$$

В силу предположения (3) для X^{-1} существует блочное LU-разложение:

$$X^{-1} = L\,U = \begin{bmatrix} I & 0 \\ L_{21} & I \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}.$$

Используя его, находим:

$$\begin{aligned} A^k\,U_k^{-1} &= (X\,\Lambda\,X^{-1})^k\,U_k^{-1} = X\,\Lambda^k\,(L\,U)\,U_k^{-1} \\ &= (X\,\Lambda^k\,L\,\Lambda^{-k})\,\Phi_k, \quad \text{где} \quad \Phi_k = \Lambda^k\,U\,U_k^{-1}. \\ \Rightarrow \quad U_k\,A^{-k} &= \Phi_k^{-1}\,(\Lambda^k\,L^{-1}\,\Lambda^{-k}\,X^{-1}) \quad \Rightarrow \\ A_k &= \Phi_k^{-1}\,(\Lambda^k\,L^{-1}\,\Lambda^{-k}\,X^{-1})\,(X\,\Lambda\,X^{-1})(X\,\Lambda^k\,L\,\Lambda^{-k})\,\Phi_k \\ &= \Phi_k^{-1}\,\Psi_k\,\Phi_k, \quad \text{где} \quad \Psi_k = \Lambda^k\,[L^{-1}\,\Lambda\,L]\,\Lambda^{-k}. \end{aligned}$$

Матрица Ψ_k — нижняя блочно треугольная: $\Psi_k = \begin{bmatrix} \Lambda_1 & 0 \\ \Psi_{21}^{(k)} & \Lambda_2 \end{bmatrix}$, и для ее поддиагонального блока справедлива оценка (докажите)

$$||\Psi_{21}^{(k)}||_2 \;\le\; \alpha\,||\Lambda_2^k||_2\,||\Lambda_1^{-k}||_2, \quad \alpha \;=\; ||A||_2\,||L||_2\,||L^{-1}||_2.$$

Поскольку матрица $A^k U_k^{-1}$ унитарная, ее 2-норма равна 1. Легко проверить, что

$$\begin{aligned} A^k U_k^{-1} &= X \Lambda^k X^{-1} U_k^{-1} = X \Lambda^k L U U_k^{-1} \\ &= X (\Lambda^k L \Lambda^{-k}) (\Lambda^k U U_k^{-1}) = X (\Lambda^k L \Lambda^{-k}) \Phi_k \quad \Rightarrow \\ ||\Phi_k||_2 &\le ||X^{-1}||_2 \, ||\Lambda^k L^{-1} \Lambda^{-k}||_2. \end{aligned}$$

Матрица L^{-1} — нижняя блочно треугольная: $L^{-1} = \begin{bmatrix} I & 0 \\ -L_{22} & I \end{bmatrix}$ (проверьте!); поэтому

$$\Lambda^k L^{-1} \Lambda^{-k} = \begin{bmatrix} \Lambda_1^k & 0 \\ 0 & \Lambda_2^k \end{bmatrix} \begin{bmatrix} I & 0 \\ -L_{22} & I \end{bmatrix} \begin{bmatrix} \Lambda_1^{-k} & 0 \\ 0 & \Lambda_2^{-k} \end{bmatrix} = \begin{bmatrix} I & 0 \\ -\Lambda_2^k L_{22} \Lambda_1^{-k} & I \end{bmatrix}.$$

В силу предположений (1), (2) все элементы этой матрицы по абсолютной величине не больше 1 $\Rightarrow$ нормы $||\Phi_k||_2$ ограничены равномерно по k. Аналогично, нормы $||\Phi_k^{-1}||_2 \le ||X||_2 \, ||\Lambda^k L \Lambda^{-k}||_2$ также равномерно ограничены по k. Пусть $\max\{||\Phi_k||_2, ||\Phi_k^{-1}||_2\} \le \Gamma$; тогда, учитывая верхний треугольный вид матрицы Φ_k, получаем

$$||A_{21}^{(k)}||_2 \le \left\| \Phi_k^{-1} \begin{bmatrix} 0 & 0 \\ \Psi_{21}^{(k)} & 0 \end{bmatrix} \Phi_k \right\|_2 \le \alpha \Gamma^2 \, ||\Lambda_2^k||_2 \, ||\Lambda_1^{-k}||_2.$$

Чтобы завершить доказательство, достаточно заметить, что для любого $\delta > 0$ при достаточно больших k

$$||\Lambda_2^k||_2 \le (|\lambda_{m+1}| + \delta)^k, \quad ||\Lambda_1^{-k}||_2 \le (\frac{1}{|\lambda_m|} + \delta)^k. \qquad \square$$

Отметим полезное следствие (докажите).

Теорема 10.4.1 *Пусть в условиях теоремы 10.3.1 матрица Λ диагональная. Тогда для некоторого $c > 0$*

$$||A_{21}^{(k)}||_2 \le c \left(\frac{|\lambda_{m+1}|}{|\lambda_m|} \right)^k, \quad k = 1, 2, \ldots.$$

10.5 *GR*-алгоритм

Предшественником QR-алгоритма был алгоритм такой же структуры, в котором вместо QR-разложения использовалось LU-разложение. Все алгоритмы такого рода можно представить в унифицированной форме, если исходить из того, что для любой невырожденной матрицы A каким-то способом определено GR-разложение $A = GR$, где R — верхняя треугольная

матрица, G — невырожденная матрица из какого-то фиксированного класса матриц.

GR-алгоритм:

$A_0 = A$ — исходная матрица;
$A_{k-1} = G_k\, R_k$ (GR-разложение), $A_k = R_k\, G_k, \quad k = 1, 2, \ldots .$

Основные соотношения для анализа GR-алгоритма аналогичны основным соотношениям для QR-алгоритма (различие в том, что теперь G_k^{-1} может не совпадать с G_k^*):

$$A_k = Z_k^{-1}\, A\, Z_k, \quad Z_k = G_1 \ldots G_k; \tag{10.5.3}$$

$$A^k \;=\; Z_k\, U_k, \quad U_k \;=\; R_k \ldots R_1. \tag{10.5.4}$$

Теорема 10.5.1 *Пусть*

$$A = X\, \Lambda\, X^{-1}, \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \tag{10.5.5}$$

$$|\lambda_1| \;>\; \ldots \;>\; |\lambda_n| \;>\; 0. \tag{10.5.6}$$

Предположим, что матрица X^{-1} строго регулярная (то есть все ее ведущие подматрицы невырожденные) и GR-алгоритм порождает последовательность матриц G_k с равномерно ограниченным по k числом обусловленности $\operatorname{cond}_2(G_1 \ldots G_k)$. *Тогда*

$$\{A_k\}_{ij} \;\to\; 0, \quad i > j; \tag{10.5.7}$$

$$\operatorname{diag}(A_k) \;\to\; \operatorname{diag}(\Lambda). \tag{10.5.8}$$

Доказательство легко строится по аналогии с доказательством теоремы 10.3.1.

Вполне возможно, что условие строгой регулярности матрицы X^{-1} не показалось Вам очень естестенным. Оно является важным атрибутом доказательства, использующего LU-разложение матрицы X^{-1}. Но существенно ли оно?

Конечно, сколь угодно малым возмущением можно добиться строгой регулярности матрицы X^{-1}. Поэтому можно было бы не тратить силы на выяснение существенности этого предположения. Но чтобы удовлетворить эстетическое чувство, давайте все же посмотрим, как можно модифицировать формулировку теоремы и доказательство с тем, чтобы не опираться на LU-разложение и связанное с ним требование строгой регулярности. Это можно сделать, используя *разложение Брюа*.

10.6 Разложение Брюа

Под (основным) *разложением Брюа* невырожденной матрицы понимается разложение

$$A = L_1 \,\Pi\, L_2,$$

где $\Pi = \Pi\,(A)$ — матрица перестановки, L_1 и L_2 — невырожденные нижние треугольные матрицы. Под *модифицированным разложением Брюа* понимается разложение

$$A = L\,P\,U,$$

где $P = P\,(A)$ — матрица перестановки, L — невырожденная нижняя треугольная, U — невырожденная верхняя треугольная матрица.

Теорема 10.6.1 *Модифицированное и основное разложения Брюа существуют для любой невырожденной матрицы A. Матрицы перестановок Π и P определяются однозначно и при этом*

$$\Pi\,(A) = P\,(A\,J)\,J, \quad P\,(A) = \Pi\,(A\,J)\,J, \quad \textit{где} \quad J = \begin{bmatrix} 0 & & 1 \\ & \cdots & \\ 1 & & 0 \end{bmatrix}.$$

Доказательство. Установим существование модифицированного разложения Брюа. Для этого будем умножать A слева на невырожденные нижние треугольные матрицы и справа на невырожденные верхние треугольные матрицы — в итоге хотим получить матрицу перестановки.

Рассмотрим первый (при движении слева направо) ненулевой элемент в первой строке матрицы A. Назовем его ведущим. С помощью умножения справа на верхнюю треугольную матрицу мы можем сделать ведущий элемент равным 1 и исключить все следующие за ним элементы в первой строке. Затем с помощью умножения слева на нижнюю треугольную матрицу мы можем исключить все элементы, расположенные ниже ведущего элемента в том же столбце. Далее, выбираем в качестве ведущего первый ненулевой элемент во второй строке, и так далее. В результате мы получим некоторую матрицу перестановки P.

Рассмотрим подстроки

$$r_i^{(j)}(A) \equiv [a_{i\,1}, \ldots, a_{ij}], \quad 1 \le i, j \le n,$$

и пусть

$$\sigma(i; A) = \min j \quad : \quad r_i^{(j)}(A) \notin \operatorname{span}\{r_1^{(j)}(A), \ldots, r_{i-1}^{(j)}(A)\}.$$

Заметим, что для любых невырожденных нижней и верхней треугольных матриц L и U

$$\sigma(i;A) = \sigma(i, LPU).$$

Отсюда вытекает, что

$$\sigma(i;A) = \sigma(P).$$

Таким образом, матрица перестановки P не зависит от способа построения разложения Брюа.

Остается установить связь между модифицированным и основным разложениями:

$$AJ = LPU \;\Leftrightarrow\; A = L\,(PJ)\,(JUJ),$$
$$AJ = L_1 \Pi L_2 \;\Leftrightarrow\; A = L_1\,(\Pi J)\,(JL_2J). \quad \square$$

10.7 Что будет, если матрица X^{-1} не является строго регулярной

Теорема 10.7.1 *Пусть для матрицы A выполнены предположения* (10.5.5), (10.5.6) *и GR-алгоритм порождает последовательность матриц G_k с равномерно ограниченным по k числом обусловленности* $\mathrm{cond}_2\,(G_1 \dots G_k)$. *Тогда*

$$\{A_k\}_{ij} \;\to\; 0, \quad i > j; \tag{10.7.9}$$
$$\mathrm{diag}\,(A_k) \;\to\; \mathrm{diag}\,(P^{-1}\Lambda P), \tag{10.7.10}$$

где P — матрица перестановки из модифицированного разложения Брюа $X^{-1} = LPU$.

Доказательство. Используя соотношения (10.5.3), (10.5.4) и действуя по аналогии с доказательством теоремы 10.3.1, получаем

$$A_k = (U_k U^{-1} P^{-1} \Lambda^{-k} P)\,\{P^{-1}\Lambda^k\,[L^{-1}\Lambda L]\,\Lambda^{-k} P\}\,(P^{-1}\Lambda^k P U U_k^{-1}).$$

Матрицы в круглых скобках являются взаимно обратными верхними треугольными матрицами. Вследствие равномерной по k ограниченности числа обусловленности $\mathrm{cond}_2\,(Z_k)$ эти матрицы будут равномерно ограничены по k (докажите). В силу (10.5.6) и нижнего треугольного вида матрицы $L^{-1}\Lambda L$

$$\Lambda^k\,[L^{-1}\Lambda L]\,\Lambda^{-k} \;\to\; \mathrm{diag}\,(L^{-1}\Lambda L) \;=\; \Lambda.$$

Следовательно, матрица в фигурных скобках имеет вид $P^{-1}\Lambda P + F_k$, где $F_k \;\to\; 0$. Остается учесть, что

$$\mathrm{diag}\,(A_k) = P^{-1}\Lambda P + (U_k U^{-1} P^{-1} \Lambda^{-k})\,F_k\,(P^{-1}\Lambda^k P U U_k^{-1}). \quad \square$$

Итак, если матрица X^{-1} не строго регулярная, то диагональные элементы матриц A_k (при достаточно больших k) аппроксимируют те же собственные значения матрицы A, но взятые в другом порядке (не таком, как в Λ, то есть не по убыванию модулей). Пронаблюдать это переупорядочение в условиях приближенных вычислений очень не просто (почему?).

10.8 *QR*-итерации и итерации подпространств

Запишем $Z_k = [z_1^{(k)}, \ldots, z_n^{(k)}]$ и рассмотрим подпространства

$$L_m^k \equiv \operatorname{span}\{z_1^{(k)}, \ldots, z_m^{(k)}\}.$$

Подпространство $L_m \equiv L_m^0$ — это линейная оболочка, натянутая на первые m столбцов единичной матрицы.

Предположим, что матрица A невырожденная. Тогда

$$L_m^k = A^k L_m, \quad m = 1, \ldots, n.$$

Следовательно, одна QR-итерация порождает (виртуально) n подпространств $L_1^k, \ldots, L_n^k$, которые возникли бы на k-ом шаге итераций с матрицей A и начальными подпространствами $L_1, \ldots, L_n$.

Мы уже знаем условия, при которых гарантируется сходимость итераций подпространства к инвариантному подпространству. В частности, пусть матрица A диагонализуема и ее собственные значения различны по модулю:

$$A = X \operatorname{diag}(\lambda_1, \ldots, \lambda_n) X^{-1}, \quad |\lambda_1| > \ldots > |\lambda_n|.$$

Пусть

$$\rho(X^{-1} L_m,\, X^{-1} M_m) < 1 \quad \forall\, m, \tag{10.8.11}$$

где M_m — инвариантное относительно A подпространство, натянутое на первые m столбцов матрицы X. Тогда

$$\forall\ m \quad L_m^k \to M_m \text{ при } k \to \infty.$$

Если подпространство L_m^k инвариантно относительно A, матрица $A_k = Z_k^* A Z_k$ будет иметь нулевой поддиагональный блок. Интуитивно ясно, что чем ближе подпространство к инвариантному, тем меньше должен быть этот поддиагональный блок. Вот строгая формулировка:

Лемма 10.8.1 *Пусть*

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \in \mathbb{C}^{n\times n}, \quad T_{11} \in \mathbb{C}^{m\times m},\ T_{22} \in \mathbb{C}^{r\times r}.$$

Пусть L — линейная оболочка, натянутая на первые m столбцов единичной матрицы, M — произвольное инвариантное относительно T подпространство в $\mathbb{C}^n$ размерности m. Тогда

$$||T_{21}||_2 \;\leq\; 3\,\rho\,(L,\,M)\,||T||_2.$$

Доказательство. Возьмем ортонормированный базис в M и достроим его до ортонормированного базиса в $\mathbb{C}^n$. Из полученных столбцов образуем унитарную матрицу

$$U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}, \quad M = \operatorname{im} \begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix}.$$

Поскольку M инвариантно относительно T, находим

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$$

для некоторых блоков R_{11}, R_{12}, R_{22}. Отсюда

$$T_{21} = [U_{21}\,U_{22}] \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \begin{bmatrix} U_{11}^* \\ U_{12}^* \end{bmatrix}.$$

Получаем $||T_{21}||_2 \;\leq\; 3\,||U_{21}||_2\,||T||_2$. Остается заметить, что $\rho\,(M,\,L) = ||U_{21}||_2 = ||U_{21}\,y||_2$ для некоторого y единичной длины (почему?). $\square$

Итак, мы получили еще одно доказательство того, что при достаточно общих условиях все поддиагональные элементы матриц A_k стремятся к нулю при $k \;\to\; \infty \quad \Rightarrow$ диагональные элементы матрицы A_k аппроксимируют собственные значения матрицы A.

Заметим, что условие (10.8.11) равносильно строгой регулярности матрицы X^{-1} (докажите).

Задачи

1. Приведите пример матрицы, для которой QR-алгоритм не сходится.

2. Докажите, что из соотношений

$$A_0 = A; \quad A_{k-1} = G_k\,R_k,\; A_k = R_k\,G_k, \quad k = 1, 2, \ldots,$$

вытекают следующие соотношения (основные для анализа GR-алгоритма):

$$A_k \;=\; Z_k^{-1} A Z_k,\; Z_k \;=\; G_1 \ldots G_k; \quad A^k \;=\; Z_k U_k,\; U_k \;=\; R_k \ldots R_1.$$

3. Пусть верхняя треугольная матрица $A \in \mathbb{C}^{n\times n}$ имеет попарно различные собственные значения $\lambda_1, \ldots, \lambda_n$ и последовательность матриц A_k такова, что

$$\{A_k\}_{ij} \;\to\; \{A\}_{ij} \quad \text{при} \quad i \;\geq\; j.$$

Докажите, что для каждого k собственные значения $\lambda_i(A_k)$ для A_k можно занумеровать таким образом, что

$$\lambda_i(A_k) \;\to\; \lambda_i, \quad i = 1, \ldots, n.$$

4. Докажите, что если $||L||_2 \;\leq\; ||M||_2$, то

$$\left\| \begin{bmatrix} I & 0 \\ L & I \end{bmatrix} \right\|_2 \leq \left\| \begin{bmatrix} I & 0 \\ M & I \end{bmatrix} \right\|_2.$$

5. Докажите, что условие (10.8.11) равносильно строгой регулярности матрицы X^{-1}.

Глава 11

11.1 QR-алгоритм со сдвигами

В общем случае QR-алгоритм в ортодоксальной форме сходится медленно. Медленное убывание поддиагонального $(n-m)\times m$-блока означает, что отношение $|\lambda_{m+1}|/|\lambda_m|$ недостаточно мало. Чтобы ускорить процесс, можно перейти к "сдвинутой" матрице $A - s\,I$, уповая на то, что

$$|\lambda_{m+1} - s|/|\lambda_m - s| \;\ll\; |\lambda_{m+1}|/|\lambda_m|.$$

QR-алгоритм со сдвигами:

$$\begin{aligned}
&A_0 = A \text{ — исходная матрица;}\\
&A_{k-1} - s_k\,I = Q_k\,R_k \quad (QR\text{-разложение}),\\
&A_k = R_k\,Q_k + s_k\,I, \quad k = 1,\,2,\,\ldots\,.
\end{aligned}$$

11.2 Обобщенный QR-алгоритм

QR-алгоритм со сдвигами — это частный случай более общего алгоритма, использующего полиномы f_k.

Обобщенный QR-алгоритм:

$$\begin{aligned}
&A_0 = A \text{ — исходная матрица;}\\
&f_k\,(A_{k-1}) = Q_k\,R_k \quad (QR\text{-разложение}),\\
&A_k = Q_k^{-1}\,A_{k-1}\,Q_k, \quad k = 1,\,2,\,\ldots\,.
\end{aligned}$$

Вот основные формулы для анализа обобщенного QR-алгоритма:

$$A_k = Z_k^{-1} A Z_k, \quad Z_k = Q_1 \ldots Q_k; \tag{11.2.1}$$

$$p_k\,(A) \equiv \prod_{i=1}^{k} f_i\,(A) = Z_k U_k, \quad U_k = R_k \ldots R_1. \tag{11.2.2}$$

Доказательство. Первое очевидно. Второе легко доказывается по индукции. Если уже установлено, что

$$f_2(A_1)\ldots f_k(A_1) = (Q_2\ldots Q_k)\,(R_k\ldots R_2),$$

то, поскольку $A_1 = Q_1^{-1} A Q_1$, находим

$$\begin{aligned} f_2(Q_1^{-1}AQ_1)\dots f_k(Q_1^{-1}AQ_1) &= (Q_1^{-1}f_2(A)Q_1)\dots(Q_1^{-1}f_k(A)Q_1) \\ &= Q_1^{-1}f_2(A)\dots f_k(A)Q_1 \\ \Rightarrow \quad Q_1^{-1}f_2(A)\dots f_k(A)Q_1 &= (Q_2\dots Q_k)\,(R_k\dots R_2). \end{aligned}$$

Умножив обе его части слева на Q_1 и справа на R_1 и вспомнив, что $f_1(A) = Q_1 R_1$, получаем (11.2.2). $\square$

Обобщенный QR-алгоритм иногда называют *QR-алгоритмом с мультисдвигами*. Под *мультисдвигом* степени $r = r\,(k)$ подразумевается полный набор корней полинома

$$f_k(x) = \prod_{i=1}^{r}(x - s_i^{(k)}).$$

Один мультисдвиг степени r эквивалентен последовательности r мультисдвигов степени 1 (почему?).

11.3 Лемма о QR-итерации

Пусть $G(m, n, X)$ — множество матриц $A \in \mathbb{C}^{n\times n}$, удовлетворяющих следующим требованиям:

(1) $$A = X\Lambda X^{-1}, \quad \Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix},$$

$$\Lambda_1 = \operatorname{diag}\{\lambda_1, \dots, \lambda_m\}, \quad \Lambda_2 = \operatorname{diag}\{\lambda_{m+1}, \dots, \lambda_{m+r}\},$$

(2) $$|\lambda_1| \geq \dots \geq |\lambda_m| \;>\; |\lambda_{m+1}| \geq \dots \geq |\lambda_{m+r}| \;>\; 0;$$

(3) ведущая подматрица порядка m в матрице X^{-1} невырожденная.

Лемма 11.3.1 *Предположим, что $A_0 \in G(m, n, X)$ и A_1 — результат QR-итерации вида*

$$A_0 = QR, \quad A_1 = RQ.$$

Тогда для поддиагональных блоков матриц

$$A_0 = \begin{bmatrix} A_{11}^{(0)} & A_{12}^{(0)} \\ A_{21}^{(0)} & A_{22}^{(0)} \end{bmatrix}, \quad A_1 = \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} \\ A_{21}^{(1)} & A_{22}^{(1)} \end{bmatrix}, \quad A_{11}^{(0)},\ A_{11}^{(1)} \in \mathbb{C}^{m\times m}.$$

выполняется неравенство

$$||A_{21}^{(1)}||_2 \leq c\,||A_{21}^{(0)}||_2\,\frac{|\lambda_{m+1}|}{|\lambda_m|},$$

где константа $c \;>\; 0$ зависит только от m и X.

Доказательство. Пусть

$$X^{-1} = LU \equiv \begin{bmatrix} I_m & 0 \\ L_{21} & I_r \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}, \; U_{11} \in \mathbb{C}^{m\times m}.$$

Матрицы L и U, конечно, невырожденные. Обратные к ним имеют следующее блочное строение (почему?):

$$L^{-1} = \begin{bmatrix} I_m & 0 \\ -L_{21} & I_r \end{bmatrix}, \quad U^{-1} = \begin{bmatrix} U_{11}^{-1} & -U_{11}^{-1}U_{12}U_{22}^{-1} \\ 0 & U_{22}^{-1} \end{bmatrix}.$$

Легко проверить, что

$$A_1 = \Phi^{-1}\Psi\Phi, \quad \text{где} \quad \Phi = \Lambda U R^{-1}, \quad \Psi = \Lambda(L^{-1}\Lambda L)\Lambda^{-1}.$$

Матрица $Q = X(\Lambda L\Lambda^{-1})\,(\Lambda U R^{-1})$ унитарная $\Rightarrow$

$$||\Phi||_2 \le ||X^{-1}||_2 \; ||\Lambda L^{-1}\Lambda^{-1}||_2.$$

Матрица Q^{-1} тоже унитарная $\Rightarrow$

$$||\Phi^{-1}||_2 \le ||X||_2 \; ||\Lambda L\Lambda^{-1}||_2.$$

Далее,

$$\Lambda L\Lambda^{-1} = \begin{bmatrix} I & 0 \\ \Lambda_2 L_{21}\Lambda_1^{-1} & I \end{bmatrix} \quad \Rightarrow$$

$$||\Phi||_2 \le ||X^{-1}||_2 \,(1 + ||L_{21}||_2), \qquad ||\Phi^{-1}||_2 \le ||X||_2 \,(1 + ||L_{21}||_2).$$

Рассмотрим блочное разбиение матрицы

$$\Psi = \begin{bmatrix} \Lambda_1 & 0 \\ \Lambda_2 \{L^{-1}\Lambda L\}_{21} \Lambda_1^{-1} & \Lambda_2 \end{bmatrix}.$$

Поскольку $L^{-1}\Lambda L = UAU^{-1}$, находим $\{L^{-1}\Lambda L\}_{21} = U_{22}A_{21}^{(0)}U_{11}^{-1}$, откуда

$$||\{L^{-1}\Lambda L\}_{21}||_2 \le (||U_{22}||_2 \; ||U_{11}^{(-1)}||_2) \; ||A_{21}^{(0)}||_2.$$

Окончательно,

$$||A_{21}^{(1)}||_2 \le \left\| \Phi^{-1} \begin{bmatrix} 0 & 0 \\ \Lambda_2 \{L^{-1}\Lambda L\}_{21} \Lambda_1^{-1} & 0 \end{bmatrix} \Phi \right\|_2 \le c \; ||A_{21}^{(0)}||_2 \; \frac{|\lambda_{m+1}|}{|\lambda_m|},$$

где

$$c = ||X||_2 \; ||X^{-1}||_2 \,(1 + ||L_{21}||_2)^2 \; ||U_{22}||_2 \; ||U_{11}^{-1}||_2. \quad \square$$

11.4 Квадратичная сходимость

Пусть на каждом шаге обобщенного QR-алгоритма используются полиномы f_k (мультисдвиги) одной и той же степени r.

Мультисдвиг называется *мультисдвигом Релея*, если f_k есть характеристический полином $r \times r$-блока $A_{22}^{(k)}$:

$$A_k = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & A_{22}^{(k)} \end{bmatrix}, \quad A_{11}^{(k)} \in \mathbb{C}^{m\times m}, \quad A_{22}^{(k)} \in \mathbb{C}^{r\times r}.$$

Теорема 11.4.1 *Пусть $A \in G(m,n,X)$, и предположим, что обобщенный QR-алгоритм с мультисдвигами Релея степени r сходится, то есть*

$$\varepsilon_k \equiv ||A_{21}^{(k)}||_2 \ \to \ 0.$$

Тогда сходимость квадратичная, то есть

$$\exists \ \delta, \ c \ > \ 0 : \varepsilon_k \le \delta \ \Rightarrow \ \varepsilon_{k+1} \le c\, \varepsilon_k^2.$$

Доказательство. Один шаг обобщенного QR-алгоритма эквивалентен одной QR-итерации для модифицированной матрицы:

$$f_k(A_{k-1}) = QR, \quad A_k = RQ.$$

Поэтому мы можем применить лемму о QR-итерации (лемма 11.3.1):

$$\varepsilon_k \le c\, \varepsilon_{k-1}\, \frac{\alpha_k}{\beta_k}, \quad \alpha_k = ||f_k(\Lambda_2)||_2, \quad \beta_k = ||(f_k(\Lambda_1))^{-1}||_2.$$

Пусть $s_1, \ldots, s_r$ — собственные значения блока $A_{22}^{(k-1)}$. В силу теоремы Бауэра–Файка и теорем Гершгорина, при достаточно малых ε_{k-1} имеем

$$\min_{1\le i\le r} \ |\lambda_{m+j} - s_i| \ \le \ \mathrm{cond}_2\,(X)\ \varepsilon_{k-1} \quad j = 1, \ldots, r.$$

Поэтому

$$|f_k(\lambda_{m+j})| = \left| \prod_{i=1}^{r} (\lambda_{m+j} - s_i) \right| \le c_1\ \varepsilon_{k-1}, \quad j = 1, \ \ldots, \ r;$$

$$|f_k(\lambda_j)| = \left| \prod_{i=1}^{r} (\lambda_j - s_i) \right| \ge c_2 \ > \ 0, \quad j = 1, \ \ldots, \ m,$$

где $c_1, c_2 \ > \ 0$ не зависят от k (почему?). Следовательно,

$$\alpha_k \le c_1\ \varepsilon_{k-1}, \quad \beta_k \ge c_2 \ > \ 0. \quad \square$$

Заметим, что квадратичная сходимость устанавливается в теореме при условии, что процесс сходится.

11.5 Кубическая сходимость

Теорема 11.5.1 *Пусть $A \in G(m,n,X)$, и предположим, что обобщенный QR-алгоритм с мультисдвигами Релея степени r сходится, то есть*

$$\varepsilon_k \equiv ||A_{21}^{(k)}||_2 \;\to\; 0,$$

и при этом для всех k

$$c_1\,||A_{21}^{(k)}||_2 \le ||A_{12}^{(k)}||_2 \le c_2\,||A_{21}^{(k)}||_2, \tag{11.5.3}$$

где $c_1, c_2 > 0$ не зависят от k. Тогда сходимость кубическая, то есть

$$\exists\;\delta,\; c\;>\;0 : \varepsilon_k \le \delta \;\Rightarrow\; \varepsilon_{k+1} \le c\,\varepsilon_k^3.$$

В силу теоремы 11.4.1 мы уже имеем квадратичную сходимость. Будем использовать логику и обозначения ее доказательства. Условие (11.5.3) позволяет нам утверждать, что

$$\alpha_k \le c_1 \varepsilon_{k-1}^2.$$

Это неравенство очевидно вытекает из следующего аналога теоремы Бауэра–Файка.

Теорема 11.5.2 (Элснер–Ваткинс) *Пусть μ — собственное значение для $A+F$, где*

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \quad F = \begin{bmatrix} 0 & F_{12} \\ F_{21} & 0 \end{bmatrix},$$

$$A_{11} = X_1 \Lambda_1 X_1^{-1}, \quad A_{22} = X_2 \Lambda_2 X_2^{-1},$$
$$\Lambda_1 = \operatorname{diag}(\lambda_1, \ldots, \lambda_m), \quad \Lambda_2 = \operatorname{diag}(\lambda_{m+1}, \ldots, \lambda_{m+r}).$$

Тогда

$$\min_{1\le i\le m} |\mu - \lambda_i| \;\min_{1\le j\le r} |\mu - \lambda_{m+j}| \;\le\; \operatorname{cond}_2(X_1)\,\operatorname{cond}_2(X_2)\,||F_{12}||_2\,||F_{21}||_2.$$

Доказательство. Если $\mu \in \lambda(A)$, то неравенство очевидным образом выполнено. Пусть $\mu \notin \lambda(A)$. Тогда следующие матрицы будут вырожденными:

$$\begin{bmatrix} \Lambda_1 - \mu I & X_1 F_{12} X_2^{-1} \\ X_2 F_{21} X_1^{-1} & \Lambda_2 - \mu I \end{bmatrix},$$

$$\begin{bmatrix} \Lambda_1 - \mu I & X_1 F_{12} X_2^{-1} \\ 0 & (\Lambda_2 - \mu I) - X_2 F_{21} X_1^{-1} (\Lambda_1 - \mu I)^{-1} X_1 F_{12} X_2^{-1} \end{bmatrix},$$

$$(I - (\Lambda_2 - \mu I)^{-1} X_2 F_{21} X_1^{-1} (\Lambda_1 - \mu I)^{-1} X_1 F_{12} X_2^{-1}).$$

В силу вырожденности последней матрицы

$$||(\Lambda_2 - \mu I)^{-1} X_2 F_{21} X_1^{-1} (\Lambda_1 - \mu I)^{-1} X_1 F_{12} X_2^{-1}||_2 \geq 1. \quad \square$$

Следствие 11.5.1 *Если в условиях теоремы 11.5.1 матрица A эрмитова, то для обобщенного QR-алгоритма с мультисдвигом Релея имеет место кубическая сходимость.*

Для доказательства достаточно заметить, что эрмитовость матрицы A обеспечивает (11.5.3).

11.6 Что делает QR-алгоритм эффективным

Сдвиги или мультисдвиги — это непременный атрибут эффективного QR-алгоритма. Но этого мало.

Дело в том, что одна QR-итерация для матрицы общего вида требует $\mathcal{O}(n^3)$ арифметических операций. Это очень большие затраты, даже если итераций не очень много (обычно их требуется не больше 5 на каждое собственное значений). К счастью, есть чрезвычайно простое средство сделать итерацию дешевой. Вспомним, что с помощью отражений или вращений матрицу A можно привести к унитарно подобной верхней хессенберговой матрице H. После этого QR-алгоритм будет применяться к хессенберговой матрице $A_0 = H$. Само приведение к хессенберговой форме требует, конечно, $\mathcal{O}(n^3)$ операций. Но после этого любая QR-итерация будет выполняться за $\mathcal{O}(n^2)$ операций!

Сокращение затрат на одну итерацию объясняется *инвариантностью хессенберговой формы* по отношению к QR-итерациям. Другими словами: если H — верхняя хессенбергова матрица, то QR-итерацию

$$f(H) = QR, \quad H_1 = Q^{-1} H Q \tag{11.6.4}$$

можно выполнить таким образом, что матрица H_1 сохранит верхнюю хессенбергову форму.

Достаточно убедиться в этом в случае мультисдвига степени 1 (почему?). В этом случае QR-разложение может строиться с помощью вращений, исключающих элементы на нижней поддиагонали: $Q^* = G_{n\,n-1} \dots G_{21}$. То, что матрица $H_1 = RQ = RG_{21}^* \dots G_{n\,n-1}^*$ будет верхней хессенберговой, проверяется непосредственно (проверьте).

В случае эрмитовой матрицы A матрица H будет трехдиагональной. Следовательно, H_1 тоже трехдиагональная (почему?). Чтобы реализовать одну QR-итерацию, теперь достаточно всего лишь $\mathcal{O}(n)$ операций!

11.7 Неявные QR-итерации

Если известны корни полинома f_k, то мультисдвиг можно выполнять как последовательность мультисдвигов степени 1. Это не всегда желательно.

Например, если матрица вещественная, то корни f_k могут оказаться комплексными, то есть придется переходить к комплексной арифметике. В то же время, в случае вещественной матрицы коэффициенты полинома f_k для мультисдвига Релея будут вещественными (докажите).

Поэтому часто предпочитают неявную реализацию QR-итерации, не сводящую ее к мультисдвигам степени 1. В данном способе действий нам нужны именно коэффициенты полинома f_k. Другими словами: *нужны коэффициенты характеристического полинома блока $A_{22}^{(k)}$, а не его корни.*

В основе неявной реализации QR-итерации лежит следующее наблюдение.

Лемма 11.7.1 *Пусть A — хессенбергова матрица, и предположим, что две хессенберговы матрицы B и C имеют ненулевую поддиагональ и таковы, что*

$$B = P^*AP, \quad C = Q^*AQ$$

для некоторых унитарных матриц P и Q, имеющих коллинеарные первые столбцы. Тогда существует унитарная диагональная матрица D такая, что

$$P = QD, \quad B = D^*CD.$$

Доказательство. Пусть уже доказано, что $p_i = q_i d_i,\ |d_i| = 1,\ 1 \le i \le k.$ $\Rightarrow \quad b_{ij} = d_i^* c_{ij} d_j,\ 1 \le i, j \le k. \quad \Rightarrow$

$$\begin{aligned} p_{k+1} b_{k+1\,k} &= Ap_k - \sum_{i=1}^{k} p_i b_{ik} = Aq_k d_k - \sum_{i=1}^{k} q_i d_i (d_i^* c_{ik} d_k) \\ &= \left(Aq_k - \sum_{i=1}^{k} q_i (d_i d_i^*) c_{ik} \right) d_k = q_{k+1} c_{k+1\,k} d_k. \quad \square \end{aligned}$$

На каждом шаге QR-алгоритма осуществляется переход от некоторой хессенберговой матрицы H к унитарно подобной хессенберговой матрице $H_1 = Q^*HQ$. В условиях машинной арифметики всегда можно считать, что

H_1 имеет ненулевую поддиагональ. Поэтому как бы ни строилась матрица H_1, достаточно проследить за тем, чтобы первый столбец в соответствующей матрице Q был таким же, как в обычной (явной) реализации QR-шага.

Неявная QR-итерация:

(1) Находим первый столбец h матрицы $f_k(H)$.

(2) Находим матрицу отражений V_0 такую, что $V_0^* h = [*, 0, \ldots, 0]^T$.

(3) Находим матрицу $W_0 = V_0^* H V_0$.

(4) С помощью отражений приводим W_0 к унитарно подобной матрице $W_1 = V_1^* W_0 V_1$ (V_1 — произведение $n-1$ матриц отражения).

(5) Полагаем $H_1 = W_1$.

Легко видеть, что $H_1 = (V_0 V_1)^* H (V_0 V_1)$. При этом первый столбец в матрице $V_0 V_1$ такой же, как в матрице V_0 (почему?). В то же время первый столбец матрицы V_0 совпадает с первым столбцом матрицы Q в некотором QR-разложении матрицы $f_k(H)$ (почему?).

11.8 Организация вычислений

"Умная" программа, реализующая QR-алгоритм, обычно начинает с *уравновешивания* матрицы A: ищется диагональная матрица D, выравнивающая (насколько это возможно) длины строк и столбцов в матрице $A_D = D^{-1} A D$. Уравновешивание может улучшить обусловленность собственных значений (иногда – на несколько порядков).

Затем матрица A_D приводится к хессенберговой форме A_0. После этого:

(1) сканируется поддиагональ в A_0 и каждый достаточно малый (в соответствии с некоторым критерием малости) элемент заменяется нулем;

(2) выбирается непустой диагональный блок в A_0, расположенный между двумя последними нулями на поддиагонали; если непустого блока не нашлось, то процесс заканчивается (матрица приведена к треугольному виду);

(3) для выбранного блока проводится одна QR-итерация (со сдвигом), после чего все повторяется, начиная с п. (1).

Заметим, что описанная организация вычислений может не дать высокой производительности на параллельном компьютере. С точки зрения

параллельных вычислений кажутся привлекательными мультисдвиги. Однако, первые эксперименты с неявной реализацией мультисдвигов показали неудовлетворительные результаты (из-за ошибок округления значительно замедлялась сходимость). В 1994 Ваткинс [1] "реабилитировал" мультисдвиги: он показал, что неявный мультисдвиг нужно реализовывать в виде конвейера неявных мультисдвигов меньшей степени. Такая реализация оказывается численно устойчивой и не замедляет сходимость.

11.9 Как найти сингулярное разложение

Прежде всего, умножая $A \in \mathbb{C}^{n\times n}$ слева и справа на унитарные матрицы, получаем верхнюю бидиагональную матрицу B. Не ограничивая общности, можно считать, что B — вещественная матрица (почему?).

Если B имеет нуль на диагонали, то, умножая ее слева и справа на унитарные матрицы, можно получить блочную матрицу вида $\begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}$ с бидиагональными блоками B_1 и B_2. В самом деле, пусть $b_{kk} = 0$. Элемент $b_{k\,k+1}$ аннулируем с помощью левостороннего вращения k-й и $k+1$-й строк. Ненулевой элемент, который может возникнуть в позиции $(k,\, k+2)$, аннулируем с помощью левостороннего вращения k-й и $k+2$-й строк. Ненулевой элемент, возникший в позиции $(k,\, k+3)$, аннулируем с помощью левостороннего вращения k-й и $k+3$-й строк, и так далее. В итоге мы получим блок B_2. Аналогичным образом с помощью правосторонних вращений можно получить блок B_1.

Итак, не ограничивая общности, можно считать, что B — невырожденная матрица с ненулевыми диагоналями. Теперь можно смело применять QR-алгоритм со сдвигами к вещественной симметричной трехдиагональной матрице $T \equiv B^TB$. Как только мы найдем разложение $B^TB = V\Sigma^2V^T$ с ортогональной V и диагональной $\Sigma > 0$, полагаем $U = BV\Sigma^{-1}$. Матрица U будет унитарной (почему?) и, очевидно, $B = U\Sigma V^T$.

В 1965 году Дж. Голуб и В. Кахан заметили, что неявная форма QR-итераций позволяет не формировать явно трехдиагональные матрицы. Рассмотрим одну QR-итерацию с трехдиагональными матрицами: $T - sI = QR, T_1 = Q^TTQ$. Запишем разложение Холецкого для T_1 в виде $T_1 = B_1^TB_1$, где B_1 — верхняя бидиагональная матрица, и попробуем вычислить B_1, не формируя T и T_1. Это можно сделать так:

- Находим первый столбец h матрицы $T - sI = B^TB - sI$.

[1] D.S.Watkins, Shifting strategies for the parallel QR-algorithm, SIAM J. Sci. Comput. 15 (4): 953–958 (1994).

- Строим матрицу вращения G такую, что $Gh = [*, 0, \ldots, 0]^T$.
- Формируем матрицу $W_0 = BG^T$ и с помощью последовательного умножения ее слева и справа на матрицы вращения получаем верхнюю бидиагональную матрицу $W_1 = ZW_0V$ (Z, V — произведения матриц вращения).
- Полагаем $B_1 = W_1$.

Ясно, что $B_1^T B_1 = (GV)^T (B^T B) (GV)$ и при некоторой реализации QR-разложения для $T - sI$ первые столбцы ортогональных матриц GV и Q совпадают.

Задачи

1. Приведите пример матрицы, для которой QR-алгоритм со сдвигом Релея (мультисдвигом степени 1) не сходится.
2. Пусть B – верхняя двухдиагональная матрица. Рассмотрим следующий процесс:

   $$B_0 = B; \quad C_k = B_{k-1}Q_k,\ B_k = Z_k C_k, \quad k = 1, 2, \ldots,$$

   где матрицы B_k — верхние двухдиагональные, матрицы C_k — нижние двухдиагональные, матрицы Q_k, Z_k — унитарные матрицы. Докажите, что матрицы B_k и C_k сходятся при $k \to \infty$ к одной и той же диагональной матрице.
3. Придумайте какой-либо алгоритм, строящий верхние и нижние двухдиагональные матрицы предыдущей задачи. Покажите, как с его помощью можно получить сингулярное разложение двухдиагональной матрицы.
4. Опишите алгоритм, вычисляющий сингулярные числа произвольной матрицы A.
5. Пусть $A = A^T \in \mathbb{C}^{n\times n}$. Докажите, что существует разложение

   $$A = V\Sigma V^T,$$

   где V — унитарная матрица, Σ — диагональная матрица с неотрицательной диагональю. Придумайте алгоритм, вычисляющий такое разложение.

Глава 12

12.1 Приближение функций

Пусть задан класс функций F и подкласс "простых" функций $\Phi \subset F$. Типичная задача: для $f \in F$ найти приближение $\phi \approx f, \quad \phi \in \Phi$.

Чтобы от интуитивного описания перейти к точной постановке, необходимо специфицировать классы F и Φ и придать строгий смысл понятию "приблизить". Будем считать, что F — линейное пространство функций с областью определения Ω. Можно выделить два общих подхода.

Минимизационный подход. Выбирается какая-либо норма (или полунорма [1]) $||\cdot||$ на F и ищется функция $\phi \in \Phi \subset F$, минимизирующая $\|f - \phi\|$.

Интерполяционный подход. Выбираются узлы $x_0, \ldots, x_n \in \Omega$ и ищется функция $\phi \in \Phi$, удовлетворяющая интерполяционным условиям

$$\phi(x_i) \;=\; f(x_i), \quad i = 0, 1, \ldots, n.$$

12.2 Полиномиальная интерполяция

Будем рассматривать функции $f(x)$ одной вещественной переменной x. Если $f(x)$ приближается полиномом

$$\phi(x) = L_n(x) \equiv a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0,$$

то интерполяционные условия принимают вид

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \ldots \\ a_n \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \ldots \\ f(x_n) \end{bmatrix}. \qquad (*)$$

[1] Полунорма отличается от нормы только тем, что элемент с нулевой полунормой не обязан быть нулевым

Матрица коэффициентов $W = W(x_0, \ldots, x_n)$ этой системы называется (транспонированной) *матрицей Вандермонда*. Известно, что (докажите)

$$\det W(x_0, \ldots, x_n) = \prod_{0 \le i < j \le n} (x_j - x_i).$$

Если узлы $x_0, \ldots, x_n$ попарно различны, то матрица коэффициентов для $(*)$ невырожденная. Таким образом, *интерполяционный полином существует и определен однозначно.*

12.3 Плохая обусловленность матрицы Вандермонда

Чтобы найти $L_n(x)$, казалось бы, можно использовать стандартные методы для решения системы $(*)$. Но так никто не делает, потому что: (1) стандартные методы не учитывают специфики матрицы Вандермонда; (2) в случае вещественных узлов матрица Вандермонда очень плохо обусловлена.

Теорема 12.3.1 [2] *Для любых попарно различных вещественных узлов* $x_0, x_1, \ldots, x_n$

$$\mathrm{cond}_2 W(x_0, x_1, \ldots, x_n) \ge 2^{n-1}/\sqrt{n+1}.$$

Доказательство. Пусть $h = [h_0, ..., h_n]^\top$ — вектор, полученный вычитанием из последнего столбца W линейной комбинации предыдущих столбцов:

$$h_i = x_i^n - (a_0 + a_1 x_i + ... + a_{n-1} x_i^{n-1}), \quad 0 \le i \le n.$$

Заметим, что $h_0, ..., h_n$ — значения полинома n-й степени со старшим коэффициентом 1 в некоторых точках $x \in [-1, 1]$. Можно доказать, что при любом n функция

$$T_n(x) = \cos(n \arccos x), \quad -1 \le x \le 1,$$

является полиномом n-й степени со старшим коэффициентом 2^{n-1} — это знаменитые *полиномы Чебышева.* $\Rightarrow$ При $a < b$ полином

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

имеет старший коэффициент 1 и, кроме того,

$$|P_n(x)| \le 2^{1-2n}(b-a)^n, \quad a \le x \le b.$$

[2] E. E. Tyrtyshnikov, How bad are Hankel matrices? – *Numer. Math.* 67: 261–269 (1994).

Пусть $b = \max\limits_{0\le i\le n} |x_i|$ и $a = -b$. Определив $a_0, ..., a_{n-1}$ равенством

$$P_n(x) = x^n - (a_0 + a_1 x + ... + a_{n-1}x^{n-1})$$

и учитывая, что расстояние от W до множества вырожденных матриц меньше или равно $||h||_2$, находим

$$|h_i| \le 2^{1-n} b^n \,\forall\, i \;\Rightarrow\; ||h||_2 \le \sqrt{n+1}\, 2^{1-n}\, b^n \;\Rightarrow\; \sigma_{n+1}(W) \le \sqrt{n+1}\, 2^{1-n}\, b^n;$$

$$\sigma_1(W) \ge b^n \;\Rightarrow\; \operatorname{cond}_2 W \ge \frac{b^n}{\sqrt{n+1}\, 2^{1-n}\, b^n} = 2^{n-1}/\sqrt{n+1}. \quad \square$$

Используя ту же технику, нетрудно показать, что если $|x_i| \le 1\ \forall\, i$, то $\operatorname{cond}_2 W(x_0, ..., x_n) \ge 2^{n-1}$. То же верно, если $|x_i| \ge 1\ \forall\, i$.

12.4 Интерполяционный полином Лагранжа

К счастью, не обязательно искать именно коэффициенты интерполяционного полинома. В действительности нужно иметь какое-либо "компактное" представление этого полинома, позволяющее вычислить его значение в любой заданной точке.

Если полиномы $l_0(x), \ldots, l_n(x)$ степени n удовлетворяют (интерполяционным) условиям

$$l_j(x_i) = \begin{cases} 1, i = j, \\ 0, i \ne j, \end{cases}$$

то, очевидно,

$$L_n(x) \equiv \sum_{j=0}^{n} f(x_j) l_j(x). \tag{12.4.1}$$

Полиномы $l_j(x)$ существуют и единственны (как решения задач полиномиальной интерполяции). Их называют *элементарными полиномами Лагранжа*. Полином L_n называют *интерполяционным полиномом Лагранжа*, а о самой задаче полиномиальной интерполяции часто говорят как о *лагранжевой интерполяции*.

Легко проверить, что

$$l_j(x) = \prod_{\substack{k=0 \\ k\ne j}}^{n} \frac{x - x_k}{x_j - x_k}. \tag{12.4.2}$$

Учитывая это, получаем полезное представление (проверьте!)

$$L_n(x) = \sum_{j=0}^{n} \frac{f(x_j)\,\omega(x)}{(x - x_j)\,\omega'(x_j)}, \qquad \omega(x) = \prod_{k=0}^{n}(x - x_k) \tag{12.4.3}$$

($\omega'(x)$ — производная от $\omega\,(x)$).

12.5 Погрешность лагранжевой интерполяции

Теорема 12.5.1 *Пусть* $x, x_0, \ldots, x_n \in [a,\, b]$ *и* $f \in C^{n+1}[a,\, b]$. *Тогда*

$$f(x) - L_n(x) = \frac{f^{(n+1)}\left(\xi\left(x\right)\right)}{(n+1)!}\,\omega\left(x\right), \quad \omega\left(x\right) = \prod_{k=0}^{n}(x - x_k),$$

где

$$\min\{x, x_0, \ldots, x_n\} \;<\; \xi\left(x\right) \;<\; \max\{x, x_0, \ldots, x_n\}.$$

Доказательство. Фиксируем $x \notin \{x_0, \ldots, x_n\}$. Тогда $\omega(x) \neq 0$ и мы имеем право рассмотреть такую функцию от t:

$$g(t) \equiv f(t) - L_n(t) - c\,\omega\left(t\right), \quad c \equiv \frac{f(x) - L_n(x)}{\omega\left(x\right)}.$$

Функция $g(t)$ обращается в нуль при $t = x,\, x_0,\, \ldots\,,\, x_n$. По теореме Ролля, $g^{(1)}(t)$ имеет $n+1$ нуль $\;\Rightarrow\;$ $g^{(2)}(t)$ имеет n нулей $\;\Rightarrow\;$ $\ldots\;\Rightarrow\;$ $g^{n+1}(t)$ обращается в нуль хотя бы в одной точке: $\exists\;\; \xi : g^{(n+1)}(\xi) = 0$. Остается заметить, что

$$g^{(n+1)}\left(t\right) = f^{(n+1)}\left(t\right) - c\,(n+1)!. \quad \square$$

12.6 Разделенные разности

Значения функции $f(x)$ в узлах будем называть ее *разделенными разностями порядка* 0. Для любой пары узлов $x_0,\; x_1$ величины

$$f\left(x_0;\, x_1\right) \equiv \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

будем называть *разделенными разностями порядка* 1. По индукции, величины

$$f\left(x_0;\, \ldots\,;\, x_k\right) \equiv \frac{f\left(x_1;\, \ldots\,;\, x_k\right) - f\left(x_0;\, \ldots\,;\, x_{k-1}\right)}{x_k - x_0}$$

будем называть *разделенными разностями порядка* k.

Лемма 12.6.1

$$f\left(x_0;\, \ldots\,;\, x_k\right) \;=\; \sum_{j=0}^{k} \frac{f(x_j)}{\prod\limits_{\substack{l=0 \\ l \neq j}} (x_j - x_l)}.$$

Доказательство. По индукции,

$$
\begin{aligned}
f\left(x_0;\,\ldots\,;x_k\right) &= \sum_{j=1}^{k}\frac{f(x_j)}{\prod\limits_{\substack{l=1\\ l\neq j}}^{k}(x_j-x_l)\,(x_k-x_0)} - \sum_{j=0}^{k-1}\frac{f(x_j)}{\prod\limits_{\substack{l=0\\ l\neq j}}^{k-1}(x_j-x_l)\,(x_k-x_0)} \\
&= \frac{f(x_0)}{\prod\limits_{\substack{l=0\\ l\neq 0}}^{k}(x_0-x_l)} + \frac{f(x_k)}{\prod\limits_{\substack{l=0\\ l\neq k}}^{k}(x_k-x_l)} \\
&+ \sum_{j=1}^{k-1}\frac{f(x_j)}{\prod\limits_{\substack{l=1\\ l\neq j}}^{k-1}(x_j-x_l)\,(x_k-x_0)}\left\{\frac{1}{x_j-x_k}-\frac{1}{x_j-x_0}\right\}.\square
\end{aligned}
$$

Следствие 12.6.1 *Значение разделенной разности $f(x_0;\,\ldots\,;x_k)$ не зависит от упорядочения узлов.*

Следствие 12.6.2 $f(x)-L_n(x) \;=\; f(x;\,x_0;\,\ldots\,;x_n)\,\omega\,(x)$.

Доказательство.

$$
f(x)-L_n(x)=\omega\,(x)\left\{\frac{f(x)}{\omega\,(x)} \;+\; \sum_{j=0}^{n}\frac{f(x_j)}{(x_j-x)\,\omega'\,(x_j)}\right\}. \quad \square
$$

Следствие 12.6.3

$$
\begin{gathered}
f(x_0;\,\ldots\,;x_k) \;=\; \frac{f^{(k)}(\xi)}{k!}, \\
\min\{x_0,\,\ldots\,,x_k\} \;<\; \xi \;<\; \max\{x_0,\,\ldots\,,x_k\}.
\end{gathered}
$$

12.7 Формула Ньютона

Теорема 12.7.1 (Формула Ньютона)

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

Доказательство. Запишем

$$
L_n \;=\; L_0 \;+\; (L_1-L_0) \;+\; (L_2-L_1) \;+\; \ldots \;+\; (L_n-L_{n-1}),
$$

где L_k — полином Лагранжа, интерполирующий функцию $f(x)$ в узлах $x_0, \ldots, x_k$. Поскольку L_{k-1} интерполирует L_k в узлах $x_0, \ldots, x_{k-1}$, находим

$$L_k(x) - L_{k-1}(x) = L_k(x;\, x_0;\, \ldots\, ; x_{k-1})\,(x - x_0)\ldots(x - x_{k-1}).$$

Если $L_k(x) = a_k\, x^k + \ldots$, то в силу следствия 12.6.3 величина

$$L_k(x;\, x_0;\, \ldots\, ; x_{k-1}) = a_k$$

не зависит от x. Поэтому мы можем взять $x = x_k$; тогда

$$L_k(x_k;\, x_0;\, \ldots\, ; x_{k-1}) = f(x_0;\, \ldots\, ; x_k). \quad \square$$

Формулу Ньютона можно считать дискретным аналогом ряда Тейлора. При этом она не содержит производных и, как мы скоро увидим, может быть много лучше по точности приближения.

12.8 Разделенные разности с кратными узлами

Теперь допустим, что среди чисел $x_0, \ldots, x_n$ могут быть одинаковые (кратные). Если $y \in \{x_0, \ldots, x_n\}$ встречается ровно m раз, то y называется *узлом кратности* m. Совокупность узлов $M = \{x_0, \ldots, x_n\}$ называется *сеткой — кратной*, если узлы кратные, и *простой*, если узлы попарно различны.

Для любой сетки $M = \{x_0, \ldots, x_n\}$ существует семейство простых сеток $M^\varepsilon = \{x_0^\varepsilon, \ldots, x_n^\varepsilon\}$, $\varepsilon > 0$, такое, что $x_i^\varepsilon \to x_i$ при $\varepsilon \to 0$ для всех i. Под разделенной разностью с кратными узлами будем понимать предел

$$f(x_0;\, \ldots\, ;\, x_n) \equiv \lim_{\varepsilon \to 0} f(x_0^\varepsilon;\, \ldots\, ;\, x_n^\varepsilon). \qquad (*)$$

Лемма 12.8.1 *Если кратность каждого узла не превосходит m и $f \in C^{m-1}$, то предел $(*)$ существует и не зависит от выбора семейства M^ε.*

Доказательство. Разделенные разности для простых сеток не зависят от упорядочения узлов. Поэтому будем считать, что узлы в простых сетках упорядочены таким образом, что если два узла имеют общий предельный узел, то и все простые узлы между ними имеют тот же предельный узел. Разделенные разности $f(x_i)$ порядка 0, очевидно, существуют. Далее,

$$f(x_i^\varepsilon;\, x_{i+1}^\varepsilon) \;=\; \begin{cases} \frac{f(x_{i+1}^\varepsilon) - f(x_i^\varepsilon)}{x_{i+1}^\varepsilon - x_i^\varepsilon}, & x_i \neq x_{i+1}, \\ \\ f^{(1)}(\xi^\varepsilon), & x_i = x_{i+1}. \end{cases}$$

В случае $x_i \neq x_{i+1}$ при всех достаточно малых ε $x_i^\varepsilon \neq x_{i+1}^\varepsilon$; при $\varepsilon \to 0$ существуют пределы числителя и знаменателя. В случае $x_i = x_{i+1}$ предел существует вследствие того, что ξ^ε находится между x_i^ε и x_{i+1}^ε.

Согласно рекурсивному определению разделенных разностей для простых сеток,

$$f(x_i^\varepsilon; \ldots; x_{i+k}^\varepsilon) = \begin{cases} \frac{f(x_{i+1}^\varepsilon; \ldots; x_{i+k}^\varepsilon) - f(x_i^\varepsilon; \ldots; x_{i+k-1}^\varepsilon)}{x_{i+k}^\varepsilon - x_i^\varepsilon}, & x_i \neq x_{i+k}, \\ f^{(k)}(\xi^\varepsilon), & x_i = x_{i+k}. \end{cases}$$

Пусть уже установлено существование пределов для разделенных разностей порядка $k-1$. Тогда при $x_i \neq x_{i+k}$ существуют пределы числителя и знаменателя. При $x_i = x_{i+k}$ предел существует, поскольку

$$\min\{x_i^\varepsilon, \ldots, x_{i+k}^\varepsilon\} < \xi^\varepsilon < \max\{x_i^\varepsilon, \ldots, x_{i+k}^\varepsilon\}$$

и, следовательно, $\xi^\varepsilon \to x_i = x_{i+1}$ при $\varepsilon \to 0$. $\square$

Следствие 12.8.1 *Если $f \in C^k$, то*

$$f(x_0; \ldots; x_k) = \frac{f^{(k)}(\xi)}{k!}, \tag{12.8.4}$$

где

$$\min\{x_0, \ldots, x_k\} \leq \xi \leq \max\{x_0, \ldots, x_k\}. \tag{12.8.5}$$

Доказательство. В случае простых узлов это вытекает из следствия 12.6.3. В случае кратных узлов $f(x_0; \ldots; x_k)$ есть предел величин

$$f(x_0^\varepsilon; \ldots; x_k^\varepsilon) = \frac{f^{(k)}(\xi^\varepsilon)}{k!}.$$

Пусть $\varepsilon = 1/N$. Вообще говоря, правая часть не обязана иметь предел $N \to \infty$. Однако, заведомо существует (почему?) подпоследовательность точек ξ^ε, сходящаяся к некоторой точке ξ, удовлетворяющей неравенствам (12.8.5). $\square$

12.9 Обобщенные интерполяционные условия

Если узел $z \in \{x_0, \ldots, x_n\}$ имеет кратность m, то под *обобщенными интерполяционными условиями* для определения *обобщенного интерполяционного полинома* $H_n(x)$ степени не выше n понимаются условия

$$H_n^{(j)}(z) = f^{(j)}(z), \quad j = 0, \ldots, m-1.$$

Если имеется ν попарно различных узлов $z_1, \ldots, z_\nu$ с кратностями $m_1, \ldots, m_\nu$, то

$$n = m_1 \ + \ \ldots \ + \ m_\nu \ - \ 1.$$

Нетрудно установить единственность полинома H_n (сделайте это!). Имеет место также следующее обобщение формулы Ньютона:

Теорема 12.9.1

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

Доказательство. Рассмотрим простые сетки M^ε и отвечающие им интерполяционные полиномы Лагранжа $L_n^\varepsilon(x)$. Поскольку $L_n^\varepsilon(x) \to H_n(x)$ при $\varepsilon \to 0$, мы получим один и тот же полином $H_n(x)$ при любом упорядочении узлов. Пусть $z = x_0 = \ldots = x_{m-1}$ — узел кратности m. Тогда

$$H_n(x) = \sum_{j=0}^{m-1} \frac{f^{(j)}(z)}{j!}\,(x - z)^j \ + \ (x - z)^m\, p_{n-m}\,(x),$$

где степень полинома $p_{n-m}\,(x)$ не выше $n - m$. Обобщенные интерполяционные условия в узле z выполняются очевидным образом. $\square$

Следствие 12.9.1 *Если $f \in C^{n+1}$, то*

$$f(x) - H_n(x) = \frac{f^{n+1}(\xi\,(x))}{(n+1)!}\,\omega\,(x), \quad \omega\,(x) = \prod_{k=0}^{n}(x - x_k),$$

$$\min\{x, x_0, \ldots, x_n\} \ \le \ \xi\,(x) \ \le \ \max\{x, x_0, \ldots, x_n\}.$$

Обобщенный интерполяционный полином называют *полиномом Эрмита*, а об обобщенной полиномиальной интерполяции часто говорят как об *эрмитовой интерполяции*.

12.10 Таблица разделенных разностей

При вычислении разделенных разностей в случае простых или кратных узлов явно или неявно строят следующую таблицу разделенных разностей:

$$
\begin{array}{lllllll}
f(x_0) & & & & & & \\
 & \ddots & & & & & \\
f(x_1) & \dots & f(x_0;\,x_1) & & & & \\
 & \ddots & & \ddots & & & \\
f(x_2) & \dots & f(x_1;\,x_2) & \dots & f(x_0;\,x_1;\,x_2) & & \\
 & \ddots & & \ddots & & \ddots & \\
f(x_3) & \dots & f(x_2;\,x_3) & \dots & f(x_1;\,x_2;\,x_3) & \dots & f(x_0;\,x_1;\,x_2;\,x_3) \\
 & & & & & & \\
\dots & \dots & \dots & & & &
\end{array}
$$

При построении таблицы в случае кратных узлов следует нумеровать узлы так, чтобы равенство $x_i = x_j$ влекло за собой $x_k = x_i$ при всех $i \le k \le j$.

Диагональ таблицы разделенных разностей содержит коэффициенты дискретного аналога ряда Тейлора. Используя их, можно с легкостью решать задачи как лагранжевой, так и эрмитовой интерполяции.

12.11 Остаточный член многомерной интерполяции

Пусть $f(x) = f(x_1, \dots, x_m)$ – функция от m координат вектора x и $p(x) = p(x_1, \dots, x_m)$ – интерполяционный полином для сетки

$$\mathcal{M} = \{x_1^{i_1}\}_{i_1=1}^{p_1} \times \dots \times \{x_m^{i_m}\}_{i_m=1}^{p_m}.$$

Положим

$$D_{i_1,\dots,i_k}(x) \equiv \frac{1}{p_{i_1}! \dots p_{i_k}!} \frac{\partial^{p_{i_1}} \dots \partial^{p_{i_k}}}{(\partial x_{i_1})^{p_{i_1}} \dots (\partial x_{i_k})^{p_{i_k}}} f(x),$$

$$\Omega_{i_1,\dots,i_k} \equiv \omega_{i_1}(x_{i_1}) \dots \omega_{i_k}(x_{i_k}), \quad \omega_i(t) = \prod_{j=1}^{p_i} (t - x_i^j).$$

Теорема 12.11.1 *Если f имеет непрерывные частные производные порядка $p_1 + \dots + p_m$, то*

$$f(x) - p(x) = \sum_{k=1}^{m} (-1)^{(k-1)} \sum_{1 \le i_1 < \dots < i_k \le m} E_{i_1,\dots,i_k},$$

$$E_{i_1,\dots,i_k} = D_{i_1,\dots,i_k}(\xi_{i_1,\dots,i_k})\, \Omega_{i_1,\dots,i_k}.$$

Точки $\xi_{i_1,\dots,i_k}$ принадлежат выпуклой оболочке, натянутой на $\mathcal{M}$ и x.

Доказательство. При $m = 1$ мы получаем классическое представление остаточного члена одномерной лагранжевой интерполяции. Далее по индукции. Чтобы избежать сложной индексации, положим

$$z = (x_1, \ldots, x_{m-1}), \quad i = (i_1, \ldots, i_{m-1}), \quad z_i = (x_1^{i_1}, \ldots, x_{m-1}^{i_{m-1}});$$

$$y = x_m, \quad j = i_m, \quad y_j = x_m^j;$$

$$l_j(y) = \prod_{k=1,\, k\neq j}^{p_m} \frac{y - y_k}{y_j - y_k}, \quad \omega(y) = \prod_{j=1}^{p_m} (y - y_j);$$

$$L_i(z) = \prod_{t=1}^{m-1} \prod_{\substack{k_t=1 \\ k_t \neq i_t}}^{p_t} \frac{x_t - x_t^{k_t}}{x_t^{i_t} - x_t^{k_t}}.$$

Тогда при некотором значении η

$$f(z, y) - p(z, y) = R + \frac{1}{p_m!}\, D_m(z, \eta)\, \omega(y),$$

где

$$R = \sum_{j=1}^{p_m} (f(z, y_j) - p(z, y_j))\; l_j(y).$$

Согласно индуктивному предположению,

$$R = \sum_i \left(\sum_{j=1}^{p_m} (f(z_i, y_j) - p(z_i, y_j) l_j(y) \right)\; L_i(z) + E,$$

$$E = \sum_{k=1}^{m-1} (-1)^{k-1} \sum_{1 \le i_1 < \cdots < i_k \le m-1} \sum_{j=1}^{p_m} D_{i_1,\ldots,i_k}(\xi_{i_1,\ldots,i_k}, y_j) l_j(y)\, \Omega_{i_1,\ldots,i_k}.$$

Остается заметить, что при некотором ζ

$$\sum_{j=1}^{p_m} D_{i_1,\ldots,i_k}(\xi_{i_1,\ldots,i_k}, y_j) l_j(y) =$$

$$= D_{i_1,\ldots,i_k}(\xi_{i_1,\ldots,i_k}, y) - \frac{1}{p_m!} D_{i_1,\ldots,i_k,m}(\xi_{i_1,\ldots,\xi_k}, \zeta)\, \omega(y).$$

Теорема доказана. $\square$

Следствие 12.11.1 *Пусть точка x принадлежит выпуклой оболочке, натянутой на $\mathcal{M}$. Предположим, $p_1 = \cdots = p_m \equiv N$ и сетка $\mathcal{M}$ является равномерной по каждому направлению с шагом $0 < h \le 1$. Тогда если все частные производные любого порядка $l \le k \le mN$ ограничены по модулю константой C_k, то*

$$|f(x) - p(x)| \le C_N\, m\, h^N + C_{mN}\, (2^m - m)\, h^{2N}.$$

Задачи

1. Для каждого n приведите пример матрицы Вандермонда порядка n, для которой спектральное число обусловленности равно 1.

2. Пусть $W(x_0, ..., x_n)$ — матрица Вандермонда с узлами $x_i \in [-1, 1]$. Докажите, что $\mathrm{cond}_2 W(x_0, ..., x_n) \geq 2^{n-1}$.

3. Дана таблица значений полинома 2-й степени:

x	-2	-1	0	1	2
$f(x)$	7	3	1	0	3

   Известно, что во второй строке содержится ровно одна ошибка. Найти ошибку, исправить ее и восстановить исходный полином.

4. Некто составляет таблицу значений для функции

$$f(x) = \frac{2}{\pi} \int\limits_0^x \mathrm{e}^{-t^2} d\,t$$

   на отрезке $[0, 1]$ с постоянным шагом h. При квадратичной интерполяции табличных значений ошибка не должна превышать 0.01. Каким должно быть h?

5. Полином $f(x) = x^n + a_1\, x^{n-1} + \ldots + a_n$ имеет попарно различные корни $x_1, \ldots, x_n$. Докажите, что

$$\sum_{j=1}^{n} \frac{x_j^k}{f'(x_j)} = \begin{cases} 0, & 0 \leq k \leq n-2, \\ 1, & k = n-1. \end{cases}$$

6. Докажите, что если $f \in C^k$, то

$$f\,(x_0; \ldots ; x_k) = \int\limits_{\substack{\xi_0, ..., \xi_k \geq 0 \\ \xi_0 + ... + \xi_k = 1}} f^{(k)}(\xi_0 x_0 + ... + \xi_k x_k) d\xi_0 ... d\xi_k$$

$$= \int\limits_0^1 \int\limits_0^{t_1} \ldots \int\limits_0^{t_{k-1}} f^{(k)}\left(x_0 + t_1\,(x_1 - x_0) + ... + t_k\,(x_k - x_{k-1})\right) dt_k \ldots dt_1.$$

7. Докажите единственность полинома, решающего задачу эрмитовой интерполяции.

8. Интерполяционный полином Лагранжа $L_n(x)$ приближает функцию $f(x)$ с точностью $\varepsilon > 0$. С какой точностью $L_n'(x)$ приближает $f'(x)$?

9. По значениям функции $f(x)$ в попарно различных точках $x_1, \ldots x_n$ требуется найти коэффициенты интерполяционного полинома. Придумайте алгоритм, делающий это за $\mathcal{O}(\log_2 n)$ параллельных шагов.

10. Пусть фиксированы попарно различные числа $\alpha_1, \ldots \alpha_n$. Докажите, что функции вида $e^{\alpha_1 x}, ..., e^{\alpha_n x}$ линейно независимы на любом интервале. Докажите также, что на любом интервале можно выбрать попарно различные узлы $x_1, \ldots x_n$, для которых задача интерполяции $c_1 e^{\alpha_1 x_j} + ... + c_n e^{\alpha_n x_j} = f(x_j)$, $1 \le j \le n$, имеет и притом единственное решение.

11. Пусть задана аналитическая функция $f(x) = \sum\limits_{k=0}^{\infty} a_k x^k$ и требуется найти функцию вида

$$\phi(x) = \frac{p_0 + p_1 x + \cdots + p_{n-1} x^{n-1}}{q_0 + q_1 x + \cdots + q_n x^n},$$

удовлетворяющую обобщенным интерполяционным условиям:

$$\phi^{(j)}(0) = f^{(j)}(0), \quad j = 0, 1, \ldots, 2n.$$

Докажите, что для существования решения достаточно, чтобы матрицы $A_k = [a_{k+i-j}]_{ij=0}^{k}$ были невырожденными при $k = n$ и $k = n-1$.

Глава 13

13.1 Сходимость интерполяционного процесса

Пусть на отрезке $[a,\ b]$ задана последовательность простых сеток

$$M_n = \{x_{n\,0},\ \ldots\ , x_{nn}\}, \quad n = 0,\ 1,\ \ldots\ ,$$

порождающих последовательность полиномов Лагранжа $L_n(x)$, где $L_n(x)$ интерполирует значения $f(x)$ на сетке M_n. Верно ли, что $L_n(x) \to f(x)$ $\forall\, x \in [a,\ b]$? Будет ли сходимость равномерной по x? Как оценить скорость сходимости?

Ответы зависят от свойств последовательности сеток и от свойств функции $f(x)$. Если $f \in C^{\infty}[a,\ b]$ и $\sup\limits_{x} |f^{(n)}(x)| \le M^n \ \ \forall\, n$, где $M > 0$ не зависит от n, то теорема 12.4.1 о погрешности лагранжевой интерполяции дает

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

Правая часть, очевидно, стремится к нулю при $n \to \infty$.

"Либерализация" требований к f связана с особыми требованиями к сеткам.

13.2 Сходимость проекторов

Пусть $F \ = \ C\,[a,\ b]$ — банахово пространство с нормой

$$||f||_{C\,[a,\ b]} \ \equiv \ \max_{a \ \le \ x \ \le \ b} |f(x)|. \tag{13.2.1}$$

Если Π_n — пространство полиномов степени не выше n, то для $f \ \in \ F$ лагранжева интерполяция порождает оператор

$$P_n :\ F \ \to \ \Pi_n, \qquad P_n\ f \ = \ L_n.$$

Оператор P_n является линейным непрерывным оператором (почему?). Кроме того, $P_n^2 = P_n \ \ \Rightarrow \ \ P_n$ является проектором. Вопрос о сходимости интерполяционного процесса можно сформулировать таким образом: при каких условиях

$$P_n\ f \ \to \ f \qquad \forall\, f \in F \quad ?$$

13.3 Линейные непрерывные операторы в банаховом пространстве

Рассмотрим произвольное банахово пространство F и последовательность линейных непрерывных операторов $P_n : F \to F$ (не обязательно проекторов).

Теорема 13.3.1 (Принцип равномерной ограниченности)

$$\sup_n ||P_n f|| \le c(f) < +\infty \quad \forall f \in F \quad \Leftrightarrow \quad \sup_n ||P_n|| < +\infty.$$

Доказательство. Часть "$\Leftarrow$" очевидна. Часть "$\Rightarrow$" докажем от противного. Допустим, что последовательность $||P_n||$ не ограничена. Тогда в любом шаре найдутся элементы u_n такие, что $||P_n u_n|| \to \infty$.

Для $m = 1, 2, \ldots$ рассмотрим множества $O_m \equiv \{f : \sup_n ||P_n f|| > m\}$. Они открыты, $O_1 \supset O_2 \supset \ldots$ и если $O_m \neq \emptyset$, то O_m плотно в любом шаре (почему?). Предположим, что $O_m \neq \emptyset \ \forall m$. Построим B_m и $\bar{B}_m$ — открытый и замкнутый шары радиуса $r_m \le 1/m$, входящие в $O_m \cap B_{m-1}$ (пусть $B_0 \equiv O_1$). Шары $\bar{B}_m$ образуют последовательность вложенных замкнутых шаров, радиус которых стремится к нулю. Поэтому

$$\exists \ f \in \bigcap_{m=1}^{\infty} \bar{B}_m \subset \bigcap_{m=1}^{\infty} O_m \ \Rightarrow \ \sup_n ||P_n f|| > m \ \forall m \quad \text{(противоречие!)} \quad \square$$

Теорема 13.3.2 (Банах–Штейнгауз) *Для того чтобы последовательность функций $P_n f$ была сходящейся для всех $f \in F$, необходимо и достаточно выполнение двух условий:*

(1) $\sup_n ||P_n|| < +\infty$;

(2) *$P_n f$ сходится при $n \to \infty$ на некотором подмножестве $\tilde{F}$, всюду плотном в F.*

Доказательство. Необходимость вытекает из принципа равномерной ограниченности. Чтобы доказать достаточность, возьмем $f \in F$ и рассмотрим ε-приближение $f_\varepsilon \in \tilde{F} : \quad ||f - f_\varepsilon|| \le \varepsilon$. Находим

$$\begin{aligned} ||P_n f - P_m f|| &\le ||P_n f - P_n f_\varepsilon|| + ||P_n f_\varepsilon - P_m f_\varepsilon|| + ||P_m f_\varepsilon - P_m f|| \\ &= \mathcal{O}(\varepsilon) \quad \text{при} \quad n, m \to \infty. \quad \square \end{aligned}$$

Если P_n — интерполяционные проекторы, то сходимость на всюду плотном множестве в $C[a, b]$ заведомо имеет место (почему?). Ключевой вопрос для анализа сходимости интерполяционного процесса: как ведут себя нормы $||P_n||$ при $n \to \infty$?

13.4 Алгебраические и тригонометрические полиномы

Если $x \in [a,\, b]$, то замена $x = \frac{a+b}{2} + \frac{b-a}{2}\, t$ позволяет перейти к функциям от $t \in [-1,\, 1]$.

Еще одна замена $t = \cos\phi$ позволяет перейти к четным 2π-периодическим функциям от ϕ, определенным на всей вещественной прямой. При этом алгебраический полином $p_n(t)$ степени не выше n превращается в четный тригонометрический полином

$$p_n(\cos\phi) = q_n(\phi) \equiv \sum_{k=0}^{n} \alpha_k\ \cos k\phi.$$

Функция $\cos k\phi$ есть алгебраический полином от $\cos\phi$ степени не выше k (докажите) $\Rightarrow$ любой четный тригонометрический полином $q_n(\phi)$ степени не выше n можно записать в виде $p_n(\cos\phi)$, где $p_n(t)$ — алгебраический полином степени не выше n.

Произвольный (не обязательно четный) тригонометрический полином степени не выше n имеет вид

$$Q_n(\phi) = \alpha_0 + \sum_{k=1}^{n} \left(\alpha_k\ \cos k\phi + \beta_k\ \cos k\phi\right).$$

Его удобно записывать также в виде

$$Q_n\left(\phi\right)\ =\ \sum_{k=-n}^{n} c_k\ \mathrm{e}^{\mathrm{i}\,k\,\phi},$$

где $c_{-k}\ =\ \bar{c}_k$ (комплексное сопряжение) для всех k (в этом и только в этом случае значения $Q_n\left(\phi\right)$ вещественны при всех ϕ).

Важный вывод: вопросы приближения функций алгебраическими полиномами сводятся к вопросам приближения четных периодических функций четными тригонометрическими полиномами.

13.5 Проекторы, связанные с рядом Фурье

Пусть $C = C[-\pi,\, \pi]$ — банахово пространство вещественных непрерывных 2π-периодических функций с нормой (13.2.1) и S_n — проектор, переводящий $f \in C$ в n-й отрезок ее ряда Фурье:

$$f(\phi) \mapsto s_n\left(\phi\right) = \sum_{k=-n}^{n} c_k\ \mathrm{e}^{\mathrm{i}\,k\,\phi}, \quad c_k = \frac{1}{2\,\pi} \int\limits_{-\pi}^{\pi} f\left(\psi\right) \mathrm{e}^{-\mathrm{i}\,k\,\psi}\, d\,\psi.$$

Лемма 13.5.1 $||S_n|| \ \geq \ c \ \ln \ n, \quad c \ > \ 0.$

Доказательство. Учитывая выражения для c_k, находим

$$s_n\,(\phi) \ = \ \frac{1}{2\,\pi}\int\limits_{-\pi}^{\pi}\left(\sum_{k=-n}^{n}\mathrm{e}^{\mathrm{i}\,k\,(\phi-\psi)}\right)\,f\,(\psi)\,d\,\psi \ = \ \frac{1}{2\,\pi}\int\limits_{-\pi}^{\pi}\frac{\sin\,(n+\frac{1}{2})\,t}{\sin\,\frac{t}{2}}\,f\,(\phi+t)\,d\,t.$$

Отсюда (используем неравенства $\frac{2}{\pi}x \leq \sin x \leq x$ при $0 \leq x \leq \frac{\pi}{2}$)

$$\begin{aligned}
||S_n|| \ &= \ \frac{1}{2\,\pi}\int\limits_{-\pi}^{\pi}\left|\frac{\sin\,(n+\frac{1}{2})\,t}{\sin\,\frac{t}{2}}\right|\,d\,t \ = \ \frac{1}{\pi}\int\limits_{0}^{\pi}\left|\frac{\sin\,N\,u}{\sin\,u}\right|\,d\,u \quad (N=2\,n+1)\\
&> \ \frac{1}{\pi}\sum_{k=0}^{N-1}\int\limits_{\frac{\pi\,k}{N}}^{\frac{\pi\,k}{N}+\frac{\pi}{2\,N}}\left|\frac{\sin\,N\,u}{\sin\,u}\right|\,d\,u \ = \ \frac{1}{\pi}\sum_{k=0}^{N-1}\int\limits_{0}^{\frac{\pi}{2\,N}}\left|\frac{\sin\,N\,v}{\sin\,(\frac{\pi\,k}{N}\,+\,v)}\right|\,d\,v\\
&> \ \frac{1}{\pi}\sum_{k=0}^{N-1}\int\limits_{0}^{\frac{\pi}{2\,N}}\frac{\frac{2}{\pi}\,N\,v}{\frac{\pi\,k}{N}\,+\,v}\,d\,v \ > \ \frac{1}{\pi}\sum_{k=0}^{N-1}\frac{\frac{2}{\pi}\,N\,\left(\frac{\pi}{2\,N}\right)^2\,\frac{1}{2}}{\frac{\pi\,k}{N}\,+\,\frac{\pi}{2\,N}}\\
&= \ \frac{1}{4\,\pi}\sum_{k=0}^{N-1}\frac{1}{k\,+\,\frac{1}{2}} \ > \ \frac{1}{2\,\pi}\sum_{k=1}^{N-1}\frac{1}{k} \ \geq \ c\,\ln\,n. \quad \square
\end{aligned}$$

Следствие 13.5.1 *Существует непрерывная периодическая функция, для которой ряд Фурье не сходится равномерно ни к какой непрерывной функции.*

Следствие 13.5.2 *Для любой точки существует непрерывная периодическая функция, для которой ряд Фурье расходится в этой точке.*

Для заданной точки $x_0 \in [-\pi,\,\pi]$ нужно рассмотреть функционалы $s_n(x_0)$ и заметить, что $||s_n(x_0)|| = ||S_n||$.

13.6 "Пессимистические" результаты

Лемма 13.6.1 *Пусть линейный непрерывный оператор $T_n : C \to C$ осуществляет проектирование на подпространство четных (нечетных) тригонометрических полиномов степени не выше n и переводит любую нечетную (четную) функцию в нуль. Тогда для любой четной (нечетной) функции f имеет место тождество*

$$\frac{1}{\pi}\int\limits_{-\pi}^{\pi}\left[T_n\,f\,(t+h)\right]\,(t-h)\,d\,h \ = \ c_0(f) \ + \ \left[S_n\,f\right]\,(t),$$

$$c_0(f) = \frac{1}{2\pi}\int\limits_{-\pi}^{\pi} f(t)dt.$$

Доказательство. Тождество легко проверить, взяв $f(t) = \cos nt$ в случае четных функций и $f(t) = \sin nt$ в случае нечетных функций. Остается заметить, что множество четных (нечетных) тригонометрических полиномов всюду плотно в пространстве непрерывных четных (нечетных) функций. □

Теорема 13.6.1 (Фабер–Бернштейн) *Для любой последовательности простых сеток на* $[a,\, b]$ *существует непрерывная на* $[a,\, b]$ *функция, для которой последовательность интерполяционных полиномов не сходится равномерно ни к какой непрерывной функции.*

Доказательство. Пусть T_n — проектор на пространство четных тригонометрических полинонов, а S_n — аналогичный проектор, связанный с рядом Фурье. Пусть f — четная функция и $||f|| = 1$. Тогда в силу леммы 13.6.1 $|[S_n\, f](t)| - 1 \le 2||T_n|| \;\Rightarrow\; c \log n \le ||T_n||$. Согласно теореме 13.3.2, последовательность интерполяционных полиномов не может сходиться равномерно для всех $f \in C\,[a,\, b]$, так как $||P_n|| \;\to\; \infty$. □

Заметим, что если $T_n \;:\; C \;\to\; C$ — линейный непрерывный проектор на пространство тригонометрических полиномов степени не выше n, то из леммы 13.6.1 вытекает *тождество Зигмунда–Марцинкевича–Бермана* $\frac{1}{2\,\pi} \int\limits_{-\pi}^{\pi} [T_n\; f\,(t\,+\,h)]\;(t\,-\,h)\;d\,h \;\;=\;\; [S_n\;f]\;(t)$ и неравенство $||T_n|| \ge ||S_n|| \ge c\,\ln n$, известное как *теорема Лозинского–Харшиладзе.*

Пример Бернштейна: если $f(x) = |x|, \quad x \in [-1,\, 1]$, то интерполяционные полиномы, построенные на равномерных сетках, не будут сходиться к $f(x)$ ни в одной точке, кроме $x \;=\; -1,\, 0,\, 1$.

13.7 Чем плохи равномерные сетки

Для равномерных сеток нормы $||P_n||$ растут экспоненциально. В самом деле, пусть

$$t_j = -1 + \frac{2}{n}\, j \in [-1,\, 1], \quad j = 0,\, 1,\, \ldots\,, n,$$

Тогда

$$||P_n|| = \max_{-1 \le t \le 1} \sum_{k=0}^{n} \left| \prod_{\substack{j=0 \\ j \ne k}}^{n} \frac{t - t_j}{t_k - t_j} \right|.$$

Для $t = -1 + \frac{2}{n}\,\theta, \quad 0 < \theta < 1$, находим

$$||P_n|| \;\geq\; \sum_{k=0}^{n} \left| \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{\theta - j}{k - j} \right| \;\geq\; \frac{\theta\,(1-\theta)}{n^2} \sum_{k=0}^{n} \frac{n!}{k!\;(n-k)!} \;=\; \frac{\theta\,(1-\theta)}{n^2}\; 2^n,$$

поскольку $2^n \;=\; \binom{n}{0} \;+\; \binom{n}{1} \;+\; \ldots \;+\; \binom{n}{n}$.

Более точная оценка

$$||P_n|| \;\geq\; \frac{\theta}{n} \sum_{k=0}^{n} \frac{n!}{k!\;(n-k)!} \prod_{k=1}^{n} (1 - \frac{\theta}{k}) \;\geq\; c\,\frac{2^n}{n^{1+\theta}}, \quad 0 < \theta < 1,$$

вытекает из известного в теории гамма-функции соотношения

$$\prod_{k=1}^{n} (1 - \frac{\theta}{k}) \;=\; \frac{n^{-\theta}}{\Gamma\,(1-\theta)} \;+\; \mathcal{O}\,(n^{-\theta-1}),$$

$$\Gamma\,(z) \;=\; \int_0^\infty x^{z-1}\, e^{-x}\, d\,x \quad \text{(представление Эйлера для гамма-функции)}.$$

13.8 Полиномы Чебышева и чебышевские сетки

Чтобы получить "оптимистические" результаты, необходимо ограничить произвол в использовании сеток. Логарифмический рост норм $||P_n||$ имеет место всегда. Чтобы иметь не более чем логарифмический рост, достаточно использовать *чебышевские сетки*: их узлы определяются как корни функций

$$T_n(t) = \cos(n\,\arccos t), \quad -1 \leq x \leq 1,$$

называемых *полиномами Чебышева*.

То, что это полиномы от t, доказывается по индукции. При $n = 0$ и $n = 1$ имеем $T_0(t) = 1$ и $T_1(t) = t$. При $n \geq 1$

$$\cos((n+1)\,\arccos\,t) + \cos((n-1)\,\arccos t) = 2\,\cos(n\,\arccos t)\cos(\arccos t)$$

$$\Rightarrow \quad T_{n+1}(t) = 2t\,T_n(t) - T_{n-1}(t), \quad n = 1,\,2,\,\ldots\,. \quad \Rightarrow$$

$T_n(t)$ — полином степени n со старшим коэффициентом 2^{1-n}. Корни $T_n(t)$ легко находятся как решения уравнения $\cos(n\,\arccos t) = 0$:

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

(Можно заметить, что t_{nj} получаются из равномерной сетки на $[0, \pi]$ при отображении $t = \cos\phi$.) Таким образом,

$$\omega_n(t) \equiv \prod_{j=0}^{n-1} (t \;-\; t_{nj}) = 2^{1-n}\;\cos\,(n\;\arccos\;t).$$

Отсюда следует

Лемма 13.8.1 *Если* $P_n : \; C[-1,\,1] \;\to\; C[-1,\,1]$ *— интерполяционный проектор для чебышевской сетки с числом узлов* $n+1$*, то*

$$||P_n|| = \max_{-1\le t\le 1} \rho(t), \quad \rho(t) \;\equiv\; \sum_{k=0}^{n} \frac{|\cos((n+1)\,\arccos\,t)|\;\sin\frac{\pi(2k+1)}{2(n+1)}}{(n+1)\,|t-t_k|}.$$

Теорема 13.8.1 (Бернштейн) *Для интерполяционного проектора на чебышевской сетке*

$$||P_n|| \;=\; \mathcal{O}(\ln\,n).$$

Доказательство. Достаточно рассмотреть $t \in [0,\,1]$. Пусть

$$t \;=\; \cos\,\frac{\pi(2m+2\theta)}{2(n+1)}, \quad 0 \;<\; \theta \;<\; \frac{1}{2}, \quad m - \text{целое}.$$

Тогда

$$\begin{aligned}
\rho(t) \;&=\; \sum_{k=0}^{n} \frac{|\cos\,\frac{\pi(2m+1+2\theta)}{2}|\;\sin\frac{\pi(2k+1)}{2(n+1)}}{(n+1)\;2\,|\sin\,\frac{\pi(k+m+1+\theta)}{2(n+1)}\,|\sin\,\frac{\pi(k-m-\theta)}{2(n+1)}} \\
&\le\; c\sum_{k=0}^{n} \frac{\theta(2k+1)}{|k+m+1+\theta|\;|k-m-\theta|}
\end{aligned}$$

(используем неравенство $c_1\,\psi \le \sin\,\psi \le \psi$ при $0 \le \psi \le c_2\,\pi$, $0 < c_1, c_2 < 1$)

$$\begin{aligned}
\le\; c \;&+\; c\sum_{k=0}^{m-1} \frac{\theta(2k+1)}{(k+m+1+\theta)\,(m-k+\theta)} \\
&+\; \sum_{k=m+1}^{n} \frac{\theta(2k+1)}{(k+m+1+\theta)\,(k-m-\theta)} \\
=\; c \;&+\; c\,\theta\sum_{k=0}^{m-1}\left(\frac{1}{m-k+\theta} \;-\; \frac{1}{k+m+1+\theta}\right) \\
&+\; c\,\theta\sum_{k=m+1}^{n}\left(\frac{1}{k+m+1+\theta} \;+\; \frac{1}{k-m-\theta}\right) \\
&=\; \mathcal{O}(\ln n). \qquad \square
\end{aligned}$$

13.9 "Оптимистические" результаты

Теорема 13.9.1 *Если* $f \in C^m[-1, 1]$, $m \geq 1$, *и* P_n *— интерполяционный проектор на чебышевской сетке с* $n+1$ *узлом, то*

$$||f - P_n f||_{C[-1,1]} = \mathcal{O}\left(\frac{\ln n}{n^m}\right). \tag{13.9.2}$$

Рассмотрим взаимно-однозначное соответствие $\Theta: f(t) \to g(\phi) \equiv f(\cos\phi)$. Важно, что если $f \in C^m[-1, 1]$, то $g \in C^m[-\infty, \infty]$. Есть много способов приблизить $g \in C^m$ четным тригонометрическим полиномом $S_n g$ степени не выше n. При этом существуют такие полиномы $S_n g$, что

$$||g - S_n g||_C = \mathcal{O}(\frac{1}{n^m}). \tag{13.9.3}$$

Это утверждение *не является тривиальным*. Однако, если $g \in C^{m+1}$, то в качестве $S_n g$ мы можем взять n-й отрезок ряда Фурье для g. Тогда (13.9.3) устанавливается элементарно.

В силу (13.9.3)

$$||f - \hat{S}_n f||_{C[-1,1]} = \mathcal{O}(\frac{1}{n^m}), \quad \hat{S}_n = \Theta^{-1} S_n \Theta.$$

Функция $\hat{S}\, f$ есть алгебраический полином степени не выше $n \quad \Rightarrow P_n \hat{S}_n f = \hat{S}_n f$. Следовательно,

$$||f - P_n f|| \leq ||f - S_n f|| + ||P_n S_n f - P_n f|| \leq (1 + ||P_n||)\ ||f - S_n f||,$$

и остается учесть оценку Бернштейна для $||P_n||$.

В случае аналитических функций можно получить "еще более оптимистические" оценки. Для этого нам понадобятся эллипсы Бернштейна.

13.10 Полиномы Чебышева и эллипсы Бернштейна

Значения $T_n(z)$ определены, конечно, при всех $z \in \mathbb{C}$. В общем случае

$$T_n(z) = \frac{1}{2}\left(z + \sqrt{z^2-1}\right)^n + \frac{1}{2}\left(z - \sqrt{z^2-1}\right)^n, \tag{13.10.4}$$

где в каждой их двух скобок $\sqrt{z^2-1}$ заменяется на одно и то же значение из двух возможных значений квадратного корня. Достаточно проверить, что правая часть (13.10.4) удовлетворяет рекуррентному соотношению для $T_n(z)$ и совпадает с $T_n(z)$ при $n = 0$ и $n = 1$.

Эллипсы Бернштейна — это эллипсы Γ_ρ с фокусами в точках ± 1 и суммой полуосей $\rho > 1$. Они получаются при отображении точек окружности $w = \rho e^{\mathbf{i}\phi} \mapsto z$ с помощью *функции Жуковского*:

$$z = \frac{1}{2}(w + w^{-1}) = \frac{1}{2}\left(\rho + \frac{1}{\rho}\right)\cos\phi + \mathbf{i}\frac{1}{2}\left(\rho - \frac{1}{\rho}\right)\sin\phi.$$

Очевидно, что данное множество точек z есть эллипс с полуосями

$$a = \frac{1}{2}\left(\rho + \frac{1}{\rho}\right), \quad b = \frac{1}{2}\left(\rho - \frac{1}{\rho}\right).$$

При этом $a + b = \rho$ и, поскольку $a^2 - b^2 = 1$, фокусы действительно находятся в точках ± 1.

Заметим, что

$$z^2 - 1 = \frac{1}{2}\left(\rho e^{\mathbf{i}\phi} - \frac{1}{\rho}e^{-\mathbf{i}\phi}\right)^2 \quad \Rightarrow \quad T_n(z) = \frac{1}{2}\left(\rho^n e^{\mathbf{i}n\phi} + \frac{1}{\rho^n}e^{-\mathbf{i}n\phi}\right).$$

Таким образом, $|T_n(z)| \sim \frac{1}{2}\rho^n$ при $\rho \to \infty$ (линии уровня для $|T_n(z)|$ асимптотически совпадают с эллипсами Бернштейна) и всегда имеет место неравенство

$$|T_n(z)| \geq \frac{1}{2}\left(\rho^n - \frac{1}{\rho^n}\right) \quad \forall\, z \in \Gamma_\rho.$$

13.11 Интерполяция аналитических функций

В практических задачах функция $f(x)$, интересующая нас при $x \in [-1, 1]$, часто может рассматриваться как след функции $f(z)$ комплексной переменной $z \in \Omega$, где Ω — открытая область на комплексной плоскости, содержащая $[-1, 1]$. Предположим, что границей области $\Omega = \Omega_\rho$ является эллипс Бернштейна Γ_ρ.

Теорема 13.11.1 *Пусть $f(z)$ — аналитическая функция в замыкании области Ω_ρ, $\rho > 1$, а полином Лагранжа $L_n(z)$ интерполирует ее значения в чебышевских узлах $x_0, \ldots, x_n \in [-1, 1]$. Тогда при некотором $c > 0$*

$$|f(x) - L_n(x)| \leq \frac{c}{\rho^n} \quad \forall\, n, \quad x \in [-1, 1].$$

Доказательство. Пусть $x \in [-1, 1]$ не совпадает ни с одним из узлов x_i. Функция

$$F(z) = \frac{f(z)}{(z - x)\prod\limits_{i=0}^{n}(z - x_i)}$$

имеет полюсы в точках $z = x, x_0, ..., x_n$. Поэтому, согласно теории аналитических функций,

$$\int\limits_{\Gamma_\rho} F(z)dz = 2\pi\mathbf{i}\left(\frac{f(x)}{\prod\limits_{j=0}^{n}(x-x_j)} - \sum_{i=0}^{n}\frac{f(x_i)}{(x-x_i)\prod\limits_{\substack{0\le j\le n\\ j\ne i}}(x_i-x_j)}\right) \quad \Rightarrow$$

$$\begin{aligned} f(x) - L_n(x) &= \frac{1}{2\pi\mathbf{i}}\prod_{j=0}^{n}(x-x_j)\int\limits_{\Gamma_\rho}\frac{f(z)}{(z-x)\prod\limits_{j=0}^{n}(z-x_j)}dz \\ &= \frac{1}{2\pi\mathbf{i}}T_{n+1}(x)\int\limits_{\Gamma_\rho}\frac{f(z)}{(z-x)T_{n+1}(z)}dz. \end{aligned}$$

Остается учесть, что $|T_{n+1}(x)| \le 1$ при $x \in [-1,\,1]$ и $|T_{n+1}(z)| \ge c_1\rho^{n+1}$ при всех $z \in \Gamma_\rho$, где $c_1 > 0$ не зависит от n. Тогда

$$|f(x) - L_n(x)| \le \frac{\gamma_\rho}{2\pi c_1 d_\rho}\max_{z\in\Gamma_\rho}|f(z)|\,\frac{1}{\rho^{n+1}},$$

$$\gamma_\rho = \int\limits_{\Gamma_\rho}|dz|, \qquad d_\rho = \min_{z\in\Gamma_\rho,\ x\in[-1,1]}|z-x|. \quad \square$$

13.12 Многомерная интерполяция на чебышевских сетках

Пусть $f(x) = f(x_1, \ldots, x_k) \in C(K)$, где $K \equiv [-1,\,1]^k$, и $L_n(x) = L_n(x_1, \ldots, x_k)$ – ее интерполяционный полином для декартова произведения одномерных чебышевских сеток

$$\mathcal{M} = \{x_1^{i_1}\}_{i_1=0}^{n} \times \ldots \times \{x_m^{i_m}\}_{i_m=0}^{n}.$$

Рассмотрим проектор P_{in}, переводящий f в ее интерполяционный полином $L_{in}(x_1, ..., x_m)$ относительно i-й переменной при фиксированных значениях остальных переменных. Тогда

$$L_n = P_{1n}...P_{kn}f.$$

Теорема 13.12.1 *Если $f(x_1, ..., x_k)$ имеет m непрерывных частных производных по каждой переменной, то*

$$||f - L_n||_{C(K)} = O\left(\frac{\ln^k n}{n^m}\right).$$

Доказательство. Согласно теореме 13.8.1 $||P_i|| = O(\ln n)$, а в силу теоремы 13.9.1 $||f - P_{in}f|| = O(\ln n/n^m)$. Отсюда

$$\begin{aligned} ||f - L_n|| &\leq ||f - P_{1n}f|| + ||P_{1n}f - P_{1n}P_{2n}f|| + ... \\ &\quad ... + ||P_{1n}...P_{k-1\,n}f - P_{1n}...P_{k-1\,n}P_{kn}f|| \\ &\leq (1 + \ln n + ... + \ln^{k-1} n) \cdot O\left(\frac{\ln n}{n^m}\right) = O\left(\frac{\ln^k n}{n^m}\right). \quad \square \end{aligned}$$

Теорема 13.12.2 *Пусть для каждого i функция $f_i(x_i) \equiv f(x_1, ..., x_k)$ является следом функции $f_i(z_i)$, аналитичной в замкнутой области, ограниченной эллипсом Бернштейна Γ_ρ, $\rho > 1$, и непрерывной по всем переменным. Тогда*

$$||f - L_n||_{C(K)} = O\left(\frac{\ln^{k-1} n}{\rho^n}\right).$$

Для доказательства достаточно принять во внимание теорему 13.11.1.

Задачи

1. На отрезке $[-\pi,\ \pi]$ заданы значения на простой сетке с $2n+1$ узлом. Докажите существование и единственность тригонометрического полинома

$$Q_n(\phi) \ = \ \alpha_0 \ + \ \sum_{k=1}^{n} (\alpha_k \ \cos \ k\,\phi \ + \ \beta_k \ \cos \ k\,\phi)\,,$$

интерполирующего эти значения.

2. Пусть $f(x) = |x|$ и $L_n(x)$ — полиномы Лагранжа, построенные по чебышевским сеткам на $[-1,\ 1]$. Докажите, что $L_n(x) \ \to \ f(x)$ равномерно по $x \ \in \ [-1,\ 1]$.

3. Пусть C — банахово пространство вещественных непрерывных 2π-периодических функций с нормой пространства $C\,[-\pi,\ \pi]$ и S_n — проектор, переводящий $f \ \in \ C$ в n-й отрезок ее ряда Фурье. Докажите, что

$$||S_n|| = \frac{1}{2\,\pi} \int\limits_{-\pi}^{\pi} \left| \frac{\sin\,(n + \frac{1}{2})\,t}{\sin\ \frac{t}{2}} \right| \ d\,t.$$

4. Пусть C — банахово пространство вещественных непрерывных 2π-периодических функций с нормой пространства $C[-\pi, \pi]$ и S_n — проектор, переводящий $f \in C$ в n-й отрезок ее ряда Фурье. Докажите, что
$$||S_n|| \leq c \ln n,$$
где $c > 0$ не зависит от n.

5. Пусть C — банахово пространство вещественных непрерывных четных 2π-периодических функций с нормой пространства $C[-\pi, \pi]$ и S_n — проектор, переводящий $f \in C$ в n-й отрезок ее ряда Фурье. Докажите, что
$$||S_n|| \geq c \ln n,$$
где $c > 0$ не зависит от n.

6. Пусть на отрезке $[-1, 1]$ выбраны попарно различные точки $x_0, ..., x_n$ и $P_n : C[-1, 1] \to C[-1, 1]$ — проектор, переводящий непрерывную функцию в ее интерполяционный полином степени не выше n. Докажите, что
$$||P_n|| = \max_{-1 \leq x \leq 1} \sum_{0 \leq i \leq n} \left| \prod_{\substack{0 \leq j \leq n \\ j \neq i}} \frac{x - x_j}{x_i - x_j} \right|.$$

7. Пусть Γ_ρ — эллипс Бернштейна для некоторого $\rho > 1$. Докажите, что
$$d_\rho \equiv \min_{z \in \Gamma_\rho,\ x \in [-1,1]} |z - x| \geq \frac{(\rho - 1)^2}{2\rho}, \qquad \gamma_\rho \equiv \int_{\Gamma_\rho} |dz| \leq 2\rho.$$

Глава 14

14.1 Сплайны

Естественный путь для получения "оптимистических" результатов по сходимости интерполяционного процесса — отказаться от использования "чистых" полиномов и интерполировать, например, с помощью кусочно-полиномиальных функций. Такие функции называются *сплайнами*. Сплайн имеет степень m, если степень каждого полинома не выше m и равна m хотя бы для одного полинома.

Если сплайн степени m имеет m непрерывных производных, то он является "чистым" полиномом. Поэтому *максимально гладким сплайном степени* m нужно считать сплайн с числом непрерывных производных $m-1$.

Для заданной простой сетки $a = x_0 < ... < x_n = b$ и значений $f_k = f(x_k)$ рассмотрим множество функций

$$\Phi \equiv \{\phi \in C^2[a,\, b] : \phi(x_k) = f_k, \quad k = 0,\, 1, ...,\, n\}. \tag{14.1.1}$$

Функция $S(x) \in \Phi$ называется *интерполяционным кубическим сплайном*, если на каждом отрезке $[x_{k-1},\, x_k]$, $k = 1, \ldots, n$, она является кубическим полиномом.

14.2 Естественные сплайны

Чтобы построить кубический сплайн, требуется найти $4n$ коэффициентов, определяющих кубические полиномы на каждом отрезке. Определение интерполяционного кубического сплайна дает $4n-2$ уравнений (проверьте). Чтобы число уравнений совпадало с числом неизвестных, обычно вводят два дополнительных условия.

Интерполяционный кубический сплайн, удовлетворяющий дополнительным условиям

$$S''(x_0) = S''(x_n) = 0, \tag{14.2.2}$$

называется *естественным* сплайном.

Приходится работать и с различными "неестественными" интерполяционными кубическими сплайнами. Например, вместо (14.2.2) можно проинтерполировать первые производные

$$S'(x_0) \;=\; f'(x_0), \quad S'(x_n) \;=\; f'(x_n), \tag{14.2.3}$$

или (в случае $f_0 = f_n$) задать условия периодичности первых и вторых производных:

$$S'(x_0) \;=\; S'(x_0), \quad S''(x_n) \;=\; S''(x_n). \tag{14.2.4}$$

14.3 Вариационное свойство естественных сплайнов

Замечательное свойство естественного сплайна: он минимизирует *функционал энергии*

$$E(\phi) \equiv \int_a^b (\phi''(x))^2 \, dx.$$

Теорема 14.3.1 *Пусть* Φ *имеет вид (14.1.1) и* $S(x) \in \Phi$ *— естественный сплайн. Тогда*

$$E(\phi) \;\geq\; E(S) \quad \forall\ \phi \;\in\; \Phi,$$

причем неравенство строгое, если $\phi \neq S$.

Доказательство. $(\phi'')^2 - (S'')^2 \;=\; (\phi'' - S'')^2 \;+\; 2\,S''(\phi'' - S'') \quad \Rightarrow$

$$E(\phi) \;-\; E(S) \;=\; E(\phi - S) \;+\; 2\int_a^b S''(\phi'' - S'')\, dx.$$

Интегрируя по частям, получаем

$$\int_a^b S''(\phi'' - S'')\, dx =$$

$$\sum_{k=1}^{n} \left(S''(\phi' - S') \,|_{x_{k-1}}^{x_k} - S'''(x_{k-1} + 0)(\phi - S) \,|_{x_{k-1}}^{x_k} \right) = 0. \quad \square$$

14.4 Построение естественных сплайнов

Рассмотрим $S(x)$ на k-ом отрезке $[x_{k-1},\, x_k]$ и положим

$$u_k \;=\; S''(x_k), \quad h_k \;=\; x_k - x_{k-1}, \quad x \;=\; x_{k-1} \;+\; t\,h_k.$$

Поскольку $S(x)$ — кубический полином при $x \in [x_{k-1},\, x_k]$, находим

$$\begin{aligned} S''(x) &= (1-t)\,u_{k-1} + t\,u_k \quad \Rightarrow \\ S'(x) &= S'(x_{k-1}) + h_k\,\{\frac{1}{2} - \frac{(1-t)^2}{2}\}\,u_{k-1} + h_k\,\frac{t^2}{2}\,u_k \quad \Rightarrow \\ S(x) &= S(x_{k-1}) + h_k\,t\,S'(x_{k-1}) \\ &+ h_k^2\,\left(\frac{t}{2} + \frac{(1-t)^3}{6} - \frac{1}{6}\right)\,u_{k-1} + h_k^2\,\frac{t^3}{6}\,u_k. \end{aligned}$$

Положим $\delta f_k = (f_k - f_{k-1})/h_k$. Тогда при $t = 1$ из последнего уравнения получаем

$$S'(x_{k-1}) \;=\; \delta f_k - \frac{h_k}{3}\,u_{k-1} - \frac{h_k}{6}\,u_k.$$

Учитывая это, полагаем $t = 1$ в выражении для $S'(x)$:

$$S'(x_k) \;=\; \delta f_k + \frac{h_k}{6}\,u_{k-1} + \frac{h_k}{3}\,u_k.$$

Приравнивая первые производные сплайна на правом конце k-го отрезка и на левом конце $k+1$-го отрезка, получаем уравнение

$$\delta f_k + \frac{h_k}{6}\,u_{k-1} + \frac{h_k}{3}\,u_k \;=\; \delta f_{k+1} - \frac{h_{k+1}}{3}\,u_k - \frac{h_{k+1}}{6}\,u_{k+1} \quad \Rightarrow$$

$$h_k\,u_{k-1} + 2\,(h_k + h_{k+1})\,u_k + h_{k+1}\,u_{k+1} \;=\; \rho_k \equiv 6\,(\delta f_{k+1} - \delta f_k)\,.$$

Поскольку $u_0 = u_n = 0$, имеем систему

$$T\,\left[\begin{array}{c} u_1 \\ \cdots \\ u_{n-1} \end{array}\right] \;=\; \left[\begin{array}{c} \rho_1 \\ \cdots \\ \rho_{n-1} \end{array}\right] \tag{14.4.5}$$

с трехдиагональной матрицей коэффициентов

$$T = \left[\begin{array}{cccc} 2\,(h_1+h_2) & h_2 & & \\ h_2 & 2\,(h_2+h_3) & h_3 & \\ & \cdots & \cdots & \cdots \\ & & h_{n-1} & 2\,(h_{n-1}+h_n) \end{array}\right]. \tag{14.4.6}$$

Теорема 14.4.1 *Естественный сплайн существует и единствен.*

Доказательство. После подстановки $S'(x_{k-1})$ в выражение для $S(x)$ будем иметь

$$S(x) \;=\; (1-t)\,f_{k-1} + t\,f_k + u_{k-1}\,h_k^2\,a(t) + u_k\,h_k^2\,b(t), \tag{14.4.7}$$

где

$$a(t) \;=\; \frac{t}{6} + \frac{(1-t)^3}{6} - \frac{1}{6}, \quad b(t) \;=\; -\frac{t}{6} + \frac{t^3}{6}. \tag{14.4.8}$$

При любых u_k формула (14.4.7) дает кусочно-полиномиальную функцию, которая интерполирует значения f_k и имеет непрерывную вторую производную. Условие непрерывности первой производной описывается системой с трехдиагональной матрицей коэффициентов T относительно u_k. Матрица T имеет диагональное преобладание и поэтому является невырожденной. $\square$

14.5 Аппроксимационные свойства естественных сплайнов

Рассмотрим простые сетки M_n: $a = x_{n0} < ... < x_{nn} = b$, и естественные сплайны, интерполирующие на них значения одной и той же функции $f(x)$.

Теорема 14.5.1 *Пусть* $1 \le j \le 4$ *и* $f \in C^j[a,b]$. *Тогда*

$$||f - S_n||_{C[a,b]} = O(h^j), \qquad h \equiv \max_k h_k.$$

Доказательство. Пусть $f \in C^2$. Поделим k-е уравнение системы (14.4.5) на $h_k + h_{k+1}$:

$$\alpha_k u_{k-1} + 2u_k + \beta_k u_{k+1} = \rho_k/(h_k + h_{k+1}), \quad 1 \le k \le n-1; \quad \alpha_0 = \beta_n = 0.$$

Для величин $d_k \equiv u_k - f_k''$ получаем уравнения

$$\alpha_k d_{k-1} + 2d_k + \beta_k d_{k+1} = b_k \equiv 6\frac{\frac{f_{k+1}-f_k}{k_{k+1}} - \frac{f_k - f_{k-1}}{h_k}}{h_k + h_{k+1}} - \alpha_k f''_{k-1} - 2f''_k - \beta_k f''_{k+1}.$$

Пусть d_l — максимальное по модулю среди d_k, $1 \le k \le n-1$. Тогда из l-го уравнения находим (заметим, что $\alpha_k + \beta_k \le 1$)

$$\max_k |d_k| = |d_l| \le 2|d_l| - \alpha_l|d_{l-1}| - \beta_l|d_{l+1}| \le |\alpha_l d_{l-1} + 2d_l + \beta_l d_{l+1}| \le \max_k |b_k|.$$

Если $f \in C^2$, то разделенная разность $f(x_{k-1}; x_k; x_{k+1})$ равна $f''(\xi)/2$ для некоторого $\xi \in [x_{k-1}, x_{k+1}]$. Поэтому

$$|b_k| \le \alpha_k|f''(\xi) - f''_{k-1}| + 2|f''(\xi) - f''_k| + \beta_k|f''(\xi) - f''_{k+1}| = o(1) \ \text{ при } \ h \to 0.$$

Если $f \in C^3$ или $f \in C^4$, то из разложений f_{k-1} и f_{k+1} в ряд Тейлора в точке x_k получаем соответственно $|b_k| = O(h)$ или $|b_k| = O(h^2)$. Следовательно,

$$|u_k - f''_k| = \begin{cases} o(1), & f \in C^2, \\ O(h), & f \in C^3, \\ O(h^2), & f \in C^4. \end{cases}$$

Далее, в силу (14.4.7) $S(x) - f(x) = Q + R$, где

$$Q = (1-t)(f_{k-1} - f(x)) + t(f_k - f(x)) + f''_{k-1}h_k^2 a(t) + f''_k h_k b(t),$$

$$R = (u_{k-1} - f''_{k-1})h_k^2 a(t) + (u_k - f''_k)h_k^2 b(t).$$

Если $f \in C^j$, то уже ясно, что $R = O(h^j)$. Для Q та же оценка получается с помощью разложений f_{k-1}, f_k и $f(x)$ в ряд Тейлора в точке x_k.

Случай $f \in C^1$ требует отдельного рассмотрения. Тот же самый сплайн можно строить с помощью параметров $v_k = S'(x_k)$. В целом аналогичные выкладки приводят к системе уравнений относительно $v_0, ..., v_n$. Подобно (14.4.5), это будет система с трехдиагональной (хотя и другой) матрицей с диагональным преобладанием. В итоге оказывается, что при $x \in [x_{k-1}, x_k]$

$$S(x) = (1-t)^2(1+2t)\, f_{k-1} + t^2(3-2t)\, f_k + v_{k-1}h_k\, t(1-t)^2 - v_k h_k\, t^2(1-t).$$

Можно доказать, что $|v_k - f'_k| = o(1)$, а затем, как и раньше, использовать разложения Тейлора для f_{k-1}, f_k и $f(x)$ в точке x_k. $\square$

Последовательность сеток M_n называется *квазиравномерной*, если отношение максимального шага $h = h(M_n) = \max\limits_k h_k$ к минимальному шагу $\delta = \delta(M_n) = \min\limits_k h_k$ равномерно ограничено при $n \to \infty$.

Теорема 14.5.2 *Для любой функции $f(x) \in C[a,b]$ последовательность естественных сплайнов $S_n(x)$ на квазиравномерных сетках M_n сходится к $f(x)$ равномерно по $x \in [a,\, b]$, если $h \to 0$.*

Доказательство. Используя строчное диагональное преобладание матрицы T, находим $u_k = O(\omega(h; f)/\delta^2)$, где

$$\omega(h;\, f) \equiv \max_{|x-y| \le h,\ x,y \in [a,b]} |f(x) - f(y)|$$

обозначает *модуль непрерывности* функции f. Важно, что $\omega(h; f) \to 0$ при $h \to 0$, если $f \in C[a,b]$. Согласно формуле (14.4.7), $|S(x) - f(x)| = O(\omega(\Delta_n;\, f)(1 + h^2/\delta^2)) = o(1)$. $\square$

14.6 *B*-сплайны и разделенные разности

Пусть сетка $x_0 < \ldots < x_n$ вложена в бесконечную простую сетку

$$M_\infty = \{\ldots x_{-1} < x_0 < \ldots < x_n < x_{n+1} < \ldots\},$$

и нас интересуют сплайны степени m максимальной гладкости (то есть класса $C^{m-1}(\mathbb{R})$). Любой такой сплайн можно представить в виде линейной комбинации базисных сплайнов, равных нулю вне отрезков вида $[x_k,\, x_{k+m+1}]$.

Чтобы построить базисные сплайны, рассмотрим *усеченные степенные функции*

$$(t-x)_+^m = \begin{cases} (t-x)^m, & t-x \geq 0, \\ 0, & t-x < 0. \end{cases}$$

Очевидно, функция$(t-x)_+^m$ непрерывно дифференцируема $m-1$ раз как по t, так и по x. Обозначим через $[x_k, ..., x_{k+m+1}]f(t,x)$ разделенную разность для $f(t,x)$ как функции от t при фиксированном значении параметра x.

B-сплайном порядка m называется функция вида

$$B_m^k(x) = [x_k, ..., x_{k+m+1}](t-x)_+^m. \tag{14.6.9}$$

Лемма 14.6.1 *$B_m^k(x)$ есть сплайн степени m класса $C^{m-1}(\mathbb{R})$.*

Доказательство. Согласно лемме 12.6.1, при $x_{k+l} < x < x_{k+l+1}$ величина $B_m^k(x)$ есть взвешенная сумма значений $(x_i - x)^m$ при $i \geq k+l+1 \;\Rightarrow\;$ это полином при $x_{k+l} < x < x_{k+l+1}$. Остается заметить, что при вычислении разделенной разности в узлах $t = x_k, ..., x_{k+m+1}$ сохраняется дифференцируемость по параметру x. $\square$

14.7 Рекуррентная формула для *B*-сплайнов

Лемма 14.7.1 *Пусть $B_0^k(x) = 1$ при $x_k \leq x < x_{k+1}$ и 0 иначе. Тогда при $m \geq 1$ имеет место тождество*

$$B_m^k(x) = \frac{x - x_k}{x_{k+m+1} - x_k} B_{m-1}^k(x) + \frac{x_{k+m+1} - x}{x_{k+m+1} - x_k} B_{m-1}^{k+1}(x).$$

Доказательство. Не ограничивая общности, положим $k = 0$. Тогда

$$B_m^0(x) = [x_0, ..., x_{m+1}](t-x)_+^m = \frac{A - B}{x_{m+1} - x_0},$$

$$A = [x_1, ..., x_{m+1}](t-x)_+^m, \quad B = [x_0, ..., x_m](t-x)_+^m.$$

Пусть

$$\omega_0(t) = \prod_{i=0}^{m}(t - x_i), \quad \omega_1(t) = \prod_{i=1}^{m+1}(t - x_i).$$

Тогда если $x_l < x < x_{l+1}$, то

$$A = \sum_{i=l+1}^{m+1} \frac{(x_i - x)^m}{\omega_1'(x_i)} = (x_{m+1} - x)B_{m-1}^1(x) + \sum_{i=l+1}^{m} (x_i - x_{m+1}) \frac{(x_i - x)^{m-1}}{\omega_1'(x_i)},$$

$$B = \sum_{i=l+1}^{m} \frac{(x_i - x)^m}{\omega_0'(x_i)} = (x_0 - x)B_{m-1}^0(x) + \sum_{i=l+1}^{m} (x_i - x_0) \frac{(x_i - x)^{m-1}}{\omega_0'(x_i)}$$

$$= (x_0 - x)B_{m-1}^0(x) + \sum_{i=l+1}^{m} (x_i - x_{m+1}) \frac{(x_i - x)^{m-1}}{\omega_1'(x_i)}$$

$$\Rightarrow \quad A - B = (x - x_0)B_{m-1}^0(x) + (x_{m+1} - x)B_{m-1}^{k+1}(x). \quad \square$$

Следствие 14.7.1 *Система $n + m$ сплайнов $B_m^k(x)$, $-m \le k \le n - 1$, линейно независима как система функций на отрезке $[x_0, x_n]$.*

Доказательство проводится по индукции. Заметим, что общее число коэффициентов полиномов степени m на n отрезках, составляющих $[x_0, x_n]$, равно $(m + 1)n$, а принадлежность сплайна классу C^{m-1} описывается линейно независимой системой $m(n-1)$ уравнений. Отсюда следует, что размерность линейного пространства сплайнов степени m, непрерывно дифференцируемых $m - 1$ раз во внутренних точках отрезка $[x_0, x_n]$, равна $n + m$.

Носителем функции, определенной на всей числовой оси, называется замыкание множества ее ненулевых значений. Обозначение: $\operatorname{supp} f$. Если $\operatorname{supp} f \subset [a, b]$, то говорят, что $f(x)$ является функцией *с конечным носителем* или *финитной* функцией. Для сплайна на сетке M_∞ конечный носитель непременно является отрезком, ограниченным узлами сетки. Можно доказать, что B-сплайны порядка m обладают носителем, содержащим минимально возможное число узлов сетки M_∞.

14.8 *B*-сплайны на равномерных сетках

Если сетка равномерная, то все базисные сплайны можно получить путем замены x на $x - x_k$ из *одной* базисной функции. Пусть M_∞ — все целые числа. Построим рекуррентно такие функции:

$$B_0(x) = \begin{cases} 1, & 0 \le x < 1, \\ 0, & \text{иначе;} \end{cases}$$

$$B_m(x) = \int_0^1 B_{m-1}(x - y)\, B_0(y)\, dy, \quad m = 1, 2, \ldots .$$

Вот их основные свойства (докажите!):

(1) $\operatorname{supp}\ B_m \ = \ [0,\, m+1]$.

(2) B_m — сплайн степени m на сетке M_∞.

(3) $B_m \in C^{m-1}$ и при этом $B'_m\,(x) \ = \ B_{m-1}\,(x) - B_{m-1}\,(x-1)$.

(4) Любой сплайн степени m класса C^{m-1} на сетке $0 < 1 < \ldots < n$ однозначно представим в виде линейной комбинации сплайнов $B_m\,(x-k)$ для целых $-m \le k \le n-1$.

(5) Функции $B_m\,(h^{-1}\,x - k)$ при целых k являются базисными сплайнами на равномерной сетке с шагом h.

14.9 Сплайны и интеграл Фурье

Удобным инструментом для анализа аппроксимационных свойств сплайнов на равномерных сетках оказывается *интегральное преобразование Фурье*:

$$\phi(\xi) = \int\limits_{-\infty}^{\infty} f(x) e^{\mathbf{i}\xi x} dx, \quad -\infty < \xi < \infty.$$

Очевидно, что интеграл существует, если $f(x)$ — интегрируемая финитная функция. В случае $f(x) = B_m(x)$ нетрудно вычислить, что

$$\phi(\xi) = \left(e^{\mathbf{i}(\xi/2)}\right)^{m+1} \left(\frac{\sin(\xi/2)}{\xi/2}\right)^{m+1}.$$

Замечательный факт, широко известный в теории связи: если функция $f(x)$ имеет конечный носитель на отрезке длины L, то $\phi(\xi)$ однозначно восстанавливается по своим значениям на бесконечной равномерной сетке

$$\xi_k = \frac{2\pi}{a} k, \quad k = 0, \pm 1, \ldots .$$

В самом деле, пусть $\operatorname{supp} f \subset [-L/2, L/2]$. Разложив $f(x)$ в ряд Фурье

$$f(x) = \sum_{k=-\infty}^{\infty} a_k e^{-\mathbf{i}\frac{2\pi}{a}x}, \quad -L/2 < x < L/2,$$

получаем

$$\phi(\xi) = \int\limits_{-a/2}^{a/2} f(x) e^{\mathbf{i}\xi x} dx = L \sum_{k=-\infty}^{\infty} a_k \operatorname{Sinc}(L(\xi - \xi_k)/2), \qquad \xi_k = \frac{2\pi k}{L},$$

где

$$\mathrm{Sinc}(\xi) \equiv \frac{\sin \xi}{\xi}, \quad \xi \neq 0, \qquad \mathrm{Sinc}(0) = 1.$$

Остается заметить, что $\phi(\xi_m) = L a_m$. [1]

Полезные аппроксимации возникают при достаточно быстром убывании значений $\phi(\xi_k)$ при $k \to \infty$.

14.10 Квазилокальность и ленточные матрицы

Естественный сплайн не обладает свойством локальности: при замене f_k на $\tilde{f}_k$ в одном узле x_k значения сплайна изменяются во всех точках. Однако, естественный сплайн обладает важным свойством *квазилокальности*: значения сплайна изменяются очень мало вне некоторой окрестности точки x_k. В основе квазилокальности — некоторое общее свойство ленточных матриц.

Теорема 14.10.1 *Пусть $A = [a_{ij}]$ — невырожденная ленточная матрица порядка n с ненулевой диагональю и полушириной ленты не больше L: $a_{ij} = 0$ при $|i-j| \geq L$. Пусть матричная норма $||\cdot||$ такова, что норма любой матрицы не может быть меньше модуля каждого из ее элементов. Тогда если*

$$q \equiv ||(\mathrm{diag}\, A)^{-1}\, \mathrm{off}\, A|| \;<\; 1,$$

то для элементов $a_{ij}^{(-1)}$ матрицы A^{-1} имеет место неравенство

$$|a_{ij}^{(-1)}| \leq ||(\mathrm{diag}\, A)^{-1}||\, \frac{q}{1-q}\, q^{|i-j|/L}, \quad i,j = 1, \ldots, n.$$

Доказательство. Положим $F \equiv (\mathrm{diag}\, A)^{-1}\, \mathrm{off}\, A$ и рассмотрим ряд Неймана $(\mathrm{diag}\, A)\; A^{-1} \;=\; I + F + F^2 \ldots$. Заметим, что F^k — ленточная матрица с полушириной ленты не больше, чем $(k-1)\, L$. Это вытекает из легко проверяемого факта: полуширина ленты для произведения AB не превосходит $L_1 + L_2 - 1$, где L_1 — полуширина ленты для A и L_2 — полуширина ленты для B. Фиксируем i, j и рассмотрим k такое, что $(k-1)\, L \leq |i-j| \;<\; k\, L$. Тогда

$$g_{ij} \equiv \{\sum_{j=0}^{\infty} F^j\}_{ij} = \{\sum_{j=k}^{\infty} F^j\}_{ij} \quad \Rightarrow \quad |g_{ij}| \leq \frac{c\, q^k}{1-q}. \qquad \square$$

[1] Обоснование проведенных нами формальных выкладок можно найти в книге: *А. И. Жуков*, Метод фурье в вычислительной математике. - М.: Наука, 1992.

Следствие 14.10.1 *Пусть естественные сплайны $S(x)$ и $\tilde{S}(x)$ интерполируют значения f_i и $\tilde{f}_i$, совпадающие при всех i, кроме $i = k$. Пусть $\varepsilon \equiv |f_k - \tilde{f}_k|$. Тогда если $x \in [x_{i-1},\, x_i]$, то*

$$|S(x) - \tilde{S}(x)| \;\le\; \hat{c}\,\varepsilon\,\frac{\Delta_n^2}{\delta_n^2}\,\frac{1}{2^{|i-k|}}, \quad \hat{c} \;>\; 0.$$

Доказательство. Для трехдиагональной матрицы T вида (14.4.6) имеем $L = 1$ и, кроме того, $q \;=\; \frac{1}{2}$ и $c = 1$, если используется норма $||\;\cdot\;||_\infty$.

Если значение $f(x)$ изменяется только в одном узле x_k, то в правой части системы $T\,u = \rho$ изменяются лишь три компоненты: ρ_{k-1}, ρ_k, ρ_{k+1}. Возмущения имеют вид $\mathcal{O}(\varepsilon/\delta_n)$, где δ_n — минимальный шаг сетки, а для компонент векторов $u \;=\; T^{-1}\,\rho$ и $\tilde{u} \;=\; T^{-1}\,\tilde{\rho}$ находим $|u_i - \tilde{u}_k| \le c_1\,\frac{\varepsilon}{\delta_n^2}\,\frac{1}{2^{|i-k|}}, \quad c_1 \;>\; 0$. Остается воспользоваться формулой (14.4.7). $\square$

Задачи

1. Докажите, что для матрицы T вида (14.4.6) $||T^{-1}||_\infty \le \frac{1}{\min\limits_{1\le k\le n-1}(h_k + h_{k+1})}$.

2. Постройте естественный кубический сплайн, который в узлах $x = -2,\, -1,\, 0,\, 1,\, 2$ принимает значения $y = 0,\, \frac{1}{6},\, \frac{2}{3},\, \frac{1}{6},\, 0$.

3. Пусть задана сетка $a = x_0 < \ldots < x_n = b$, и пусть для функции f с периодом $b - a$ строится интерполяционный кубический сплайн, удовлетворяющий дополнительным условиям периодичности

   $$S'(x_0) \;=\; S'(x_0), \quad S''(x_n) \;=\; S''(x_n).$$

   Какой вид имеет алгебраическая система относительно величин $u_k \;=\; S''(x_k)$? Докажите, что она имеет единственное решение.

4. Пусть на сетке $a = x_0 < \ldots < x_n = b$ заданы значения $f_0,\, \ldots,\, f_{n-1}$, $f_n = f_0$, и пусть $S(x)$ — интерполяционный кубический сплайн, удовлетворяющий условиям периодичности. Докажите, что если функция $\phi \in C^2\,[-\infty,\, \infty]$ с периодом $b - a$ интерполирует те же значения, то

   $$\int_a^b (\phi''(x))^2\,dx \;\ge\; \int_a^b (S''(x))^2\,dx.$$

5. Пусть $\phi \;\in\; C^r\,[a,\, b]$ и $\phi(x_k) \;=\; f_k, \quad 0 \;\le\; k \;\le\; n$. Докажите, что минимум функционала $E(\phi) \;\equiv\; \int_a^b (f^{(r)}(x))^2\;dx$ достигается на функции ϕ, являющейся сплайном степени $2r - 1$, интерполирующим значения f_k и удовлетворяющим дополнительным условиям $f^{(j)}(x_0) \;=\; f^{(j)}(x_n) \;=\; 0, \quad r \le j \le 2r - 2$.

6. Естественный сплайн S интерполирует значения $f \in C^4[a,b]$ на сетке с максимальным шагом h. Докажите, что $||S^{(j)} - f^{(j)}||_{C[a,b]} = \mathcal{O}(h^{4-j})$, $0 \le j \le 3$.

7. На сетке с узлами $x = -2,\ -1,\ 0,\ 1,\ 2$ построить кубический сплайн $B(x) \in C^2$, подчиненный условиям

$$B^{(r)}(\pm 2) = 0, \quad r = 0,\ 1,\ 2.$$

Будет ли такой сплайн единственным? Может ли такой сплайн быть нечетной функцией?

8. Фиксирована сетка $a = x_0 < \ldots < x_n = b$. Доказать, что функция, минимизирующая на $C^2[a,\ b]$ функционал

$$J(f) \equiv \int_a^b (f'')^2\, dx + \sum_{k=0}^{n} (f(x_k) - y_k)^2,$$

где y_k — заданные значения, с необходимостью является является естественным сплайном.

9. Верно ли, что любой сплайн степени m на сетке $0 < 1 < ... < n$ представим в виде линейной комбинации сплайнов $B_m(x-k)$ для целых $-m \le k \le n-1$?

10. Рассматриваются равномерные сетки $x_k = kh$ с шагом $h = b/n$ и функция $f \in C^2(\mathbb{R})$ аппроксимируется на отрезке $[0,\ b]$ с помощью функций S_h вида

$$S_h(x) = \sum_{k=-1}^{n+1} \alpha_k\, B_3(h^{-1}x - (k-2)).$$

Пусть $\alpha_k = f(x_k)$ при $-1 \le k \le n+1$. Докажите, что

$$||f - S_h||_{C[0,b]} = O(h^2).$$

Является ли S_h интерполяционным сплайном?

11. Докажите, что вывод предыдущей задачи остается в силе, если $\alpha_k = f(x_k)$ лишь при $0 \le k \le n$, а значения α_{-1} и α_{n+1} определяются соответственно по $f(x_0)$, $f(x_1)$ и по $f(x_n)$, $f(x_{n-1})$ с помощью линейной интерполяции.

12. Докажите, что $\int\limits_{-\infty}^{\infty} B_m^k(x) = (m+1)/(x_{k+m+1} - x_k)$.

13. Докажите, что для любого целого $0 \le l \le m$

$$x^l \;=\; \frac{1}{C_m^l} \sum_k \sigma_l(x_{k+1}, ..., x_{k+m})\, B_m^k(x),$$

где суммирование ведется по всем целым k таким, что $x \in \operatorname{supp} B_m^k$; функция

$$\sigma_l(z_1, ..., z_m) = \sum_{1 \le i_1 < ... < i_l \le m} z_{i_1} ... z_{i_l}$$

представляет собой элементарную симметрическую функцию степени l от m переменных; C_m^l обозначает число сочетаний из m по l.

Глава 15

15.1 Минимизация нормы

Теория и методы минимизации нормы $||f - \phi||$ на всем множестве "простых" функций $\phi \in \Phi$ существенно зависят от типа используемой нормы. Наиболее интересны следующие два случая.

Равномерное приближение. $||f|| \equiv \sup\limits_x |f(x)|$ (норма не порождается никаким скалярным произведением).

Метод наименьших квадратов. Норма порождается некоторым скалярным произведением, например, $||f|| \equiv \left(\int_a^b |f(x)|^2 \, dx\right)^{\frac{1}{2}}$.

15.2 Равномерные приближения

Пусть $f(x) \in C[a, b]$ и нас интересует полином $p_n(x)$ степени не выше n, минимизирующий $||f - p_n||_{C[a,b]}$. Такой полином называется полиномом *наилучшего равномерного приближения* для f.

В теории равномерных приближений основную роль играет понятие *точек чебышевского альтернанса* для функции $R(x) = f(x) - p_n(x)$. Точки альтернанса степени m — это узлы простой сетки

$$a \le x_1 \le \ldots \le x_m \le b$$

такие, что:

(1) $|R(x_i)| = \max\limits_{a \le x \le b} |R(x)|, \quad i = 1, \ldots, m.$

(2) $R(x_i) R(x_{i+1}) < 0, \quad i = 1, \ldots, m-1.$

Множество всех таких простых сеток на $[a, b]$ обозначим $\mathcal{A}(m, a, b, R)$.

Теорема 15.2.1 (П. Л. Чебышев) *Полином p_n степени не выше n является полиномом наилучшего равномерного приближения для $f \in C[a, b]$ тогда и только тогда, когда $\mathcal{A}(n+2, a, b, R)$ не пусто.*

Доказательство достаточности. Предположим, что для некоторого полинома q_n степени n

$$||f - q_n|| \ < \ ||f - p_n||.$$

Тогда в точках чебышевского альтернанса

$$|f(x_i) - q_n(x_i)| \ < \ |f(x_i) - p_n(x_i)| \quad \Rightarrow$$

функция $g(x) \equiv (f(x) - p_n(x)) - (f(x) - q_n(x))$ имеет тот же знак в точках x_i, что и функция $R(x) = f(x) - p_n(x)$. Поскольку знаки $R(x_i)$ чередуются, внутри каждого отрезка $[x_i,\ x_{i+1}]$ есть нуль функции $g(x)$. Значит, $g(x)$ имеет $n+1$ нуль на отрезке $[a,\ b] \quad \Rightarrow \quad g(x)$ не может быть ненулевым полиномом степени не выше $n \quad \Rightarrow \quad g(x) \equiv 0. \quad \square$

15.3 Полиномы, наименее уклоняющиеся от нуля

Полином $q_n(x) = x^n + a_{n-1}x^{n-1} + ... + a_0$ называется наименее уклоняющимся от нуля на отрезке $[a,\ b]$, если он имеет наименьшую норму в $C[a,\ b]$ среди всех полиномов такого вида.

Согласно этому определению,

$$||q_n(x)||_{C[a,\,b]} \ \leq \ ||x^n - p_{n-1}(x)||_{C[a,\,b]}$$

для любого полинома $p_{n-1}(x)$ степени не выше $n-1$. Поэтому разность $x^n - q_n(x)$ есть наилучшее равномерное приближение к функции x^n на отрезке $[a,\ b]$ среди всех полиномов степени не выше $n-1$.

Легко проверить (сделайте это), что полином $2^{1-n}T_n(x)$ имеет $n+1$ точку чебышевского альтернанса на отрезке $[-1,\ 1]$. Полином $x^n - 2^{1-n}T_n(x)$ имеет, очевидно, степень не выше $n-1$, и в силу теоремы Чебышева он является наилучшим равномерным приближением к x^n на $[-1,\ 1]$.

Следовательно, полином $2^{1-n}T_n(x)$ является наименее уклоняющимся от нуля на отрезке $[-1,\ 1]$.

В случае произвольного отрезка $[a,\ b]$ рассмотрим преобразование $x = \frac{a+b}{2} + t\,\frac{b-a}{2}$ (оно отображает $[-1,\ 1]$ на $[a,\ b]$) и обратное преобразование:

$$t \ = \ \frac{2x - a - b}{b - a}.$$

Очевидно, старший коэффициент полинома $T_n(\frac{2x-a-b}{b-a})$ равен $\frac{2^{2n-1}}{(b-a)^n}$. Таким образом, мы получаем следующее утверждение.

Теорема 15.3.1 *Полином*

$$Q_n(x) \equiv 2^{1-2n}(b-a)^n\, T_n\left(\frac{2x-a-b}{b-a}\right) \qquad (15.3.1)$$

является наименее уклоняющимся от нуля на отрезке $[a,\,b]$ *и при этом*

$$||Q_n\,(x)||_{C\,[a,\,b]} \;=\; 2^{1-2n}\,(b-a)^n. \qquad (15.3.2)$$

15.4 Ряд Тейлора и его дискретный аналог

Функцию $f\,(x) \in C^{n+1}\,[a,\,b]$ можно приблизить n-м отрезком ряда Тейлора

$$P_n\,(x) \;=\; \sum_{k=0}^{n} \frac{f^{(k)}\,(c)}{k!}\,(x-c)^k, \quad c = \frac{a+b}{2}.$$

Еще одно приближение для f — это полином Лагранжа $L_n\,(x)$ на чебышевской сетке $x_i = \frac{a+b}{2} \;+\; \frac{b-a}{2}\,t_i,\;\; 0 \le i \le n,\;$ где t_i — корни полинома T_{n+1}. Запись $L_n(x)$ с помощью разделенных разностей имеет внешнее сходство с рядом Тейлора и может считаться его дискретным аналогом. Для двух способов приближения имеем оценки

$$||f - P_n||_{C\,[a,\,b]} \;\le\; \frac{||f^{(n+1)}||_{C\,[a,\,b]}}{(n+1)!}\,\left(\frac{b-a}{2}\right)^{n+1}; \qquad (15.4.3)$$

$$||f - L_n||_{C\,[a,\,b]} \;\le\; \frac{||f^{(n+1)}||_{C\,[a,\,b]}}{2^n(n+1)!}\,\left(\frac{b-a}{2}\right)^{n+1}. \qquad (15.4.4)$$

Мы видим, что при использовании чебышевских узлов интерполяционное приближение может иметь существенное преимущество по точности.

15.5 Квазиоптимальность интерполяционных приближений

Теорема 15.5.1 *Пусть* $f \in C[-1,\,1]$ *и* $E_n(f) = \min\limits_{\deg\phi\le n} ||f-\phi||_C$ *— погрешность наилучшего равномерного приближения для* f *среди всех полиномов степени не выше* n*. Тогда если* $L_n(x)$ *— интерполяционный полином степени не выше* n *на чебышевской сетке, то*

$$||f - L_n||_C \;\le\; c\ln n\,E_n(f),$$

где $c > 0$ *не зависит от* n*.*

Доказательство. Пусть P_n — проектор, переводящий f в L_n, и $\phi_n(x)$ — наилучшее равномерное приближение для f среди полиномов степени не выше n. Учитывая, что $P_n\phi_n = \phi_n$, находим

$$||f - L_n||_C \le ||f-\phi_n||_C + ||P_n\phi_n - P_nf||_C \le (1 + ||P_n||)E_n(f).$$

Оценка для $||P_n||$ дается теоремой 13.8.1. $\square$

15.6 Принцип наибольших объемов

Простое условие квазиоптимальности интерполяционных приближений в многомерном случае дает *принцип наибольших объемов*. Под объемом квадратной матрицы понимается модуль ее определителя.

Пусть K — компактное множество в $\mathbb{R}^k$ и $C(K)$ — множество функций, непрерывных на K. Предположим, что для $f \in C(K)$ строится приближение вида

$$\Phi(x) = \alpha_1\phi_1(x) + ... + \alpha_n\phi_n(x),$$

где коэффициенты $\alpha_1, ..., \alpha_n$ выбираются из интерполяционных условий

$$\Phi(x_i) = f(x_i), \quad 1 \le i \le n,$$

для некоторой системы попарно различных точек $x_1, ..., x_n \in K$.

Рассмотрим матрицы $M_i(x)$, $1 \le i \le n$, полученные из матрицы

$$M = M(x_1, ..., x_n) = \begin{bmatrix} \phi_1(x_1) & ... & \phi_1(x_n) \\ ... & ... & ... \\ \phi_n(x_1) & ... & \phi_n(x_n) \end{bmatrix}$$

заменой i-го столбца на $[\phi_1(x), ..., \phi_n(x)]^\top$. Предположим, что $\det M \ne 0$. Легко проверяется, что функции $\det M_i(x)/\det M$ играют роль элементарных полиномов Лагранжа:

$$\frac{\det M_i(x_j)}{\det M} = \begin{cases} 1, & i = j, \\ 0, & i \ne j. \end{cases}$$

Поэтому

$$\Phi(x) = \sum_{i=1}^{n} f(x_i)\,\frac{\det M_i(x)}{\det M}.$$

Теорема 15.6.1 *Пусть точки $x_1, ..., x_n \in K$ выбраны таким образом, что*

$$|\det M(x_1, ..., x_n)| = \max_{z_1,...,z_n \in K} |\det M(z_1, ..., z_n)|.$$

Тогда

$$||f - \Phi_n||_{C(K)} \le (1 + n)E(f),$$

где $E(f)$ — погрешность наилучшего равномерного приближения $f(x)$ функциями вида $\alpha_1\phi_1(x) + ... + \alpha_n\phi_n(x)$.

Доказательство. Обозначим через $\Psi(x) = \beta_1\phi_1(x) + ... + \beta_n\phi_n(x)$ наилучшее равномерное приближение: $E(f) = ||f - \Psi||_{C(K)}$. Очевидно,

$$\Psi(x) = \sum_{i=1}^{n} \Psi(x_i)\frac{\det M_i(x)}{\det M} \quad \Rightarrow$$

$$|f(x)-\Phi_n(x)| \le |f(x)-\Psi(x)| + \sum_{i=1}^{n} |\Psi(x_i)-f(x_i)| \left| \frac{\det M_i(x)}{\det M} \right|$$
$$\le (1+n)||f-\Psi||_{C(K)} = E(f).$$

Важно, что по условию теоремы $|\det M_i(x)| \le |\det M| \ \ \forall\, x \in K$. $\square$

Если K — компактное множество на комплексной плоскости, то в случае полиномиальной интерполяции в точках $z_1, ..., z_n \in C$ матрица M есть матрица Вандермонда и ее объем вычисляется по формуле

$$|M(z_1, ..., z_n)| = \prod_{1 \le i < j \le n} |z_j - z_i|.$$

Точки, максимизирующие правую часть, называются *узлами Фекете*.

15.7 Метод наименьших квадратов

Пусть норма порождается скалярным произведением. В этом случае теория и методы поиска наилучшего приближения в подпространстве кажутся совершенно ясными.

Теория сводится к теореме о том, что для любого вектора f в гильбертовом пространстве (полном пространстве со скалярным произведением) и любого замкнутого подпространства Φ существует и единственно разложение

$$f \ = \ u \ + \ \phi, \quad \phi \in \Phi, \quad u \perp \Phi.$$

Если подпространство Φ конечномерно, то данный факт элементарен. Пусть $\Phi = \mathrm{span}\,\{v_1, \ldots, v_n\}$. Тогда элемент наилучшего приближения $\phi = \alpha_1 v_1 + ... + \alpha_n v_n$ можно найти, решив систему с матрицей Грама

$$\begin{bmatrix} (v_1, v_1) & \ldots & (v_n, v_1) \\ \ldots & \ldots & \ldots \\ (v_1, v_n) & \ldots & (v_n, v_n) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \ldots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} (v_1, f) \\ \ldots \\ (v_n, f) \end{bmatrix}.$$

Лучше всего иметь в Φ ортонормированный базис. Тогда матрица Грама будет единичной. Ортонормированный базис можно построить с помощью процесса ортогонализации Грама–Шмидта.

15.8 Ортогональные полиномы

Рассмотрим пространство $\mathcal{P}$ всех алгебраических вещественных полиномов со скалярным произведением

$$(f, g) \equiv \int_a^b f(x)\, g(x) w(x)\, dx, \qquad (15.8.5)$$

где $w(x)$ — неотрицательная функция с положительным интегралом по отрезку $[a,\, b]$, или, в более общем случае,

$$(f,\, g) \equiv \int_a^b f(x)\, g(x)\, dW(x), \tag{15.8.6}$$

где $W(x)$ — монотонно неубывающая функция с бесконечным числом точек роста, а интеграл понимается в смысле Стилтьеса. Функцию $w(x)$ обычно называют *весом*, или весовой функцией.

Очевидно, в каждом из случаев (15.8.5) и (15.8.6) скалярное произведение обладает следующим свойством:

$$(xu(x), v(x)) = (u(x), xv(x)), \quad u(x), v(x) \in \mathcal{P}. \tag{15.8.7}$$

Проведя процесс ортогонализации Грама–Шмидта, мы получаем последовательность полиномов $L_n(x)$ степени $n = 0, 1, \ldots$, для которых выполнены условия ортогональности

$$(L_i(x), L_j(x)) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{15.8.8}$$

Эти условия определяют $L_n(x)$ однозначно с точностью до множителя ± 1.

Имеет место очевидная явная формула для произвольных ортогональных полиномов (докажите!):

$$L_n(x) = \gamma_n \det \begin{bmatrix} h_0 & h_1 & \ldots & h_{n-1} & h_n \\ h_1 & h_2 & \ldots & h_n & h_{n+1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ h_{n-1} & h_n & \ldots & h_{2n-2} & h_{2n-1} \\ 1 & x & \ldots & x^{n-1} & x^n \end{bmatrix}, \qquad h_k = (x^k, 1). \tag{15.8.9}$$

Коэффициент γ_n определяется условием нормировки.

Часто от $L_n(x)$ переходят к полиномам $P_n(x) = c_n L_n(x)$ с более удобным условием нормировки: например, $P_n(1) = 1$, если полиномы строятся на отрезке $[-1,\, 1]$.

Замечательно, что ортогональные полиномы обладают некоторыми общими свойствами независимо от конкретного вида весовой функции и даже в общем случае скалярного произведения со свойством (15.8.7).

15.9 Трехчленные рекуррентные соотношения

Лемма 15.9.1 *Предположения (15.8.7), (15.8.8) гарантируют выполнение трехчленных соотношений*

$$xL_n(x) = \beta_{n-1} L_{n-1}(x) + \alpha_n L_n(x) + \beta_n L_{n+1}(x), \quad n = 0, 1, \ldots, \tag{15.9.10}$$

где

$$\beta_0 = 0; \qquad \beta_n \neq 0,\ n > 0.$$

Доказательство. Запишем

$$xL_n(x) = s_{n0}L_0(x) + \ldots + s_{nn}L_n(x) + s_{n\ n+1}L_{n+1}(x).$$

Тогда

$$s_{nj} = (xL_n(x), L_j) = (L_n(x), xL_j(x)) = 0 \quad \text{при} \quad j \leq n-2.$$

Положим $\alpha_n = s_{nn}$, $\beta_n = s_{n\ n+1} = (xL_n(x), L_{n+1}(x))$. Тогда, согласно (15.8.7), получаем $s_{n\ n-1} = \beta_{n-1}$ и в итоге (15.9.10). $\square$

Следствие 15.9.1

$$x \begin{bmatrix} L_0(x) \\ L_1(x) \\ \cdots \\ L_{n-2}(x) \\ L_{n-1}(x) \end{bmatrix} = \begin{bmatrix} \alpha_0 & \beta_1 & & & \\ \beta_1 & \alpha_1 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{n-2} & \alpha_{n-2} & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_{n-1} \end{bmatrix} \begin{bmatrix} L_0(x) \\ L_1(x) \\ \cdots \\ L_{n-2}(x) \\ L_{n-1}(x) \end{bmatrix} + \beta_n L_n(x) \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ 1 \end{bmatrix}. \tag{15.9.11}$$

Следствие 15.9.2 *Полином $L_n(x)$ имеет n простых корней $x_1, \ldots, x_n$, совпадающих с собственными значениями вещественной симметричной трехдиагональной матрицы из (15.9.11). Отвечающие им собственные векторы имеют вид*

$$[L_0(x_j), \ldots, L_{n-1}(x_j)]^T, \quad 1 \leq j \leq n.$$

Доказательство. Пусть T_n — симметричная трехдиагональная матрица в равенстве (15.9.11). Заметим, что $L_0(x) \neq 0$ для всех x (почему?). Поэтому очевидно, что если $L_n(x) = 0$ при $x = \lambda$, то λ будет собственным значением для T_n. Если λ — кратный корень полинома $L_n(x)$, то $L_n'(\lambda) = 0$. Дифференцируя (15.9.11) по x и затем подставляя $x = \lambda$, находим

$$T_n \begin{bmatrix} L_0'(\lambda) \\ \cdots \\ L_{n-1}'(\lambda) \end{bmatrix} = \lambda \begin{bmatrix} L_0'(\lambda) \\ \cdots \\ L_{n-1}'(\lambda) \end{bmatrix} + \begin{bmatrix} L_0(\lambda) \\ \cdots \\ L_{n-1}(\lambda) \end{bmatrix}, \quad T_n \begin{bmatrix} L_0(\lambda) \\ \cdots \\ L_{n-1}(\lambda) \end{bmatrix} = \lambda \begin{bmatrix} L_0(\lambda) \\ \cdots \\ L_{n-1}(\lambda) \end{bmatrix}.$$

Отсюда

$$\begin{bmatrix} L_0'(\lambda) \\ \cdots \\ L_{n-1}'(\lambda) \end{bmatrix} \in \ker(T_n - \lambda I)^2 = \ker(T_n - \lambda I).$$

Равенство ядер $\ker(T_n - \lambda I)^2 = \ker(T_n - \lambda I)$ справедливо для любой эрмитовой (даже для любой нормальной) матрицы.

В нашем случае $\beta_i \neq 0$ для всех $i \quad \Rightarrow \quad T_n$ имеет отличный от нуля минор порядка $n-1 \quad \Rightarrow \quad \dim\ker(T_n - \lambda I) = 1$. Поэтому для некоторого α

$$\begin{bmatrix} L_0'(\lambda) \\ \cdots \\ L_{n-1}'(\lambda) \end{bmatrix} = \alpha \begin{bmatrix} L_0(\lambda) \\ \cdots \\ L_{n-1}(\lambda) \end{bmatrix} \;\Rightarrow\; \alpha = 0 \;\Rightarrow\; \begin{bmatrix} L_0'(\lambda) \\ \cdots \\ L_{n-1}'(\lambda) \end{bmatrix} = \begin{bmatrix} L_0(\lambda) \\ \cdots \\ L_{n-1}(\lambda) \end{bmatrix} = 0,$$

что невозможно, так как $L_0(x)$ — ненулевая константа. $\square$

Важный пример: полиномы Чебышева ортогональны на отрезке $[-1,\, 1]$ с весом $w\,(x) = 1/\sqrt{1-x^2}$. Трехчленные соотношения для них имеют вид

$$xL_n(x) \;=\; \frac{1}{2}L_{n-1}(x) \,+\, 0 \cdot L_n(x) \,+\, \frac{1}{2}L_{n+1}(x).$$

Удивительный (не очень простой) факт: при весьма слабом ограничении на вес в формуле 15.8.5 трехчленные соотношения для *любых* ортогональных полиномов на отрезке $[-1,\, 1]$ асимптотически одни и те же:

$$\beta_n \to \frac{1}{2}, \quad \alpha_n \to 0.$$

Эти соотношения выполняются, если вес подчинен *условию Сеге*

$$\int\limits_{-1}^{1} \frac{\ln w(x)}{\sqrt{1-x^2}} dx \;>\; -\infty.$$

15.10 Корни ортогональных полиномов

Связь с эрмитовыми трехдиагональными матрицами дает, как минимум, разумный алгоритм вычисления корней ортогональных полиномов (например, применяем QR-алгоритм для вычисления собственных значений трехдиагональных матриц T_n). Кроме того, она позволяет установить еще одно важное свойство корней ортогональных полиномов — так называемые *соотношения разделения.*

Теорема 15.10.1 *Для корней $\lambda_1 \;>\; \ldots \;>\; \lambda_n$ полинома L_n и корней $\mu_1 \;>\; \ldots \;>\; \mu_{n-1}$ полинома L_{n-1} выполняются следующие соотношения разделения:*

$$\lambda_k \;>\; \mu_k \;>\; \lambda_{k+1}, \quad k = 1,\, \ldots\,,\, n-1. \tag{15.10.12}$$

Доказательство. Достаточно вспомнить теорему 5.7.1, в которой устанавливаются соотношения разделения для собственных значений эрмитовых матриц, и заметить, что полиномы L_n и L_{n-1} не могут иметь общих корней (в силу трехчленных соотношений общий корень должен быть корнем не имеющего корней полинома L_0). $\square$

Теорема 15.10.2 *Пусть ортогональные полиномы порождены скалярным произведением вида (15.8.5) с весовой функцией, определенной на отрезке* $[a, b]$. *Тогда при* $n \geq 1$ *все корни полинома* L_n *вещественны, попарно различны и расположены внутри отрезка* $[a,\ b]$.

Доказательство. При $n \geq 1$ запишем

$$q_n(x) \ = \ (x - \zeta_1) \ldots (x - \zeta_m)\, p_{n-m}(x),$$

где $\zeta_1, \ldots, \zeta_m$ — попарно различные корни полинома $q_n(x)$, расположенные внутри отрезка $[a,\ b]$ и имеющие нечетную кратность. Будем считать, что m — максимально взможное число корней с такими свойствами. Тогда полином $p_{n-m}(x)$ имеет один и тот же знак при всех $x \in [a,\ b]$.

Если $m \ < \ n$, то в силу ортогональности q_n ко всем полиномам меньшей степени находим

$$\int_a^b (x - \zeta_1)^2 \ldots (x - \zeta_m)^2\, p_{n-m}(x)\, w(x)\, dx \ = \ 0.$$

Это равенство невозможно (почему?). Поэтому $m \ = \ n$. $\square$

15.11 Разложение интерполяционного полинома

Ортогональные полиномы образуют удобный базис для представления интерполяционных полиномов.

Лемма 15.11.1 *Матрица*

$$Q_n = \begin{bmatrix} L_0(x_1) & \ldots & L_{n-1}(x_1) \\ \ldots & \ldots & \ldots \\ L_0(x_n) & \ldots & L_{n-1}(x_n) \end{bmatrix}$$

имеет ортогональные строки, а матрица $D_n^{-1} Q_n$, *где*

$$D_n = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix}, \qquad d_j^2 = L_0^2(x_j) + \ldots + L_{n-1}^2(x_j), \qquad (15.11.13)$$

является ортогональной.

Доказательство. Данное утверждение вытекает из ортогональности системы собственных векторов вещественной симметричной матрицы для попарно различных собственных значений. $\square$

Теорема 15.11.1 *Полином $\Pi_{n-1}(x)$, интерполирующий значения $f_1, \ldots, f_n$ в узлах $x_1, \ldots, x_n$, допускает представление*

$$\Pi_{n-1}(x) = \gamma_1 L_0(x) + \ldots + \gamma_n L_{n-1}(x), \qquad (15.11.14)$$

где

$$\begin{bmatrix} \gamma_1 \\ \ldots \\ \gamma_n \end{bmatrix} = Q_n^\top D_n^{-2} \begin{bmatrix} f_1 \\ \ldots \\ f_n \end{bmatrix}. \qquad (15.11.15)$$

Доказательство. Поскольку $(D_n^{-1} Q_n)^{-1} = (D_n^{-1} Q_n)^\top = Q_n^\top D_n^{-1}$, находим $Q_n^{-1} = Q_n^\top D_n^{-2}$, откуда и вытекает соотношение (15.11.15). $\square$

Удобство записи интерполяционного полинома в виде (15.11.14) заключается в том, что коэффициенты γ_1, ..., γ_n определяются устойчивым образом (умножением на диагональную и ортогональную матрицы). Для их вычисления достаточно $O(n^2)$ операций.

Для умножения на Q_n существуют и более быстрые алгоритмы (сложности $O(n \log^2 n)$ или даже $O(n \log n)$ в специальных случаях), но их численная устойчивость, вообще говоря, требует изучения.

15.12 Ортогональные полиномы и разложение Холецкого

Запишем $q_i(x) = q_{i0} + q_{i1} x + \ldots + q_{ii} x^i$ и рассмотрим нижнюю треугольную матрицу

$$L_n = \begin{bmatrix} q_{00} & & & \\ q_{10} & q_{11} & & \\ \ldots & \ldots & \ldots & \\ q_{n0} & q_{n1} & \ldots & q_{nn} \end{bmatrix}.$$

Соотношения ортогональности для полиномов $q_i(x)$ и $q_j(x)$ при $0 \le i, j \le n$ можно записать таким образом:

$$\int_a^b \left\{ L_n \begin{bmatrix} 1 \\ x \\ \ldots \\ x^n \end{bmatrix} [1 \;\; x \;\; x^2 \;\; \ldots \;\; x^n] \, L_n^T \right\}_{ij} w(x)\, dx = \begin{cases} 1, & i = j, \\ 0, & i \ne j. \end{cases}$$

Введем *матрицу моментов*

$$M_n \equiv [(x^i, x^j)]_{ij=0}^n.$$

Тогда $L_n M_n L_n^T = I$, и следовательно,

$$M_n = L_n^{-1} L_n^{-T}.$$

Таким образом, матрица L_n, составленная из коэффициентов ортогональных полиномов, является матрицей, обратной к нижней треугольной матрице из разложения Холецкого для матрицы моментов M_n.

Задачи

1. Докажите единственность полинома наилучшего равномерного приближения для $f \in C[a, b]$.

2. Для любой функции $f \in C[a, b]$ докажите существование полинома наилучшего равномерного приближения.

3. Пусть функция $f \in C[-1, 1]$ четная. Докажите, что полином наилучшего равномерного приближения для нее должен быть четной функцией. Верно ли, что для нечетной функции он должен быть нечетной функцией?

4. Докажите, что полиномы Чебышева — это ортогональные полиномы на отрезке $[-1, 1]$ с весом $w(x) = 1/\sqrt{1-x^2}$.

5. Докажите, что произвольные ортогональные полиномы представляются формулой (15.8.9).

6. Полиномы, ортогональные на отрезке $[-1, 1]$ с весом $w(x) = 1$, называются *полиномами Лежандра*. Докажите, для при условии нормировки $P_n(1) = 1$ справедлива *формула Родрига*

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

7. Докажите, что для полиномов Лежандра с нормировкой $P_n(1) = 1$

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

8. Найдите коэффициенты трехчленных рекуррентных соотношений для полиномов Лежандра $P_n(x)$, нормированных условием $P_n(1) = 1$.

9. Докажите, что для полиномов Лежандра с условием нормировки $P_n(1) = 1$ справедливы соотношения

$$(2n+1)P_n(x) = \frac{d}{dx}(P_{n+1}(x) - P_{n-1}(x)), \quad n = 1, 2, \ldots.$$

10. Докажите, что значения полиномов Лежандра с условием нормировки $P_n(1) = 1$ удовлетворяют неравенству $|P_n(x)| \leq 1$ при $-1 \leq x \leq 1$.

11. Полиномы, ортогональные на отрезке $[0, +\infty)$ с весом $w(x) = e^{-x}$, называются *полиномами Лагерра*. Докажите, что при условии нормировки $P_n(0) = 1$ имеет место формула

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

12. Докажите, что разложение по степеням x для полиномов Лагерра с условием нормировки $P_n(0) = 1$ имеет вид

$$P_n(x) = \frac{1}{n!}\left(x^n - \frac{n^2}{1!}x^{n-1} + \frac{n^2(n-1)^2}{2!}x^{n-2} - ... + (-1)^n n!\right).$$

13. Найдите коэффициенты трехчленных рекуррентных соотношений для полиномов Лагерра $P_n(x)$, нормированных условием $P_n(0) = 1$.

14. Докажите, что множество корней всех полиномов Лагерра не может принадлежать какому-либо конечному отрезку.

15. Функции

$$U_n(x) \;=\; \frac{\sin((n+1)\arccos x))}{\sqrt{1-x^2}}, \qquad -1 < x < 1,$$

называются *полиномами Чебышева второго рода*. Докажите, что $U_n(x)$ являются полиномами от x, ортогональными на $[-1, 1]$ с весом $w(x) = \sqrt{1-x^2}$. Получите для них трехчленные рекуррентные соотношения.

Глава 16

16.1 Численное интегрирование

Очевидная идея численного интегрирования: приблизить f "простой" функцией ϕ и положить

$$I(f) \equiv \int_a^b f(x)\,dx \approx S(f) \equiv \int_a^b \phi(x)\,dx.$$

В качестве "простых" обычно берутся функции, интегрируемые аналитически, например, полиномы или сплайны.

Обширный класс методов численного интегрирования описывается *квадратурной формулой*

$$S(f) \;=\; \sum_{i=1}^{n} d_i\, f(x_i), \tag{16.1.1}$$

определяемой *узлами* x_i и *весами* d_i. Часто веса представляют в виде $d_i \;=\; \frac{b-a}{2}\, D_i$, где D_i не зависит от a и b.

16.2 Интерполяционные квадратурные формулы

Рассмотрим стандартный отрезок $[-1,\,1]$ и отобразим его на $[a,\,b]$:

$$x \;=\; x(t) \;=\; \frac{a+b}{2} + \frac{b-a}{2}\, t.$$

Выберем узлы $t_1,\,\ldots\,,\,t_n \;\in [-1,\,1]$ и положим $x_i \;=\; x(t_i)$.

Если узлы попарно различны, построим интерполяционный полином Лагранжа

$$L_{n-1}(x) \;=\; \sum_{i=1}^{n} \prod_{\substack{j=1 \\ j\neq i}}^{n} f(x_i)\, \frac{x-x_j}{x_i-x_j}.$$

Интегрируя его по $[a,\, b]$, полагаем

$$S\,(f) \equiv \int_a^b L_{n-1}\,(x)\,d\,x \;=\; \sum_{i=1}^{n} d_i\,f\,(x_i), \quad d_i \;=\; \frac{b-a}{2}\;\int_{-1}^{1} \prod_{\substack{j=1\\ j\neq i}}^{n} \frac{t-t_j}{t_i-t_j}\,d\,t. \tag{16.2.2}$$

Такие квадратурные формулы называют *формулами Ньютона–Котеса.*

Пусть $f \;\in\; C^n\,[a,\, b]$. Тогда, используя оценку погрешности для лагранжевой интерполяции, получаем

$$|I\,(f) - S\,(f)| \;\leq\; \frac{||f^n||_{C\,[a,\,b]}}{n!}\;\left(\frac{b-a}{2}\right)^{n+1}\;\int_{-1}^{1} |\prod_{j=1}^{n}(t-t_j)|\,d\,t. \tag{16.2.3}$$

16.3 Алгебраическая точность квадратурной формулы

Если $I\,(f) = S\,(f)$ для всех полиномов f степени не выше m и $I\,(f) \neq S\,(f)$ хотя бы для одного полинома степени $m+1$, то говорят, что квадратурная формула S имеет алгебраическую точность m.

Теорема 16.3.1 *Квадратурная формула с n узлами имеет алгебраическую точность $m \geq n-1$ тогда и только тогда, когда она является интерполяционной квадратурной вида* (16.2.2).

Доказательство. Алгебраическая точность формулы (16.2.2) не меньше $n-1$ — это очевидно. Если же формула вида (16.1.1) точна для полиномов степени $n-1$, то чтобы найти d_i, достаточно с ее помощью проинтегрировать элементарный полином Лагранжа $l_i\,(x)$. $\square$

16.4 Популярные квадратурные формулы

Положим $h \equiv b-a$ и $M_m \equiv ||f^m||_{C\,[a,\,b]}$.

Формула прямоугольников $(t_1 = 0)$: $S\,(f) \;=\; f\,(\frac{a+b}{2})\,h$. Оценка погрешности (проверьте): $f \in C^1 \quad \Rightarrow \quad |I-S| \leq \frac{1}{4}\,M_1\,h^2$.

Любопытно, что эту же формулу мы можем получить, интегрируя интерполяционный полином Эрмита для кратного узла $t_1 \;=\; t_2 \;=\; 0$. Для стандартного отрезка формально имеем $H_1\,(t) = f\,(0) + f'\,(0)\,t$, но в силу нечетности член с производной дает при интегрировании нуль. Теперь мы получаем такую оценку погрешности (проверьте):

$$f \in C^2 \quad \Rightarrow \quad |I-S| \leq \frac{1}{24}\,M_2\,h^3.$$

Формула трапеций ($t_1 = -1,\ t_2 = 1$): $S(f) = \frac{1}{2}(f(a)+f(b))\,h$. Оценка погрешности (проверьте): $f \in C^2 \quad \Rightarrow \quad |I - S| \le \frac{1}{12} M_2 h^3$.

Формула Симпсона: $t_1 = -1,\ t_2 = 1,\ t_3 = t_4 = 0$. Для стандартного отрезка полином Эрмита имеет вид

$$H_3(t) = f(-1) + f(-1;\,1)(t+1) + f(-1;\,1;\,0)(t+1)(t-1) + f(-1;\,1;\,0;\,0)(t+1)(t-1)t.$$

В силу нечетности последний член при интегрировании дает нуль.

Доведите построение до конца: найдите веса и оцените погрешность квадратурной формулы Симпсона.

16.5 Формулы Гаусса

При заданном числе узлов n попытаемся найти квадратурную формулу вида (16.1.1) с максимально возможной алгебраической точностью m. Такие формулы называются *формулами Гаусса*.

Теорема 16.5.1 *Для любого числа узлов n квадратурная формула Гаусса существует, единственна и имеет алгебраическую точность $2n-1$.*

Доказательство. Положим

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

Если бы существовала формула с алгебраической точностью $2n$, то мы имели бы равенство $I(\omega_n^2) = S(\omega_n^2) = 0$, что невозможно $\Rightarrow\ m \le 2n-1$.

Предположим, что формула (16.1.1) имеет алгебраическую точность $m = 2n-1$. Тогда

$$I(\omega_n(x)\,r_{n-1}(x)) = S(\omega_n(x)\,r_{n-1}(x)) = 0$$

для любого полинома $r_{n-1}(x)$ степени не выше $n-1$. Следовательно, полином $\omega_n(x)$ есть n-й полином из последовательности ортогональных полиномов на отрезке $[a,\,b]$ с весом 1. Мы знаем, что такой полином определяется однозначно с точностью до нормировки. Мы знаем также, что такой полином должен иметь n попарно различных корней внутри $[a,\,b]$. Эти корни мы и будем использовать в качестве узлов x_i. В силу теоремы 16.3.1 искомая квадратурная формула является интерполяционной. Поэтому она имеет вид (16.2.2).

Докажем, что полученная квадратурная формула действительно имеет алгебраическую точность $m = 2n - 1$. Возьмем произвольный полином $p_{2n-1}(x)$ и разделим его с остатком на $\omega_n(x)$:

$$p_{2n-1}(x) \;=\; q_{n-1}(x)\,\omega_n(x) + r_{n-1}(x).$$

В силу линейности, ортогональности и вида полученной квадратуры

$$\begin{aligned} I(p_{2n-1}) \;&=\; I(q_{n-1}\,\omega_n) + I(r_{n-1}) \;=\; I(r_{n-1}) \\ &=\; S(r_{n-1}) \;=\; S(q_{n-1}\,\omega_n) + S(r_{n-1}) \;=\; S(p_{2n-1}). \quad \square \end{aligned}$$

16.6 Составные квадратурные формулы

Рассмотренные нами квадратурные формулы дают приемлемую погрешность при небольших $h = b - a$. В общем случае применяют *составные* квадратурные формулы: разбивают отрезок интегрирования на какое-то число элементарных отрезков, на каждом из них вычисляют значение "элементарной" квадратуры, а интеграл приближают их суммой.

Оценки погрешности для составных квадратур легко вытекают из оценок для "элементарных" квадратур. Например, если отрезок $[a, b]$ разбивается на какое-то число отрезков длины h и на каждом из них используется формула трапеций, то погрешность такой составной квадратуры имеет вид $\mathcal{O}\left(\frac{b-a}{h}\,h^3\right) \;=\; \mathcal{O}(h^2)$.

Полезно обратить внимание на то, что при интегрировании по периоду гладких периодических функций составная формула прямоугольников, несмотря на простоту, превосходит многие более сложные квадратурные правила: по n узлам она *точно* интегрирует тригонометрические полиномы степени меньше $n/2$.

16.7 Правило Рунге для оценки погрешности

При разбиении отрезка интегрирования на элементарные отрезки важно учитывать поведение интегрируемой функции. Если о функции заранее ничего не известно, то мы можем проводить разбиение "постепенно", двигаясь, например, слева направо. Для очередного элементарного отрезка длины h мы должны каким-то образом оценить погрешность интегрирования на нем и принять решение об уменьшении или увеличениии шага интегрирования.

"Наивный" способ оценки погрешности: взять две разные квадратурные формулы S_1 и S_2 и судить о погрешности по разности значений $S_1 - S_2$. Иногда можно провести более аккуратное оценивание.

Пусть на отрезке длины h используется квадратурная формула S_1, точная для полиномов степени не выше $n-1$. Разложим функцию $f(x)$ в ряд Тейлора в середине отрезка . Тогда

$$I(f) - S_1(f) = \alpha\, f^{(n)}(c)\, h^{n+1} + \mathcal{O}(h^{n+2}).$$

Обозначим через S_2 составную формулу, полученную применением формулы S_1 для двух половинок отрезка длины h. Тогда с тем же α находим:

$$I(f) - S_2(f) = \alpha\, f^{(n)}(c)\, \frac{h^{n+1}}{2^n} + \mathcal{O}(h^{n+2})$$

(докажите!). Следовательно, с точностью до членов $\mathcal{O}(h^{n+2})$ получаем следующее *правило Рунге*:

$$I(f) - S_2(f) \approx \frac{S_2 - S_1}{2^n - 1}. \tag{16.7.4}$$

Если мы хотим найти интеграл с точностью ε, то каждый шаг следует выбирать таким образом, чтобы выполнялось неравенство

$$\frac{|S_1 - S_2|}{2^n - 1} \;\le\; \frac{h}{b-a}\,\varepsilon.$$

16.8 Как интегрировать "плохие" функции

Рассмотренные выше квадратурные формулы не очень хороши для недостаточно гладких функций. Даже в тех случаях, когда интегрирование с автоматическим выбором шага позволяет получить ответ, оно может требовать очень большого количества вычислений. Отметим два основных подхода к численному интегрированию "плохих" функций.

- Записать $f = w + g$, где w — "плохая" функция; проинтегрировать w отдельно (лучше всего, аналитически), а квадратуру использовать для более гладкой функции g.
- Записать $f = wg$ и попытаться получить квадратурные формулы для фиксированной "плохой" функции w. Функция g предполагается уже достаточно гладкой. Она приближается полиномом p (например, интерполяционным полиномом), и чтобы получить метод интегрирования, остается предъявить достаточно точный способ вычисления интегралов от функций вида $w\,p$. Если функция w знакопостоянна, то ее можно рассматривать как вес, и следовательно, можно строить формулы типа формул Гаусса.

Конечно, есть и другие рецепты. Например, если функция имеет в нуле особенность вида x^α, где $0 < \alpha < \frac{1}{2}$, то можно перейти к более гладкой функции с помощью замены переменной $x = y^m$, где $m \geq 2$.

16.9 Интегралы от быстроосциллирующих функций

Для вычисления интегралов от быстроосциллирующих функций вида

$$A(p) = \int_{-1}^{1} f(x)\cos(px)\ dx, \quad B(p) = \int_{-1}^{1} f(x)\sin(px)\ dx \tag{16.9.5}$$

естественно приблизить $f(x)$ интерполяционным полиномом, а затем проинтегрировать произведение полинома и тригонометрической функции аналитически. В случае полинома первой степени получается так называемая *формула Филона*.

Однако, для полиномов произвольной степени необходимо иметь способ аналитического интегрирования, избегающий вычисления коэффициентов интерполяционного полинома (почему?). Это можно сделать с помощью теоремы 15.11.1 о разложении интерполяционного полинома по какой-либо системе ортогональных полиномов. [1]

16.10 Применение полиномов Лежандра

Для получения алгоритмов интегрирования быстросциллирующих функций удобной оказывается система *полиномов Лежандра* —- полиномов, ортогональных на $[-1, 1]$ с весом $w(x) = 1$. Пусть $P_n(x)$ — полиномы с условием нормировки $P_n(1) = 1$.

Лемма 16.10.1 *Пусть q — отличное от нуля комплексное число. Тогда для величин вида*

$$\Phi_n = \int_{-1}^{1} P_n(x)\mathrm{e}^{-qx}\ dx,$$

имеют место рекуррентные соотношения

$$\Phi_{n+1} = \frac{2n+1}{q}\Phi_n + \Phi_{n-1}, \quad n = 1, 2, \ldots. \tag{16.10.6}$$

[1] Е. Е. Тыртышников, Модификации методов вычисления интегралов Чебышева–Лагерра и Гаусса–Лежандра, *ЖВМ и МФ*, том 44, N 7, 1187–1195 (2004).

Доказательство. Для полиномов $P_n(x)$ известно, что $P'_{n+1}(x) = (2n+1)P_n(x) + P'_{n-1}(x), \quad n = 1, 2, \ldots$. Учитывая также, что $P_n(1) = 1$ и $P_n(-1) = (-1)^n$, отсюда получаем

$$\Phi_{n+1} = \int_{-1}^{1} P_{n+1}(x)\mathrm{e}^{-qx}\,dx = -\frac{\mathrm{e}^{-qx}}{q}P_{n+1}(x)\Big|_{-1}^{1} + \frac{1}{q}\int_{-1}^{1}\mathrm{e}^{-qx}T'_{n+1}(x)\,dx =$$

$$-\frac{\mathrm{e}^{-qx}}{q}P_{n+1}(x)\Big|_{-1}^{1} + \frac{2n+1}{q}\int_{-1}^{1}P_n(x)\mathrm{e}^{-qx}\,dx + \frac{1}{q}\int_{-1}^{1}\mathrm{e}^{-qx}T'_{n-1}(x)\,dx =$$

$$-\frac{\mathrm{e}^{-qx}}{q}P_{n+1}(x)\Big|_{-1}^{1} + \frac{2n+1}{q}\Phi_n + \frac{\mathrm{e}^{-qx}}{q}P_{n-1}(x)\Big|_{-1}^{1} + \Phi_{n-1} = \frac{2n+1}{q}\Phi_n + \Phi_{n-1}. \quad \square$$

Можно доказать, что $\lim\limits_{n\to\infty}\Phi_n = 0$. Но несмотря на это, прямое вычисление по формулам (16.10.6) будет численно неустойчивым. Чтобы получить устойчивый алгоритм вычисления значений $\Phi_0, \ldots, \Phi_n$, возьмем достаточно большое $N > n$, положим

$$a_j = \frac{2j+1}{q}, \quad j = 1, \ldots\ N,$$

и запишем соотношения (16.10.6) в матрично-векторном виде:

$$\begin{bmatrix} a_1 & -1 & & & \\ 1 & a_2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & a_{N-1} & -1 \\ & & & 1 & a_N \end{bmatrix} \begin{bmatrix} \Phi_1 \\ \Phi_2 \\ \ldots \\ \Phi_{N-1} \\ \Phi_N \end{bmatrix} = \begin{bmatrix} -\Phi_0 \\ 0 \\ \ldots \\ 0 \\ \Phi_{N+1} \end{bmatrix} \approx \begin{bmatrix} -\Phi_0 \\ 0 \\ \ldots \\ 0 \\ 0 \end{bmatrix}.$$

Очевидная вариация метода исключения элементов позволяет получить следующее разложение:

$$\begin{bmatrix} a_1 & -1 & & & \\ 1 & a_2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & a_{n-1} & -1 \\ & & & 1 & a_n \end{bmatrix} = \begin{bmatrix} 1 & -\beta_1 & & & \\ & 1 & -\beta_2 & & \\ & & \ddots & \ddots & \\ & & & 1 & -\beta_{n-1} \\ & & & & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 & & & & \\ 1 & \alpha_2 & & & \\ & \ddots & \ddots & & \\ & & 1 & \alpha_{n-1} & \\ & & & 1 & \alpha_n \end{bmatrix}.$$

Расчетные формулы имеют вид

$$\alpha_n = a_n; \quad \beta_j = \frac{1}{\alpha_{j+1}}, \quad \alpha_j = a_j + \beta_j, \qquad j = n-1, n-2, \ldots, 1. \tag{16.10.7}$$

$$\Phi_j = -\Phi_{j-1}/\alpha_j, \quad j = 1, 2, \ldots, n.$$

В отличие от исходных трехчленных соотношений (16.10.6), вычисления по формулам (16.10.7) оказываются численно устойчивыми и позволяют найти $\Phi_0, \ldots, \Phi_n$ с нужной точностью при подходящем выборе $N > n$ (хорошие результаты получаются, например, при $N \geq n + |p| + 10$).

Задачи

1. Пусть имеется последовательность квадратурных формул

$$S_n(f) \;=\; \sum_{i=1}^{n} d_{ni}\, f(x_{ni}), \quad x_{ni} \;\in\; [a,\, b].$$

   Докажите, что если $\sum\limits_{i=1}^{n} |d_{ni}| \to \infty$ при $n \to \infty$, то существует функция $f \in C[a,\, b]$, для которой $S_n(f)$ не сходится к интегралу от f по $[a,\, b]$.

2. Пусть формула Ньютона–Котеса с нечетным числом узлов n используется для вычисления интеграла по отрезку $[a,\, b]$ от функции $f \in C^{n+1}[a,\, b]$. Докажите, что погрешность по абсолютной величине не превосходит $c\,||f^{n+1}||_{C[a,\,b]}\, h^{n+2}$, где $c > 0$ не зависит от f или $h = b-a$.

3. Докажите, что веса в квадратурной формуле Гаусса положительны.

4. Пусть формула Гаусса с n узлами применяется к $f \in C^{2n}[a,\, b]$. Докажите, что погрешность по абсолютной величине не превосходит $c\,||f^{2n}||_{C[a,\,b]}\, h^{2n+1}$, где $c > 0$ не зависит от f и $h = b - a$.

5. Оцените погрешность для составной квадратурной формулы, использующей на каждом элементарном отрезке длины h формулу Симпсона.

6. Пусть на отрезке $[a,\, b]$ рассматривается последовательность S_n составных формул трапеций с шагом $h = (b-a)/n$. Докажите, что если $f \in C^4[a,\, b]$, то

$$\int_a^b f(x)\, dx \;=\; S_n(f) - \frac{1}{12}(f''(b) - f''(a))\, h^2 + \mathcal{O}(h^4).$$

Глава 17

17.1 Нелинейные уравнения

Итерационные методы вычисления *изолированного* (отделенного от других корней) корня z уравнения $f(x) = 0$ обычно требуют указания какой-либо области D (желательно, малой), локализующей z.

Если f — непрерывная функция, то z принадлежит любому отрезку, на концах которого $f(x)$ имеет разные знаки. Деля отрезок пополам, мы получаем универсальный метод вычисления корня. Он будет работать, если корень локализован на некотором (возможно, большом) отрезке, на концах которого $f(x)$ имеет разные знаки.

Метод деления пополам замечателен тем, что не требует хорошего приближения к корню. Если таковое имеется, то для гладких функций есть смысл переключиться на методы с более высокой скоростью сходимости.

Мы хотим приблизить z с точностью ε и не хотим итерировать слишком долго. Когда останавливать итерации?

Обратим внимание на то, что малость $f(x_k)$ является весьма сомнительным критерием остановки (почему?). Если доступно вычисление производной, то более разумным критерием остановки может служить неравенство ($y = x_k$ или $y \approx x_k$)

$$|f(x_k)/f'(y)| \le \varepsilon. \qquad (*)$$

Утверждение. *Если производная непрерывна и неравенство* $(*)$ *выполняется для всех* $y \in [x_k - \varepsilon,\ x_k + \varepsilon]$, *то* $f(z) = 0$ *для некоторого* $z \in [x_k - \varepsilon,\ x_k + \varepsilon]$.

Доказательство. По теореме Лагранжа $f(x_k + t) = f(x_k) + f'(y)t \quad \Rightarrow$ $f(x_k+t)/f'(y) = f(x_k)/f'(y)+t \ge 0$ при $t = \varepsilon$ и ≤ 0 при $t = -\varepsilon$. Поскольку производная не меняет знак на $[-\varepsilon, \varepsilon]$, приходим к выводу о том, что если оба значения $f(x_k + \varepsilon)$ и $f(x_k - \varepsilon)$ отличны от нуля, то они имеют разные знаки. $\square$

На практике неравенство $(*)$ проверяется лишь при $y = x_k$, а такой критерий также не идеален: пусть $g(t) = f(\alpha t)$; тогда при $t_k = x_k/\alpha$

отношение $|g(t_k)/g'(t_k)|$ может быть сделано сколь угодно малым за счет выбора α (независимо от того, насколько t_k близко к решению уравнения $g(t) = 0$). Конечно, в этом случае неравенство $(*)$ нарушается для y в ε-окрестности точки x_k.

17.2 Метод простой итерации

Перепишем уравнение $f(x) = 0$ в виде $x = F(x)$ (например, можно взять $F(x) = x - f(x)$), выберем начальное приближение x_0 и рассмотрим *метод простой итерации*

$$x_{k+1} \;=\; F(x_k), \quad k = 0, 1, \ldots \,.$$

Решение z уравнения $z = F(z)$ называется *неподвижной точкой* отображения F.

Теорема 17.2.1 *Пусть M — полное метрическое пространство с расстоянием ρ и отображение $F : M \to M$ является сжимающим:*

$$\rho(F(x), F(y)) \;\le\; q\,\rho(x, y) \quad \forall\; x, y \;\in\; M, \tag{17.2.1}$$

где $0 < q < 1$ не зависит от x и y. Тогда уравнение $x = F(x)$ имеет решение z, это решение единственно и для любого начального приближения x_0 метод простой итерации сходится к z со скоростью геометрической прогрессии:

$$\rho(x_k, z) \;\le\; \frac{q^k}{1-q}\,\rho(x_1, x_0). \tag{17.2.2}$$

Доказательство. При $m \ge k$ находим

$$\rho(x_m, x_k) \;\le\; \sum_{i=k}^{m-1} \rho(x_{i+1}, x_i) \;\le\; \sum_{i=k}^{m-1} q^i\,\rho(x_1, x_0) \;\le\; \frac{q^k}{1-q}\,\rho(x_1, x_0).$$

$\Rightarrow$ x_k — последовательность Коши $\Rightarrow$ в силу полноты M она сходится к некоторому $z \in M$. Ясно, что $F(z) = z$ (почему?). Переходом к пределу при $m \to \infty$ получаем (17.2.2). $\square$

17.3 Сходимость и расходимость метода простой итерации

Пусть $F \in C^1[z-\delta, z+\delta]$, где z — единственная неподвижная точка для F. Тогда если $|F'(z)| < 1$, то при некотором $\delta > 0$ будем иметь

$$q \equiv \max_{|x-z|\le\delta} |F'(x)| < 1 \quad \Rightarrow \quad |F(x)-F(y)| \le q\,|x-y| \;\forall\; x, y \in [z-\delta, z+\delta].$$

В данном случае F — сжимающее отображение на полном метрическом пространстве $M = [z - \delta,\, z + \delta]$. Поэтому метод простой итерации будет сходиться для любого начального приближения $x_0 \in M$.

Если $|F'(z)| > 1$, то метод простой итерации расходится для любого начального приюближения $x_0 \neq z$ (докажите!).

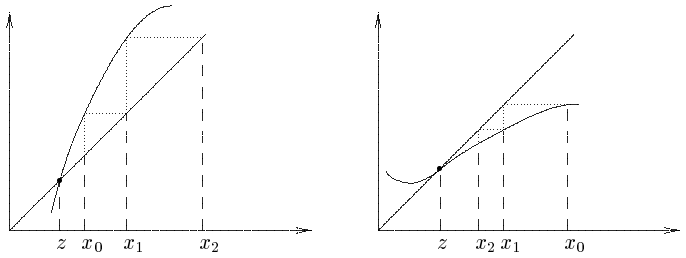


Рисунок 17.1. Расходимость и сходимость метода простой итерации.

Аналогичные утверждения можно получить и в многомерном случае. Пусть $F : \mathbb{R}^n \to \mathbb{R}^n$ — непрерывно дифференцируемое отображение в окрестности единственной неподвижной точки $z = F(z)$. Запишем

$$F(x) = [f_1(x), \ldots, f_n(x)]^T, \quad x = [x_1, \ldots, x_n]^T,$$

и рассмотрим матрицу

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

Матрица $F'(x)$ называется *якобианом* отображения F в точке x.

Непрерывная дифференцируемость отображения F в точке x означает существование и непрерывность элементов якобиана (всех частных производных) в точке x.

Теорема 17.3.1 *Пусть отображение $F : \mathbb{R}^n \to \mathbb{R}^n$ имеет единственную неподвижную точку $z = F(z)$ и непрерывно дифференцируемо в некоторой ее окрестности. Тогда если спектральный радиус якобиана $F'(z)$ меньше 1, то для всех начальных приближений x_0 из некоторой окрестности точки z метод простой итерации сходится к z.*

Доказательство проводится по аналогии с одномерным случаем.

17.4 Оптимизация метода простой итерации

От уравнения $f(x) = 0$ к равносильному уравнению $x = F(x)$ можно перейти многими способами. Например, так:

$$F(x) = x - \alpha(x) f(x), \quad \text{где} \quad \alpha(x) \neq 0 \; \forall \; x.$$

В частности, α может быть любой ненулевой константой. Если z — искомый изолированный корень и $f'(z) \neq 0$, то α всегда можно выбрать так, чтобы выполнялось достаточное условие сходимости метода простой итерации:

$$|F'(z)| \; = \; |1 - \alpha f'(z)| \; < \; 1.$$

Чтобы ускорить сходимость, нужно уменьшить значение $|F'(z)|$. Нулевое значение получается при $\alpha = 1/f'(z)$. Однако, близкое значение может дать выбор

$$\alpha \; = \; \frac{1}{f'(x_k)} \approx \frac{1}{f'(z)}.$$

В итоге возникает *метод Ньютона*:

$$x_{k+1} \; = \; x_k - \frac{f(x_k)}{f'(x_k)}. \tag{17.4.3}$$

Это не что иное, как метод простой итерации для функции

$$F(x) \; = \; x - \frac{f(x)}{f'(x)}.$$

17.5 Метод Ньютона и эрмитова интерполяция

Другое название метода Ньютона — *метод касательных* — связано с очевидной геометрической интерпретацией метода. Мы будем опираться на еще одну — интерполяционную интерпретацию метода Ньютона.

Имея x_k, определим x_{k+1} как единственный корень интерполяционного полинома Эрмита $H(x) \; = \; f(x_k) + f'(x_k)(x - x_k)$. Легко проверить, что он выражается формулой (17.4.3).

Будем считать, что

$$f \in C^2 \quad \text{и} \quad f'(z) \neq 0, \tag{17.5.4}$$

и рассмотрим следующие два равенства:

$$f(z) - H(z) \; = \; \tfrac{f''(\xi_k)}{2}(z - x_k)^2 \qquad \text{(погрешность эрмитовой интерполяции)},$$

$$H(x_{k+1}) - H(z) \; = \; f'(x_k)(x_{k+1} - z) \qquad \text{(тождество Лагранжа)}.$$

Отсюда получаем ($f(z) = H(x_{k+1}) = 0$)

$$e_{k+1} \;=\; -\frac{f''(\xi_k)}{2\,f'(x_k)}\,e_k^2, \quad e_k \equiv z - x_k. \tag{17.5.5}$$

17.6 Сходимость метода Ньютона

Важный вывод: если f удовлетворяет условиям (17.5.4) и метод метод Ньютона для f сходится , то он сходится *квадратично.*

По определению, последовательность x_k сходится к z с порядком p, если

$$\limsup_{k\,\to\,\infty} \left|\,\frac{e_{k+1}}{e_k^p}\,\right| \le c \;<\; +\infty.$$

При $p = 1$ сходимость называется *линейной*; при $p \;>\; 1$ — *сверхлинейной*; при $p = 2$ — *квадратичной.*

Условие $f'(z) \neq 0$ означает, что корень z является *простым.* В общем случае, корень z называется *корнем кратности* m, если $f^{(j)}(z) = 0$ при $0 \le j \le m-1$ и $f^{(m)}(z) \neq 0$. Метод Ньютона может сходиться и для кратного корня, но сходимость не обязана быть квадратичной. Например, для $f(x) = x^2$ имеем $e_{k+1} = e_k/2$, то есть сходимость линейная.

Теорема 17.6.1 *Пусть z — простой корень уравнения $f(x) = 0$, и предположим, что*

$$f \;\in\; C^2[z-\delta,\, z+\delta], \qquad f'(x) \neq 0 \quad \textit{при} \quad x \;\in\; [z-\delta,\, z+\delta],$$

$$\gamma \equiv \max_{|x-z|\,\le\,\delta} |f''(x)| \;/\; \min_{|x-z|\,\le\,\delta} |f'(x)| \;\neq\; 0.$$

Фиксируем любое $0 < \varepsilon < \min\{\delta,\, \gamma^{-1}\}$. Тогда метод Ньютона сходится для любого начального приближения $x_0 \in [z-\varepsilon,\, z+\varepsilon]$ и для всех k выполняются следующие неравенства:

$$(a) \quad |e_{k+1}| \;\le\; \gamma\,|e_k|^2; \qquad (b) \quad |e_k| \;\le\; \gamma^{-1}\,(\gamma\,|e_0|)^{2^{2^k}}.$$

Доказательство. Пусть $x_k \in [z-\varepsilon,\, z+\varepsilon]$. Тогда в силу (17.5.5) и определения γ получаем

$$|e_{k+1}| \;\le\; \gamma\,|e_k|^2 \;\le\; (\gamma\,\varepsilon)\,\varepsilon \;\le\; \varepsilon \quad \Rightarrow \quad x_{k+1} \in [z-\varepsilon,\, z+\varepsilon].$$

Итак, (a) имеет место для всех k. Умножив обе его части на γ и положив $d_k \equiv \gamma\,|e_{k+1}|$), находим $d_{k+1} \le d_k^2 \quad \Rightarrow \quad d_k \le d_0^{2^k}$. Согласно условиям на начальное приближение, $d_0 < 1 \quad \Rightarrow \quad e_k \to 0$. $\square$

Следствие 17.6.1 *В условиях теоремы* 17.6.1

$$\lim_{k\,\to\,\infty} \frac{e_{k+1}}{e_k^2} \;=\; -\frac{f''(z)}{2\,f'(z)}.$$

17.7 Всюду Ньютон

Метод Ньютона применяется чаще, чем можно бы думать: он помогает делить числа на компьютере. Операция $c = a/b$ обычно выполняется в два приема:

$$(1) \quad z = 1/b, \quad (2) \quad c = a \cdot z.$$

При этом z вычисляется по методу Ньютона — как корень уравнения

$$\frac{1}{x\,a} - 1 \;=\; 0.$$

Важно, что берется именно такое уравнение — для него ньютоновы итерации не требуют операций деления:

$$x_{k+1} \;=\; x_k - \frac{(\frac{1}{x_k\,a} - 1)}{-\frac{1}{a\,x_k^2}} \;=\; 2\,x_k - a\,x_k^2.$$

17.8 Многомерное обобщение

Метод Ньютона легко перенести на случай, когда требуется решить систему нелинейных уравнений вида

$$\left\{ \begin{array}{rcl} f_1(x_1, \ldots, x_n) & = & 0, \\ & \ldots & \\ f_n(x_1, \ldots, x_n) & = & 0. \end{array} \right. \quad \Leftrightarrow \quad \begin{array}{c} f(x) \;=\; 0, \\ x,\; f(x) \;\in\; \mathbb{R}^n, \\ f : \mathbb{R}^n \to \mathbb{R}^n. \end{array}$$

В этом случае $1/f'(x_k)$ заменяется матрицей, обратной к якобиану отображения f в точке x_k:

$$x_{k+1} \;=\; x_k - [f'(x_k)]^{-1}\, f(x_k). \tag{17.8.6}$$

Теорема 17.8.1 *Пусть z — решение уравнения $f(x) = 0$, и предположим, что в замкнутом шаре $\{x : \; ||x - z||_\infty \le \delta\}$ якобиан отображения $f : \mathbb{R}^n \to \mathbb{R}^n$ существует, является невырожденным, и для любых x, y из этого шара*

$$||f'(x) - f'(y)||_\infty \;\le\; c\,||x - y||_\infty, \quad c > 0 \quad \text{(условие Липшица)}.$$

Пусть $\gamma \equiv c \max\limits_{||z-x||_\infty \le \delta} ||[f'(x)]^{-1}||_\infty$ и $0 < \varepsilon < \min\{\delta,\, \gamma^{-1}\}$. Тогда для любого начального приближения $x_0 \in \{x : \; ||x - z||_\infty \le \varepsilon\}$ метод Ньютона сходится и для погрешностей $e_k \equiv z - x_k$ выполняются следующие неравенства:

$$(a) \quad ||e_{k+1}||_\infty \;\le\; \gamma\,||e_k||_\infty^2; \qquad (b) \quad ||e_k||_\infty \;\le\; \gamma^{-1}\,(\gamma\,||e_0||_\infty)^{2^{2^k}}.$$

Доказательство. Если якобиан непрерывен на отрезке, соединяющем точки x, $y \in \mathbb{R}^n$, то в силу тождества Лагранжа на этом отрезке найдутся точки $\xi_1, \ldots, \xi_n$ такие, что

$$f(x) - f(y) = J_k(x-y), \quad J_k = \begin{bmatrix} \frac{\partial f_1(\xi_1)}{\partial x_1} & \ldots & \frac{\partial f_n(\xi_1)}{\partial x_n} \\ \ldots & \ldots & \ldots \\ \frac{\partial f_1(\xi_n)}{\partial x_1} & \ldots & \frac{\partial f_n(\xi_n)}{\partial x_n} \end{bmatrix} \quad \Rightarrow$$

$$\begin{aligned} e_{k+1} &= e_k - [f'(x_k)]^{-1}(f(z) - f(x_k)) \\ &= e_k - [f'(x_k)]^{-1} J_k e_k \\ &= [f'(x_k)]^{-1}(f'(x_k) - J_k)\, e_k \quad \Rightarrow (a). \end{aligned}$$

Согласно условию Липшица,

$$||f'(x_k) - J_k||_\infty \le c \max_{1\le j\le n} ||x_k - \xi_j||_\infty \le c\,||x_k - z||_\infty.$$

Следовательно, если $||z - x_k||_\infty \le \varepsilon$, то $||z - x_{k+1}||_\infty \le (\gamma\,\varepsilon)\,\varepsilon \le \varepsilon$. Зная, что (a) имеет место для каждого k, мы немедленно получаем (b). $\square$

17.9 Прямая и обратная интерполяция

Интерполяционная трактовка метода Ньютона интересна тем, что подсказывает общий метод построения итерационных алгоритмов.

Прямая интерполяция. Имея точки x_k, x_{k-1}, $\ldots$, x_{k-m}, мы можем построить интерполяционный полином $L(x)$ степени m и в качестве x_{k+1} взять один из его корней. В случае попарно различных точек это будет полином Лагранжа, в случае кратных точек (как в методе Ньютона) — полином Эрмита. Чтобы успешно реализовать эту идею, нужно иметь хороший метод отбора "подходящего" корня полинома $L(x)$.

Обратная интерполяция. Имея точки x_k, x_{k-1}, $\ldots$, x_{k-m} и значения y_k, y_{k-1}, $\ldots$, y_{k-m} функции $f(x)$ в этих точках, мы можем построить полином $P(y)$ степени m, интерполирующий в точках y_k, y_{k-1}, $\ldots$, y_{k-m} значения обратной функции $f^{-1}(y)$ (т.е. x_k, x_{k-1}, $\ldots$, x_{k-m}). После этого естественно взять $x_{k+1} = P(0)$. В случае попарно различных значений y_k, y_{k-1}, $\ldots$, y_{k-m} строится полином Лагранжа, в противном случае — полином Эрмита.

Известное нам представление погрешности лагранжевой и эрмитовой интерполяции дает естественную основу для анализа таких методов. Теоретически можно получить метод с любым порядком сходимости.

17.10 Метод секущих

Прямая и обратная лагранжева интерполяция полиномом степени 1 приводят к одному и тому же методу — *методу секущих*.

Имея разные точки x_{k-1} и x_k, строим интерполяционный полином Лагранжа

$$L(x) = f(x_{k-1}) \frac{x - x_k}{x_{k-1} - x_k} + f(x_k) \frac{x - x_{k-1}}{x_k - x_{k-1}}$$

и находим его единственный корень:

$$x_{k+1} = x_k - f(x_k) \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}. \tag{17.10.7}$$

В то же время, имея разные значения $f(x_{k-1})$ и $f(x_k)$ и интерполируя обратную функцию, находим

$$P(y) = x_{k-1} \frac{y - f(x_k)}{f(x_{k-1}) - f(x_k)} + x_k \frac{y - f(x_{k-1})}{f(x_k) - f(x_{k-1})}.$$

Легко проверить, что $P(0) = x_{k+1}$.

Пусть z — искомый корень и $e_k \equiv z - x_k$. Пусть $f \in C^2$. Для одного шага метода секущих получаем такие соотношения:

$$\begin{aligned} f(z) - L(z) &= \frac{f''(\xi_k)}{2}(z - x_k)(z - x_{k-1}), \\ L(x_{k+1}) - L(z) &= \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}(x_{k+1} - z) = f'(\zeta_k)(x_{k+1} - z). \quad \Rightarrow \end{aligned}$$

$$e_{k+1} = -\frac{f''(\xi_k)}{2 f'(\zeta_k)} e_k e_{k-1}. \tag{17.10.8}$$

17.11 Что лучше: метод секущих или метод Ньютона?

Пусть $f \in C^2$, $f'(z) \neq 0$, и предположим, что метод секущих сходится. Тогда, согласно (17.10.8), для некоторого $\gamma > 0$ будем иметь

$$|e_{k+1}| \leq \gamma |e_k| |e_{k-1}|.$$

Введем величины $d_k \equiv \gamma |e_k|$ и предположим, что $d_0 \leq d < 1$, $d_1 \leq d < 1$. Тогда $d_2 \leq d_1 d_0 \leq d^2$, $d_3 \leq d_2 d_1 \leq d^5$, и так далее. В общем случае, очевидно,

$$d_k \leq d^{\phi_k}, \tag{17.11.9}$$

где

$$\begin{aligned} &\phi_0 = \phi_1 = 1; \\ &\phi_k = \phi_{k-1} + \phi_{k-2}, \quad k = 2, 3, \ldots . \end{aligned} \tag{17.11.10}$$

Числа ϕ_k, определяемые рекуррентным соотношением (17.11.10), называются *числами Фибоначчи*. Легко проверяется, что

$$\phi_k \;=\; \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^{k+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{k+1}\right) \quad\Rightarrow\quad \phi_k \;=\; O\left(\left(\frac{1+\sqrt{5}}{2}\right)^{k}\right).$$

Для погрешностей в методе Ньютона справедлива оценка вида d^{2^k}. Поскольку

$$\frac{1+\sqrt{5}}{2} \approx 1.618 \;<\; 2,$$

приходим к выводу, что метод Ньютона сходится быстрее. Однако, на одной итерации он требует "двойной" работы: нужно вычислить значение функции и значение производной. Поэтому с точки зрения общих вычислительных затрат метод Ньютона уступает методу секущих.

Задачи

1. Пусть $F \in C^1[z-\delta,\, z+\delta]$, где z — единственная неподвижная точка для F. Может ли метод простой итерации сходиться к z, если $|F'(z)| \;=\; 1$? Может ли он расходиться в этом случае?

2. Пусть отображение $F : \mathbb{R}^n \;\to\; \mathbb{R}^n$ имеет единственную неподвижную точку $z = F(z)$ и непрерывно дифференцируемо в некоторой ее окрестности. Докажите, что если все собственные значения его якобиана в точке z по модулю больше 1, то метод простой итерации расходится.

3. Отображение $F : \mathbb{R}^n \;\to\; \mathbb{R}^n$ имеет единственную неподвижную точку $z = F(z)$ и непрерывно дифференцируемо в некоторой ее окрестности. Известно, что хотя бы одно собственное значение якобиана $F'(z)$ по модулю больше 1. Может ли метод простой итерации сходиться для всех начальных приближений x_0, достаточно близких к z?

4. Выяснить сходимость метода простой итерации при различных начальных приближениях для следующих уравнений:

$$(1)\quad x \;=\; e^{2x} - 1; \quad (2)\quad x + \ln x \;=\; \frac{1}{2}; \quad (3)\quad x \;=\; \operatorname{tg} x.$$

5. В полном метрическом пространстве M с расстоянием $\rho(x,y)$ задано отображение $f : M \to M$ с неподвижной точкой z и свойством

$$\rho(f(x), f(y)) \le \rho(x,y)/(1+\rho(x,y)) \;\; \forall\, x, y \in M.$$

Докажите, что простые итерации $x_{k+1} = f(x_k)$ сходятся к z.

6. Проверьте, что $z = [1,\ 1,\ 1]^\top$ — одно из решений уравнения $f(x) = 0$, где $f : \mathbb{R}^3 \to \mathbb{R}^3$ имеет вид

$$f(x) = \begin{bmatrix} x_1 x_2^3 + x_2 x_3 - x_1^4 - 1 \\ x_2 + x_2^2 + x_3 - 3 \\ x_2 x_3 - 1 \end{bmatrix}.$$

Будет ли метод Ньютона сходиться к z при достаточно близких к z начальных приближениях?

7. Согласно преданию, метод Ньютона впервые был опробован на уравнении $f(x) \equiv x^5 - 2x - 5 = 0$. Возьмите $x_0 = 2$ и проведите две итерации по методу Ньютона. Докажите, что уравнение имеет единственный вещественный корень z и что $|z - x_2| \leq 10^{-4}$.

8. Приведите пример бесконечно дифференцируемой функции f, для которой уравнение $f(x) = 0$ имеет корень z такой, что метод Ньютона не сходится к z для всех $x_0 \neq z$.

9. При $1 \leq a \leq 4$ решается уравнение $x^2 = a$. В качестве начального приближения x_0 берется значение $p_1(a)$, где $p_1(t)$ — полином степени 1 наилучшего равномерного приближения к функции $\sqrt{t}$ на отрезке $[1,\ 4]$, а затем применяется метод Ньютона. Найдите вид полинома $p_1(t)$ и докажите, что $|x_4 - \sqrt{a}| \leq \frac{1}{2}\, 10^{-25}$.

10. Функция $f \in C^{p+1}$ имеет изолированный нуль z кратности p. Рассмотрите итерационный процесс

$$x_{k+1} = x_k - p\,\frac{f(x_k)}{f'(x_k)}$$

и докажите, что если он сходится к z, то сходимость квадратичная и для погрешностей $e_k \equiv z - x_k$ справедливо предельное соотношение

$$\lim_{k \to \infty} \frac{e_{k+1}}{e_k^2} = \frac{f^{(p+1)}(z)}{p\,(p+1)\,f^{(p)}(z)}.$$

11. Матрица $A \in \mathbb{R}^{n \times n}$ невырожденная, а обратная для нее матрица ищется как решение уравнения $A - X^{-1} = 0$. Докажите, что итерации метода Ньютона в данном случае имеют вид

$$X_{k+1} = 2X_k - X_k A X_k.$$

Докажите, что $X_k \to A^{-1}$ квадратично для любой начальной матрицы X_0 такой, что $\rho(I - AX_0) < 1$ ($\rho(\cdot)$ — спектральный радиус).

Глава 18

18.1 Методы минимизации

Трудно (практически невозможно) придумать разумную задачу, которая не сводилась бы к поиску минимума функционала в заданной области. По этой причине разработка и анализ методов для "экстремальных" задач — это обширнейшая область, из которой мы возьмем для обсуждения лишь некоторые полезные идеи и методы.

Если минимальное значение функционала $f \in C^2$ достигается в точке z, то

$$f(z+\delta) = f(z) + \mathcal{O}(||\delta||^2)$$

(почему?). Отсюда заключаем, что если минимальное значение f вычисляется с точностью ε, то соответствующая минимизирующая точка z вычисляется, в лучшем случае, с точностью порядка $\sqrt{\varepsilon}$.

18.2 Снова Ньютон

Пусть $x = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$ и функционал $f(x) \in C^2$ имеет единственную точку минимума z. Тогда z удовлетворяет уравнению

$$f'(x) \equiv \operatorname{grad} f(x) \equiv \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \ldots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} = 0\,.$$

Уравнение $\operatorname{grad} f(x) = 0$ можно решать, например, по методу Ньютона:

$$x_{k+1} = x_k - [f''(x_k)]^{-1} \operatorname{grad} f(x_k), \qquad (18.2.1)$$

где $f''(x) = [\operatorname{grad} f(x)]'$ — якобиан отображения $\operatorname{grad} f(x)$, называемый также *гессианом*.

Чтобы получить теорию сходимости для метода (18.2.1), достаточно вспомнить то, что мы уже знаем о методе Ньютона для решения нелинейных уравнений.

18.3 Релаксация

В методе Ньютона, к сожалению, нужно вычислять вторые производные и нужно иметь хорошее начальное приближение. Существуют методы, в которых этого не требуется.

Поиск минимума можно уподобить спуску с горы, когда в каждой точке x_k нужно выбрать направление спуска p_k и величину шага — малый шаг заставляет двигаться медленно, а слишком большой шаг опасен, так как может завести на склон соседней горы. Положим

$$x_{k+1} \;=\; x_k + \alpha_k\, p_k, \tag{18.3.2}$$

где p_k должно быть направлением убывания функционала $f \in C^1$:

$$(p_k,\, -f'(x_k)) \;\geq\; c\,||p_k||_2\,||f'(x_k)||_2, \quad c \;>\; 0, \tag{18.3.3}$$

а $\alpha_k \;\geq\; 0$ должно удовлетворять условию *релаксации*:

$$f\,(x_k + \alpha_k\, p_k) \;\leq\; f\,(x_k) - \tau\,\alpha_k\,(p_k,\, -f'(x_k)), \tag{18.3.4}$$

где $0 \;<\; \tau \;<\; 1$ — константа условия релаксации.

Условие (18.3.3) навеяно фактом локального скорейшего убывания функционала в направлении антиградиента. Оно выполняется очевидным образом при $p_k \;=\; -f'(x_k)$. В этом случае метод называется *градиентным* методом.

18.4 Дробление шага

Если (18.3.3) выполнено, $f \in C^1$ и $f'(x_k) \neq 0$, то условие (18.3.4) имеет место для всех достаточно малых $\alpha = \alpha_k$ (докажите!).

Поэтому подходящее α всегда можно найти путем *дробления шага*: берем $\alpha = 1$, проверяем условие релаксации, если оно не выполнено, берем $\alpha/2$, и так далее до выполнения условия релаксации. Если же при $\alpha = 1$ условие релаксации выполнено, то проверяем его для 2α; в случае выполнения заменяем α на 2α и повторяем проверку для удвоенного значения, в противном случае завершаем процедуру выбора.

Лемма 18.4.1 *Пусть* $f \;\in\; C^2\,(\mathbb{R}^n)$ *и* $||f''(x)||_2 \;\leq\; M \quad \forall\; x \in \mathbb{R}^n$, *и пусть в любой точке* x_k *направление спуска* p_k *удовлетворяет условию (18.3.3) с одним и тем же* $c \;>\; 0$. *Тогда в любой точке* x *условие релаксации с константой* τ *выполняется для всех* $0 \leq \alpha_k \leq \hat{\alpha}$, *где*

$$\hat{\alpha} = (1-\tau)\,\frac{2\,c\,||f'(x_k)||_2}{M\,||p_k||_2}. \tag{18.4.5}$$

Доказательство. Согласно формуле Тейлора с остаточным членом в форме Лагранжа,

$$f(x_k + \alpha p_k) = f(x_k) + \alpha (f'(x_k), p_k) + \frac{\alpha^2}{2} (f''(\xi) p_k, p_k) \quad \Rightarrow$$

$$\begin{aligned} f(x_k) - f(x_k + \alpha p_k) &\geq \tau \alpha(-f'(x_k), p_k) \\ &+ \alpha (-f'(x_k), p_k)\{1 - \tau - \tfrac{\alpha M ||p_k||}{2 c ||f'(x_k)||}\}. \end{aligned}$$

Если выражение в фигурных скобках неотрицательно, то, отбросив его, мы и получаем условие релаксации. □

Следствие 18.4.1 *При выборе α_k путем дробления шага*

$$\alpha_k \geq (1 - \tau) \frac{c ||f'(x_k)||}{M ||p_k||}. \tag{18.4.6}$$

Теорема 18.4.1 *Пусть метод (18.3.2) – (18.3.4) с релаксацией путем дробления шага применяется для минимизации ограниченного снизу функционала f в условиях леммы 18.4.1. Тогда для любого начального приближения $x_0 \in R^n$*

$$\lim_{k \to \infty} ||f'(x_k)||_2 = 0. \tag{18.4.7}$$

Доказательство. В силу условия релаксации и следствия 18.4.1

$$f(x_k) - f(x_{k+1}) \geq \tau \alpha_k (-f'(x_k), p_k) \geq \frac{\tau (1 - \tau)}{2} c^2 ||f'(x_k)||^2.$$

Левая часть стремится к нулю, так как последовательность $f(x_k)$ монотонно убывает и ограничена снизу. □

Ключевую роль играет поведение относительного шага $\alpha ||p_k|| / ||f'_k||$. Условие $||f''||_2 \leq M$ нужно лишь для того, чтобы обеспечить его ограниченность снизу.

Если последовательность x_k или какая-либо ее подпоследовательность сходится к z, то z — точка минимума f (почему?).

18.5 Существование и единственность точки минимума

Лемма 18.5.1 *Пусть $f \in C^2(\mathbb{R}^n)$ и существуют положительные константы m и M такие, что для всех $x, y \in \mathbb{R}^n$*

$$m ||y||^2 \leq (f''(x) y, y) \leq M ||y||^2. \tag{18.5.8}$$

Тогда f имеет единственную точку минимума z.

Доказательство. Если функционал f не ограничен снизу, то существует последовательность точек y_k такая, что $f(y_k) \to -\infty$. При этом последовательность y_k не может быть ограниченной (почему?). Учитывая (18.5.8), получаем

$$f(y_k) \;\geq\; f(y_1) - ||f'(y_1)||\,||y_k - y_1|| + \frac{m}{2}\,||y_k - y_1||^2. \qquad (*)$$

Отсюда $f(y_k) \to +\infty$. Полученное противоречие означает ограниченность f снизу.

В силу $(*)$ множество $\{x : \; f(x) \;\leq\; f(x_0)\}$ компактно (почему?). Поэтому точка минимума существует. Чтобы доказать единственность, предположим, что точки y_1 и y_k — точки минимума. Поскольку $f'(y_1) = 0$, из $(*)$ получаем $||y_k - y_1|| = 0 \quad \Rightarrow \quad y_1 = y_k$. $\square$

18.6 Градиентный метод с дроблением шага

Для краткости будем писать $f_k \equiv f(x_k)$, $f'_k \equiv f'(x_k)$ и примем обозначения

$$\varepsilon_k \equiv f_k - f(z), \quad e_k \equiv x_k - z,$$

где z — точка минимума для f. В градиентном методе $p_k = -f_k \quad \Rightarrow$

$$x_{k+1} \;=\; x_k - \alpha_k\, f'_k, \quad f_{k+1} \;\leq\; f_k - \tau\,\alpha_k\,||f'_k||. \qquad (18.6.9)$$

Теорема 18.6.1 *В условиях леммы (18.5.1) градиентный метод с дроблением шага сходится для любого начального приближения x_0 со скоростью геометрической прогрессии:*

$$\varepsilon_k \;\leq\; q^k\,\varepsilon_0, \quad ||e_k|| \;=\; \mathcal{O}(q^{k/2}), \qquad 0 < q < 1. \qquad (18.6.10)$$

Доказательство. В силу следствия 18.4.1 для случая $p_k = -f'_k$ имеем $\alpha_k \;\geq\; \alpha \equiv (1-\tau)/M$, и поэтому, в соответствии с условием релаксации,

$$\varepsilon_{k+1} \;\leq\; \varepsilon_k - \tau\,\alpha\,||f'_k||^2.$$

Чтобы получить неравенство вида $\varepsilon_{k+1} \leq q\,\varepsilon_k$, достаточно иметь неравенство вида

$$||f'_k||^2 \;\geq\; \gamma\,\varepsilon_k \qquad (18.6.11)$$

при условии $\quad 0 < q \equiv 1 - \tau\,\alpha\,\gamma < 1$.

Легко вывести (18.6.11) с константой $\gamma = 2\,m^2/M$. В самом деле,

$$(f'_k,\, e_k) = (f'_k - f'(z),\, e_k) = (f''(\xi)\,e_k,\, e_k) \quad \Rightarrow$$

$$m\,||e_k||^2 \le ||f'_k||\,||e_k|| \quad \Rightarrow \quad ||f'_k||^2 \ge m^2\,||e_k||^2 \ge \frac{2\,m^2}{M}\,\varepsilon_k.$$

Более тонкий результат: (18.6.11) справедливо при $\gamma = \frac{m^2}{M} + m$. Действительно, запишем ряд Тейлора в точке z:

$$f(z) = f_k + (f'_k, -e_k) + \frac{1}{2}\,(f''(\zeta)\,e_k,\, e_k) \quad \Rightarrow$$

$$\varepsilon_k \le \frac{||f'_k||^2}{m} - \frac{m}{2}\,||e_k||^2 \le \frac{||f'_k||^2}{m} - \frac{m}{M}\,\varepsilon_k \quad \Rightarrow ||f'_k||^2 \ge (\frac{m^2}{M} + m)\,\varepsilon_k.$$

Итак, $\varepsilon_{k+1} \le q\,\varepsilon_k$ при $q = 1 - \frac{\tau\,(1-\tau)}{M}\,(\frac{m^2}{M} + m)$. $\square$

Мы доказали, что $q < 1$. Но заметим, к сожалению, что выбор α_k путем деления пополам не дает возможности получить $q < \frac{1}{2}$ (докажите!).

18.7 Метод скорейшего спуска

Метод скорейшего спуска — это градиентный метод, в котором α_k минимизирует $f(x_k - \alpha\,f'_k)$ по всем $\alpha \in \mathbb{R}$. В условиях леммы (18.5.1) он сходится со скоростью геометрической прогрессии (докажите!).

Теорема 18.7.1 *Если метод скорейшего спуска применяется для минимизации квадратичного функционала*

$$f(x) \equiv \frac{1}{2}\,(A\,x,\, x) - (b,\, x), \quad A = A^T \in \mathbb{R}^{n\times n},\ b \in \mathbb{R}^n, \tag{18.7.12}$$

при условии

$$0 < m\,||x||^2 \le (Ax,\, x) \le M\,||x||^2 \quad \forall\, x \ne 0,$$

то справедлива оценка

$$\varepsilon_{k+1} \le \left(\frac{M-m}{M+m}\right)^2 \varepsilon_k. \tag{18.7.13}$$

Доказательство. В данном случае $f'(x) = Ax - b$ (проверьте) и точка минимума имеет вид $z = A^{-1}\,b$. Легко показать также, что

$$f(x) - f(z) = \frac{1}{2}\,(A\,(x-z),\, x-z).$$

Рассмотрим градиентный метод с постоянным шагом α:

$$x_{k+1} = x_k - \alpha\,(A\,x_k - b) \quad \Rightarrow \quad e_{k+1} = (I - \alpha\,A)\,e_k.$$

В то же время (проверьте!)

$$\begin{aligned}\varepsilon_{k+1} \;=\; \tfrac{1}{2}\,(A\,e_{k+1},\,e_{k+1}) \;&=\; \tfrac{1}{2}\,(A\,(I-\alpha\,A)\,e_k,\,(I-\alpha\,A)\,e_k)\\ &=\; \tfrac{1}{2}\,((I-\alpha\,A)\,A\,e_k,\,(I-\alpha\,A)\,e_k)\\ &\le\; ||I-\alpha\,A||_2^2\,(A\,e_k,\,e_k) \;=\; ||I-\alpha\,A||_2^2\,\varepsilon_k.\end{aligned}$$

Понятно, что (почему?)

$$||I-\alpha\,A||_2 \;=\; \max\{|1-\alpha\,m|,\,|1-\alpha\,M|\}.$$

Выражение в правой части минимально при $\alpha = \frac{2}{M+m}$ и равно $\frac{M-m}{M+m}$ (проверьте). Поэтому если мы возьмем именно такое α, то для градиентного метода с постоянным шагом будет справедливо неравенство (18.7.13).

Если градиентный метод с шагом α и метод скорейшего спуска стартуют с одной и той же точки x_k, то, очевидно, последний имеет меньшую погрешность ε_{k+1}. $\square$

Менее тривиальным способом оценка (18.7.13) была получена в 1947 г. Л. В. Канторовичем (см., например, книгу Д. К. и В. Н. Фаддевых "Вычислительные методы линейной алгебры").

18.8 Сложность простого вычисления

"Недостатком" градиентного метода является, конечно, то, что он требует вычислять градиенты. Кажется, что вычисление градиента в одной точке должно быть минимум в n раз дороже каждого вычисления значения функционала в одной точке. В начале 80-х Баур и Штрассен [1] показали, что это не так. Хотя "стоимость" градиента и выше "стоимости" вычисления функционала, но лишь в конечное, не зависящее от n, число раз!

Для строгой формулировки нужно, конечно, точно определить понятие стоимости вычисления.

Пусть имеется некоторый запас $\mathcal{O}$ *элементарных* бинарных операций, то есть операций вида $w \;=\; a\,(u,\,v)$ (например, $w = u+v$ или $w = uv$), и предположим, что задана последовательность

$$\begin{aligned} y_i \;&=\; x_i, \quad 1 \;\le\; i \;\le\; n;\\ y_i \;&=\; a_i\,(y_{i'},\,y_{i''}), \quad n+1 \;\le\; i \;\le\; n+m,\end{aligned}$$

где $1 \;\le\; i' \;\le\; i'' \;<\; i$ для всех $n+1 \;\le\; i \;\le\; n+m$. Такую последовательность будем называть *простым вычислением*. Число m называется *сложностью* (стоимостью) простого вычисления.

[1] W.Baur, V.Strassen, The complexity of partial derivatives, Theor. Comput. Sci. 22: 317–330 (1983).

Простое вычисление можно рассматривать как алгоритм вычисления любой из величин $y_{n+k},\ k = 1, \ldots, m$ или любой их совокупности.

Если элементарные операции имеют вид

$$a_i\,(u,\,v)\ =\ c_i\,u + d_i\,v,$$

где $c_i,\ d_i$ — фиксированные константы, то соответствующее простое вычисление называется *линейным вычислением.*

18.9 Быстрое вычисление градиента

Потребуем, чтобы $\mathcal{O}$ содержало операции $u \pm v$, $u\,v$ и чтобы для любой операции $a\,(u,\,v)\ \in\ \mathcal{O}$ существовали простые вычисления сложности не выше c, вычисляющие частные производные $a'(u,\,v)\ \equiv\ \frac{\partial a}{\partial u}\,(u,\,v)$ и $a''(u,\,v)\ \equiv\ \frac{\partial a}{\partial v}\,(u,\,v)$. Тогда имеет место

Теорема 18.9.1 *Если функционал f от n переменных определяется простым вычислением сложности m, то n компонент градиента от f и значение f в одной точке определяются общим простым вычислением сложности не выше $(5 + 2\,c)\,m$.*

Доказательство. Введем величины

$$u_{ij} \equiv \frac{\partial y_i}{\partial x_j}$$

и заметим, что при любом фиксированном j величины u_{ij} удовлетворяют следующим линейным уравнениям:

$$\begin{aligned}
&u_{1j} = \cdots = u_{j-1\,j} = 0,\\
&u_{jj} = 1,\\
&u_{j+1\,j} = \cdots = u_{nj} = 0;\\
&-a'_i\,u_{i'\,j} - a''_i\,u_{i''\,j} + u_{ij}\ =\ 0, \quad n+1\ \le\ i\ \le\ n+m.
\end{aligned}$$

Запишем их в матричном виде:

$$V\,U\ =\ Z,$$

$$V\ =\ \begin{bmatrix} I_n & 0_{n\times m} \\ V_{21} & V_{22} \end{bmatrix}\ \in\ \mathbb{R}^{(n+m)\times(n+m)},$$

$$U\ =\ [u_{ij}]\ \in\ \mathbb{R}^{(n+m)\times n}, \qquad Z\ =\ \begin{bmatrix} I_n \\ 0_{m\times n} \end{bmatrix}\ \in\ \mathbb{R}^{(n+m)\times n}.$$

Матрица V является нижней треугольной с единицами на главной диагонали; ее первые n строк те же, что в единичной матрице; в каждой i-строке при $i > n$ не более двух ненулевых элементов, кроме элемента главной диагонали.

Не ограничивая общности, можно считать, что $f(x) = y_{n+m}$. Тогда, очевидно,

$$\operatorname{grad} f = [u_{n+m\,1}, \ldots, u_{n+m\,n}].$$

Поэтому нас интересуют лишь первые n компонент последней строки матрицы V^{-1}, или, первые n компонент решения линейной системы

$$[u_{n+m\,1} \ \ldots \ u_{n+m\,n}, \ \ldots\,]\,V = [0 \ \ldots \ 0 \ 1].$$

Это решение находится обратной подстановкой, которая реализуется с помощью умножений и сложений. Число умножений совпадает с числом сложений и равно числу ненулевых элементов под диагональю. К этим операциям нужно добавить операции для вычисления значений $u_{i'}$, $u_{i''}$, $a'_i(u_{i'}, u_{i''})$, $a''(u_{i'}, u_{i''})$. $\square$

Теорема легко обобщается на случай, когда элементарные операции имеют p аргументов: вместо $(5+2\,c)\,m$ будем иметь $(1+(2+c)\,p)\,m$ (почему?).

Доказательство теоремы дает идею конвертирования программы, вычисляющей значение функционала, в программу, вычисляющую одновременно значение функционала и его градиента.

18.10 Полезные идеи

По-видимому, каждый метод минимизации имеет свой недостаток. Обладая глобальной сходимостью, градиентные методы сходятся довольно медленно. Вот, например, картина сходимости метода скорейшего спуска при минимизации квадратичного функционала в $\mathbb{R}^2$:

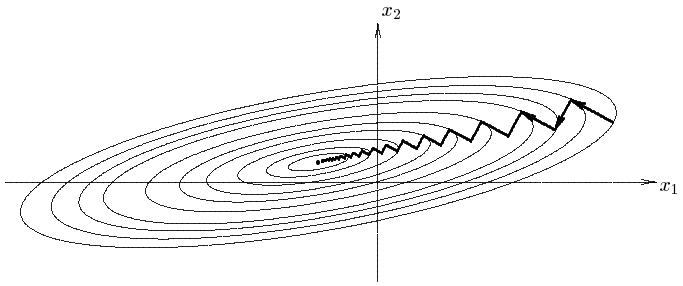


Рисунок 18.1. Итерации метода скорейшего спуска.

Если линии уровня для f сильно вытянуты, то направление на искомую точку минимума может быть почти перпендикулярно градиенту. Можно

попытаться выбирать "более разумные" направления спуска. Например, по двум точкам x_{k-1} и x_k можно найти некоторую новую точку $y_k = x_k + (x_k - x_{k-1})\,\beta_k$ и затем сделать шаг в направлении антиградиента в этой точке: $x_{k+1} = x_k - \alpha_k\, f'(y_k)$. Эта идея приводит к так называемому *овражному* методу.

Другая идея — искать x_{k+1} в виде

$$x_{k+1} \;=\; x_k + \alpha\, f'(x_k) + \beta\,(x_k - x_{k-1}),$$

выбирая α и β так, чтобы минимизировать $f\,(x_k + \alpha\, f'(x_k) + \beta\,(x_k - x_{k-1}))$. Вместо одномерной минимизации в градиентных методах мы получаем двумерную минимизацию на каждом шаге. Эта идея ведет к методам *сопряженных направлений.*

Почти очевидная идея — отталкиваясь от метода Ньютона, получить метод без вторых производных путем какой-либо их аппроксимации. Эта идея может привести к *квазиньютоновским* методам.

Наконец, отметим идею "глобализации" сходимости: это знакомая нам идея релаксации. В частности, релаксация возможна в методе Ньютона:

$$x_{k+1} \;=\; x_k - \alpha_k\,[f''(x_k]^{-1}\, f'(x_k).$$

Если выбирать α_k из условия релаксации типа (18.3.4) с константой $0 < \tau < 1/2$, то в условиях леммы 18.5.1 можно получить факт сходимости (сверхлинейной) независимо от начального приближения. Если потребовать дополнительно, чтобы гессиан удовлетворял условию Липшица

$$||f''(x) - f''(y)|| \;\le\; L\,||x - y|| \quad \forall\; x,\, y \in \mathbb{R}^n,$$

то для любого начального приближения можно гарантировать квадратичную сходимость (попробуйте это сделать!).

18.11 Квазиньютоновские методы

Пусть ищется минимум функционала $f(x),\ x \in \mathbb{R}^n$, и предположим, что в точках $x_0, x_1, ..., x_k$ найдены значения градиента $f'_0, f'_1, ..., f'_k$. Примем обозначения

$$c_i = x_i - x_{i-1}, \quad d_i = f'_i - f'_{i-1}, \qquad 1 \le i \le k.$$

В случае *квадратичного функционала*

$$f(x) = \frac{1}{2}(Ax, x) - (b, x) + c, \qquad A \in \mathbb{R}^{n\times n}, \quad b \in \mathbb{R}^n, \quad c \in \mathbb{R}, \quad (18.11.14)$$

находим $f'(x) = Ax - b$ и $f''(x) = A$. Поэтому при $x = x_i$

$$f''(x_i)\, c_i \;=\; d_i, \quad 1 \le i \le k.$$

В общем случае можно ожидать, что эти равенства выполняются приближенно.

Пусть x_{k+1} получается как решение задачи одномерной минимизации в направлении вектора $H_k^{-1} f'_k$:

$$x_{k+1} = x_k - \alpha_k H_k^{-1} f'_k, \qquad c_{k+1} = x_{k+1} - x_k \perp f'_{k+1}, \tag{18.11.15}$$

где H_k рассматривается как приближение к гессиану f''_k в точке x_k, удовлетворяет *квазиньютоновскому условию*

$$H_k c_k = d_k \tag{18.11.16}$$

и строится с помощью малоранговой поправки уже известной матрицы H_{k-1}. Обычно используют поправки ранга 2, гарантирующие симметрию матриц H_k. Наиболее простой и эффективной признается следующая *формула Бройдена–Флетчера–Гольдфарба–Шанно*:

$$H_k = H_{k-1} - \frac{H_{k-1} c_k c_k^\top H_{k-1}}{c_k^\top H_{k-1} c_k} + \frac{d_k d_k^\top}{d_k^\top c_k}. \tag{18.11.17}$$

Вид формулы очевидно дает нам свойство (18.11.16). По определению, $H_0 = I$.

Лемма 18.11.1 *Если матрица H_{k-1} положительно определенная, то для положительной определенности H_k необходимо и достаточно, чтобы $d_k^\top c_k > 0$.*

Доказательство. Легко проверить, что симметричная матрица

$$M = H_{k-1} - \frac{H_{k-1} c_k c_k^\top H_{k-1}}{c_k^\top H_{k-1} c_k}$$

имеет собственное значение 0 (для собственного вектора c_k). В силу соотношений разделения 0 есть минимальное собственное значение $M \;\Rightarrow$ отрицательность минимального собственного значения H_k равносильна условию $d_k^\top c_k < 0$. $\square$

18.12 Сходимость для квадратичных функционалов

Условие (18.11.16) не означает, что $H_k c_i = d_i$ при $i < k$. Однако, для квадратичных функционалов это верно!

Теорема 18.12.1 *Пусть точки $x_0, x_1, ..., x_k$ получены согласно предписаниям (18.11.15) и (18.11.16) для квадратичного функционала (18.11.14) и $f_i' \neq 0$ при $0 \leq i \leq k-1$. Тогда*

$$(H_{k-1}c_i, c_j) = 0, \qquad i \neq j, \quad 1 \leq i, j \leq k,$$
$$H_k c_i = d_i, \qquad 1 \leq i \leq k.$$

Доказательство. Согласно условию теоремы $\alpha_i \neq 0$ при $0 \leq i \leq k-1$. Для $k = 2$ находим $c_1 \perp f_1'$ и $H_1 c_2 = -\alpha_1 f_1' \quad \Rightarrow \quad (H_1 c_1, c_2) = 0$. Кроме того, $(d_2, c_1) = (Ac_2, c_1) = (c_2, Ac_1) = (c_2, H_1 c_1) = 0 \quad \Rightarrow \quad H_2 c_1 = d_1$.

Далее по индукции. Как и раньше, $c_{k-1} \perp f_{k-1}'$ и $H_{k-1}c_k = -\alpha_{k-1} f_{k-1}'$. Поэтому $(H_{k-1}c_k, c_{k-1}) = 0$. Кроме того, при $i \leq k-2$

$$\begin{aligned}(H_{k-1}c_k, c_i) &= -\alpha_{k-1}(d_{k-1} + f_{k-2}', c_i) \\ &= -\alpha_{k-1}(H_{k-1}c_{k-1}, c_i) + \frac{\alpha_{k-1}}{\alpha_{k-2}}(H_{k-2}c_{k-1}, c_i) = 0,\end{aligned}$$

так как оба скалярных произведения в правой части равны нулю в силу индуктивного предположения. Отсюда уже легко следуют равенства $H_k c_i = d_i$, $1 \leq i \leq k$. $\square$

Следствие 18.12.1 *Квазиньютоновский метод (18.11.15), (18.11.17) в применении к квадратичному функционалу находит точку минимума не более чем за n шагов.*

Доказательство. От противного, пусть $f_i' \neq 0$ при $0 \leq i \leq n$. Тогда

$$(H_n c_i, c_j) = (Ac_i, c_j) = 0, \qquad i \neq j, \quad 1 \leq i, j \leq n+1.$$

В силу положительной определенности A, векторы $c_1, ..., c_{n+1}$ образуют линейно независимую систему, что невозможно. $\square$

При численной реализации матрицы H_k в явном виде строить не нужно. Вместо этого можно получать сразу матрицу H_k^{-1} (она легко получается из H_{k-1}^{-1} с помощью симметричной поправки ранга 2).

Однако, лучше использовать разложения Холецкого для матриц H_k: они столь же эффективно пересчитываются при переходе от H_{k-1} к H_k и, кроме того, позволяют естественным образом следить за сохранением положительной определенности. Последнее важно: при утрате положительной определенности теряется связь матриц H_k с гессианом, то есть вся работа по его приближению оказывается напрасной.

Задачи

1. Покажите, что метод скорейшего спуска в общем случае не может сходиться быстрее, чем со скоростью геометрической прогрессии.

2. Пусть для f выполнены условия леммы 18.5.1. Докажите, что метод скорейшего спуска сходится к точке минимума со скоростью геометрической прогрессии для любого начального приближения.

   Верно ли, что
   $$\varepsilon_{k+1} \;\leq\; \left(\frac{M-m}{M+m}\right)^2 \varepsilon_k \;?$$

3. Множество элементарных операций содержит лишь операции сложения, вычитания, умножения и деления. Докажите, что если значение функционала $f(x)$ в точке x находится простым вычислением сложности m, то $f(x)$ и $\operatorname{grad} f(x)$ в точке x можно найти с помощью некоторого простого вычисления сложности не выше $5\,m$.

4. Множество элементарных операций содержит лишь операции сложения, вычитания и умножения. Докажите, что если значение функционала $f(x)$ в точке x находится простым вычислением сложности m, то $f(x)$ и $\operatorname{grad} f(x)$ в точке x можно найти с помощью некоторого простого вычисления сложности не выше $3\,m$.

5. Линейное вычисление сложности m находит компоненты вектора $y = A\,x, \quad A \in \mathbb{R}^{k\times n}$. Верно ли, что компоненты вектора $z = A^T y$ могут быть найдены с помощью некоторого линейного вычисления сложности m ?

6. Предположим, что A и $B = A - uu^\top + vv^\top$ — вещественные симметричные положительно определенные $n \times n$-матрицы и $u, v \in \mathbb{R}^n$. Докажите, что разложение Холецкого для B можно получить из разложения Холецкого для A с помощью $O(n^2)$ операций.

Глава 19

19.1 Квадратичные функционалы и линейные системы

Если $f(x) = \frac{1}{2}(Ax, x) - \mathrm{Re}(b, x)$, $A = A^* \in \mathbb{C}^{n\times n}$, то ограниченность функционала f снизу равносильна неотрицательной определенности матрицы A (докажите!). Будем считать, что $A > 0$.

В этом случае система $Ax = b$ имеет единственное решение z, и легко проверить, что для любого x

$$f(x) - f(z) = \frac{1}{2}(A(x-z), x-z) \equiv E(x).$$

Отсюда ясно, что z — единственная точка минимума для $f(x)$. $\Rightarrow$ Любой метод минимизации квадратичного функционала f может использоваться как метод решения линейной системы с эрмитовой положительно определенной матрицей коэффициентов.

Функционал $E(x)$ часто называют *функционалом ошибки* для системы $Ax = b$. Он отличается от f лишь на константу. Поэтому любой метод минимизации f одновременно минимизирует E.

Если A — произвольная невырожденная матрица (не обязательно эрмитова), то решение системы $Ax = b$ можно получить, минимизируя квадратичный функционал $r(x) = ||b - Ax||_2^2$. Его называют *функционалом невязки* для системы $Ax = b$.

19.2 Минимизация и проекционные методы

Подобно методу скорейшего спуска, многие методы минимизации f требуют на каждой итерации решать локальную задачу минимизации. Локальная минимизация может проводиться на некотором подпространстве фиксированной размерности (как в методе скорейшего спуска). Однако, чем шире подпространство, тем ближе можно подобраться к искомому решению.

Пусть в $\mathbb{C}^n$ строится цепочка подпространств

$$L_1 \subset L_2 \subset \ldots \subset L_k \subset \mathbb{C}^n, \quad \dim L_i = i, \quad i = 1, \ldots, k,$$

и на k-й итерации находится $x_k = \operatorname{argmin}\{f(x) : \ x \in L_k\}$. Для ограниченного снизу функционала f такой подход дает решение не позже, чем на n-й итерации (почему?).

Развивая описанную идею, мы можем строить различные *проекционные методы*. Пусть требуется решить систему $A\,x = b$. Каким-либо способом введем проекторы Q_k, P_k ранга k и предположим, что *проекционное уравнение*

$$(Q_k\,A\,P_k)\,x \ = \ Q_k\,b, \quad x \ = \ P_k\,x,$$

имеет единственное решение x_k. Можно надеяться на то, что x_k окажется неплохим приближением к x.

Проекционное уравнение часто является эквивалентной формулировкой задачи минимизации f на подпространстве $L_k = \operatorname{im} L_k$.

19.3 Подпространства Крылова

Подпространства вида

$$\mathcal{K}_i \ \equiv \ \mathcal{K}_i\,(b,\,A) \ \equiv \ \operatorname{span}\{b,\,A\,b,\,\ldots\,,\,A^{i-1}\,b\}, \quad i = 1,\,2,\,\ldots\,,$$

называются *подпространствами Крылова*.

Если мы хотим решить систему $A\,x = b$, то идея минимизации на подпространствах может приобрести такую форму:

$$x_i \ = \ \operatorname{argmin}\{||b - A\,x||_2 : \ x \in \mathcal{K}_i\}, \quad i = 1,\,2\,,\ \ldots\ . \tag{19.3.1}$$

Если матрица A невырожденная, то векторы x_i определяются однозначно (докажите!). Метод, получающий последовательность x_i, называется *методом минимальных невязок*.

Почему метод минимальных невязок обязательно приведет к решению системы $A\,x = b$? Рано или поздно будем иметь $\mathcal{K}_i = \mathcal{K}_{i+1} \ \Rightarrow \ A\,\mathcal{K}_i \subset \mathcal{K}_i$. Матрица A невырожденная $\Rightarrow \ A\,\mathcal{K}_i \ = \ \mathcal{K}_i \ \Rightarrow$ поскольку $b \in \mathcal{K}_i$, для некоторого $x \in \mathcal{K}_i$ имеем $A\,x = b$.

Конкретные реализации метода минимальных невязок мы обсудим позже.

19.4 Оптимальные подпространства

Пусть система $A\,x = b$ с невырожденной матрицей A решается путем минимизации невязки на подпространствах. Как выбирать подпространства?

Рассмотрим произвольный алгоритм Φ генерации подпространств L_i:

$$L_{i+1} \ = \ \Phi\,(b,\,L_i,\,A\,L_i).$$

Если $L_i = \mathrm{span}\,\{p_1, \ldots, p_i\}$, то новое подпространство определяется вектором

$$p_{i+1} \;=\; \phi_{i+1}\,(b,\, p_1, \ldots, p_i,\, Ap_1, \ldots, Ap_i).$$

Это означает, что для генерации нового подпространства достаточно выполнить одну операцию умножения матрицы A на вектор p_i. Результат этой же операции может использоваться при минимизации невязки на L_i. Зафиксировав $\varepsilon > 0$, стоимость алгоритма Φ определим таким образом:

$$m\,(\Phi,\, A,\, b) \;\equiv\; \min\{i : \min_{y \in L_i} ||b - A\,y||_2 \;\le\; \varepsilon\}.$$

Плохому алгоритму Φ не возбраняется иметь $m\,(\Phi,\, A,\, b) = +\infty$.

Для индивидуальной матрицы и правой части лучше тот алгоритм генерации, который имеет меньшую стоимость. Но один и тот же алгоритм генерации обычно применяется к разным матрицам, и о его оптимальности, по-видимому, следует судить по его поведению на всех интересующих нас матрицах. Под стоимостью алгоритма Φ на классе матриц $\mathcal{A}$ понимается его стоимость в наихудшем случае:

$$m\,(\Phi,\, b) \;\equiv\; \sup_{A \in \mathcal{A}} m\,(\Phi,\, A,\, b).$$

Алгоритм $\hat{\Phi}$ называется *оптимальным* на классе матриц $\mathcal{A}$, если

$$m\,(\hat{\Phi},\, b) \;\le\; m\,(\Phi,\, b) \quad \forall\ \Phi, \quad \forall\ b.$$

В конце 1970-х отечественные математики Немировский и Юдин обнаружили, что информация о линейной системе, содержащаяся в подпространствах Крылова, является почти оптимальной с точки зрения любого способа ее использования. В наших рассмотрениях этот способ фиксирован вполне определенным образом: имея подпространство (другими словами, его базис), мы находим на нем минимальное значение невязки.

19.5 Оптимальность подпространств Крылова

Класс матриц $\mathcal{A}$ называется *унитарно инвариантным*, если для всякой входящей в него матрицы он содержит все унитарно подобные ей матрицы: $A \in \mathcal{A} \;\Rightarrow\; Q^*AQ \in \mathcal{A}$ для любой унитарной матрицы Q. Алгоритм $\mathcal{K}$, генерирующий подпространства Крылова, является почти оптимальным на любом унитарно инвариантном классе матриц $\mathcal{A}$ в том смысле, что

$$m\,(\mathcal{K},\, b) \;\le\; 2\,m\,(\Phi,\, b) + 1 \quad \forall\ \Phi, \quad \forall\ b. \tag{19.5.2}$$

Это сразу же вытекает из следующей теоремы.

Теорема 19.5.1 *Для любой невырожденной матрицы A и любого алгоритма генерации подпространств Φ существует унитарная матрица Q такая, что*

$$m\,(\mathcal{K},\,A,\,b)\;\leq\;2\,m\,(\Phi,\,Q^*AQ,b)+1. \tag{19.5.3}$$

Доказательство. Мы построим унитарную матрицу Q такую, что для $S=Q^*AQ$ либо $m\equiv m\,(\Phi,\,S,b)\;=\;+\infty$, либо

$$m\,(\mathcal{K},\,A,\,b)\;\leq\;\dim\operatorname{span}\{b,\;L_m,\;SL_m\}, \tag{19.5.4}$$

где L_m — подпространство, полученное алгоритмом Φ для S. [1]

Рассмотрим ортонормированный базис $v_1,\,\ldots\,,v_n$, для которого $\operatorname{span}\{v_1,\,\ldots\,,v_i\}\;=\;\mathcal{K}_i\,(A,\,b)$ до тех пор, пока $\dim\mathcal{K}_i\;=\;i$. Используя подпространства, порождаемые алгоритмом Φ, мы будем строить некоторую последовательность ортонормированных векторов $u_1,u_2,\ldots$ и унитарных матриц Q_i, произвольных во всем, кроме условий

$$Q_i\,u_j\;=\;v_j,\quad 1\;\leq\;j\;\leq\;r'\,(i), \tag{19.5.5}$$

где целочисленная неубывающая функция $r'(i)$ определяется ниже.

Положим $r'(1)\;=\;1$, $Q_1\;=\;I$ и $u_1\;=\;v_1$. Предположим, что каким-то образом уже определены матрицы $S_j\;=\;Q_j^*\,A\,Q_j$, $1\;\leq\;j\;\leq\;i$, такие, что алгоритм Φ порождает для S_j подпространства $L_j\;=\;\operatorname{span}\{p_1,\,\ldots\,,p_j\}$ (важно, что те же подпространства получаются для S_i). Рассмотрим следующие подпространства:

$$\begin{aligned}\mathcal{M}_i\;&\equiv\;\operatorname{span}\{b,\;L_i,\;S_i\,L_i\}\;=\;\operatorname{span}\{u_1,\,\ldots\,,u_{r\,(i)}\},\\ \mathcal{M}'_{i+1}\;&\equiv\;\operatorname{span}\{\mathcal{M}_i,\,p_{i+1}\}\;=\;\operatorname{span}\{u_1,\,\ldots\,,u_{r'\,(i+1)}\}.\end{aligned}$$

Полагаем, что p_{i+1} генерируется алгоритмом Φ для S_i. Матрица Q_{i+1} определяется как произвольная унитарная матрица, удовлетворяющая условиям (19.5.5).

Теперь предположим, что $\exists\;\;y\in L_m:\;\;||b-S_m\,y||_2\;\leq\;\varepsilon$. По построению, $Q_mb=b$ и $L_m\subset\mathcal{M}'_m$. Поэтому

$$||b-S_m\,y||_2\;\geq\;\min_{u\in\mathcal{M}'_m}||b-A\,Q_m\,u||\;=\;\min_{v\in\mathcal{K}_{r'(m)}}||b-A\,v||,$$

и чтобы получить (19.5.4), остается заметить, что $r'(m)\;\leq\;\dim\mathcal{M}_m$.

Если ни для какого m точность ε для S_m не достигается на m-м шаге, то рано или поздно будем иметь $Q_m\;=\;Q_{m+1}\;=\;\ldots$. Поэтому для S_m точность ε не достигается ни на каком шаге $\Rightarrow\;m\,(\Phi,\,S_m,\,b)\;=\;+\infty$. $\square$

[1] Схема рассуждения взята из работы: A.W.Chou, On the optimality of Krylov information, *J. of Complexity*, 3: 26–40 (1987).

19.6 Метод минимальных невязок

Чтобы решить систему $Ax = b$ с невырожденной матрицей A, мы выбираем (произвольное) начальное приближение x_0 и затем фактически решаем редуцированную систему $Au = r_0$, где $r_0 = b - Ax_0$, $x = x_0 + u$. Подпространства Крылова строятся для редуцированной системы (в предположении, что $r_0 \neq 0$).

Пусть $x_i = x_0 + y$, где $y \in \mathcal{K}_i$. Невязка имеет вид $r_i = r_0 - Ay$, а ее длина минимальна, в силу теоремы Пифагора, в том и только том случае, когда

$$r_i \perp A\mathcal{K}_i.$$

Таким образом, для реализации i-го шага требуется опустить перпендикуляр из вектора r_0 на подпространство $A\mathcal{K}_i$. [2] Проще всего это сделать, если в данном подпространстве уже найден ортогональный базис.

Геометрическая реализация метода минимальных невязок заключается в построении последовательности векторов $q_1, q_2, ...$ таким образом, что $q_1, ..., q_i$ дают базис в подпространстве Крылова $\mathcal{K}_i$ и при этом векторы $p_1 = Aq_1, ..., p_i = Aq_i$ образуют ортогональный базис в подпространстве $A\mathcal{K}_i$. Вектор $q_{i+1} \notin \mathcal{K}_i$ должен обладать следующими свойствами:

$$q_{i+1} \in \mathcal{K}_{i+1}, \quad p_{i+1} = Aq_{i+1} \perp A\mathcal{K}_i.$$

Очевидно, его можно получить с помощью рассмотренного нами ранее процесса ортогонализации, примененного к вектору $p = Aq$, где $q = Aq_i$. Заметим, что в геометрической реализации нужно хранить две системы векторов: $q_1, ..., q_i$ и $p_1, ..., p_i$.

Алгебраическая реализация метода минимальных невязок использует лишь одну систему векторов, образующих ортогональные базисы в подпространствах $\mathcal{K}_i$. Она предложена в 1986 г. Саадом и Шульцем. [3] Именно с их легкой руки за методом закрепилась аббревиатура GMRES (обобщенный метод минимальных невязок).

Пусть $q_1 = r_0/||r_0||_2$. Чтобы получить ортонормированный базис $q_1, \ldots, q_{i+1}$ в $\mathcal{K}_{i+1} = \mathcal{K}_{i+1}(r_0, A)$, проводим ортогонализацию вектора $A\,q_i$ к векторам $q_1, \ldots, q_i$. Если $Q_i \equiv [q_1 \ldots q_i] \in \mathbb{C}^{n\times i}$, то получаем

$$A\,Q_i = Q_{i+1}\,\hat{H}_i, \quad \hat{H}_i = \begin{bmatrix} & & H_i & \\ 0 & \ldots & 0 & h_{i+1\,i} \end{bmatrix},$$

[2] Метод минимальных невязок в такой интерпретации впервые описан в книге: Г. И. Марчук, Ю. А. Кузнецов, Итерационные методы и квадратичные функционалы, *Методы вычислительной математики*, Новосибирск, 1975, с. 4–143.

[3] Y.Saad, M.H.Schultz, GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Scientific and Stat. Comp.* 7: 856–869 (1986).

где H_i — верхняя хессенбергова матрица порядка i.

Далее рассмотрим QR-разложение прямоугольной матрицы $\hat{H}_i$:

$$\hat{H}_i \;=\; U_i\, R_i, \quad R_i \in \mathbb{C}^{i\times i}.$$

Тогда минимум невязки $||r_0 - AQ_i y||_2$ по всем y будет достигаться в том случае, если $y = y_i$ удовлетворяет уравнению (докажите!)

$$R_i y_i \;=\; z_i \;\equiv\; ||r_0||_2\, U_i^*\, e_1, \quad \text{где} \quad e_1 = [1\,0\,\ldots\,0]^T.$$

Матрица R_i невырожденная (почему?). Следовательно,

$$x_i \;=\; x_0 + Q_i\, y_i \;=\; x_0 + Q_i\, R_i^{-1}\, z_i.$$

Заметим, что хессенберговы матрицы H_i являются ведущими подматрицами "самой большой" хессенберговой матрицы, отвечающей последней итерации. Главные затраты i-й итерации связаны с ортогонализацией. Помимо этого, в силу хессенберговости матриц H_i вычисление векторов z_i и затем y_i можно осуществить с затратой лишь $\mathcal{O}\,(i^2)$ арифметических операций (придумайте, как это сделать!).

Под "необобщенным" методом минимальных невязок обычно подразумевается метод минимальных невязок в применении к эрмитовым матрицам. В этом случае матрицы H_i оказываются трехдиагональными (почему?).

При большом числе итераций вычислительные затраты и память для хранения векторов q_i могут оказаться непозволительно большими. В таких случаях применяют метод минимальных невязок с *рестартами*: задают максимально допустимую размерность подпространства Крылова, и если точность не достигнута, берут полученное приближение в качестве начального приближения для новой генерации подпространств Крылова.

19.7 A-норма и A-ортогональность

Пусть A —эрмитова положительно определенная матрица порядка n. Тогда для любой пары векторов $x,\ y \in \mathbb{C}^n$ положим

$$(x,\, y)_A \;\equiv\; (Ax,\, y).$$

Проверьте, что это есть скалярное произведение на $\mathbb{C}^n$.

Норма $||x||_A \;\equiv\; \sqrt{(x,\, y)_A}$ называется A-нормой. Векторы x и y называются A-ортогональными, если $(x,\, y)_A = 0$. Для них справедлива теорема Пифагора: $||x+y||_A^2 \;=\; ||x||_A^2 + ||y||_A^2$.

19.8 Метод сопряженных градиентов

Пусть $A = A^* > 0$. Чтобы решить систему $Ax = b$, выбираем (произвольное) начальное приближение x_0 и переходим, по сути, к редуцированной системе $Au = r_0$, где $r_0 = b - Ax_0$, $x = x_0 + u$. Подпространства Крылова строятся для редуцированной системы (в предположении, что $r_0 \neq 0$).

Метод сопряженных градиентов — это метод, в котором на подпространствах Крылова $\mathcal{K}_i(r_0, A)$ минимизируется A-норма ошибки $e = x - z$, где z — решение системы $Ax = b$.

Итак, $x_i = x_0 + y_i$, где $y_i = \operatorname{argmin}\{||x_0 + y - z||_A : \quad y \in \mathcal{K}_i\}$. В силу теоремы Пифагора,

$$(x_i, y)_A = (z, y)_A \ \forall\ y \in \mathcal{K}_i \quad \Leftrightarrow \quad r_i \equiv b - A x_i \ \perp \ \mathcal{K}_i.$$

Если имеется A-ортогональный базис $p_1, \ldots, p_i$ в $\mathcal{K}_i$, то, очевидно, коэффициент α_j, $1 \leq j \leq i$, разложения

$$y_i \ = \ \alpha_1 p_1 \ + \ \ldots \ + \ \alpha_i p_i$$

не зависит от i. Поэтому (не забудем, что $r_i \perp \mathcal{K}_i$)

$$x_i \ = \ x_{i-1} + \alpha_i p_i \ \Rightarrow \ r_i \ = \ r_{i-1} - \alpha_i A p_i \ \Rightarrow \ \alpha_i \ = \ \frac{(r_{i-1}, p_i)}{(A p_i, p_i)}.$$

Если $r_i \neq 0$, то будем искать p_{i+1} в виде $p_{i+1} = r_i + \beta_{i1} p_1 + \cdots + \beta_{ii} p_i$. $r_i \perp \mathcal{K}_i \ \Rightarrow \ \beta_{ij} = 0$ при $j < i$ (докажите!). Вместо β_{ii} будем писать $\beta_i \ \Rightarrow$

$$p_{i+1} \ = \ r_i + \beta_i p_i, \quad \beta_i \ = \ \frac{(r_i, Ap_i)}{(Ap_i, p_i)}.$$

Поскольку $p_i = r_{i-1} + \beta_{i-1} p_{i-1}$, находим: $\alpha_i = \frac{(r_{i-1}, r_{i-1})}{(Ap_i, p_i)}$. Если $r_{i-1} \neq 0$, то $\alpha_i \neq 0$. $\Rightarrow$

$$(r_i, Ap_i) = (r_i, \frac{r_{i-1} - r_i}{\alpha_i}) = -\frac{(r_i, r_i)}{\alpha_i} \ \Rightarrow \ \beta_i \ = \ \frac{(r_i, r_i)}{(r_{i-1}, r_{i-1})}.$$

Окончательно, получаем следующие расчетные формулы *метода сопряженных градиентов*:

$$\begin{aligned} \alpha_i &= (r_{i-1}, r_{i-1})/(Ap_i, p_i), \\ x_i &= x_{i-1} + \alpha_i p_i, \\ r_i &= r_{i-1} - \alpha_i A p_i, \\ \beta_i &= (r_i, r_i)/(r_{i-1}, r_{i-1}), \\ p_{i+1} &= r_i + \beta_i p_i. \end{aligned} \tag{19.8.6}$$

Замечательно, что рекуррентные соотношения оказались "короткими": при минимизации на подпространстве $\mathcal{K}_i$ нам не потребовалось иметь его полный базис!

19.9 От матричных разложений к итерационным методам

Метод сопряженных градиентов можно получить также как следствие разложения Холецкого. Пусть $A = A^* > 0$. Тогда $A_j = Q_j^* A Q_j$ — эрмитова трехдиагональная матрица и, более того, $A_j > 0$ (почему?). Рассмотрим ее разложение Холецкого: $A_j = R_j^* R_j$. В силу трехдиагональности A_j верхняя треугольная матрица R_j будет бидиагональной:

$$R_j = \begin{bmatrix} \gamma_1 & \delta_1 & & & & \\ & \gamma_2 & \delta_2 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & \gamma_{j-1} & \delta_{j-1} \\ & & & & & \gamma_j \end{bmatrix}.$$

Возьмем (пока произвольную) невырожденную диагональную матрицу D_j и положим

$$P_j = [p_1 \ \dots \ p_j] = Q_j R_j^{-1} D_j \quad \Rightarrow \quad P_j^* A P_j = D_j^* D_j. \quad \Rightarrow$$

Столбцы матрицы P_j являются A-ортогональными и при этом

$$p_{j+1} \frac{\gamma_{j+1}}{d_{j+1}} + p_j \frac{\delta_j}{d_j} = q_{j+1} \quad \Rightarrow \quad p_{j+1} = \frac{d_{j+1}}{\gamma_{j+1}} q_{j+1} + \frac{-\delta_j}{\gamma_{j+1}} \frac{d_{j+1}}{d_j} p_j.$$

Мы хотим решить систему $Ax = b$. Потребуем, чтобы $x_j = x_0 + P_j y$ и при этом $r_j \perp \operatorname{im} P_j$ (это равносильно минимизации A-нормы ошибки). Отсюда $x_j = x_{j-1} + \alpha_j p_j$, $r_j = r_{j-1} - \alpha_j A p_j$. Кроме того, если $r_j \neq 0$, то q_{j+1} коллинеарен r_j. Поэтому если мы строим векторы $q_1, \dots, q_{j+1}$, стартуя с $q_1 = r_0/||r_0||_2$, то ничто не мешает выбирать $q_{j+1} = r_j/||r_j||_2$ и специфицировать D_j таким образом:

$$d_1 = ||r_0||_2; \quad d_{j+1} = \gamma_{j+1}/||r_j||_2, \ j = 0, 1, \dots .$$

После этого выражение для p_{j+1} приобретает привычный вид

$$p_{j+1} = r_j + \beta_j p_j.$$

19.10 Формальное скалярное произведение

Замечательно, что метод сопряженных градиентов и метод минимальных невязок в случае эрмитовых матриц могут быть реализованы с помощью "коротких" рекуррентных соотношений, позволяющих хранить в памяти

компьютера лишь небольшое, не зависящее от порядка матрицы число векторов. Заменив задачу локальной минимизации решением некоторого проекционного уравнения, можно получить "короткие" соотношения и для неэрмитовых матриц.

Чтобы вывести подходящее проекционное уравнение, удобно перейти от скалярных произведений к более общим билинейным или полуторалинейным формам. Взяв матрицу $D \in \mathbb{C}^{n\times n}$, введем *формальное скалярное произведение* одним из двух способов:

$$\langle x, y\rangle = y^T D x, \quad x, y \in \mathbb{C}^n, \tag{$*$}$$

либо

$$\langle x, y\rangle = y^* D x, \quad x, y \in \mathbb{C}^n. \tag{$**$}$$

Для матрицы $A \in \mathbb{C}^{n\times n}$ обозначим через A' *дуальную матрицу*, определяемую соотношением

$$\langle Ax, y\rangle = \langle x, A'y\rangle \quad \forall\ x, y \in \mathbb{C}^n.$$

Если α — скалярная величина, то α' будет обозначать дуальную величину, для которой, по определению,

$$\langle \alpha x, y\rangle = \langle x, \alpha' y\rangle \quad \forall\ x, y \in \mathbb{C}^n.$$

Далее будем писать $x \perp y$, если $\langle x, y\rangle = 0$.

Будем считать пока, что $D = I$. Тогда $A' = A^T$ в случае $(*)$ и $A' = A^*$ в случае $(**)$.

19.11 Метод биортогонализации

Чтобы решить систему $Ax = b$ с невырожденной матрицей $A \in \mathbb{C}^{n\times n}$, выберем начальный вектор x_0 и попытаемся приблизить x вектором вида $x_i = x_0 + y_i$, где $y_i \in \mathcal{K}_i(r_0, A) = \operatorname{span}\{p_1, \ldots, p_i\}$, $r_0 = b - Ax_0$, потребовав, чтобы $r_i \equiv b - Ax_i \perp \mathcal{K}_i(r'_0, A') = \operatorname{span}\{p'_1, \ldots, p'_i\}$, где r'_0 выбирается в начале процесса таким образом, что $\langle r_0, r'_0\rangle \neq 0$.

Предположим, что векторы $p_1, \ldots, p_i$ и $p'_1, \ldots, p'_i$ являются *формально A-биортогональными* в том смысле, что $\langle Ap_j, p'_k\rangle = 0$ для всех $j \neq k$ и $\langle Ap_j, p'_k\rangle \neq 0$ при $j = k$. Тогда из выражения для невязки $r_i = r_0 - Ax = r_0 - \sum\limits_{j=1}^{i} \alpha_{ji}\, Ap_j$ вытекает, что

$$\alpha_{ji} = \alpha_j = \langle r_0, p'_j\rangle \ /\ \langle Ap_j, p_j\rangle.$$

Отсюда $x_i = x_{i-1} + \alpha_i p_i \;\Rightarrow\; r_i = r_{i-1} - \alpha_i A p_i$. Далее мы будем считать, что $r_{i-1} \perp \mathcal{K}_{i-1}(r'_0, A')$. Тогда $\alpha_i = \langle r_{i-1}, p'_i\rangle \,/\, \langle Ap_i, p'_i\rangle$.

Введем в рассмотрение также векторы $r'_j = r'_{j-1} - \tilde\alpha'_j A' p'_j$, удовлетворяющие условиям $\mathcal{K}_j(r_0, A) \perp r'_j$ для всех $j \le i$. Если эти условия выполнены для всех $j < i$, то на очередном шаге нам нужно выбрать $\tilde\alpha_i = \langle p_i, r'_{i-1}\rangle \,/\, \langle Ap_i, p'_i\rangle$.

Предположим, что $\alpha_j, \tilde\alpha_j \ne 0$ для всех $j \le i$. Тогда модно считать, что

$$p_{i+1} = r_i + \beta_i p_i, \qquad p'_{i+1} = r'_i + \tilde\beta'_i\, p'_i$$

и, следовательно, $\alpha = \tilde\alpha = \langle r_{i-1}, r'_{i-1}\rangle \,/\, \langle Ap_i, p'_i\rangle$. Поскольку $\langle Ar_i, p'_j\rangle = \langle r_i, A'p'_j\rangle = 0$ для $j < i$, биортогональность будет поддерживаться при выборе $\beta_i = -\langle Ar_i, p'_i\rangle \,/\, \langle Ap_i, p'_i\rangle$. Далее,

$$\langle Ar_i, p'_i\rangle = \langle r_i, A'p'_i\rangle = \langle r_i, \frac{r'_{i-1} - r'_i}{\alpha_i}\rangle = \frac{-\langle r_i, r'_i\rangle}{\alpha_i} \quad\Rightarrow\quad \beta_i = \frac{\langle r_i, r'_i\rangle}{\langle r_{i-1}, r'_{i-1}\rangle}.$$

Легко проверяется, что $\tilde\beta_i = \beta_i$.

В итоге мы получаем следующий *метод биортогоналоизации*:

$$\begin{aligned}
&r_0 = b - Ax_0, \quad p_1 = r_0, \quad \text{выбрать } r'_0, \quad p'_1 = r'_0;\\
&\alpha_i = \langle r_{i-1}, r'_{i-1}\rangle \,/\, \langle Ap_i, p'_i\rangle,\\
&x_i = x_{i-1} + \alpha_i\, p_i, \quad r_i = r_{i-1} - \alpha_1 A p_i, \quad r'_i = r'_{i-1} - \alpha'_1 A' p'_i,\\
&\beta_i = \langle r_i, r'_i\rangle \,/\, \langle r_{i-1}, r'_{i-1}\rangle,\\
&p_{i+1} = r_i + \beta_i p_i, \quad p'_{i+1} = r'_i + \beta'_i p'_i, \quad i = 1, 2, \ldots .
\end{aligned}$$

Заметим, что $\alpha'_i = \alpha_i$, $\beta'_i = \beta_i$ для формального скалярного произведения типа $(*)$ и $\alpha'_i = \bar\alpha_i$, $\beta'_i = \bar\beta_i$ (комплексно сопряженные величины) в случае $(**)$.

В описанном процессе возможны *аварийные случаи* двух типов:

- $\langle Ap_i, p'_i\rangle = 0$ для некоторого i;
- $\langle r_{i-1}, r'_{i-1}\rangle = 0 \;\Rightarrow\; \alpha_i = 0$, но при этом $r_i \ne 0$.

В каждом из этих случаев продолжение процесса в прежнем виде невозможно. Для борьбы с аварийными случаями используются различные *блочные версии* метода биортогонализации.

19.12 Метод квазиминимальных невязок

В неэрмитовом случае платой за сохранение “коротких” рекуррентных соотношений является отказ от минимизации невязки (или другого связанного с ней функционала). В результате невязка может падать или расти

от шага к шагу весьма хаотически и, как следствие, по ее величине трудно судить о том, надо ли процесс останавливать или продолжать. Поэтому может возникнуть желание перейти от векторов x_i к некоторым другим векторам $\hat{x}_i$, для которых невязка убывает монотонно или, по крайней мере, имеет "более регулярное" поведение.

Предположим, что имеется произвольный итерационный процесс, в котором вычисленные векторы $x_0, x_1, ..., x_i$ удовлетворяют соотношениям $x_j = x_{j-1} + \alpha_j p_j$, $\alpha_j \neq 0$, $1 \leq j \leq i$. Для невязок получаем $r_j = r_{j-1} - \alpha_j A p_i$. Пусть $P_i = [p_1, \ldots, p_i]$, $\rho_i = \|r_i\|_2$, и $R_i = [r_0/\rho_0, \ldots, r_i/\rho_i]$. Тогда

$$[r_0, Ap_1, ..., Ap_i] = R_i Z_i, \quad Z_i = \begin{bmatrix} \rho_0 & \rho_0/\alpha_1 & & & \\ & -\rho_1/\alpha_1 & \rho_1/\alpha_2 & & \\ & & ... & ... & \\ & & & -\rho_{i-1}/\alpha_{i-1} & \rho_{i-1}/\alpha_i \\ & & & & -\rho_i/\alpha_i \end{bmatrix}.$$

Согласно этому равенству, если $\hat{x}_i = x_0 + P_i y$, то

$$\hat{r}_i = r(y) \equiv r_0 - AP_i y = R_i Z_i \begin{bmatrix} 1 \\ -y \end{bmatrix}.$$

Отсюда $\hat{r}_i = R_i v$, где вектор $v = v(y)$ имеет вид

$$v = [\rho_0(1 - \xi_1),\ \rho_1(\xi_1 - \xi_2),\ \ldots,\ \rho_{i-1}(\xi_{i-1} - \xi_i),\ \rho_i \xi_i]^T,$$

$$\xi_j = y_j/\alpha_j, \quad 1 \leq j \leq i.$$

Давайте выберем $\xi_1, ..., \xi_i$ таким образом, чтобы минимизироаать $\|v\|_2$. Поскольку $\|R_i\|_2 \leq \sqrt{i+1}$ (почему?), находим

$$\|\hat{r}_i\|_2 \leq \sqrt{i+1} \min_y \|v(y)\|_2.$$

Это неравенство можно рассматривать как *свойство квазиминимизации*, так как вместо нормы невязки $r(y)$ минимизируется норма вектора $v(y)$, однозначно определяемого равенством $r(y) = R_i v(y)$. В качестве приближенного решения системы берется $\hat{x}_i$.

Пусть $\eta_0 = 1 - \xi_1$, $\eta_1 = \xi_1 - \xi_2$, $\ldots$, $\eta_{i-1} = \xi_{i-1} - \xi_i$, $\eta_i = \xi_i$. Нам нужно минимизировать функционал $f = \|v\|_2^2 = \rho_0^2 \eta_0^2 + \cdots + \rho_i^2 \eta_i^2$ при ограничении $\eta_0 + \cdots + \eta_i = 1$. Используя множители Лагранжа, [4] находим

$$\eta_j = \frac{\sigma_j}{s_i},$$

[4] L. Zhou and H. F. Walker. Residual smoothing techniques for iterative methods. *SIAM J. on Sci. Comput.* 15 (2): 297–312 (1994).

где

$$\sigma_j = \frac{1}{\rho_j^2}, \qquad s_i = \sum_{j=1}^{i} \frac{1}{\rho_j^2}. \tag{19.12.7}$$

Следовательно,

$$\hat{r}_i = \sum_{j=0}^{i} \frac{\sigma_j}{s_i} r_j \quad \Rightarrow \quad \hat{x}_i = \sum_{j=0}^{i} \frac{\sigma_j}{s_i} x_j.$$

Из этих формул можно вывести, что

$$\hat{x}_i = \left(1 - \frac{\sigma_i}{s_i}\right) \hat{x}_{i-1} + \frac{\sigma_i}{s_i} x_i, \qquad \hat{r}_i = \left(1 - \frac{\sigma_i}{s_i}\right) \hat{r}_{i-1} + \frac{\sigma_i}{s_i} r_i. \tag{19.12.8}$$

Пусть в качестве основного итерационного процесса используется метод биортогонализации. Дополнив его вычислениями по формулам (19.12.7) и (19.12.8), мы получаем *метод квазиминимальных невязок.* [5]

Задачи

1. Пусть $Az = b$ и $f(x) = \frac{1}{2}(Ax, x) - \mathrm{Re}\,(b, x)$, $A = A^* \in \mathbb{C}^{n\times n}$, $b \in \mathbb{C}^n$. Докажите, что $f(x) - f(z) = \frac{1}{2}(A(x-z), x-z)$.

2. Пусть $A = A^* \in \mathbb{C}^{n\times n}$ и $b \in \mathbb{C}^n$. Докажите, что ограниченность функционала $f(x) = \frac{1}{2}(Ax, x) - \mathrm{Re}\,(b, x)$ снизу равносильна неотрицательной определенности матрицы A.

3. Какие векторы являются одновременно ортогональными и A-ортогональными?

4. Докажите, что в методе сопряженных градиентов i-я невязка A-ортогональна любой j-й невязке, если $|j - i| > 1$.

5. Объясните, почему после рестарта в методе минимальных невязок обычно наблюдается увеличение невязки.

6. Возможно ли возрастание невязки в методе квазиминимальных невязок?

[5] Идея квазиминимизации появилась в работе: R. W. Freund and N. M. Nachtigal. QMR: a quasi-minimal residual method for non-Hermitian linear systems, *Numer. Math.* 60: 315–339 (1991). Расчетные формулы метода QMR отличаются от полученных нами формул, но в точной арифметике приближенные решения будут такими же.

Глава 20

20.1 Сходимость метода минимальных невязок

Метод минимизации невязок на подпространствах Крылова за конечное число шагов приводит к точному решению и по этой причине является прямым методом. Но чаще его рассматривают как итерационный метод — с типичными для итерационных методов *оценками сходимости.*

Заметим, что получение оценок требует каких-либо *дополнительных предположений* относительно матрицы коэффициентов: в общем случае для матрицы порядка n число шагов может оказаться равным n, а норма невязки может оставаться равной норме начальной невязки на всех шагах, кроме последнего. Вот пример системы $Ax = b$:

$$\begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \dots & \dots & \\ & & & 0 & 1 \\ 1 & & & & \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 1 \end{bmatrix}.$$

Если $x_0 = 0$, то $r_0 = b$. Подпространства Крылова получаются такие:

$$\mathcal{K}_i = \operatorname{span}\{e_n, e_{n-1}, ..., e_{n-i+1}\}, \quad 1 \le i \le n.$$

Здесь e_j обозначает j-й столбец единичной матрицы. Точное решение системы имеет вид $x = e_1$, а на итерациях получаются следующие векторы (докажите!):

$$x_0 = x_1 = ... = x_{n-1} = 0, \quad x_n = e_1.$$

20.2 Условие строгой эллиптичности

Проще всего оценки сходимости метода минимальных невязок получаются при условии *строгой эллиптичности* (коэрцитивности):

$$\operatorname{Re}(Ax, x) \ge \tau(x, x) \quad \forall\, x,$$

$\tau > 0$ — константа, одинаковая для всех x.

Матрица A называется *строго эллиптической* (коэрцитивной).

Вектор x_i минимизирует невязку на множестве векторов вида $x_0 + y$, $y \in \mathcal{K}_i$. Отсюда, в частности,

$$\begin{aligned} ||r_i||_2 &\leq \min_{\alpha} ||r_{i-1} - \alpha A r_{i-1}||_2 \\ &= |\alpha|^2 (Ar_{i-1}, Ar_{i-1}) - 2\mathrm{Re}\ (\alpha(Ar_{i-1}, r_{i-1})) + (r_{i-1}, r_{i-1}). \end{aligned}$$

В случае вещественных α минимум правой части достинается при

$$\alpha = \frac{\mathrm{Re}\,(Ar_{i-1}, r_{i-1})}{(Ar_{i-1}, Ar_{i-1})} \quad \Rightarrow$$

$$||r_i||_2 \leq \sqrt{1 - \frac{\left(\mathrm{Re}\,(Ar_{i-1}, r_{i-1})/(r_{i-1}, r_{i-1})\right)^2}{(Ar_{i-1}, Ar_{i-1})/(r_{i-1}, r_{i-1})}}\ ||r_{i-1}||_2.$$

Следовательно, при условии строгой эллиптичности имеет место неравенство [1]

$$||r_i||_2 \leq \sqrt{1 - \frac{\tau^2}{||A||_2^2}}\ ||r_{i-1}||_2. \tag{20.2.1}$$

Обратим внимание на то, что оценка (20.2.1) опирается на то, что норма i-й невязки в методе минимальных невязок не может быть больше нормы невязки любого вектора вида $x_{i-1} + \alpha r_{i-1}$, так как $r_{i-1} \in \mathcal{K}_i$. По существу, это оценка для метода, в котором на i-м шаге решается задача одномерной минимизации невязки, и никак не учитывается то, что длина невязки в действительности имеет минимальное значение на всем подпространстве Крылова размерности i.

20.3 Оценки с помощью полиномов

Исходной точкой для получения более точных оценок служит следующее наблюдение:

$$r_i \in r_0 + A\mathcal{K}_i \quad \Rightarrow \quad r_i = f_i(A) r_0, \tag{20.3.2}$$

где $f_i(\zeta)$ — полином степени не выше i со свободным членом $f_i(0) = 1$. Обозначим через $\mathcal{F}_i$ множество всех таких полиномов. Тогда минимальность длины невязки r_i означает, что

$$||r_i||_2 \ \leq\ ||f_i(A)||_2\ ||r_0||_2 \quad \forall\ f_i \in \mathcal{F}_i. \tag{20.3.3}$$

[1] Данная оценка по форме совпадает с оценкой, полученной в 1952 году М. А. Красносельским и С. Г. Крейном для положительно определенных матриц. В 1982 году Элман заметил, что эта же оценка справедлива для произвольных строго эллиптических матриц.

20.4 Полиномы и резольвента

Пусть $f(z)$ — произвольная функция от комплексного переменного z, аналитическая в открытой ограниченной односвязной области Ω с кусочно-гладкой границей Γ. Известно, что в этом случае $f(z)$ выражается с помощью *интеграла Коши*: [2]

$$f(z) = \frac{1}{2\pi \mathbf{i}} \int\limits_{\Gamma} \frac{f(\zeta)}{z-\zeta} d\zeta, \quad z \in \Omega.$$

Пусть $f(z) = a_0 + a_1 z + ... + a_n z^n$ и $f(A) = a_0 I + a_1 A + ... + a_n A^n$ — произвольный полином и порождаемый им полином от матрицы A. Матрица $(A - \zeta I)^{-1}$ называется *резольвентой* матрицы A; очевидно, что она определена для всех ζ, отличных от собственных значений A.

Лемма 20.4.1 *Если все собственные значения матрицы A принадлежат ограниченной области Ω с границей Γ, то для произвольного полинома $f(z)$*

$$f(A) = \frac{1}{2\pi \mathbf{i}} \int\limits_{\Gamma} (A - \zeta I)^{-1} f(\zeta) d\zeta, \quad z \in \Omega.$$

Доказательство. Кривая Γ является общей границей для двух открытых областей комплексной плоскости — ограниченной области Ω и дополнительной неограниченной области Ω'. При достаточно большом $R > 0$ элементы матрицы $(A - \zeta I)^{-1}$ являются аналитическими функциями в "кольце" — пересечении области Ω' и круга $\{|\zeta| < R\}$. Поэтому

$$\frac{1}{2\pi \mathbf{i}} \int\limits_{\Gamma} (A - \zeta I)^{-1} f(\zeta) d\zeta = \frac{1}{2\pi \mathbf{i}} \int\limits_{|\zeta|=R} (A - \zeta I)^{-1} f(\zeta) d\zeta.$$

При достаточно большом $|\zeta| = R$ находим

$$(A - \zeta I)^{-1} = -\sum_{k=0}^{\infty} \zeta^{-k-1} A^k.$$

Прямое вычисление показывает, что

$$-\frac{1}{2\pi \mathbf{i}} \int\limits_{|\zeta|=1} A^k \zeta^{j-k-1} d\zeta = \begin{cases} A^j, & k = j, \\ 0, & k \neq j. \end{cases}$$

[2] Направление обхода границы Γ выбирается таким образом, чтобы область Ω находилась слева.

Следовательно,

$$-\frac{1}{2\pi\mathbf{i}}\int\limits_{|\zeta|=R}\sum_{k=0}^{\infty}A^k\zeta^{-k-1}\sum_{j=0}^{n}a_j\zeta^j d\zeta \;=\; \sum_{j=0}^{n}a_jA^j \;=\; f(A). \qquad \square$$

20.5 Предельная скорость сходимости

Из проведенных рассмотрений сразу же вытекает следующая

Теорема 20.5.1 *Если все собственные значения матрицы A принадлежат ограниченной области Ω с границей Γ, то для невязок r_i метода минимальных невязок справедливы оценки*

$$\frac{||r_i||_2}{||r_0||_2} \;\le\; \frac{|\Gamma|}{2\pi}\max_{\zeta\in\Gamma}||(A-\zeta I)^{-1}||_2\,\min_{f\in\mathcal{F}_i}\max_{\zeta\in\Gamma}|f(\zeta)|,$$

где $|\Gamma|$ — длина кривой Γ, а $\mathcal{F}_i$ — множество всех полиномов $f(\zeta)$ степени не выше i со свободным членом $f(0)=1$.

Таким образом, получение оценок сходимости метода минимальных невязок сводится к оценке величин

$$T_i(\Gamma) \;\equiv\; \min_{f\in\mathcal{F}_i}\max_{\zeta\in\Gamma}|f(\zeta)|. \tag{20.5.4}$$

Предположим, что решаются системы с матрицами A_n порядка n при различных $n\to\infty$. Обозначим через $\mathcal{A}$ множество всевозможных последовательностей матриц A_n таких, что все собственные значения каждой из матриц A_n принадлежат Ω и норма резольвенты на Γ для каждой из матриц A_n не превышает не зависящей от n константы $R(\mathcal{A},\Gamma)$. Далее, пусть r_i обозначает i-ю невязку для произвольно выбранной матрицы A_n при $n\ge i$. Тогда

$$\frac{||r_i||_2}{||r_0||_2} \;\le\; \frac{|\Gamma|R(\mathcal{A},\Gamma)}{2\pi}T_i(\Gamma) \quad\Rightarrow$$

$$\overline{\lim_{i\to\infty}}\left(\frac{||r_i||_2}{||r_0||_2}\right)^{1/i} \;\le\; q,$$

где

$$q=q(\Gamma)\equiv\overline{\lim_{i\to\infty}}(T_i(\Gamma))^{1/i}. \tag{20.5.5}$$

Величина $q=q(\Gamma)$ называется *предельной скоростью сходимости* метода минимальных невязок на множестве $\mathcal{A}$. Название оправдано, так как оказывается, что всегда можно выбрать последовательность матриц $\{A_n\}\in\mathcal{A}$

и невязок r_i таким образом, что

$$\overline{\lim_{i\to\infty}}\left(\frac{||r_i||_2}{||r_0||_2}\right)^{1/i} = q. \qquad (20.5.6)$$

20.6 Числовая область матрицы

В полученной нами оценке сходимости метода минимальных невязок норма резольвенты на кривой Γ определяет лишь коэффициент, не зависящий от номера итерации i. Но для конкретной матрицы A порядка n этот коэффициент может оказаться большим (например, величиной порядка q^{-n}) и сделать оценку бесполезной. К счастью, при специальном выборе кривой Γ норма резольвенты легко оценивается величиной, не зависящей от n.

Это можно сделать, если область Ω содержит *числовую область* матрицы A, определяемую как множество точек вида

$$\Phi(A) = \{z \in \mathbb{C} : z = (Ax, x), ||x||_2 = 1\}.$$

Множество $\Phi(A)$ замкнуто (докажите). Кроме того, справедлива

Теорема 20.6.1 (Теплиц–Хаусдорф) *Числовая область матрицы является выпуклым множеством.*

Доказательство. Множество Φ содержит отрезок, соединяющий точки z_1 и z_2 в том и только том случае, когда множество $a\Phi + b \equiv \{az + b, z \in \Phi\}$ содержит отрезок, соединяющий точки $az_1 + b$ и $az_2 + b$ — при условии, конечно, что $a \neq 0$. Если $||z_1||_2 = ||z_2||_2 = 1$ и $(Az_1, z_1) \neq (Az_2, z_2)$, то можно рассмотреть матрицу $B = aA + b$ и выбрать a и b таким образом, что $(Bz_1, z_1) = 0$ и $(Bz_2, z_2) = 1$. Кроме того, умножив z_1 на число, равное по модулю 1, можно считать, что число $(Bz_1, z_2) + (Bz_2, z_1)$ является вещественным. Достаточно доказать, что $\Phi(B)$ содержит все точки отрезка $[0, 1]$. Векторы z_1 и z_2 линейно независимы (почему?) $\Rightarrow$ при $0 \leq t \leq 1$ функция

$$\phi(t) \equiv \frac{(B(tz_1 + (1-t)z_2), tz_1 + (1-t)z_2)}{||tz_1 + (1-t)z_2||_2^2}$$

является непрерывной и принимает вещественные значения, причем $\phi(0) = 1$ и $\phi(1) = 0$ $\Rightarrow$ по теореме Ролля $\phi(t)$ принимает все значения между 0 и 1. $\square$

Заметим, что числовая область $\Phi(A)$ содержит все собственные значения матрицы A (докажите).

20.7 Оценка резольвенты

Теорема 20.7.1 *Пусть ζ не принадлежит числовой области $\Phi(A)$. Тогда*

$$||(A-\zeta I)^{-1}||_2 \;\le\; \frac{1}{d(\zeta,\Phi(A))}, \tag{20.7.7}$$

где $d(\zeta,\Phi(A)) = \min\limits_{\xi\in\Phi(A)} ||\zeta-\xi||_2$ *— расстояние от точки ζ до $\Phi(A)$.*

Доказательство. Для некоторых векторов x и y единичной длины имеет место равенство (почему?)

$$(A-\zeta I)^{-1}y \;=\; ||(A-\zeta I)^{-1}||_2\; x.$$

Значит,

$$|(Ax,x)-\zeta(x,x)| = |((A-\zeta I)x,x)| = \frac{|(y,x)|}{||(A-\zeta I)^{-1}||_2} \le \frac{1}{||(A-\zeta I)^{-1}||_2}$$

$$\Rightarrow \quad ||(A-\zeta I)^{-1}||_2 \le \frac{1}{|(Ax,x)-\zeta|} \le \frac{1}{d(\zeta,\Phi(A))}. \quad \square$$

Следствие 20.7.1 *Пусть расстояние от любой точки кривой Γ до числовой области $\Phi(A)$ не меньше d. Тогда*

$$\max_{\zeta\in\Gamma} ||(A-\zeta I)^{-1}||_2 \;\le\; \frac{1}{d}.$$

20.8 Сходимость в случае нормальных матриц

Нормальность матрицы A означает существование ортонормированного базиса из ее собственных векторов $\Rightarrow$

$$A = Q\Lambda Q^{-1},$$

где Q — унитарная матрица, а Λ — диагональная матрица из собственных значений A. Если f — полином, то в силу унитарной инвариантности спектральной нормы

$$||f(A)||_2 = ||Qf(\Lambda)Q^{-1}||_2 = ||f(\Lambda)||_2.$$

Пусть все собственные значения нормальной матрицы A принадлежат области Ω с границей Γ. Тогда

$$\frac{||r_i||_2}{||r_0||_2} \;\le\; \min_{f\in\mathcal{F}_i}\; \max_{\zeta\in\Omega\cup\Gamma} |f(\zeta)| \;=\; \min_{f\in\mathcal{F}_i}\; \max_{\zeta\in\Gamma} |f(\zeta)|. \tag{20.8.8}$$

Теорема 20.8.1 *Пусть собственные значения нормальной матрицы A принадлежат области Ω, граница которой Γ есть эллипс с центром в точке c, большей полуосью a и расстоянием между фокусами $2d$, причем $0 \notin \Omega \cup \Gamma$. Тогда для метода минимальных невязок при решении системы с матрицей A справедливы оценки $||r_i||_2 \le \left|\frac{T_i(a/d)}{T_i(c/d)}\right| ||r_0||_2$, где T_i — полином Чебышева степени i.*

Доказательство получается, если в оценке (20.8.8) взять $f(x) = \frac{T_i\left(\frac{c-z}{d}\right)}{T_i\left(\frac{c}{d}\right)}$ и учесть наблюдения, сделанные нами ранее при использовании эллипсов Бернштейна.

20.9 Минимальные невязки и уравнение Лапласа

Для некоторых кривых (в частности, для эллипсов) предельную скорость сходимости метода минимальных невязок можно найти аналитически. В случае произвольной гладкой (и даже кусочно-гладкой) кривой ее можно вычислить приближенно.

Не требуя высокой точности, это можно сделать просто и быстро с помощью решения следующей внешней краевой задачи для уравнения Лапласа:

$$\begin{gathered} \Delta g(z) = 0, \quad z \in \Omega', \\ g(z) = 0, \quad z \in \Gamma, \\ g(z) = \ln|z| + \gamma + o(1), \quad z \to \infty. \end{gathered} \tag{20.9.9}$$

Поясним смысл обозначений:

$g(z)$ рассматриваеся как функция $z = x + \mathbf{i}y$, где x и y — вещественные переменные;

$\Delta \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ — оператор Лапласа;

γ — константа, которая называется *постоянной Робэна* и находится одновременно с функцией $g(z)$. [3]

Известно, что задача (20.9.9) имеет единственное решение. Оно позволяет дать точное выражение для предельной скорости сходимости метода минимальных невязок. Чтобы упростить обоснование, полагаем в дальнейшем, что Γ — кривая класса C^1.

Теорема 20.9.1 *Пусть $0 \in \Omega$ и $g(z)$ — решение задачи (20.9.9). Тогда*

$$q(\Gamma) \;=\; \mathrm{e}^{-g(0)}. \tag{20.9.10}$$

[3] Заметим также, что величина $e^{-\gamma}$ называется *логарифмической емкостью* кривой Γ.

20.10 Метод логарифмического потенциала

Доказательство теоремы 20.9.1 и удобный приближенный метод решения задачи (20.9.9) используют *логарифмический потенциал* — функцию вида

$$v(z) = \int\limits_{\Gamma} \ln|z-\zeta|\,\sigma(\zeta)\,|d\zeta|, \tag{20.10.11}$$

где $\sigma(\zeta)$ — принимающая вещественные значения *плотность потенциала.*

Довольно просто проверяется, что при любой плотности

$$\Delta v(z) = 0 \quad \text{при } x \in \Omega' \ \text{ и при } \ z \in \Omega.$$

В теории потенциала доказывается, что функция $v(z)$ определена и непрерывна при всех z.

Рассмотрим следующую систему интегральных уравнений относительно функции $\sigma(\zeta)$ и скалярной величины γ:

$$\int\limits_{\Gamma} \ln|z-\zeta|\,\sigma(\zeta)\,|d\zeta| \;=\; -\gamma, \tag{20.10.12}$$

$$\int\limits_{\Gamma} \sigma(\zeta)\,|d\zeta| \;=\; 1. \tag{20.10.13}$$

Лемма 20.10.1 *Пусть $\sigma(\zeta)$ и γ удовлетворяют соотношениям (20.10.12) и (20.10.13). Тогда функция $g(z) = v(z) + \gamma$ является решением внешней краевой задачи (20.9.9).*

Доказательство.

$$v(z) = \ln|z| \int\limits_{\Gamma} \sigma(\zeta)\,|d\zeta| \;+\; \int\limits_{\Gamma} \ln|1-\zeta/z|\,\sigma(\zeta)\,|d\zeta| \;=\; \ln|z| + o(1). \quad \square$$

Для вычисления $\sigma(\zeta)$ и γ можно действовать по такой схеме: найти решение интегрального уравнения

$$\int\limits_{\Gamma} \ln|z-\zeta|\,\phi(\zeta)\,|d\zeta| = -1,$$

вычислить

$$s = \int\limits_{\Gamma} \phi(\zeta)\,|d\zeta|$$

и положить

$$\sigma(\zeta) = \phi(\zeta)/s, \quad \gamma = 1/s;$$

если приведенное выше интегральное уравнение не имеет решения, то найти нетривиальное решение $\phi(\zeta)$ аналогичного уравнения с тождественно нулевой правой частью, в этом случае $\gamma = 0$, а $\sigma(\zeta)$ определяется из $\phi(\zeta)$ путем нормировки, как и раньше.

Следующий раздел — для тех, кто хотел бы познакомиться с обоснованием описанного метода и доказательством теоремы 20.9.1.

20.11 Обоснование метода

Лемма 20.11.1 *Если $g(z)$ — решение задачи (20.9.9), то $g(z) > 0$ в любой точке $z \in \Omega'$.*

Доказательство. Множество $M \subset \Omega'$ точек z, в которых $g(z) = 0$, является замкнутым. Если M не пусто, то в случае ограниченности оно является компактным $\Rightarrow$ существует точка $z \in M$ с максимальным модулем $|z|$ $\Rightarrow$ для любой окружности γ с центром в точке z, целиком принадлежащей области Ω', имеет место равенство

$$\int\limits_{\gamma} g(\zeta)|d\zeta| = 0 \quad \Rightarrow \quad g(\zeta) = 0 \text{ при } \zeta \in \gamma.$$

Это противоречит максимальности модуля $|z|$ $\Rightarrow$ множество M не может быть ограниченным $\Rightarrow$ $g(z)$ не стремится к ∞ при $z \to \infty$. Значит, множество M пусто. $\square$

Лемма 20.11.2 *В случае кривой Γ класса C^1 существуют непрерывная плотность потенциала $\sigma(\zeta)$ и число γ, удовлетворяющие системе уравнений (20.10.12), (20.10.13). При этом $\sigma(\zeta) \geq 0$, а интеграл от $\sigma(\zeta)$ по любой конечной части кривой Γ положителен.*

Доказательство можно получить, используя теорию Рисса–Фредгольма (см. главу 22) и известные свойства логарифмического потенциала: его непрерывность во всех точках плоскости и свойства скачка производной по нормали к Γ при переходе через границу Γ.

Доказательство теоремы 20.9.1. [4] Фиксируем произвольное достаточно малое $\varepsilon > 0$ и рассмотрим кривую Γ_ε, определенную уравнением $g(z) = \varepsilon$.

[4] Мы адаптируем к нашему случаю схему доказательства для родственной, но формально другой задачи из книги: Г. М. Голузин, *Геометрическая теория функций комплексного переменного*, М., Наука, 1966.

Очевидно, что $\Gamma_\varepsilon \subset \Omega'$ и расстояние между Γ_ε и Γ положительно. Выбрав попарно различные точки $\zeta_1, ..., \zeta_n \in \Gamma$, построим полином

$$p_n(z) = \prod_{i=1}^{n}(z - \zeta_i)$$

такой,что

$$|\ln|p_n(z)|^{1/n} - v(z)| \ \leq \ \varepsilon \quad \forall \ \ z \in \Gamma_\varepsilon.$$

Согласно лемме 20.11.2, при обходе кривой Γ можно разбить ее на части $\gamma_1, ..., \gamma_n$ таким образом, что

$$\int\limits_{\gamma_i} \sigma(\zeta)\,|d\zeta| = \frac{1}{n}.$$

Кроме того, существует число $N = N(\varepsilon)$ такое, что при $n \geq N$ для всех $1 \leq i \leq n$ выполняются неравенства

$$|\ln|z - \zeta| - \ln|z - \eta|| \leq \varepsilon \quad \forall \ \ \zeta, \eta \in \gamma_i.$$

Чтобы определить полином $p_n(z)$, выберем $\zeta_i \in \gamma_i$. Тогда

$$\left|\frac{1}{n}\sum_{i=1}^{n}\ln|z - \zeta_i| - v(z)\right| \ \leq \ \varepsilon \quad \forall \ \ z \in \Gamma_\varepsilon \quad \Rightarrow$$

$$v(z) - \varepsilon \leq \ln|p_n(z)|^{1/n} \leq v(z) + \varepsilon, \quad z \in \Gamma_\varepsilon \quad \Rightarrow$$

$$-\gamma \leq \ln|p_n(z)|^{1/n} \leq 2\varepsilon - \gamma, \quad z \in \Gamma_\varepsilon \quad \Rightarrow$$

$$e^{-\gamma} \leq |p_n(z)|^{1/n} \leq e^{2\varepsilon - \gamma}, \quad z \in \Gamma_\varepsilon.$$

Можно считать, что ε настолько мало, что 0 принадлежит неограниченной области с границей Γ_ε. Тогда

$$|\ln|p_n(0)|^{1/n} - v(0)| \leq \varepsilon \quad \Rightarrow \quad \ln|p_n(0)|^{1/n} \geq g(0) - \gamma - \varepsilon \quad \Rightarrow$$

$$\max_{z\in\Gamma}\left|\frac{p_n(z)}{p_n(0)}\right|^{1/n} \leq \max_{z\in\Gamma_\varepsilon}\left|\frac{p_n(z)}{p_n(0)}\right|^{1/n} \leq \frac{e^{2\varepsilon-\gamma}}{e^{g(0)-\gamma-\varepsilon}} \quad \Rightarrow$$

$$\max_{z\in\Gamma}\left|\frac{p_n(z)}{p_n(0)}\right|^{1/n} \leq e^{3\varepsilon}e^{-g(0)}.$$

В силу произвольности ε получаем

$$q(\Gamma) = \overline{\lim_{n\to\infty}}(T_n(\Gamma))^{1/n} \ \leq \ e^{-g(0)}.$$

Остается доказать, что в действительности здесь имеет равенство. Пусть

$$m_n(\Gamma) = \max_{z\in\Gamma} |p_n(z)|,$$

и предположим, что последовательность полиномов такова, что

$$(T_n(\Gamma))^{1/n} = \left(\frac{m_n(\Gamma)}{|p_n(0)|}\right)^{1/n} + o(1) \to q_0.$$

Не ограничивая общности, можно считать, что корни полинома $p_n(z)$ не лежат на Γ. Вокруг каждого из корней в области Ω' (если таковые есть) опишем окружность радиуса $\delta > 0$. При достаточно малом δ все полученные круги принадлежат той же области Ω'. Обозначим через $\Omega'(\delta)$ часть области Ω' за вычетом данных кругов.

При достаточно малом δ функция

$$u(z) = g(z) - \frac{1}{n}\ln\frac{|p_n(z)|}{m_n(\Gamma)}$$

удовлетворяет уравнению Лапласа $\Delta u(z) = 0$ при всех $z \in \Omega'(\delta)$, неотрицательна на Γ и построенных окружностях радиуса δ и, кроме того, имеет конечный предел при $z \to \infty$. Отсюда вытекает ее неотрицательность при всех $z \in \Omega'(\delta)$. Будем считать, что δ настолько мало, что $0 \in \Omega'(\delta) \quad \Rightarrow$ $u(0) \geq 0 \quad \Rightarrow$

$$\left(\frac{m_n(\Gamma)}{|p_n(0)|}\right)^{1/n} \geq e^{-g(0)} \quad \Rightarrow \quad q_0 \geq e^{-g(0)} \quad \Rightarrow \quad q(\Gamma) \geq e^{-g(0)}. \qquad \square$$

Следствие 20.11.1 *Последовательность $(T_n(\Gamma))^{1/n}$ является сходящейся и ее предел равен*

$$q(\Gamma) = \lim_{n\to\infty} (T_n(\Gamma))^{1/n}. \tag{20.11.14}$$

Задачи

1. Пусть метод минимальных невязок применяется для решения системы $Ax = b$ и известно, что для некоторого комплексного числа ζ с модулем $|\zeta| = 1$ матрица A является строго эллиптической. Докажите, что в этом случае также выполняется неравенство (20.2.1).

2. Пусть все собственные значения матрицы A принадлежат сектору $\operatorname{Re}\lambda(A) \geq \tau > 0$, $|\lambda(A)| \leq \rho$. Докажите, что оценка (20.2.1) имеет место, если матрица A нормальная, и может не выполняться в противном случае (приведите пример).

3. При некотором $\tau > 0$ для всех векторов x выполняется неравенство $|\mathrm{Re}\,(Ax, x)| \geq \tau(x, x)$ и существует вектор x_0 такой, что $\mathrm{Re}\,(Ax_0, x_0) \geq \tau(x_0, x_0)$. Докажите, что $\mathrm{Re}\,(Ax, x) \geq \tau(x, x)$ для всех x.

4. Пусть x — точное решение системы $Ax = b$ с невырожденной эрмитовой матрицей A, а $x_0, x_1, \ldots$ — приближенные решения, полученные методом минимальных невязок. Докажите, что если $x_{i-1} \neq x$, то $||x_i - x||_2 < ||x_{i-1} - x||_2$.

5. Докажите, что числовая область нормальной матрицы совпадает с выпуклой оболочкой множества ее собственных значений. (Выпуклой оболочкой множества называется минимальное содержащее его выпуклое множество.)

6. Пусть $0 < r < a$ и Γ — окружность радиуса r с центром в точке a на вещественной оси. Пусть $g(z)$ — решение внешней краевой задачи (20.9.9). Докажите, что
$$g(0) = \ln \frac{a}{r}. \tag{20.11.15}$$

7. Кривая Γ является эллипсом с центром в точке a на вещественной оси и полуосями r_1 и r_2,причем $0 < r_1 < a$. Пусть $g(z)$ — решение внешней краевой задачи (20.9.9). Докажите, что
$$g(0) = \ln \frac{\sqrt{a^2 - r_1^2 + r_2^2} + a}{r_1 + r_2}. \tag{20.11.16}$$

8. Докажите теорему 20.8.1.

Глава 21

21.1 Сходимость метода сопряженных градиентов

Рассмотрим разложение невязки $r_0 = \sum\limits_{i=1}^{k} \xi_i \, z_i$ по ортонормированной подсистеме собственных векторов матрицы $A = A^* > 0$ порядка n. Если в разложении участвуют лишь k векторов, то метод сопряженных градиентов получает решение системы $Ax = b$ не позже, чем на k-м шаге. Если отвечающие этим векторам собственные значения попарно различны, то решение получается в точности на k-м шаге (докажите!).

Замечательно, что приближение с интересующей нас точностью часто получается раньше, чем на k-м шаге. Как и в методе минимальных невязок, ключевое наблюдение для получения оценок сходимости заключается в том, $x_i = x_0 + \phi_{i-1}(A)\, r_0$, где $\phi_{i-1}(\lambda)$ — полином степени $i-1$. Отсюда $r_i = \psi_i(A)\, r_0$, где $\psi_i(\lambda) = 1 - \lambda\, \phi_{i-1}(\lambda)$ — однозначно определенный данным методом полином степени i со свободным членом $\psi_i(0) = 1$.

Невязка r для любого вектора вида $x_0 + y$, где y берется в том же подпространстве Крылова, записывается с помощью некоторого, вообще говоря, другого полинома $f(\lambda)$ степени не выше i со свободным членом $f(0) = 1$. Напомним, что множество таких полиномов уже встречалось при изучении метода минимальных невязок и обозначалось $\mathcal{F}_i$. Таким образом, $r = f(A) r_0$, где $f \in \mathcal{F}_i$.

В методе сопряженных градиентов минимальная A-норма ошибки i-го приближения $e_i = A^{-1} r_i$ получается при выборе $f(\lambda) = \psi_i(\lambda) \quad \Rightarrow$

$$\begin{aligned} ||e_i||_A^2 &= (Ae_i,\, e_i) = (r_i,\, A^{-1}\, r_i) \\ &= \sum_{j=1}^{k} \frac{\psi_i^2(\lambda_j)}{\lambda_j} \xi_j^2 \;\le\; \left(\max_{1 \le j \le k} |\psi_i(\lambda_j)| \right)^2 ||e_0||_A^2. \end{aligned}$$

Если $\lambda(A) \subset [m,\, M]$, то, очевидно,

$$||e_i||_A \;\le\; \min_{f \in \mathcal{F}_i} \max_{m \le \lambda \le M} |f(\lambda)|\; ||e_0||_A. \qquad (21.1.1)$$

21.2 Классическая оценка

Запишем $\lambda \in [m, M]$ в виде $\lambda = \frac{M+m}{2} + \frac{M-m}{2}\, t,\ t \in [-1, 1]$, возьмем полином Чебышева $T_i(t)$ для отрезка $[-1, 1]$ и рассмотрим полином

$$f(\lambda) = T_i\left(\frac{\lambda - (M+m)/2}{(M-m)/2}\right) / T_i\left(-\frac{M+m}{M-m}\right).$$

Поскольку $f \in \mathcal{F}_i$, в силу (21.1.1) находим

$$||e_i||_A \ \leq\ \frac{1}{|T_i\,(-\frac{M+m}{M-m})|}\,||e_0||_A. \tag{21.2.2}$$

Эта оценка обнадеживает, так как полиномы Чебышева растут экспоненциально при $|t| > 1$:

$$T_i\,(t) \ =\ \frac{1}{2}\,\left(t + \sqrt{t^2-1}\right)^i + \frac{1}{2}\,\left(t - \sqrt{t^2-1}\right)^i.$$

Отсюда очевидно, например, что $|T_i\,(t)| \geq |t|^i/2 \quad \Rightarrow$

$$||e_i||_A \ \leq\ 2\,\left(\frac{M-m}{M+m}\right)^i\,||e_0||_A. \tag{21.2.3}$$

Эта оценка, заметим, мало отличается от оценки для метода скорейшего спуска. Но не будем спешить с выводами.

Взяв $t = -(1+\nu)/(1-\nu)$ при $\nu = m/M$, получаем

$$\Rightarrow \quad t^2 - 1 \ =\ \frac{(1+\nu)^2 - (1-\nu)^2}{(1-\nu)^2} \ =\ \frac{4\,\nu}{(1-\nu)^2} \quad \Rightarrow$$

$$\sigma \equiv |t| + \sqrt{t^2-1} \ =\ \frac{1+\nu+2\,\sqrt{\nu}}{1-\nu} \ =\ \frac{(1+\sqrt{\nu})^2}{1-\nu} \ =\ \frac{1+\sqrt{\nu}}{1-\sqrt{\nu}}.$$

Учитывая, что $2\,|T_i\,(t)| \ \geq\ 1/\sigma^i$, находим

$$||e_i||_A \ \leq\ 2\,\left(\frac{1-\sqrt{\frac{m}{M}}}{1+\sqrt{\frac{m}{M}}}\right)^i\,||e_0||_A. \tag{21.2.4}$$

Оценка (21.2.4) существенно лучше оценки (21.2.3). Таким образом, мы получаем количественное подтверждение тому, что кажется интуитивно ясным: метод сопряженных градиентов сходится быстрее метода скорейшего спуска.

21.3 Более точные оценки

Оценка (21.2.4) показывает, что сходимость ухудшается для плохо обусловленных матриц. Попробуем объяснить, почему она может оставаться весьма быстрой и в таких случаях. [1]

Пусть $\lambda_1 \geq \ldots \geq \lambda_n$ — собственные значения матрицы $A = A^* > 0$, и предположим, что $\lambda_1 \gg \lambda_2$. В этом случае метод сопряженных градиентов ведет себя так, как будто число обусловленности матрицы равно λ_2/λ_n.

В самом деле, воспользуемся оценкой (21.1.1) и в качестве $f \in \mathcal{F}_i$ возьмем полином

$$f(\lambda) = \frac{T_{i-1}\left(\frac{2\,\lambda-\lambda_2-\lambda_n}{\lambda_2-\lambda_n}\right)}{T_{i-1}\left(-\frac{\lambda_2+\lambda_n}{\lambda_2-\lambda_n}\right)} \left(1-\frac{\lambda}{\lambda_1}\right) \quad \Rightarrow$$

$$||e_i||_A \;\leq\; 2\left(\frac{1-\sqrt{\frac{\lambda_n}{\lambda_2}}}{1+\sqrt{\frac{\lambda_n}{\lambda_2}}}\right)^{i-1} ||e_0||_A. \tag{21.3.5}$$

Если $\lambda_n \ll \lambda_{n-1}$, то картина сходимости такая же: метод "игнорирует" младшее собственное значение и ведет себя так, будто число обусловленности равно λ_1/λ_{n-1}. В этом случае, правда, появляется не зависящий от i коэффициент, равный числу обусловленности матрицы:

$$||e_i||_A \;\leq\; 2\,\frac{\lambda_1}{\lambda_n}\left(\frac{1-\sqrt{\frac{\lambda_{n-1}}{\lambda_1}}}{1+\sqrt{\frac{\lambda_{n-1}}{\lambda_1}}}\right)^{i-1} ||e_0||_A. \tag{21.3.6}$$

21.4 Метод Арнольди и метод Ланцоша

Чтобы лучше понять свойства метода сопряженных градиентов, полезно обратить внимание на его тесную связь с важным для практики методом вычисления собственных значений и векторов эрмитовых матриц, известным как *метод Ланцоша*. В неэрмитовом случае метод минимальных невязок связывается с методом вычисления собственных значений и векторов неэрмитовых матриц, известным как *метод Арнольди*.

Начнем с описания метода Арнольди. Фиксируем вектор $r_0 \neq 0$ и предположим, что векторы $r_0, Ar_0, ..., A^{i-1}r_0$ линейно независимы, а вектор $A^i r_0$ уже является их линейной комбинацией $\Rightarrow\ \mathcal{K}_i = \mathcal{K}_i\,(r_0,\,A) = \mathcal{K}_{i+1}\,(r_0,\,A)$.

В подпространстве $\mathcal{K}_i$ рассмотрим такой ортонормированный базис $q_1, \ldots, q_i$, для которого $\operatorname{span}\{q_1, \ldots, q_j\} = \mathcal{K}_j$ для всех $1 \leq j \leq i$.

[1] O. Axelsson, G. Lindskog, The rate of convergence of the conjugate gradient method, *Numer. Math.*, 48: 499–523 (1986).

Вектор q_{j+1} можно построить путем ортогонализации вектора $A\,q_j$ к уже найденным векторам $q_1, \ldots, q_j$. Положим $Q_j = [q_1 \ \ldots \ q_j]$. Тогда матрица

$$A_j \ = \ Q_j^* \, A \, Q_j$$

называется *проекционным сужением* A на $\mathcal{K}_j$. При $1 \le j \le i$ матрица A_j является ведущей подматрицей в A_i.

В общем случае матрица A_i является верхней хессенберговой (почему?). Если A эрмитова, то A_i — эрмитова трехдиагональная матрица. Процесс построения матриц A_j и Q_j связывается с именем Арнольди в общем случае и с именем Ланцоша — в эрмитовом случае.

При сделанных предположениях $\mathcal{K}_i$ инвариантно относительно A. Поэтому $\lambda\,(A_i) \subset \lambda\,(A)$. Однако, собственные значения проекционных сужений A_j при $j \ < \ i$ также могут неплохо аппроксимировать собственные значения матрицы A. Указанный способ вычисления собственных значений называется методом Арнольди в общем случае и методом Ланцоша — в эрмитовом случае.

В методе Арнольди и в методе Ланцоша строятся ортонормированные базисы в подпространствах Крылова. Заметим, что те же базисы можно получить, анализируя известные матричные разложения и ничего не зная о подпространствах Крылова. Рассмотрим унитарно подобное приведение матрицы $A \in \mathbb{C}^{n\times n}$ к верхней хессенберговой матрице H:

$$A\,[q_1 \ \ldots \ q_n] \ = \ [q_1 \ \ldots \ q_n]\,H.$$

Выбрав q_1 произвольным образом (только одно условие: $||q_1||_2 = 1$), приравниваем первые столбцы:

$$A\,q_1 \ = \ q_1\,h_{11} + q_2\,h_{21}.$$

Далее, $q_1 \perp q_2 \quad \Rightarrow \quad h_{11} = (Aq_1,\, q_1)$. Вектор q_2 получается нормировкой вектора $A\,q_1 - q_1\,h_{11}$, если он не равен нулю. После этого приравниваем вторые столбцы и находим q_3. И так далее. Процесс обрывается на i-м столбце, если претендент на роль q_{i+1} оказался нулем. В этом случае

$$A\,[q_1 \ \ldots \ q_i] \ = \ [q_1 \ \ldots \ q_i]\,H_i,$$

где H_i — ведущая $i \times i$-подматрица в H.

Итак, мы имеем естественный способ генерации подпространств

$$L_j \ = \ \mathrm{span}\,\{q_1, \ldots, q_j\},$$

дающих основу для построения различных проекционных методов. Формально мы обошлись без определения подпространств Крылова. Но если все-таки вспомнить это определение, то придется констатировать, что $L_j = \mathcal{K}_j\,(q_1,\, A)$.

21.5 Числа Ритца и векторы Ритца

Пусть в методе сопряженных градиентов получены ненулевые невязки $r_0, ..., r_{i-1}$. Поскольку $r_j \perp \mathcal{K}_j$, ненулевые невязки образуют ортогональную систему, а итерации, естественно, прекращаются, как только получена нулевая невязка. Проекционное сужение A на $\mathcal{K}_i(r_0, A)$ определяется следующим образом:

$$A_i = Q_i^* A Q_i, \quad Q_i = [q_1, \ldots, q_i] = \left[\frac{r_0}{||r_0||_2}, \ldots, \frac{r_{i-1}}{||r_{i-1}||_2}\right]. \quad (21.5.7)$$

Пусть $\theta_1 \geq \ldots \geq \theta_i$ — собственные значения матрицы A_i. Обычно их называют *числами Ритца*. Если $A_i v_j = \theta_j v_j$, $||v_j||_2 = 1$, то вектор $y_j = Q_i v_j$ называется *вектором Ритца*, отвечающим θ_j.

Находим: $0 = Q_i^*(A Q_i v_j - \theta_j Q_j v_j) = Q_i^*(A y_j - \theta_j y_j)$. Следовательно,

$$A y_j - \theta_j y_j \perp \mathcal{K}_i, \quad 1 \leq j \leq i.$$

Заметим, что матрица A_i является трехдиагональной (докажите!), поэтому ее собственные значения легко вычисляются с помощью известных методов (особенно в случае, когда $i \ll n$).

21.6 Сходимость чисел Ритца

Почему, например, $\theta_1 \approx \lambda_1$? Чтобы ответить на вопрос, используем представление

$$\theta_1 = \max_{\substack{y \in \mathcal{K}_i \\ y \neq 0}} \frac{(Ay, y)}{(y, y)} = \max_{\phi_{i-1}} \frac{(A\phi_{i-1}(A) r_0, \phi_{i-1}(A) r_0)}{(\phi_{i-1}(A) r_0, \phi_{i-1}(A) r_0)},$$

где максимум берется по всем полиномам ϕ_{i-1} степени не выше $i-1$, таким, что $\phi_{i-1}(A) r_0 \neq 0$.

Очевидно, $\theta_1 \leq \lambda_1$. Используя разложение невязки $r_0 = \sum_{j=1}^{n} \xi_j z_j$ по ортонормированным собственным векторам матрицы A, находим:

$$\lambda_1 - \theta_1 \leq \lambda_1 - \frac{(A\phi_{i-1}(A) r_0, \phi_{i-1}(A) r_0)}{(\phi_{i-1}(A) r_0, \phi_{i-1}(A) r_0)} = \lambda_1 - \frac{\sum_{k=1}^{n} \lambda_k |\phi_{i-1}(\lambda_k)|^2 |\xi_k|^2}{\sum_{k=1}^{n} |\phi_{i-1}(\lambda_k)|^2 |\xi_k|^2}$$

$$= \frac{\sum_{k=2}^{n} (\lambda_1 - \lambda_k) |\phi_{i-1}(\lambda_k)|^2 |\xi_k|^2}{|\phi_{i-1}(\lambda_1)|^2 |\xi_1|^2 + \sum_{k=2}^{n} |\phi_{i-1}(\lambda_k)|^2 |\xi_k|^2} \leq (\lambda_1 - \lambda_n) \frac{\max_{2 \leq k \leq n} |\phi_{i-1}(\lambda_k)|^2}{|\phi_{i-1}(\lambda_1)|^2} \gamma,$$

где

$$\gamma = \gamma(r_0) = \frac{\sum_{k=2}^{n} |\xi_k|^2}{|\xi_1|^2}.$$

Полученное неравенство остается в силе для любого полинома ϕ_{i-1} степени не выше $i-1$. Идея: взять такой полином, для которого значение в точке λ_1 много больше значений в точках $\lambda_2, \ldots, \lambda_n$. Таким свойством обладает, например, полином Чебышева для отрезка $[\lambda_n, \lambda_2]$. Выбор

$$\psi_{i-1}(\lambda) = T_{i-1}\left(\frac{2\lambda - \lambda_2 - \lambda_n}{\lambda_2 - \lambda_n}\right) / T_{i-1}\left(-\frac{\lambda_2 + \lambda_n}{\lambda_2 - \lambda_n}\right)$$

приводит к следующей *оценке Пейджа–Каниэля* (проверьте):

$$\lambda_1 - \theta_1 \;\le\; \frac{(\lambda_1 - \lambda_n)}{T_{i-1}^2(1 + 2\,\mu)}\,\gamma(r_0), \quad \mu \;=\; \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n}. \tag{21.6.8}$$

21.7 Важное свойство

Лемма 21.7.1 *Пусть $A = A^* \in \mathbb{C}^{n\times n}$ с помощью унитарной матрицы $Q = [q_1 \ldots q_n]$ приводится к эрмитовой трехдиагональной матрице*

$$A_i \;=\; Q^* A Q \;=\; \begin{bmatrix} \alpha_1 & \beta_1^* & & & & \\ \beta_1 & \alpha_2 & \beta_2^* & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \beta_{n-2} & \alpha_{n-1} & \beta_{n-1}^* \\ & & & & \beta_{n-1} & \alpha_n \end{bmatrix}.$$

Тогда если $\beta_j \neq 0$ при $1 \le j \le i$, то $q_{i+1} = \pi_i(A)\,q_1$ для некоторого полинома π_i степени i, который с точностью до нормировки совпадает с характеристическим полиномом для A_i — ведущей подматрицы порядка i в матрице A.

Доказательство. Легко видеть, что

$$A[q_1 \ldots q_i] \;=\; [q_1 \ldots q_i]\,A_i + \beta_i\,q_{i+1}\,e_i^T, \qquad e_i^T = [0 \ldots 0\;1].$$

Отсюда очевидно существование полиномов $\pi_j(\lambda)$ таких, что $q_{j+1} = \pi_j(A)\,q_1$ при $1 \le j \le i$. Положим $\pi_0(\lambda) = 1$. Тогда имеем

$$\lambda[\pi_0(\lambda), \ldots, \pi_{i-1}(\lambda)] \;=\; [\pi_0(\lambda), \ldots, \pi_{i-1}(\lambda)]\,A_i + \beta_i\,\pi_i(\lambda)\,e_i^T,$$

или, эквивалентно,

$$[\pi_0(\lambda), \ldots, \pi_{i-1}(\lambda)]\,(A_i - \lambda I) \;=\; -\beta_i\,\pi_i(\lambda)\,e_i^T. \qquad \square$$

Если $r_i = \psi_i(A)\,r_0 \neq 0$, то корни полинома $\psi_i(\lambda)$ совпадают с собственными значениями матрицы A_i, то есть с числами Ритца для проекционного сужения A на $\mathcal{K}_i(r_0, A)$.

21.8 "Сверхлинейная сходимость" и "исчезающие" собственные значения

Понятие сверхлинейной сходимости в строгом смысле, конечно, неприменимо к процессу с конечным числом итераций. Но метод сопряженных градиентов обладает некоторыми ее чертами: отношение $\omega_i \equiv ||e_i||_A/||e_{i-1}||_A$ обычно (не монотонно) убывает (для метода с линейной сходимостью $\omega_i \;\approx\; \text{const}$).

Оказывается, в точной арифметике метод сопряженных градиентов через некоторое время начинает вести себя так, как будто в матрице A "исчезли" крайние собственные значения λ_1 и λ_n, после этого через некоторое время он начинает вести себя так, как будто "исчезли" дополнительно λ_2 и λ_{n-1}, и так далее. [2]

"Исчезновение" собственного значения происходит в тот момент, когда оно хорошо аппроксимируется некоторым собственным значением проекционного сужения, порождаемого (виртуально) методом Ланцоша, стартующим с вектора $q_1 \;=\; r_0/||r_0||_2$.

Рассмотрим разложение невязки

$$r_i = \sum_{j=1}^{n} \bar{\xi}_j\, z_j$$

по ортонормированному базису собственных векторов матрицы A и, продолжая итерации, мысленно запустим еще один итерационный процесс, стартующий с невязки

$$\bar{r}_0 = \sum_{j=2}^{n} \bar{\xi}_j\, z_j.$$

Таким образом, наряду с приближениями x_{i+j}, появляются приближения $\bar{x}_j$, порождаемые новым итерационным процессом. Положим

$$e_{i+j} = x_{i+j} - x, \quad \bar{e}_j = \bar{x}_j - x,$$

где x — искомое решение системы $Ax = b$.

[2] A. van der Sluis, H.A. van der Vorst, The rate of convergence of conjugate gradients, *Numer. Math.*, 48: 543–560 (1986).

Теорема 21.8.1 (Вандерслюис–Вандерворст). *Пусть* $\lambda_1 \geq \ldots \geq \lambda_n$ — *собственные значения матрицы* $A = A^* > 0$ *и* θ_1 — *старшее собственное значение проекционного сужения* A_i. *Тогда при* $j = 0, 1, \ldots$ *имеет место неравенство*

$$||e_{i+j}||_A \;\leq\; c_i \, ||\bar{e}_j||_A, \tag{21.8.9}$$

где

$$c_i \;=\; \max_{2\leq k\leq n} \left| \frac{(\lambda_1 - \lambda_k)\,\theta_1}{(\theta_1 - \lambda_k)\,\lambda_1} \right|. \tag{21.8.10}$$

Доказательство. Пусть $r_0 = \sum_{k=1}^n \xi_k \, z_k$. Тогда

$$\begin{aligned} r_i \;&=\; \sum_{k=1}^n \psi_i\,(\lambda_k)\,\xi_k\,z_k, \quad r_{i+j} = \sum_{k=1}^n \psi_{i+j}\,(\lambda_k)\,\xi_k\,z_k, \\ \bar{r}_0 \;&=\; \sum_{k=2}^n \psi_i\,(\lambda_k)\,\xi_k\,z_k, \quad \bar{r}_j \;= \sum_{k=2}^n \psi_i\,(\lambda_k)\,\bar{\psi}_j\,(\lambda_k)\xi_k\,z_k, \end{aligned}$$

где $\bar{\psi}_j\,(\lambda)$ — полином степени не выше j.

Согласно лемме 21.7.1, θ_1 есть корень полинома $\psi_i\,(\lambda)$, и следовательно, выражение

$$\Psi_i\,(\lambda) \;=\; \frac{1 - \frac{\lambda}{\lambda_1}}{1 - \frac{\lambda}{\theta_1}}\,\psi_i\,(\lambda)$$

представляет собой полином степени $i-1$, удовлетворяющий условиям $\Psi_i\,(0) = 1$ и $\Psi_i\,(\lambda_1) = 0$. Вследствие того, что метод сопряженных градиентов минимизирует A-норму ошибки, находим:

$$\begin{aligned} ||e_{i+j}||_A^2 \;&=\; \sum_{k=1}^n \frac{|\psi_{i+j}\,(\lambda_k)|^2}{\lambda_k}\,|\xi_k|^2 \;\leq\; \sum_{k=2}^n \frac{1}{\lambda_k} |\Psi_i\,(\lambda_k)|^2\,|\bar{\psi}_j\,(\lambda_k)|^2\,|\xi_k|^2 \\ &\leq\; \max_{2\leq k\leq n} \left| \frac{1 - \frac{\lambda_k}{\lambda_1}}{(1 - \frac{\lambda_k}{\theta_1})} \right|^2 \sum_{k=2}^n \frac{|\psi_i\,(\lambda_k)|^2}{|\lambda_k|}\,|\bar{\psi}_j\,(\lambda_k)|^2\,|\xi_k|^2 = c_i\,||\bar{e}_j||_A^2. \;\square \end{aligned}$$

21.9 Явные и неявные предобусловливатели

Если итерационный метод для системы $Ax = b$ не спешит сходиться, то обычно пытаются применить его к некоторой равносильной системе $AC^{-1}y = b$. Это и называется *предобусловливанием*.

При умножении на матрицу AC^{-1} последовательно выполняются два действия: решается некоторая система с матрицей C; выполняется умножение на A.

Лучше всего взять $C = A$ (это говорит, кстати, о том, что с помощью предобусловливания сходимость можно ускорить всегда). Но этот выбор не имеет практического смысла.

Матрица C называется *неявным* предобусловливателем. Если переходят к системе $AMy = b$, то M называется *явным* предобусловливателем. При использовании явного предобусловливателя вместо решения системы с матрицей C будет выполняться умножение на M. С точки зрения построения предобусловливателей важно иметь в виду, что $C \approx A$ (в определенном смысле), в то время как $M \approx A^{-1}$.

21.10 Предобусловливание эрмитовых матриц

В случае эрмитовых матриц A и C матрица AC^{-1}, как правило, не будет эрмитовой. Однако, она остается *эрмитовой в обобщенном смысле.*

Матрицы можно рассматривать как операторы в пространстве со скалярным произведением, выбираемым специальным образом, например, так:

$$(x, y)_D \equiv (Dx, y) \quad \text{для некоторой матрицы} \quad D = D^* > 0.$$

Матрица M называется D-*эрмитовой*, если

$$(Mx, y)_D = (x, My)_D \quad \forall\, x, y \in \mathbb{C}^n,$$

и положительно D-определенной, если $(Mx, x)_D > 0$ для всех $x \neq 0$.

Пусть A и C — эрмитовы положительно определенные матрицы. Тогда легко проверить, что матрица AC^{-1} является D-эрмитовой и положительно D-определенной при выборе $D = C^{-1}$. Поэтому систему $AC^{-1}x = b$ можно решать с помощью обычного метода сопряженных градиентов, заменив всюду обычное скалярное произведение $(\cdot, \cdot)$ на $(\cdot, \cdot)_{C^{-1}}$:

$$\begin{aligned}
\bar r_0 &= b - AC^{-1}\bar x_0, \quad \bar p_1 = \bar r_0;\\
\alpha_i &= (C^{-1}\bar r_{i-1},\, \bar r_{i-1}) \;/\; (C^{-1}AC^{-1}\bar p_i,\, \bar p_i),\\
\bar x_i &= \bar x_{i-1} + \alpha_i\, \bar p_i,\\
\bar r_i &= \bar r_{i-1} - \alpha_i\, AC^{-1}\bar p_i,\\
\beta_i &= (\bar r_i,\, \bar r_i) \;/\; (\bar r_{i-1},\, \bar r_{i-1}),\\
\bar p_{i+1} &= \bar r_i + \beta_i\, \bar p_i.
\end{aligned}$$

Для удобства сделаем такую замену:

$$\bar x_i = Cx_i, \quad \bar r_i = r_i, \quad \bar p_i = C^{-1}p_i.$$

В итоге получается следующий *предобусловленный метод сопряженных градиентов*:

$$\begin{array}{rcl} r_0 & = & b - Ax_0,\ p_1 = C^{-1}r_0; \\ \alpha_i & = & (r_{i-1},\, C^{-1}r_{i-1})/(Ap_i,\, p_i), \\ x_i & = & x_{i-1} + \alpha_i\, p_i, \\ r_i & = & r_{i-1} - \alpha_i\, Ap_i, \\ \beta_i & = & (r_i,\, C^{-1}r_i)/(r_{i-1},\, C^{-1}r_{i-1}), \\ p_{i+1} & = & C^{-1}r_i + \beta_i\, p_i. \end{array} \qquad (21.10.11)$$

Заметим, что x_i — приближенное решение, а вектор $r_i = b - Ax_i$ — его невязка по отношению к исходной системе.

Если C и A — эрмитовы положительно определенные матрицы, то все собственные значения предобусловленной матрицы AC^{-1} положительны (почему?). Обычно стремятся к тому, чтобы они попали на отрезок $[m, M]$ как можно меньшей длины. Однако, в силу изученной нами теории быстрая сходимость будет наблюдаться и в том случае, когда большинство (не все, но почти все!) предобусловленных собственных значений находится на малом отрезке $[m, M]$. В таких случаях для анализа сходимости используется изученное в главе 5 понятие *кластеров* собственных значений.

21.11 Оценки числа итераций

Согласно оценке (21.2.4), неравенство $||e_|||_A \le \varepsilon ||e_0||_A$ будет выполнено при условии

$$2\left(\frac{1-\kappa^{-1/2}}{1+\kappa^{-1/2}}\right)^i \le \varepsilon \quad \Leftrightarrow \quad i \ge \ln(2\varepsilon^{-1}) / \ln\frac{1+\kappa^{-1/2}}{1-\kappa^{-1/2}},$$

где $\kappa = M/m$ — спектральное число обусловленности матрицы $A = A^* > 0$. Заметим, что

$$\ln\frac{1+\kappa^{-1/2}}{1-\kappa^{-1/2}} = \ln\left(1 + \frac{2\kappa^{-1/2}}{1-\kappa^{-1/2}}\right) \le \frac{2\kappa^{-1/2}}{1-\kappa^{-1/2}}.$$

Значит,

$$i \ge \frac{1}{2}\ln(2\varepsilon^{-1})\kappa^{1/2} \quad \Rightarrow \quad ||e_i||_A \le \varepsilon ||e_0||_A. \qquad (21.11.12)$$

Данная оценка числа итераций точна на множестве всех матриц с заданным числом обусловленности, но, как мы уже знаем, для многих частных матриц из этого класса она сильно завышена. Поэтому при построении предобусловливателей совсем не обязательно стремиться к минимизации

именно числа обусловленности. Более полезным может оказаться, например, следующее *K-число обусловленности*:

$$K(A) = \frac{(\mathrm{tr}A/n)^n}{\det A}. \qquad (21.11.13)$$

Теорема 21.11.1 [3] *Для невязок предобусловленного метода сопряженных градиентов (21.10.11) имеют место оценки*

$$||r_i||_{C^{-1}} \le \left((K(A))^{1/i} - 1\right)^{i/2} ||r_0||_{C^{-1}}.$$

Следствие 21.11.1 $\quad i \ge \log_2 K(A) + \log_2 \varepsilon^{-1} \quad \Rightarrow \quad ||r_i||_{C^{-1}} \le \varepsilon ||r_0||_{C^{-1}}.$

Доказательство. $t-1 \le (t/2)^2 \ \ \forall\, t \ \ \Rightarrow \ \ (K(A)^{1/i}-1)^{i/2} \le K(A)/2^i.$ $\square$

Доказательство теоремы. Для простоты рассмотрим случай $C = I$. Пусть невязки $r_0, ..., r_i$ ненулевые. Тогда из расчетных формул метода сопряженных градиентов (19.8.6) получаем

$$Ap_i = \frac{1}{\alpha_i}(r_{i-1} - r_i), \quad Ap_{i+1} = \frac{1}{\alpha_{i+1}}(r_i - r_{i+1}), \quad Ap_{i+1} = Ar_i + \beta_i Ap_i \quad \Rightarrow$$

$$AR_{i+1} = R_{i+1}T_{i+1} - \frac{1}{\alpha_{i+1}} r_{i+1}[0...01]^\top, \quad R_{i+1} = [r_0, r_1, ..., r_i],$$

$$T_{i+1} = \begin{bmatrix} \frac{1}{\alpha_1} & -\frac{\beta_1}{\alpha_1} & & & & & \\ -\frac{1}{\alpha_1} & \frac{1}{\alpha_2} + \frac{\beta_1}{\alpha_1} & -\frac{\beta_2}{\alpha_2} & & & & \\ & -\frac{1}{\alpha_2} & \frac{1}{\alpha_3} + \frac{\beta_2}{\alpha_2} & -\frac{\beta_3}{\alpha_3} & & & \\ & & \dots & \dots & \dots & & \\ & & & -\frac{1}{\alpha_{i-1}} & \frac{1}{\alpha_i} + \frac{\beta_{i-1}}{\alpha_{i-1}} & -\frac{\beta_i}{\alpha_i} & \\ & & & & -\frac{1}{\alpha_i} & \frac{1}{\alpha_{i+1}} + \frac{\beta_i}{\alpha_i} \end{bmatrix}, \quad i \ge 1.$$

В силу ортогональности невязок матрица T_{i+1} имеет те же собственные значения, что и проекционное сужение A_{i+1} вида (21.5.7). Пусть $m = i+1$. Находим:

$$\det T_m = \prod_{j=1}^{m} \frac{1}{\alpha_j}, \quad \prod_{j=1}^{i} \beta_j = \frac{||r_i||_2^2}{||r_0||_2^2}.$$

Далее, обозначив через $f(a_1, ..., a_m)$ отношение среднего арифметического к среднему геометрическому чисел $a_1, ..., a_m$ и выбрав произвольное число $0 < \theta < 1$, для $B \equiv (K(T_m))^{1/m}$ получаем

$$B = (1-\theta)^{1/m} f\left(\frac{1}{\alpha_1}, ..., \frac{1}{\alpha_i}, \frac{1-\theta}{\alpha_m}\right) + \theta^{1/m} f\left(\frac{\beta_1}{\alpha_1}, ..., \frac{\beta_i}{\alpha_i}, \frac{\theta}{\alpha_m}\right) \left(\frac{||r_i||_2^2}{||r_0||_2^2}\right)^{1/m},$$

[3] I. E. Kaporin, New convergence results and preconditioning strategies for the conjugate gradient method, *Numer. Linear Algebra with Appl.*, vol. 1, no. 2, 179–210 (1994).

откуда следует, что

$$B \geq (1-\theta)^{1/m} + \theta^{1/m} \left(\frac{||r_i||_2^2}{||r_0||_2^2} \right)^{1/m} \quad \Rightarrow \quad \frac{||r_i||_2^2}{||r_0||_2^2} \leq \left(\frac{B - (1-\theta)^{1/m}}{\theta^{1/m}} \right)^m .$$

После минимизации правой части по θ получаем неравенство

$$||r_i||_2 \leq \left((K(T_m)^{1/i} - 1 \right)^{i/2} ||r_0||_2.$$

Остается доказать, что $K(T_m) \leq K(A)$. Для этого заметим, что $T = T_m$ является ведущей подматрицей порядка $m = i+1$ в матрице $\tilde{A} = \begin{bmatrix} T & S \\ S^* & H \end{bmatrix}$, унитарно подобной исходной матрице A. Для блочно диагональной матрицы $D = \begin{bmatrix} T & 0 \\ 0 & H \end{bmatrix}$ находим: $\mathrm{tr} A = \mathrm{tr} D$ и $\det D^{-1}A \leq 1 \ \Rightarrow \ K(A) \geq K(D)$. Обозначим $\mu_1, ..., \mu_m$ и $\mu_{m+1}, ..., \mu_n$ собственные значения блоков T и H. Пусть $s = (\mu_1 + ... + \mu_m)/m$. Тогда

$$K(D) = (f(\mu_1, ..., \mu_n))^n = ((f(s, ..., s, \mu_{m+1}, ..., \mu_n))^n (f(\mu_1, ..., \mu_m))^m \geq$$

$$\geq (f(\mu_1, ..., \mu_m))^m = K(T). \quad \square$$

Задачи

1. Пусть $0 < m \leq a \leq b \leq M$ и известно, что все собственные значения эрмитовой положительно определенной матрицы A принадлежат отрезку $[a, b]$, за исключением k собственных значений на отрезке $[m, a]$ и l собственныхзначений на отрезке $[b, M]$. Покажите, что для A-норм ошибок в методе сопряженных градиентов выполняются неравенства

$$||e_i||_A \ \leq \ 2 \left(\frac{M}{m} \right)^k \left(\frac{1 - \sqrt{\frac{a}{b}}}{1 + \sqrt{\frac{a}{b}}} \right)^{i-k-l} ||e_0||_A.$$

2. Пусть A — эрмитова положительно определенная матрица порядка n с единичной главной диагональю. Докажите, что для любой положительно определенной диагональной матрицы D

$$\mathrm{cond}_2 (A) \leq n \, \mathrm{cond}_2 (DAD).$$

3. Докажите, что для итераций метода сопряженных градиентов неравенство $||r_i||_2 \leq \varepsilon ||r_0||_2$ выполняется при всех $i \geq k(\varepsilon)$, где $k(\varepsilon) \leq c \ln \varepsilon^{-1} / \ln \ln \varepsilon^{-1}$ и c не зависит от ε. (Используйте теорему 21.11.1 и неравенство $t - 1 \leq (\sigma - 1)^{\sigma-1} (t/\sigma)^\sigma, \ \ \sigma, t > 1.$)

Глава 22

22.1 Операторные уравнения

Рассмотрим общее уравнение $Au = f$, где $A : U \to F$ — заданный линейный оператор, а U и F — нормированные (в общем случае бесконечномерные) пространства.

Прежде чем говорить о методах приближенного решения, необходимо прояснить вопросы существования и единственности решений. Начнем с примера:

$$-u''(x) = f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0. \tag{22.1.1}$$

Как определить U и F? Очевидный выбор пространства U — функции класса $C^2[0,1]$ с дополнительными краевыми условиями $u(0) = u(1) = 0$. Пространство F обязано содержать образ AV — можно взять, например, $F = C[0,1]$. В качестве нормы в U и F можно взять, например, C-норму. [1]

Единственность решения следует из единственности решения однородного уравнения $u'' = 0, \ 0 < x < 1, \ u(0) = u(1) = 0$. В данном случае решение существует для любой правой части $f \in F$ (докажите).

Чтобы построить метод приближенного решения, выберем на $[0,1]$ равномерную сетку $0 = x_0 < x_1 < ... < x_n < x_{n+1} = 1$ с шагом $h = 1/(n+1)$. Если $u(x) \in C^4$, то, используя ряд Тейлора, находим

$$u''(x_i) = \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1})}{h^2} + O(h^2).$$

Для получения приближенных значений $u_i = u_i(n) \approx u(x_i)$, $x_i = x_i(n)$, естественно рассмотреть систему уравнений

$$u_0 = u_{n+1} = 0, \quad -u_{i-1} + 2u_i - u_{i+1} = h^2 f(x_i), \ 1 \le i \le n. \tag{22.1.2}$$

Погрешности $z_i = z_i(n) = u_i(n) - u(x_i)$, $x_i = x_i(n)$, удовлетворяют системе

$$z_0 = z_{n+1} = 0, \quad -z_{i-1} + 2z_i - z_{i+1} = O(h^4).$$

[1] При этом U не будет банаховым -почему?

Отсюда $z_i(n) = O(1/n^2)$ (докажите!).

Заметим, что для многих $f \in F$ условие $u \in C^4$ не выполняется. Будет ли полученный выше метод сходиться и в этом случае? Кроме того, по значениям в узлах x_i можно проинтерполировать значения в других точках отрезка $[0,1]$, получив приближение $\tilde{u}^n(x) \approx u(x)$. Естественно потребовать, чтобы $\tilde{u}^n \in U \ \forall\ n$. Можно ли утверждать, что $\tilde{u}^n \to u$?

22.2 Слабые решения

Ответы на вопросы такого рода можно получать в рамках общей схемы, формирующей *метод конечных элементов* и позволяющей рассматривать более общий класс решений — так называемые *слабые решения.*

Идея заключается в следующем. Пусть $(u,v) = \int\limits_0^1 u(x)v(x)dx$. Тогда в случае уравнения (22.1.1) очевидно, что $(-u'',v) = (u',v')$, если функция $v(x)$ подчиняется краевым условиям $v(0) = v(1) = 0$ и имеет кусочно-непрерывную производную. Пространство "тестовых" функций v обозначим через $\mathcal{V}$. Таким образом, из (22.1.1) вытекает, что

$$(u',v') = (f,v) \quad \forall\ \ v \in \mathcal{V}. \tag{22.2.3}$$

Уравнение (22.2.3) относительно неизвестной функции $u \in \mathcal{V}$ можно рассматривать как *основное уравнение*, дающее более широкий класс решений, чем исходное уравнение. Его решения и называются *слабыми решениями* исходной задачи.

В общем случае задача о слабых решениях ставится таким образом: задана билинейная форма $a(u,v)$, $u,v \in \mathcal{V} \subset \tilde{\mathcal{V}}$ и для $f \in \tilde{\mathcal{V}}$ требуется найти u из уравнения

$$a(u,v) = (f,v), \quad \forall\ \ v \in \mathcal{V}. \tag{22.2.4}$$

Выбор пространств $\mathcal{V}$ и $\tilde{\mathcal{V}}$ определяется особенностями рассматриваемой задачи.

22.3 Метод конечных элементов

Пусть область определения искомой функции разбивается на подобласти S_i без общих внутренних точек, на каждой подобласти S_i фиксируется конечная система линейно независимых "простых" функций и приближение $\tilde{u}(x)$ к искомой функции на S_i ищется в виде их линейной комбинации, коэффициенты которой определяются условиями гладкости функции $\tilde{u}(x)$

на всей области определения и уравнением вида (22.2.4) при выборе некоторого конечного множества функций v. В таких случаях о подобласти S_i обычно говорят как о *конечном элементе*, а вся схема называется *методом конечных элементов*.

Для задачи (22.2.3) в качестве конечных элементов можно взять отрезки $S_i = [x_i, x_{i+1}]$, на которых приближение к решению $u(x)$ ищется в виде

$$\tilde{u}(x) = \alpha_i P_i(x) + \beta_i Q_i(x), \quad P_i(x) = \frac{b_i - x}{b_i - a_i}, \quad Q_i(x) = \frac{x - a_i}{b_i - a_i}.$$

В силу непрерывности функции $\tilde{u}(x)$ получаем, что $\alpha_i = \beta_{i-1} = \tilde{u}_i \equiv u_i$. Кроме того, ясно, что

$$\tilde{u}(x) = \sum_{i=1}^{n} u_i \phi_i(x), \quad \phi_i(x) = \begin{cases} Q_{i-1}(x), & x_{i-1} \le x \le x_i, \\ P_i(x), & x_i \le x \le x_{i+1}, \end{cases} \quad 1 \le i \le n. \tag{22.3.5}$$

В данном случае коэффициенты u_i определяются системой линейных алгебраических уравнений вида

$$\sum_{j=1}^{n} (\phi_j', \phi_i') u_j = (f, \phi_i), \quad 1 \le i \le n, \tag{22.3.6}$$

получаемой из (22.2.3) при $v = v_i$, $1 \le i \le n$. Матрица коэффициентов данной системы лишь скалярным множителем отличается от матрицы коэффициентов системы уравнений (22.1.2) относительно $u_1, ..., u_n$.

22.4 Аппроксимация, устойчивость, сходимость

Пусть U и F — банаховы пространства и $A : U \to F$ — непрерывно обратимый линейный оператор. Рассмотрим операторное уравнение $Au = f$, последовательность конечномерных проекторов

$$P_n : U \to L_n = \operatorname{im} P_n, \quad Q_n : F \to L_n' = \operatorname{im} Q_n$$

и *проекционное уравнение*

$$(Q_n A P_n) u_n = Q_n f, \; u_n = P_n u_n. \tag{22.4.7}$$

Проекционный метод — это метод получения приближенных решений уравнения $Au = f$ путем решения уравнений вида (22.4.7). Метод называется *сходящимся*, если для всех достаточно больших n проекционные уравнения однозначно разрешимы для любой правой части $f \in F$ и $u_n \to A^{-1} f$ при $n \to \infty$. При исследовании сходимости важнейшими являются следующие два свойства:

- *аппроксимация*: $P_n u \;\to\; u \,\forall\, u \in U, \quad Q_n f \;\to\; f \,\forall\, f \in F.$
- *устойчивость*: для всех достаточно больших n

$$||(Q_n A P_n)u||_F \;\geq\; c\,||P_n u||_U \quad \forall\, u \in U, \quad c > 0.$$

Лемма 22.4.1 *Пусть $z \in U$ — решение операторного уравнения $Au = f$. Если проекционный метод обладает свойством устойчивости, то для всех достаточно больших n выполняется неравенство*

$$||u_n - z||_U \;\leq\; (1 + c^{-1}\,||Q_n||\,||A||)\,||z - P_n z||_U.$$

Доказательство. Вычтем $Q_n A P_n z = Q_n A P_n z$ из $Q_n A P_n u_n = Q_n f$. Тогда $(Q_n A P_n)\,(u_n - P_n z) = Q_n A\,(z - P_n z)$. Вследствие устойчивости, для всех достаточно больших n оператор $Q_n A P_n$ осуществляет обратимое отображение подпространства $\operatorname{im} P_n$ на подпространство $\operatorname{im} Q_n$. Поэтому

$$u_n - z = -(z - P_n z) + (Q_n A P_n)^{-1}\,Q_n A\,(z - P_n z). \;\square$$

Следствие 22.4.1 *Пусть проекционный метод является устойчивым и $||Q_n||\,||u - P_n u|| \to 0$. Тогда $u_n \to u$.*

Теорема 22.4.1 *Если проекционный метод обладает свойствами устойчивости и аппроксимации, то он является сходящимся.*

Доказательство. По теореме Банаха–Штейнгауза, $||Q_n|| \;\leq\; M \;\leq\; +\infty$. Остается учесть следствие 22.4.1. $\square$

Отметим также следующее свойство *квазиоптимальности*:

$$||u_n - u|| \;\leq\; C \inf_{v \in L_n} ||v - u||, \quad C > 0.$$

22.5 Метод Галеркина

На практике проекционный метод для уравнения $Au = f$ чаще всего реализуется как *метод Галеркина*: выбираются линейно независимые функции $\phi_1, ..., \phi_n$, а приближенное решение вида $u_n = \alpha_1\phi_1 + ... + \alpha_n\phi_n$ находится из *уравнений Галеркина*:

$$\sum_{j=1}^{n} (A\phi_j, \phi_i)\alpha_j = (f, \phi_i), \quad 1 \leq i \leq n. \tag{22.5.8}$$

В общем случае скобки обозначают билинейную или полуторалинейную форму (v, u), определенную при всех $u \in U$ и $v \in F$. Матрица с элементами $a_{ij} = (A\phi_j, \phi_i)$ называется *матрицей моментов*.

Билинейная или полуторалинейная форма (v,u) называется *невырожденной*, если $\forall\ u\ \exists\ v : (v,u) \neq 0$ и $\forall\ v\ \exists\ u : (v,u) \neq 0$. Пара нормированных пространств U, F, снабженная невырожденной билинейной (полуторалинейной) формой $(v,u), u \in U, v \in F$, называется *дуальной парой*.

В случае дуальной пары пространств U, F можно доказать,что для любой линейно независимой системы $\phi_1, ..., \phi_n \in U$ существует *дуальная* система $\psi_1, ..., \psi_n \in F$ — дуальность означает, что $(\psi_i, \phi_j) = \delta_{ij}$ (1 при $i = j$ и 0 при $i \neq j$).

Проекторы P_n и Q_n, соответствующие методу Галеркина, можно определить, например,таким образом:

$$P_n u = \sum_{j=1}^{n} \phi_j (\psi_j, u), \quad Q_n v = \sum_{i=1}^{n} \psi_i (v, \phi_i).$$

Действительно, запишем $u_n = P_n u_n = \sum\limits_{j=1}^{n} \alpha_j \phi_j$. Тогда

$$Q_n A P_n u_n = \sum_{i=1}^{n} \psi_i \sum_{j=1}^{n} \alpha_j \, (A\phi_j, \phi_i), \quad Q_n f = \sum_{i=1}^{n} \psi_i (f, \psi_i) \quad \Rightarrow \quad (22.5.8).$$

Лемма 22.5.1 *Пусть оператор $A : U \to F$ и дуальная пара U, F таковы, что*

$$c_1 ||u||_U^2 \leq (Au, u) \leq c_2 ||Au||_F ||u||_U \quad \forall \ \ u \in U, \qquad c_1, c_2 > 0.$$

Тогда соответствующий методу Галеркина проекционный метод является устойчивым.

Доказательство. Достаточно заметить, что

$$(Q_n A P_n u, P_n u) = (A P_n u, P_n u). \qquad \square$$

22.6 Компактные возмущения

Напомним, что линейный непрерывный оператор $K : U \to F$ называется *компактным*, если для любой ограниченной последовательности $u_n \in U$ из последовательности $Au_n \in F$ можно выделить сходящуюся подпоследовательность.

Теорема 22.6.1 *Пусть U и F — банаховы пространства, $A : U \to F$ и $B : U \to F$ — линейные непрерывно обратимые операторы и при этом оператор $K = B - A$ является компактным. Тогда любой проекционный метод со свойствами аппроксимации и устойчивости для оператора A обладает теми же свойствами и для оператора B.*

Доказательство. Пусть $P_n : U \to U$ и $Q_n : F \to F$ — проекторы данного метода, $A_n = Q_n A P_n$ и $K_n = Q_n K P_n$. Согласно условию теоремы, $||A_n u|| \geq ||P_n u|| \; \forall\, u \in U$. От противного, допустим, что существует последовательность $u_n = P_n u_n$ такая, что

$$||(A_n + K_n)u_n|| \to 0, \quad ||u_n|| = 1. \tag{22.6.9}$$

Не ограничивая общности, можно считать, что последовательность Ku_n сходится. Пусть $Ku_n \to v \;\Rightarrow\; A^{-1}Ku_n \to A^{-1}v \;\Rightarrow$

$$A_n^{-1}K_n u_n \to A^{-1}v. \tag{22.6.10}$$

Последнее проверяется непосредственно:

$$A_n^{-1}K_n u_n - A^{-1}v = (A_n^{-1}Q_n K_n u_n - A_n^{-1}Q_n v) + (A_n^{-1}Q_n v - A^{-1}v).$$

Последовательность в первых скобках стремится к нулю, потому что норма $||A_n^{-1}Q_n||$ ограничена равномерно по n. Последовательность во вторых скобках сходится к нулю, так как проекционный метод для уравнения $Au = v$ является сходящимся.

Вследствие (22.6.9) и в силу устойчивости $(I + A_n^{-1}K_n)u_n \to 0$. Из (22.6.10) вытекает, что $u_n \to u = -A^{-1}v$. Следовательно,

$$(A + K)u = 0 \quad \Rightarrow \quad u = 0,$$

что противоречит равенству $||u|| = 1$. $\square$

22.7 Формы и операторы

Задачу (22.2.4) о слабых решениях можно записать также как некоторое операторное уравнение.

Пусть $\hat{U}$ — пространство линейных функционалов на $\mathcal{V}$ вида $\hat{u}(v) = (u, v)$, где $u \in \mathcal{V}$, и пусть $\hat{F}$ — пространство линейных функционалов на $\mathcal{V}$ вида $\hat{f}(v) = (f, v)$, где $f \in \tilde{\mathcal{V}}$. Тогда оператор $\mathcal{A} : \hat{U} \to \hat{F}$ определяется правилом $(u, v) \mapsto a(u, v)$, а уравнение (22.2.4) принимает операторный вид $\mathcal{A}\hat{u} = \hat{f}$.

Таким образом, теория проекционных методов применима и к задаче о слабых решениях. В возникающих естественным образом постановках нормированные пространства обычно неполные, но их всегда можно вложить в некоторое банахово пространство в силу известной *теоремы о пополнении пространства*: любое нормированное пространство является подпространством некоторого банахова пространства.

22.8 Существование решений

При изучении задачи (22.2.4) полезно иметь в виду следующую теорему Рисса.

Теорема 22.8.1 *Пусть H — гильбертово пространство со скалярным произведением $(\cdot,\cdot)$. Тогда для любого ограниченного линейного функционала $f(x)$ на H существует элемент $u \in H$ такой, что $(x,u) = f(x)$.*

Доказательство вытекает из теоремы о том, что в случае замкнутого линейного подпространства $L \subset H$ для любого вектора $x \in H$ существует и единственно разложение $x = z+h$, где $z \in L$ и $h \perp L$ (h — перпендикуляр, опущенный из x на L). В силу ограниченности линейного функционала f подпространство $L = \{v \in H : f(v) = 0\}$ является замкнутым. Пусть $0 \neq h \perp L$. Тогда $f(h) \neq 0$ и $x - (f(x)/f(h))h \in L \ \ \forall\ x \in H$. Поэтому $x = z + \alpha h$, $\alpha = (z,h)/(h,h)$. Положим $u = \frac{\overline{f(h)}}{(h,h)}h$. Тогда

$$(x,u) = \frac{f(h)}{(h,h)}(z,h) = \alpha f(h) = f(x). \quad \square$$

Следствие 22.8.1 *Пусть билинейная форма задачи (22.2.4) такова, что для некоторых положительных констант $c_1, c_2 > 0$*

$$c_1||v||_{\mathcal{V}}^2 \le a(v,v) \le c_2||v||_{\mathcal{V}}^2. \tag{22.8.11}$$

Предположим также, что для $c_3 > 0$

$$|(f,v)| \le c_3||v||_{\mathcal{V}} \quad \forall \ \ v \in \mathcal{V} \tag{22.8.12}$$

и пространство $\mathcal{V}$ полное. Тогда задача (22.2.4) имеет и притом единственное решение.

Для доказательства достаточно заметить, что $a(u,v)$ задает на $\mathcal{V}$ скалярное произведение, превращающее $\mathcal{V}$ в гильбертово пространство.

В задаче (22.2.3) $a(u,v) = (u',v')$, а $\mathcal{V}$ — пополнение пространства функций $u(x) \in C^1$ с краевыми условиями $u(0) = u(1) = 0$ относительно нормы

$$||u||_{\mathcal{V}} \equiv \sqrt{(u,u)} + \sqrt{(u',u')}.$$

Если $u \in C^1$, то из краевых условий вытекает, что $u(x) = \int\limits_0^x u'(t)dt \quad \Rightarrow$

$u^2(x) \le \int\limits_0^1 (u'(t))^2 dt \quad \Rightarrow \quad (u,u) \le (u',u') \quad \Rightarrow \quad$ (22.8.11).

Аналогичные построения позволяют доказать существование слабых решений в более общей задаче

$$(a(x)u')' = f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0. \tag{22.8.13}$$

Соответствующая задача о слабых решениях имеет вид

$$(a(x)u', v') = (f, v) \quad \forall \ \ v \in \mathcal{V}, \tag{22.8.14}$$

и рассматривается в предположении, что $a(x)$ — кусочно-непрерывная функция, удовлетворяющая неравенству $a(x) \geq a_0 > 0$ при $0 \leq x \leq 1$.

22.9 Теория Рисса–Фредгольма

Если A — квадратная матрица, то уравнение $Ax = b$ имеет решение для любой правой части b в том и только том случае, когда $\ker A = 0$. То же верно и для широкого класса операторных уравнений — в случае операторов вида $A = I - K$, где $K : X \to X$ — компактный оператор на банаховом пространстве X.

Прежде всего заметим, что $\ker A$ — замкнутое подпространство конечной размерности. В самом деле, $\ker A \subset \mathrm{im} K$, а сфера в бесконечномерном пространстве не является компактным множеством. Поскольку произведение компактного и ограниченного операторов является компактным оператором, каждое из ядер $\ker A^k$ при $k = 1, 2, ...$ конечномерно. Очевидно, что $\ker A \subset \ker A^2 \subset ...$

Допустим, что можно выбрать бесконечную последовательность векторов $x_k \in \ker A^k$. Тогда ее можно выбрать таким образом, что $||x_k|| = 1$ $\forall\ k$ и $||x_i - x_j|| \geq 1$ при $i \neq j$. При этом $x_k \in \mathrm{im} K$, а значит, должна быть какая-то сходящая подпоследовательность x_{k_i}, что невозможно в силу предыдущих соотношений. Следовательно, для некоторого номера m выполняется равенство $\ker A^m = \ker A^{m+1}$. Тогда элементарно проверяется, что $\ker A^k = \ker A^m$ для всех $k \geq m$. Будем считать, что m — наименьший номер с таким свойством; число m называется *индексом* оператора A.

Далее, можно доказать, что подпространства $\mathrm{im} A^k$ замкнуты, а цепочка вложений $\mathrm{im} A \supset \mathrm{im} A^2 \supset ...$ такова, что $\mathrm{im} A^n = \mathrm{im} A^k$ при $k \geq n$. Пусть n — наименьший номер с таким свойством. Тогда непременно $m = n$. В самом деле, если $m < n$, то любой вектор вида $A^{n+1}z$ можно записать в виде

$$A^{n+1}z = A^n y \ \Rightarrow \ A^n(Az - y) = 0 \ \Rightarrow \ A^{n-1}(Az - y) = 0 \ \Rightarrow \ \mathrm{im} A^n \subset \mathrm{im} A^{n-1},$$

что противоречит минимальности n при условии $\mathrm{im}A^n = \mathrm{im}A^{n+1}$. Если $m > n$, то любой вектор $y \in \mathrm{ker}A^{m+1}$ удовлетворяет уравнению $A^{m+1}y = 0$ и для некоторого z имеем $A^{m+1}y = A^m z$. Отсюда получаем

$$A^m(Ay - z) = 0 \;\Rightarrow\; A^{m-1}(Ax - z) = 0 \;\Rightarrow\; \mathrm{ker}A^m \subset \mathrm{ker}A^{m-1},$$

что противоречит минимальности n. Итак, $m = n$.

Теорема 22.9.1 *Пусть m — индекс оператора $A = I - K$, где K — компактный оператор на банаховом пространстве X. Тогда X является прямой суммой замкнутых подпространств:*

$$X = \mathrm{ker}A^m + \mathrm{im}A^m.$$

Доказательство. Пусть $x \in X$. Тогда $A^m x = A^{2m}y$ для некоторого $y \in X$ $\Rightarrow \quad z \equiv x - A^m y \in \mathrm{ker}A^m$. Таким образом, имеет место разложение

$$x = v + z, \quad v \in \mathrm{im}A^m, \;\; z \in \mathrm{ker}A^m.$$

Кроме того, v и z определены однозначно: если $x \in \mathrm{im}A^m \cap \mathrm{ker}A^m$, то $x = A^m y \;\Rightarrow\; A^{2m}y = 0 \;\Rightarrow\; A^m y = 0 \;\Rightarrow\; x = 0$. $\square$

Следствие 22.9.1 *Оператор $A = I - K$ непрерывно обратим тогда и только тогда, когда* $\mathrm{ker}A = 0$.

Доказательство. Если $\mathrm{ker}A = 0$, то $\mathrm{ker}A = \mathrm{ker}A^2$. Поэтому индекс оператора A равен 1. Согласно теореме 22.9.1, $\mathrm{im}A = X$, поэтому обратный оператор A^{-1} существует. Остается доказать его ограниченность.

От противного, пусть $||x_k|| = 1$ и $||A^{-1}x_k|| \to \infty$. Тогда векторы $z_k = A^{-1}x_k/||A^{-1}x_k||$ принадлежат единичной сфере и $Az_k = z_k - Kz_k \to 0$. В силу компактности K можно выбрать сходящуюся подпоследовательность Kz_{k_i}. Значит, $z_{k_i} \to z$. При этом $||z|| = 1$ и $Az = 0$, что противоречит условию $\mathrm{ker}A = 0$. $\square$

22.10 Сопряженные операторы

Итак, если $A = I - K$, то уравнение $Au = f$ имеет решение при условии $\mathrm{ker}A = 0$. В теории Фредгольма этот результат дополняется полезным утверждением о равенстве размерностей ядер оператора A и его сопряженного оператора A'. Пусть банаховы пространства X, Y образуют дуальную пару с невырожденной билинейной формой $\langle u, v\rangle$, $u \in X$, $v \in Y$.

Определение. Линейные операторы $A : X \to X$ и $A' : Y \to Y$ называются *сопряженными* относительно невырожденной билинейной формы $\langle \cdot, \cdot \rangle$, если $\langle Au, v\rangle = \langle u, A'v\rangle \;\forall\; u \in X,\; v \in Y$.

Теорема 22.10.1 *Пусть $K : X \to X$ и $K' : Y \to Y$ — компактные операторы, сопряженные относительно невырожденной билинейной формы $\langle \cdot, \cdot \rangle$. Тогда*

$$\dim \ker(I - K) = \dim \ker(I - K').$$

Доказательство. Пусть $u_1, ..., u_m$ — базис в ядре оператора $A = I - K$ и $v_1, ..., v_m \in Y$ — произвольная система векторов со следующими свойствами биортогональности (существование такой системы гарантируется невырожденностью билинейной формы):

$$\langle u_i, v_j \rangle = \delta_{ij}, \quad 1 \le i, j \le m.$$

Рассмотрим также базис $v'_1, ..., v'_n$ в ядре оператора $A' = I - K'$ и систему векторов $u'_1, ..., u'_n \in X$ с аналогичными свойствами биортогональности:

$$\langle u'_i, v'_j \rangle = \delta_{ij}, \quad 1 \le i, j \le n.$$

Предположим, что $m \le n$, и определим линейный оператор $\tilde{A} : X \to X$ правилом

$$\tilde{A}u = Au + \sum_{i=1}^{m} \langle u, v_i \rangle u'_i.$$

Пусть $\tilde{A}u = 0$. Тогда $Au = \sum\limits_{i=1}^{m} \alpha_i u'_i \quad \Rightarrow \quad 0 = \langle Au, v'_j \rangle + \alpha_j = 0 + \alpha_j \quad \Rightarrow$ $\alpha_j = 0 \; \forall \, j$. Значит, $Au = 0 \;\Rightarrow\; u = \sum\limits_{i=1}^{m} \beta_i u_i \;\Rightarrow\; \langle \tilde{A}u, v'_j \rangle = \langle Au, v'_j \rangle + \beta_j = \beta_j$ $\Rightarrow \quad \beta_j = 0 \; \forall \; j$. Окончательно, $u = 0$.

Мы доказали, что $\ker \tilde{A} = 0$. Легко видеть, что к оператору $\tilde{A}$ применима теорема 22.9.1. Поэтому оператор $\tilde{A}$ является обратимым. Если $m < n$, то для некоторого $u \in X$ находим

$$Au + \sum_{i=1}^{m} \langle u, v_i \rangle u'_i = u'_{m+1}.$$

Отсюда $1 = \langle Au, v'_{m+1} \rangle + \sum\limits_{i=1}^{m} \langle u, v_i \rangle \langle u'_i, v'_{m+1} \rangle = 0$, что приводит к очевидному противоречию. Следовательно, $m \ge n$.

Если допустить, что $m > n$, то противоречие возникает при рассмотрении оператора $\tilde{A}' : Y \to Y$, определяемого аналогичным образом; поэтому $m \le n$. $\square$

Следствие 22.10.1 *Уравнение $Au = f$ разрешимо в том и только случае, когда $\langle f, v \rangle \;\; \forall \; v \in \ker A'$.*

В качестве еще одного следствия можно отметить *альтернативу Фредгольма*: либо уравнение $Au = f$ имеет единственное решение при любой правой части f, либо однородное сопряженное уравнение $A'v = 0$ имеет нетривиальное решение.

22.11 Интегральные уравнения

Теория Рисса–Фредгольма особенно полезна при изучении интегральных операторов.

Пусть Γ — гладкий замкнутый контур на комплексной плоскости и рассматривается уравнение вида

$$u(z) - \int\limits_{\Gamma} K(z,\zeta)u(\zeta)|d\zeta| = f(z), \quad z \in \Gamma. \tag{22.11.15}$$

В качестве невырожденной билинейной формы естественным образом выбирается

$$\langle u, v\rangle = \int\limits_{\Gamma} u(\zeta)v(\zeta)|d\zeta|.$$

Тогда сопряженное уравнение имеет вид (проверьте!)

$$v(\zeta) - \int\limits_{\Gamma} K(z,\zeta)v(z)|dz| = g(\zeta), \quad z \in \Gamma. \tag{22.11.16}$$

Функция $K(z,\zeta)$ называется *ядром интегрального уравнения*. Типичный пример:

$$K(z,\zeta) = \frac{(\zeta - z, \vec{n}_\zeta)}{\pi\,|\zeta - z|^2}, \tag{22.11.17}$$

где $\vec{n}_\zeta$ обозначает внешнюю нормаль к Γ в точке ζ и $(\cdot,\cdot)$ — скалярное произведение векторов на плоскости. В этом случае интеграл в уравнении (22.11.15) называется *потенциалом двойного слоя*.

В теории потенциала доказывается, что

$$\int\limits_{\Gamma} \frac{(\zeta - z, \vec{n}_\zeta)}{|\zeta - z|^2}|d\zeta| = \begin{cases} 2\pi, & z \text{ внутри } \Omega, \\ \pi, & z \in \Gamma, \\ 0, & z \text{ вне } \Gamma. \end{cases}$$

Отсюда следует,что решением однородного уравнения для потенциала двойного слоя является любая константа.

Можно доказать также, что любые два решения линейно зависимы. Поэтому ядро одномерно. Значит, ядро однородного сопряженного уравнения

(22.11.16) также одномерно. Вывод: однородное сопряженное уравнение имеет нетривиальное решение. Отсюда вытекает, в частности, что уравнение для логарифмического потенциала [2]

$$\int_\Gamma \ln|z-\zeta|\, u(\zeta)|d\zeta| = c, \quad z \in \Gamma,$$

имеет нетривиальное непрерывное решение $u(\zeta)$ для некоторой константы c. Для доказательства нужно учесть, что при дифференцировании обеих частей по нормали получается уравнение типа (22.11.16).

22.12 Функциональные пространства

При изучении операторного уравнения $\mathcal{A}u = f$ важно правильно выбрать функциональные пространства для u и v. Часто это можно сделать многими способами, при этом от выбора пространств зависит, будет ли оператор непрерывно обратим или нет.

В случае интегральных уравнений на замкнутых кривых можно считать, что кривые заданы функциями от параметра $t \in [0, 2\pi]$. Тогда u и v можно рассматривать как 2π-периодические функции, определенные на всей прямой. Рассмотрим ряд Фурье

$$u(t) = \sum_{k=-\infty}^{\infty} u_k \, \exp(\mathbf{i}kt),$$

и, выбрав любое вещественное число s, положим

$$||u||_s^2 \equiv |u_0|^2 \; + \; \sum_{k\neq 0} |k|^{2s} \, |u_k|^2.$$

Эти величины определены корректно, если $u \in C^\infty$, и обладают всеми свойствами нормы.

Обозначим через H^s пространство, в котором каждый элемент ассоциируется с классом эквивалентных последовательностей Коши функций из C^∞ по отношению к норме $||\cdot||_s$ (две последовательности называются эквивалентными, если их разность стремится к нулю). Пространства H^s известны как *пространства Соболева*. Заметим, что для отрицательных s некоторые элементы из H^s не очень похожи на обычные функции — их иногда называют *распределениями*.

[2] Данное уравнение возникло в главе 20 при изучении предельной скорости сходимости метода минимальных невязок.

После очевидного переобозначения интегральное уравнение с логарифмическим ядром принимает вид

$$\int_0^{2\pi} K(x(\tau), y(t))\, u(t)\, dt = f(\tau), \quad 0 \le \tau \le 2\pi; \qquad K(x,y) = \frac{1}{\pi}\, \ln|x-y|.$$

Пусть кривая есть окружность радиуса a с центром в начале координат. Для точек $x = a\, e^{i\tau}$ и $y = a\, e^{it}$ находим

$$K(x,y) \equiv k(\tau, t) = \frac{1}{\pi}\, \ln(a\,|1 - e^{i(\tau - t)}|).$$

При $|z| < 1$ имеет место разложение $\ln(1-z) = -\sum\limits_{k=1}^{\infty} \frac{z^k}{k}$. Взяв $z = e^{i\phi}$, получаем формальное разложение $\ln|1 - e^{i\phi}| = -\sum\limits_{k=1}^{\infty} \frac{\cos k\phi}{k}$. Далее,

$$\begin{aligned} [\mathcal{K}u](\tau) &= \frac{1}{\pi} \int_0^{2\pi} \left(\ln a - \sum_{k\neq 0} \frac{e^{ik(\tau - t)}}{2\,|k|} \right) \left(\sum_m u_m\, e^{imt} \right) dt \\ &= 2\, u_0 \ln a \; - \; \sum_{k\neq 0} \frac{u_k}{|k|}\, e^{ik\tau}. \end{aligned}$$

Отсюда вытекает следующее предложение.

Лемма 22.12.1 [3] *Если $a \neq 1$, то оператор $\mathcal{K}$ с логарифмическим ядром осуществляет взаимно-однозначное непрерывно обратимое отображение $\mathcal{K}:\ H^s \to H^{s+1}\ \ \forall\ s \in \mathbb{R}$.*

Если оператор $\mathcal{K}$ рассматривать как оператор из H^0 в H^0, то уравнение $\mathcal{K}u = f$ может и не иметь решения для заданной правой части $f \in H^0$. Если же решение есть, то в норме H^0 оно может сильно измениться при малом возмущении f в той же норме. В то же время возможен выбор пространств для u и v, при котором оператор оказывается непрерывно обратимым.

Задачи

1. Докажите, что для любой непрерывной на $[0,1]$ функции уравнение $u''(x) = f(x),\ 0 < x < 1,\ \ u(0) = u(1) = 0$, имеет решение класса C^2.

[3] В случае произвольного гладкого контура то то же самое верно "почти для любого" контура.

2. Чтобы решить уравнение $u''(x) = f(x),\ 0 < x < 1,\ \ u(0) = u(1) = 0$, на $[0,1]$ выбираются равномерные сетки $0 = x_0 < x_1 < ... < x_n < x_{n+1} = 1$ с шагом $h = 1/(n+1)$ и решается система линейных алгебраических уравнений

$$u_0 = u_{n+1} = 0, \quad -u_{i-1} + 2u_i - u_{i+1} = -h^2 f(x_i),\ 1 \le i \le n.$$

Величины $u_i = u_i(n)$ рассматриваются как приближения к $u(x_i)$. Докажите, что если $f \in C^2$, то погрешности $z_i(n) = u_i(n) - u(x_i)$ имеют вид $O(1/n^2)$. Верно ли, что $z_i(n) \to 0$ при $n \to \infty$, если f — произвольная непрерывная функция f?

3. Пусть $\mathcal{V}$ — пополнение пространства функций $u(x) \in C^1[0,1]$ с краевыми условиями $u(0) = u(1) = 0$ относительно нормы $||u||_{\mathcal{V}} \equiv \sqrt{(u,u)} + \sqrt{(u',u')}$. Пусть $a(x) \in C[0,1]$ и $a(x) \ge a_0 > 0$ при $0 \le x \le 1$. Докажите существование слабого решения $u \in \mathcal{V}$ в задаче

$$(a(x)u', v') = (f, v) \quad \forall \ \ v \in \mathcal{V},$$

где f — кусочно-непрерывная функция на $[0,1]$.

4. Пусть $K(z,\zeta)$ имеет вид (22.11.17). Докажите, что уравнение (22.11.15) имеет непрерывное решение $v(z)$ для любой непрерывной правой части $g(z)$, подчиненной условию $\int\limits_\Gamma g(z)|dz| = 0$.

5. Если $k(\tau,t)$ — непрерывная функция от τ и t, то интегральный оператор $[\mathcal{K}u](\tau) = \int\limits_0^{2\pi} k(\tau,t)\,u(t)\,dt$ является компактным оператором из $C[0,2\pi]$ в $C[0,2\pi]$ (пространства с чебышевской нормой).

6. Пусть в условиях предыдущей задачи оператор $I + \mathcal{K} : C[0,2\pi] \to C[0,2\pi]$ обратим, а для решения уравнения $(I+\mathcal{K})u = f$ применяется проекционный метод, в котором $P_n = Q_n$ — интерполяционный проектор на чебышевской сетке на отрезке $[0,2\pi]$. Докажите, что $||u_n - u||_{C[0,2\pi]} \to 0$, если u — непрерывно дифференцируемая функция.

7. Дан интегральный оператор $[\mathcal{K}u](x) = \int\limits_0^1 K(x,y)\,u(y)\,dy$ с непрерывным ядром $K(x,y)$. Пусть M_n — матрица моментов для разложения по системе кусочно-постоянных функций на равномерной сетке с n узлами. Докажите, что сингулярные числа и собственные значения последовательности матриц M_n имеют кластер в нуле. Будет ли кластер собственным?

Глава 23

23.1 Многосеточный метод

Рассмотрим простейшую модельную задачу (22.1.1) и предположим, что для ее решения используется метод конечных элементов, в котором приближение к решению ищется в виде $\tilde{u}(x) = \sum_{i=1}^{n} u_{ni}\phi_i(x)$ с "пирамидальными" функциями $\phi_i(x)$ вида (22.3.5) на равномерной сетке с шагом $h_n = 1/(n+1)$. Матрица A_n алгебраической системы для определения коэффициентов u_i в данном случае имеет вид

$$A_n = \frac{1}{h_n}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ \dots & \dots & \dots & \dots & \dots \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}. \qquad (23.1.1)$$

Пусть уже найдено какое-то приближение u_n^0 к решению u алгебраической системы $A_n u_n = f_n$. Тогда следующее приближение u_n^1 можно искать в виде $u_n^1 = u_n^0 + v_n^0$, где $v_n^0 \approx v_n$ и v_n есть решение системы

$$A_n v_n = r_n \equiv f_n - A_n u_n^1.$$

Во многих типичных случаях невязка r_n соответствует функции "более гладкой", чем правая часть исходного уравнения. Тогда для того, чтобы найти приближение к v_n, можно взять "более грубую" сетку и функцию, соответствующую вектору v_n, приблизить некоторой линейной комбинацией меньшего числа "пирамидальных" функций на "более грубой" сетке путем решения алгебраического уравнения, аналогичного исходному, но с меньшим числом неизвестных.

Эта замечательная идея была предложена в 1960-х годах Р. П. Федоренко и Н. С. Бахваловым. Они же дали ее первоначальное обоснование на примере уравнения Пуассона. В 1970-х годах та же идея развивалась в работах А. Брандта, а впоследствии стала основой исключительно популярного класса *многосеточных методов*.

23.2 Алгебраическая формулировка

Мы постараемся обойтись без функций и операторов. Надо решить систему $Au = f$ порядка n. Мы намерены сделать это, используя (приближенные) решения специально подбираемых "меньших" систем $A_j u_j = f_j$ порядка $n_0 < \cdots < n_k = n$.

Все матрицы будут симметричными положительно определенными. Собственно, таковой предполагается исходная матрица A. Для остальных матриц это уже следует из способа построения.

Матрица $A_{k-1} = \hat{A}$ определяется по $A_k = A$ с помощью *интерполяционной матрицы* Q следующим образом:

$$\hat{A} = Q^T A Q,$$

где Q – прямоугольная матрица размера $n_k \times n_{k-1}$ полного ранга.

Предположим, что

$$c_1 h_j \le \frac{(A_j u_j, u_j)}{(u_j, u_j)} \le c_2 h_j^{-1} \quad \forall\, u_j \ne 0, \tag{23.2.2}$$

где h_j – особый параметр, зависящий от n_j, а c_1 и c_2 - константы. [1]

В дальнейшем договоримся обозначать одной и той же буквой c любые положительные константы (в том числе и разные даже в пределах одного выражения). Положим $h = h_k$.

По определению, $(u, v)_A \equiv (Au, v)$, $||u||_A \equiv \sqrt{(u,u)_A}$. Кроме того,

$$||M||_A \equiv ||A^{\frac{1}{2}} M A^{-\frac{1}{2}}||_2 \;=\; \max_{u \ne 0} \frac{||Mu||_A}{||u||_A}.$$

Матрица M называется A-симметричной, если $(Mu, v)_A = (u, Mv)_A$ для любых векторов u и v. Это равносильно обычной симметричности матрицы $A^{\frac{1}{2}} M A^{-\frac{1}{2}}$.

Согласно определению матрицы $\hat{A}$, имеем:

$$||Qv||_A = ||v||_{\hat{A}}, \tag{23.2.3}$$

$$||QM||_A = ||M||_{\hat{A}}. \tag{23.2.4}$$

Важную роль будут играть матрица $P \equiv \hat{A}^{-1} Q^T A$ и связанные с ней свойства:

$$(QP)^2 = QP, \quad (I - QP)^2 = I - QP; \tag{23.2.5}$$

[1] Для матрицы вида (23.1.1) эти неравенства проверяются непосредственно, при этом в случае равномерной сетки h_j — это ее шаг.

$$(A^{\frac{1}{2}}QPA^{-\frac{1}{2}})^T = A^{\frac{1}{2}}QPA^{-\frac{1}{2}}; \tag{23.2.6}$$

$$((I-QP)u, QPv)_A = 0; \tag{23.2.7}$$

$$||(I-QP)u + QPv||_A^2 = ||(I-QP)u||_A^2 + ||QPv||_A^2. \tag{23.2.8}$$

Проверяется без проблем. Как видим, QP - это A-ортогональный проектор (в обычном же смысле это косой проектор). Заметим также, что $PQ = \hat{I}$ (единичная матрица порядка n_{k-1}).

23.3 Сглаживатель

Пусть R – симметричная положительно определенная матрица, дающая сходимость классического метода простой итерации:

$$u^0 = 0; \quad u^l = u^{l-1} + R(f - Au^{l-1}), \quad l = 1, \ldots, m.$$

Сглаживателем называется матрица $S = I - RA$.

Очевидно, все собственные значения $\lambda(S)$ вещественные. Заметим, что S является A-симметричной матрицей: $(Su, v)_A = (u, Sv)_A$.

23.4 Основные предположения

1. Все собственные значения матрицы $S = I - RA$ строго больше 0 и строго меньше 1.

2. Минимальное собственное значение матрицы R ограничено снизу величиной ch.

Эквивалентное утверждение. *В смысле скалярного произведения* $(u, v)_A$ *матрица* A *мажорируется матрицей* $\frac{1}{ch}(I - S)$:

$$(Au, u)_A \leq \frac{1}{ch}((I-S)u, u)_A \tag{23.4.9}$$

3. При аппроксимации u вектором вида Qv имеет место оценка

$$\inf_v ||u - Qv||_2 \leq ch||Au||_2. \tag{23.4.10}$$

Лемма 23.4.1 $||(I-QP)u||_A^2 \leq ch||Au||_2^2$.

Доказательство. $(I-QP)u = (I-QP)(u-Qv) \Rightarrow$

$$||(I-QP)u||_A^2 \leq ||u-Qv||_A^2 < \frac{c}{h}||u-Qv||_2^2 \leq ch||Au||_2^2.$$

Лемма 23.4.2 *Для целого* $m > 0$

$$((I - S)S^m u, u)_A \leq \frac{1}{m}((I - S^m)u, u)_A. \tag{23.4.11}$$

Доказательство. В силу основного предположения 1, для любого $j \leq m$ имеем

$$((I - S)S^m u, u)_A \leq ((I - S)S^j u, u)_A.$$

Остается лишь заметить, что

$$I - S^m = \sum_{j=0}^{m-1}(I - S)S^j.$$

Лемма 23.4.3

$$||(I - QP)S^m u||_A^2 \leq \frac{c}{m}((I - S^{2m})u, u)_A.$$

Доказательство. По лемме 1, левая часть неравенства не превышает $ch(AS^m u, S^m u)_A$, а основное предположение 2 дает

$$(AS^m u, S^m)_A \leq \frac{c}{h}((I - S)S^m, S^m)_A.$$

Остается принять во внимание A-симметричность матрицы S и применить лемму 2.

23.5 Общая схема многосеточного метода

Пусть решается уравнение $Au = f$ и в результате получается вектор $Bf \approx u$. Чтобы его найти, мы строим "меньшую"систему $\hat{A}\hat{u} = \hat{f}$ и вектор $\hat{B}\hat{f} \approx \hat{u}$. Таким образом, B определяется через $\hat{B}$ рекурсивно. Вот алгоритм получения Bf.

(1) Предварительное сглаживание:

$$u^0 = 0; \quad u^l = u^{l-1} + R(f - Au^{l-1}), \quad l = 1, \ldots, m.$$

(2) Приближенное решение "меньшей"системы $\hat{A}\hat{u} = \hat{f}$:

$$\hat{A} = Q^T AQ, \quad \hat{f} = Q^T(f - Au^m);$$

$$v^0 = 0; \quad v^l = v^{l-1} + \hat{B}(\hat{f} - \hat{A}v^{l-1}), \quad l = 1, \ldots, s;$$

(3) Пост-сглаживание:

$$w^0 = u^m + Qv^s; \quad w^l = w^{l-1} + R(f - Aw^{l-1}), \quad l = 1, \ldots, m.$$

23.6 Основное уравнение и неравенство

Обозначим через $e(w) \equiv A^{-1}f - w$ отклонение вектора w от точного решения системы $Au = f$ и через $\hat{e}(v) = \hat{A}^{-1}\hat{f} - v$ отклонение вектора v от точного решения системы $\hat{A}\hat{u} = \hat{f}$. Легко видеть, что

$$e(u^m) = S^m u, \quad \hat{e}(v^s) = (\hat{I} - \hat{B}\hat{A})^s(\hat{A}^{-1}Q^T A)S^m u = (\hat{I} - \hat{B}\hat{A})^s PS^m u.$$

Следовательно,

$$v^s = PS^m u - (\hat{I} - \hat{B}\hat{A})^s PS^m u$$

В результате пост-сглаживания получаем

$$e(w^m) = (I - BA)u = S^m(S^m u - Qv^s) = S^m(S^m - QPS^m + Q(\hat{I} - \hat{B}\hat{A})^s PS^m)u.$$

Это и дает *основное уравнение мультигрида*:

$$I - BA = S^m(I - QP)S^m + S^m Q(\hat{I} - \hat{B}\hat{A})^s PS^m. \tag{23.6.12}$$

По построению, матрицы B и $\hat{B}$ симметричны. Поэтому все собственные значения матриц $I - BA$ и $\hat{I} - \hat{B}\hat{A}$ вещественны.

Предположим, что для $\hat{I} - \hat{B}\hat{A}$ они неотрицательны. Тогда, в силу A-симметричности этой матрицы, существует A-симметричная матрица $\hat{F}$ с неотрицательными собственными значениями и такая, что

$$\hat{F}^2 = \hat{I} - \hat{B}\hat{A}. \tag{23.6.13}$$

Вследствие A-симметричности, для любого целого s имеем

$$||\hat{F}||_A^s = ||\hat{F}^s||_A. \tag{23.6.14}$$

Легко проверить, что матрица $Q\hat{F}P$ является A-симметричной и $(Q\hat{F}P)^s = Q\hat{F}^s P$. Учитывая также, что S и $I - QP$ являются A-симметричными матрицами, находим

$$\begin{aligned}((I - BA)u, u)_A &= ((I - QP)S^m u, (I - QP)S^m u)_A + \\ &\quad ((Q\hat{F}^s P)S^m u, (Q\hat{F}^s P)S^m u)_A.\end{aligned}$$

Отсюда видно, кстати, что собственные значения матрицы $I - BA$ тоже будут неотрицательны.

Принимая во внимание (23.2.4) и (23.6.14), получаем *основное неравенство*:

$$(I - BA)u, u)_A \le ||(I - QP)S^m u||_A^2 + ||\hat{I} - \hat{B}\hat{A}||_{\hat{A}}^s ||QPS^m u||_A^2 \tag{23.6.15}$$

23.7 Анализ V-цикла

V-цикл соответствует $s = 1$. Пусть c – константа из неравенства леммы 3, и предположим, что

$$||\hat{I} - \hat{B}\hat{A}||_{\hat{A}} \leq \gamma < 1. \tag{23.7.16}$$

Выберем такой вектор u, для которого

$$((I - BA)u, u)_A = ||I - BA||_A, \quad ||u||_A = 1. \tag{23.7.17}$$

Примем во внимание, что

$$||QPS^m u||_A^2 = ||S^m u||_A^2 - ||(I - QP)S^m||_A^2.$$

Тогда, согласно основному неравенству (23.6.15) и оценке леммы 3,

$$||I - BA||_A \leq (1 - \gamma)\frac{c}{m}((I - S^{2m})u, u)_A + \gamma(S^{2m}u, u)_A. \tag{23.7.18}$$

Мы хотим получить оценку

$$||I - BA||_A \leq \gamma. \tag{23.7.19}$$

Чтобы сохранить одно и то же γ в (23.7.16) и (23.7.19), нам, возможно, придется его увеличить (при этом, очевидно, (23.7.16) останется в силе). Потребуем, чтобы

$$(1 - \gamma)\frac{c}{m} \leq \gamma.$$

Это означает, что

$$\gamma \geq \frac{c}{c + m} \tag{23.7.20}$$

и, с учетом (23.7.17), правая часть неравенства (23.7.18) оценивается сверху величиной γ. Очевидно, мы доказали следующую важную теорему.

Теорема 23.7.1 *Обозначим через B_j матрицу мультигрида для системы порядка n_j и предположим, что*

$$||I_j - B_j A_j||_{A_j} \leq \gamma,$$

где γ имеет вид (23.7.20). Тогда при всех $j \leq i \leq k$ имеет место неравенство

$$||I - BA||_A \leq \gamma.$$

Если система с матрицей A_0 решается точно, то это неравенство справедливо при $0 \leq i \leq k$.

23.8 Анализ W-цикла

W-цикл соответствует $s = 2$. Пусть имеет место (23.8.22). Тогда, согласно основному неравенству (23.6.15) и лемме 3, при выборе u в соответствии с (23.7.17) получаем

$$||I - BA||_A \leq (1 - \gamma^2)\frac{c}{m}((I - S^{2m})u, u)_A + \gamma^2(S^{2m}u, u)_A. \qquad (23.8.21)$$

Теперь предположим дополнительно, что

$$(1 - \gamma^2)\frac{c}{m} \leq \gamma^2.$$

Это означает, что

$$\gamma^2 \geq \frac{c}{c + m} \qquad (23.8.22)$$

Таким образом, *теорема 23.7.1 остается в силе и для W-цикла* – в предположении, что γ удовлетворяет неравенству (23.8.22). Теперь уже ясно, что аналогичный результат справедлив для любого s, если

$$\gamma^s \geq \frac{c}{c + m}.$$

23.9 Интересные наблюдения

Проведенный анализ замечателен тем, что в нем допускается *любое число сглаживаний*. Когда говорят, что W-цикл изучать проще, чем V-цикл, то имеют в виду следующее утверждение: *для любого $0 < \gamma < 1$ оценка (23.7.19) всегда имеет место для достаточно большого числа сглаживаний m*. В случае W-цикла из основного неравенства (23.6.15) мгновенно получается (более грубая, чем (23.8.21)) оценка

$$||(I - BA)||_A \leq \frac{c}{m} + \gamma^2,$$

и достаточно потребовать, чтобы

$$\frac{c}{m} + \gamma^2 \leq \gamma.$$

Последнее имеет место, если $m \geq \frac{c}{\gamma(1-\gamma)}$. Для V-цикла столь грубый анализ, конечно, не проходит.

Можно сделать также в какой-то степени неожиданное заключение в пользу V-цикла: помимо того, что в нем вычислительные затраты меньше, чем в соответствующем W-цикле, нижняя граница для γ при фиксированном числе сглаживаний m в V-цикле тоже меньше!

Заметим еще, что наши рассуждения легко модифицировать для анализа мультигрида без пост-сглаживания. В этом случае, конечно, матрица B не будет симметричной и мы не вправе опираться на (23.6.13) и (23.6.14). Основное уравнение теперь принимает вид

$$I - BA = (I - QP)S^m + Q(\hat{I} - \hat{B}\hat{A})^s PS^m.$$

Отсюда легко вывести сдедующее неравенство:

$$||(I - BA)u||_A^2 \le ||(I - QP)S^m u||_A^2 + ||\hat{I} - \hat{B}\hat{A}||_{\hat{A}}^{2s} ||QPS^m u||_A^2.$$

Пусть $||\hat{I} - \hat{B}\hat{A}||_{\hat{A}} \le \gamma < 1$. Тогда

$$||(I - BA)u||_A^2 \le (1 - \gamma^{2s})\frac{c}{m}((I - S^{2m})u, u)_A + \gamma^{2s}(S^{2m}u, u)_A.$$

Если предположить дополнительно, что

$$\gamma^{2s} \ge \frac{c}{c + m},$$

то, очевидно,

$$||I - BA||_A \le \gamma^s \le \gamma.$$

Забавно, что в случае $s \ge 2$ может иметь смысл мультигрид с пост-сглаживаниями, но без предварительных сглаживаний. (Это вроде бы уже не очень вписывается в изначальную концепцию перехода к более грубым сеткам!) Вот основное уравнения для "забавного" мультигрида:

$$I - BA = S^m(I - QP) + S^m Q(\hat{I} - \hat{B}\hat{A})^s P.$$

Запишем второй член справа в виде $(S^m QP)Q(\hat{I} - \hat{B}\hat{A})^s P$. Поскольку $||S^m(I - QP)||_A = ||(I - QP)S^m||_A$ и $||QP||_A \le 1$, имеем

$$||I - BA||_A \le ||(I - QP)S^m||_A + ||\hat{I} - \hat{B}\hat{A}||_{\hat{A}}^s.$$

Достаточно потребовать, чтобы

$$\sqrt{\frac{c}{m}} + \gamma^s \le \gamma,$$

а это, конечно, выполняется при достаточно большом m.

23.10 Простейший пример

Для применения полученной выше теории требуется проверка основных предположений. Для простейшего примера, рассмотренного нами в начале главы, это делается очень легко.

Введем равномерные сетки с шагом $h_j = 2^{-j}$ и будем получать дискретные уравнения по методу Галеркина с аппроксимацией кусочно-линейными функциями. Чтобы на более грубых сетках иметь аналогичные по смыслу уравнения, интерполяционная матрица Q должна иметь вид

$$Q = \begin{bmatrix} 1/2 & 0 & \dots \\ 1 & 0 & \dots \\ 1/2 & 1/2 & \dots \\ 0 & 1 & \dots \\ 0 & 1/2 & \dots \\ 0 & 0 & \dots \\ \dots & \dots & \dots \end{bmatrix}.$$

Основные предположения 1 и 2 легко выполняются при выборе

$$R = \tau I, \quad \tau = \frac{1}{\lambda_{\max}(A)}.$$

Предположение 3 вообще оказывается очевидным: ошибка интерполяции в i-ом узле является нулем или же в точности *равна* i-ой компоненте вектора $\frac{h}{2}A$.

23.11 Коррекции на подпространствах

Многосеточные методы органично связаны с идеей коррекции приближения на заданной системе подпространств.

Пусть решается операторное уравнение $Au = f$ с самосопряженным положительно определенным оператором A на вещественном n-мерном евклидовом пространстве V, разложенном в (не обязательно прямую!) сумму подпространств:

$$V = V_1 + ... + V_m.$$

Рассмотрим проекционные уравнения $(Q_i A P_i)u_i = Q_i f$, где $u_i \in V_i$, а Q_i и P_i — ортогональный и A-ортогональный проекторы на V_i. Тогда оператор $A_i = Q_i A P_i$ удовлетворяет равенству $A_i P_i = Q_i A$:

$$(A_i P_i x, y) = (A P_i x, Q_i y) = (Ax, P_i Q_i y) = (Ax, Q_i y) = (Q_i A x, y) \ \ \forall\ x, y \in V.$$

Пусть R_i — предобусловливатель для A_i при решении i-го проекционного уравнения. Для практического применения нужно определить действие R_i лишь на векторах из инвариантного для него подпространства V_i. Но для теоретического анализа удобно считать, что R_i определен на всех векторах из V: можно условиться, например, что он оставляет на месте все векторы из ортогонального дополнения к V_i.

Для решения исходного уравнения $Au = f$ *метод параллельной коррекции на подпространствах* предлагает использовать предобусловливатель следующего вида:

$$B = R_1 Q_1 + ... + R_m Q_m. \tag{23.11.23}$$

Легко доказывается, что если операторы R_i являются самосопряженными и положительно определенными, то оператор B будет таким же.

В связи с многосеточными методами естественно рассматривать систему вложенных подпространств

$$V_1 \subset ... \subset V_m = V.$$

Если S_i — сглаживатель, применяемый последовательно k раз, то предобусловливатель вида (23.11.23), где $R_i = (I - S_i^k)A_i^{-1}$, называется *BPX-предобусловливателем.* [2] Если A_i — матрица i-го проекционного уравнения, то типичен выбор $S_i = I - \alpha D_i^{-1} A_i$, где $D_i = \mathrm{diag} A_i$.

Данный предобусловливатель применяется для ускорения метода сопряженных градиентов. При определенных условиях он приводит к оценке числа итераций, зависящей только от предписанной точности, но не от порядка матрицы. Следующая теорема показывает, как выглядят эти условия (их проверка обычно требует изрядной работы). [3]

Теорема 23.11.1 *Пусть A и $R_1, ..., R_m$ — самосопряженные положительно определенные операторы и выполнены следующие два условия:*

- *любой вектор $v \in V$ допускает разложение $v = v_1 + ... + v_m$, $v_i \in V_i$, такое что*

$$\sum_{i=1}^{m} (R_i^{-1} v_i, v_i) \le K_0 (Av, v); \tag{23.11.24}$$

- *для операторов $T_i = R_i A_i P_i$ и любых векторов $x, y \in V$*

$$\sum_{1 \le i,j \le m} (T_i x, T_j y) \le K_1 \left(\sum_{i=1}^{m} (T_i x, x)_A \right)^{1/2} \left(\sum_{i=1}^{m} (T_i y, y)_A \right)^{1/2}. \tag{23.11.25}$$

Тогда отношение максимального и минимального собственных значений BA не превосходит $K_0 K_1$.

[2] По имени авторов: Bramble, Pasciak, Xu.

[3] Более подробное рассмотрение с приложением к решению уравнения Пуассона, а также обсуждение связей с конструкциями многосеточных методов можно найти в книге: *М. А. Ольшанский.* Лекции и упражнения по многосеточным методам. — М.: Физматлит, 2005.

Задачи

1. Найдите собственные значения и собственные векторы матрицы A вида (23.1.1). Объясните, почему сглаживатель вида $S = I - \alpha A$ делает невязку "более гладкой".

2. Проверьте неравенства (23.2.2) для матрицы (23.1.1).

3. Пусть $Q_1, ..., Q_m$ — ортогональные проекторы пространства V на подпространства $V_1, ..., V_m$, дающие в сумме V, и пусть $R_1, ..., R_m$ — самосопряженные положительно определенные операторы на V. Докажите, что оператор $B = R_1 Q_1 + ... + R_m Q_m$ будет самосопряженным и положительно определенным.

4. Докажите теорему 23.11.1.

5. Для некоторой матричной нормы $||A|| \leq 1$. Докажите, что

$$\lim_{k \to \infty} 2^{-k} \, ||(I - A)(I + A)^k|| = 0.$$

Глава 24

24.1 Матрицы специального вида

Матрица моментов $M = [\mathcal{A}\phi_j, \phi_i]_{n\times n}$ в методе Галеркина обычно бывает *разреженной* в случае дифференциального оператора $\mathcal{A}$ и *плотной* в случае интегрального оператора $\mathcal{A}$. В приложениях размеры матрицы M могут быть порядка нескольких миллионов и даже выше. В таких случаях эффективные методы решения системы с матрицей M обязаны использовать какие-либо особые свойства, выделяющие ее среди матриц общего вида.

Например, диагональные, трехдиагональные и "более широкие" ленточные матрицы — это примеры специфики в разреженных матрицах, позволяющие строить очень эффективные методы — по сравнению с тем, что удается получить при произвольном расположении ненулей, пусть даже и относительно малого их числа. При решении уравнения Пуассона

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) u(x,y) = f(x,y)$$

в прямоугольнике $(x,y) \in [0,1] \times [0,1]$ естественным образом возникают блочно-трехдиагональные матрицы с трехдиагональными блоками. В трехмерном случае возникают блочно-трехдиагональные матрицы, в которых каждый из блоков является блочно-трехдиагональной матрицей. Такого рода "вложенные" блочные разбиения с блочными матрицами специального вида приводят к концепции *многоуровневой матрицы*.

Матрица $uv^\top$ ("строка на столбец") — важный пример специфики в плотных матрицах. Кроме того, это пример специфики, которая уже не описывается линейными связями между элементами матрицы.

Конечно, есть и другие важные классы матриц специального вида. Кроме того, нужно иметь в виду, что при решении практических задач специфика матриц чаще всего оказывается *неявной* — не имея явно выраженной специфики, матрица может приближаться матрицами из хорошо изученных классов матриц специального вида.

Для плотных матриц аппроксимации дают возможность строить эффективные методы быстрого приближенного умножения матрицы на вектор — это нужно при реализации итерационных методов. Матрицы специального вида используются также как предобусловливатели для уменьшения числа итераций.

24.2 Циркулянты и теплицевы матрицы

Матрица называется *циркулянтной матрицей* или *циркулянтом*, если она имеет вид

$$C = \begin{bmatrix} c_0 & c_{n-1} & c_{n-2} & \dots & c_1 \\ c_1 & c_0 & c_{n-1} & \cdots & c_2 \\ c_2 & c_1 & c_0 & \cdots & c_3 \\ \dots & \dots & \dots & \dots & \dots \\ c_{n-1} & c_{n-2} & c_{n-3} & \dots & c_0 \end{bmatrix}.$$

Матрица T называется *теплицевой* (в честь немецкого математика О. Теплица), если она имеет вид

$$T = \begin{bmatrix} t_0 & t_{-1} & t_{-2} & \dots & t_{-1} \\ t_1 & t_0 & t_{-1} & \cdots & t_{-2} \\ t_2 & t_1 & t_0 & \cdots & t_{-3} \\ \dots & \dots & \dots & \dots & \dots \\ t_{n-1} & t_{n-2} & t_{n-3} & \dots & t_0 \end{bmatrix}.$$

Заметим, что элементы T одинаковы на каждой линии $i - j = k$ и полностью определяются элементами ее первого столбца и первой строки. Циркулянтная матрица является теплицевой матрицей, полностью определяемой своим первым столбцом.

Циркулянтные матрицы возникают при дискретизации интегрального уравнения на окружности, если его ядро $K(x, y)$ зависит лишь от расстояния между точками x и y (например, уравнение для логарифмического потенциала), а сетка предполагается равномерной. Теплицевы матрицы получаются при дискретизации интегральных уравнений с такого же типа ядрами на отрезке.

24.3 Циркулянты и матрицы Фурье

Циркулянты и теплицевы матрицы определяются малым числом независимых параметров: n и $2n-1$ в случае матриц порядка n. В случае произвольной матрицы порядка n имеется, очевидно, n^2 независимых параметров.

Для записи малопараметрических матриц и при построении алгоритмов для них появляется замечательная возможность *не использовать* массивы размеров $n \times n$. При этом экономия памяти оказывается прямо связанной с существенным сокращением числа арифметических операций.

В быстрых алгоритмах линейной алгебры ключевая роль принадлежит циркулянтным матрицам и обеспечивается их связью с *дискретным преобразованием Фурье*. Последнее определяется как операция умножения вектора на *матрицу Фурье* — матрицу вда

$$F_n = [w^{kl}], \quad 0 \le k, l \le n-1, \quad w = e^{-i\frac{2\pi}{n}}.$$

Теорема 24.3.1 *Для циркулянтной матрицы $C \in \mathbb{C}^{n\times n}$ с первым столбцом $c \in \mathbb{C}^n$ имеет место разложерние*

$$C = \frac{1}{n} F_n^* \operatorname{diag}(F_n c)\, F_n.$$

Доказательство. Возтмем $\zeta = w^k$ и положим

$$\lambda_k = \sum_{j=0}^{n-1} \zeta^j c_j \qquad \Rightarrow \qquad \zeta^l \lambda_k = \sum_{j=0}^{n-1} \zeta^j c_{j-l \,(\mathrm{mod}\, n)}.$$

Собирая вместе эти уравнения при всех l и k, мы получаем матричное равенство $F_n C = \operatorname{diag}(F_n c) F_n$. Остается проверить, что $F_n^{-1} = \frac{1}{n} F_n^*$. $\square$

Таким образом, умножение циркулянтной матрицы на вектор сводится к трем операциям умножения на матрицу Фурье. Сложность решения системы с циркулянтной матрицей коэффициентов определяется сложностью операции умножения на матрицу Фурье.

При умножении на вектор теплицевой матрицы T порядка n мы можем свести задачу к умножению на вектор циркулянтной матрицы C, содержащей T как подматрицу:

$$C = \begin{bmatrix} T & \cdots \\ \cdots & \cdots \end{bmatrix} \in \mathbb{C}^{N\times N}.$$

Такую циркулянтную матрицу C можно построить для любого предписанного порядка $N \ge 2n-1$ (докажите).

24.4 Быстрое преобразование Фурье

Вектор $y = F_n x$ для заданного x можно вычислить за $O(n \log n)$ операций. Возможно,что в связи с этим фактом следует упоминать Гаусса; сейчас мы

точно знаем, что алгоритмы быстрого преобразования Фурье были известны Рунге и Ланцошу. В любом случае они оказались в центре внимания вычислителей и инженеров благодаря работе Кули и Тьюки, опубликованной в 1965 г. [1] Давайте посмотрим, как это делается.

Предположим, что $n = 2m$, и обозначим через P_n матрицу перестановки, составленную из строк единичной матрицы I с номерами

$$1, 3, \ldots, 2m-1,\ 2, 4, \ldots, 2m.$$

Тогда нетрудно проверить, что

$$P_n F_n = \begin{bmatrix} [w^{2kl}] & [w^{2k(m+l)}] \\ [w^{(2k+1)l}] & [w^{(2k+1)(m+l)}] \end{bmatrix}, \qquad 0 \le k, l \le m-1. \qquad \Rightarrow$$

$$P_n F_n = \begin{bmatrix} F_m & 0 \\ 0 & F_m \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & W_m \end{bmatrix} \begin{bmatrix} I_m & I_m \\ I_m & -I_m \end{bmatrix}, \qquad (24.4.1)$$

$$W_m = \mathrm{diag}\{w^0, w^1, \ldots, w^{m-1}\}.$$

Итак, задача умножения вектора на F_{2m} сводится к двум аналогичным задачам для матрицы F_m. Если n есть степень числа 2, то потребуется в итоге $\frac{1}{2} n \log_2 n$ комплексных умножений и $n \log_2 n$ комплексных сложений-вычитаний (докажите).

24.5 Циркулянтные предобусловливатели

Рассмотрим интегральное уравнение с логарифмическим ядром на гладком контуре $\Gamma = \{\gamma(t) : \ 0 \le t \le 2\pi\}$. Используя равномерную сетку на $[0,\, 2\pi]$ и метод Галеркина с функциями вида $U(t) \equiv u(t)\,|\gamma'(t)|$, мы можем расщепить матрицу моментов $A = C + R$ таким образом, что C будет циркулянтной матрицей, соответствующей главной части оператора. Матрица $C^{-1}R$ может рассматриваться как дискретный аналог некоторого компактного оператора, поэтому нужно ожидать, что C будет хорошим предобусловливателем для A — вследствие кластеризации собственных значений предобусловленной матрицы $C^{-1}A$ в точке 1.

При построении циркулянтного предобусловливателя для заданной матрицы A порядка n можно и забыть о какой-либо связи с расщеплением оператора. Например, потребуем, чтобы циркулянт C минимизировал норму $||A - C||_F$ на множестве всех циркулянтных матриц порядка n. Такая

[1] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* 19 (90): 297–301 (1965).

матрица C называется *оптимальным циркулянтом* [2] и легко находится по элементам матрицы A без каких-либо предположений о ее специфике. Впервые анализ циркулянтных предобусловливателей при решении теплицевых систем был дан Р. Ченом и Г. Стрэнгом. [3]

Пусть C — оптимальный циркулянт для матрицы A порядка n. Согласно теореме 24.3.1, $C = \hat{F}^* D \hat{F}$, где $\hat{F} = n^{-1/2} F_n$, F_n — матрица Фурье порядка n, а D — диагональная матрица из собственных значений матрицы C. Нетрудно проверить, что матрица $\hat{F}$ унитарная. Поэтому

$$||A - C||_F = ||D - \hat{F} A \hat{F}^*||_F \geq ||\hat{F} A \hat{F}^* - \mathrm{diag}(\hat{F} A \hat{F}^*)||_F,$$

откуда получаем

$$C = \hat{F}^* \mathrm{diag}(\hat{F} A \hat{F}^*) \hat{F}. \tag{24.5.2}$$

Отсюда ясно, что если $A = A^*$, то и $C = C^*$, а из положительной определенности A следует положительная определенность ее оптимального циркулянта C.

24.6 Оптимальные циркулянты для теплицевых систем

Рассмотрим вещественную интегрируемую 2π-периодическую функцию $f(t)$ и ее формальное разложение в ряд Фурье

$$f(t) = \sum_{k=-\infty}^{\infty} a_k e^{\mathbf{i}kt}. \tag{24.6.3}$$

Пусть $A_n = [a_{k-l}]$ — теплицева матрица порядка n, составленная из коэффициентов этого ряда Фурье.

Лемма 24.6.1 *Для любого вектора* $x = [x_0, x_1, ..., x_{n-1}]^\top$ *справедливо равенство*

$$(A_n x, x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \left| \sum_{k=0}^{n-1} x_k e^{\mathbf{i}kt} \right|^2 dt. \tag{24.6.4}$$

Доказательство. Достаточно вспомнить, что $a_k = \frac{1}{2\pi} \int\limits_{-\pi}^{\pi} f(t) e^{-\mathbf{i}kt} dt$. Тогда правая часть равенства (24.6.4) записывается в виде

$$\sum_{k=0}^{n-1} \bar{x}_k \sum_{l=0}^{n-1} x_l \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-\mathbf{i}(k-l)t} dt = \sum_{k=0}^{n-1} \bar{x}_k \sum_{l=0}^{n-1} a_{k-l}\, x_l = (A_n x, x). \quad \square$$

[2] T. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM J. Sci. Statist. Comput.* 9: 766–771 (1988).

[3] R. Chan and G. Strang. Toeplitz equations constructed by conjugate gradients with circulant preconditioners. *SIAM J. Stat. Comput.* 10: 104–119 (1989).

Следствие 24.6.1 *Если* $c_{\min} \le f(t) \le c_{\max}$ *для всех* t, *то все собственные значения теплицевых матриц* A_n *и построенных для них оптимальных циркулянтов* C_n *находятся на отрезке* $[c_{\min}, c_{\max}]$.

Лемма 24.6.2 *Пусть* $\sum\limits_{k=-\infty}^{\infty} |a_k| \le c < +\infty$. *Тогда для любого* $\varepsilon > 0$ *существуют целые* $r = r(\varepsilon)$ *и* $N = N(\varepsilon)$ *такие, что при* $n \ge N$ *имеет место расщепление*

$$A_n - C_n = R_n + E_n, \tag{24.6.5}$$

где

$$\mathrm{rank} R_n \le r, \quad ||E_n||_\infty \le \varepsilon. \tag{24.6.6}$$

Доказательство. Фиксируем произвольное $\varepsilon > 0$ и выберем $s = s(\varepsilon)$ таким образом,что $\sum\limits_{|k|\ge s} |a_k| \le \varepsilon/2$. Будем считать, что $c_{-k} = c_{n-k}$. Тогда

$$\sum_{k=0}^{n-s} |c_{-k} - a_{-k}| \;\le\; \sum_{k=0}^{n-s} \frac{k}{n} |a_{n-k} - a_{-k}| \;\le\; \sum_{k\ge s} (|a_k| + |a_{-k}|) + \frac{cs}{n} \;\le\; \frac{\varepsilon}{2} + \frac{cs}{n} \;\le\; \varepsilon$$

при достаточно больших n. Аналогично, при достаточно больших n находим

$$\sum_{k=0}^{n-s} |c_k - a_k| \le \varepsilon.$$

В качестве R_n возьмем матрицу с элементами $(R_n)_{ij} = (A_n - C_n)_{ij}$ при $|i-j| \ge n-s$ и нулями при $|i-j| < n-s$. Очевидно, $\mathrm{rank} R_n \le r \equiv 2s$. $\square$

Теорема 24.6.1 *Пусть* $f(t) \ge c_{\min} > 0$ *для всех* t *и* $\sum\limits_{k=-\infty}^{\infty} |a_k| \le c < +\infty$. *Тогда собственные значения последовательности матриц* $C_n^{-1} A_n$ *расположены на отрезке* $\left[\frac{c_{\min}}{c}, \frac{c}{c_{\min}}\right]$ *и имеют собственный кластер в точке* 1.

Доказательство. Собственные значения $C_n^{-1} A_n$ совпадают с собственными значениями эрмитовой матрицы $C_n^{-1/2} A_n C_n^{-1/2}$ и поэтому вещественны и положительны. Согласно (24.6.4) и (24.5.2), собственные значения матриц A_n и C_n расположены на отрезке $[c_{\min}, c]$. Отсюда

$$\lambda(C_n^{-1} A_n) \le ||C_n^{-1} A_n||_2 \le \frac{c}{c_{\min}}.$$

Аналогично,

$$\frac{1}{\lambda(C_n^{-1} A_n)} = \lambda(A_n^{-1} C_n) \le ||A_n^{-1} C_n||_2 \le \frac{c}{c_{\min}}.$$

Далее, по лемме 24.6.2, для любого $\varepsilon > 0$ при всех достаточно больших n матрица $C_n^{-1}A_n$ получается из матрицы вида $I + C_n^{-1}E_n$ модификацией ранга не выше r. При этом не более r собственных значений могут оказаться вне отрезка $[1 - c\varepsilon, 1 + c\varepsilon]$. $\square$

Пусть в условиях данной теоремы система с теплицевой матрицей A_n решается по методу сопряженных градиентов с оптимальным циркулянтным предобусловливателем. В силу результатов главы 21 число итераций для получения невязки с нормой $||r_j||_2 \le \varepsilon ||r_0||_2$ не зависит от n. Более того, для любого $0 < q < 1$ при достаточно больших n выполняется неравенство $||r_j||_2 \le c(q)q^j||r_0||_2$. В таких случаях говорят, что метод сходится *суперлинейно*.

24.7 Строение обратных матриц

Матрица, обратная к циркулянтной, также является циркулянтной (почему?). Для теплицевых матриц обратная матрица в общем случае не является теплицевой (приведите пример). Тем не менее, обратные матрицы допускают полезные малопараметрические представления, обнаруживающие их связь с теплицевыми матрицами.

Теорема 24.7.1 (Гохберг–Семенцул) *Пусть $A = [a_{i-j}]$ — теплицева матрица порядка $n+1$, для которой две системы*

$$A \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}, \quad A \begin{bmatrix} y_0 \\ \dots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} 0 \\ \dots \\ 0 \\ 1 \end{bmatrix} \tag{24.7.7}$$

совместны и при этом $x_0 \ne 0$ (или $y_n \ne 0$). Тогда матрица A обратима и представима в виде разности произведений теплицевых треугольных матриц:

$$A^{-1} = x_0^{-1} \begin{bmatrix} x_0 & & & \\ x_1 & x_0 & & \\ \dots & \dots & \dots & \\ x_n & \dots & \dots & x_0 \end{bmatrix} \begin{bmatrix} u_0 & u_1 & \dots & u_n \\ & u_0 & \dots & u_{n-1} \\ & & \dots & \dots \\ & & & u_0 \end{bmatrix} -$$

$$- x_0^{-1} \begin{bmatrix} 0 & & & & \\ y_0 & 0 & & & \\ y_1 & y_0 & 0 & \dots & \\ \dots & \dots & \dots & \dots & \\ y_{n-1} & \dots & \dots & y_0 & 0 \end{bmatrix} \begin{bmatrix} 0 & v_0 & v_1 & \dots & v_{n-1} \\ & 0 & v_0 & \dots & v_{n-2} \\ & & \dots & \dots & \dots \\ & & & 0 & v_0 \\ & & & & 0 \end{bmatrix},$$

$$u_i = y_{n-i}, \quad w_i = x_{n-1}, \qquad 0 \le i \le n.$$

Доказательство. Заметим, что любая теплицева матрица является *персимметричной*:

$$A^\top = JAJ, \quad J = \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}.$$

Если персимметричная матрица обратима, то обратная матрица также персимметрична: $(A^{-1}) = JA^{-1}J$. Отсюда следует, что u_i и v_i — элементы первой и последней строк A^{-1}.

Из равенства $Ay = e_0$ вытекает, что $JAJ(Jy) = Je_n = y_0$, где e_0 и e_n — первый и последний столбцы единичной матрицы. Учитывая персимметричность матрицы A, получаем $(Jy)^\top A = e_0^\top$. Отсюда

$$y_n = (Jy)^\top Ax = x_0.$$

Рассмотрим равную нулю линейную комбинацию строк матрицы A: $[\alpha_0 \dots \alpha_n]A = [0 \dots 0]$. Умножая обе части справа на $x = [x_0 \dots x_n]^\top$, находим $\alpha_0 = 0$. Далее, теплицев вид матрицы приводит к выводу о том, что $[\alpha_1 \dots \alpha_n]\hat{A} = 0$, где $\hat{A}$ — ведущая подматрица порядка n. Из второго уравнения (24.7.7) и условия $y_n \neq 0$ следует, что последний столбец матрицы $\tilde{A}$, полученной из A исключением последней строки, есть линейная комбинация предыдущих столбцов. Поэтому $[\alpha_1 \dots \alpha_n]\tilde{A} = 0$. Те же рассуждения можно повторить для прямоугольной теплицевой матрицы $\tilde{A}$: умножая обе части справа на x, находим $\alpha_1 = 0$, и т. д. В итоге $\alpha_0 = .. = \alpha_n = 0$ $\Rightarrow$ строки матрицы A линейно независимы.

Пусть $A^{-1} = [c_{kj}]$. Тогда $\sum\limits_{k=0}^{n} a_{i-k}c_{kj} = \delta_{ij}$, $0 \le i, j \le n$. Учитывая равенства $a_i = -x_0^{-1} \sum\limits_{k=1}^{n} a_{i-k}x_k$, $1 \le i \le n$, и принимая во внимание, что $c_{0j} = u_j$, получаем

$$\sum_{k=1}^{n} a_{i-k}(c_{kj} - x_0^{-1}x_k u_j) = \delta_{ij}, \quad 1 \le i, j \le n. \qquad (*)$$

Используя равенства $a_{i-n} = -x_0^{-1} \sum\limits_{k=0}^{n-1} a_{i-k}y_k$, находим

$$\sum_{k=0}^{n-1} a_{i-k}(c_{kj} - x_0^{-1}y_k v_j) = \delta_{ij}, \quad 0 \le i, j \le n-1.$$

Заменяя i на $i-1$ и j на $j-1$, получаем

$$\sum_{k=1}^{n} a_{i-k}(c_{k-1\,j-1} - x_0^{-1}y_{k-1}v_{j-1}) = \delta_{ij}, \quad 1 \le i, j \le n. \qquad (**)$$

Системы $(*)$ и $(**)$ имеют общую матрицу коэффициентов, равную $\hat{A}$, и одну и ту же правую часть. В силу правила Крамера $x_0 = \det\hat{A}/\det A \neq 0$ $\Rightarrow$ $\det\hat{A} \neq 0$ $\Rightarrow$ решения систем $(*)$ и $(**)$ равны. Значит,

$$c_{kj} = c_{k-1\,j-1} + x_0^{-1}x_k u_j - x_0^{-1}y_{k-1}v_{j-1}, \quad 1 \le k, j \le n. \qquad (24.7.8)$$

Но точно такие же соотношения получаются для элементов матрицы в правой части доказываемой формулы. Остается заметить, что первый столбец и первая строка в ее правой части совпадают с первым столбцом и первой строкой матрицы A^{-1}. $\square$

Формула Гохберга–Семенцула дает матричную интерпретацию соотношений (24.7.8) между элементами обратной матрицы. Такого типа соотношения были впервые получены В. Тренчем в 1965 г. в предположении строгой регулярности матрицы A. Их же можно вывести из формул Кристоффеля–Дарбу, известных в теории ортогональных полиномов.

24.8 Теплицевы ранги

Обозначим через $L(u)$ теплицеву нижнюю треугольную матрицу, первый столбец которой есть u. Тогда матрица, обратная к теплицевой, имеет вид

$$L(g_1)L^\top(h_1) + L(g_2)L^\top(h_2)$$

для некоторых векторов g_1, g_2, h_1, h_2. Транспонированная матрица (в силу персимметричности) имеет вид

$$(JL(g_1)J)(JL^\top(h_1)J) + (JL(g_2)J)(JL^\top(h_2)J) = L^\top(g_1)L(h_1) + L^\top(g_2)L(h_2).$$

Эти матрицы можно рассматривать как примеры матриц более общего вида:

$$A = \sum_{k=1}^{r} L(g_k)L^\top(h_k) \quad \text{и} \quad B = \sum_{k=1}^{r} L^\top(g_k)L(h_k). \tag{24.8.9}$$

Пусть $G = [g_1, ..., g_r]$ и $H = [h_1, ..., h_r]$. Тогда нетрудно проверить, что равенства (24.8.9) выполняются в том и только том случае, когда

$$A - ZAZ^\top = GH^\top, \quad B - Z^\top BZ = (JG)(JH)^\top, \tag{24.8.10}$$

$$Z = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}.$$

Если матрица A обратима, то $\operatorname{rank}(A - ZAZ^\top) = \operatorname{rank}(A^{-1} - Z^\top A^{-1}Z)$ (докажите!). Отсюда следует, что обе матрицы A и A^{-1} представимы в виде суммы одного и того же числа произведений теплицевых треугольных матриц. Совершенно произвольная матрица тоже допускает представление такого же типа, но число произведений может быть равно ее порядку n. Практический интерес возникает, когда $r \ll n$. Ранги матриц $A - ZAZ^\top$ и $A - Z^\top AZ$ иногда называются *теплицевыми рангами* матрицы A.

Под Z-рангом (сдвиговым рангом, рангом смещения) матрицы A понимается ранг матрицы $ZA - AZ$. А вот далеко идущее обобщение: под (P, Q)-рангом матрицы A понимается ранг матрицы $PA - AQ$. Для обратимой матрицы ее Z-ранг равен Z-рангу обратной матрицы, а (P, Q)-ранг равен (Q, P)-рангу обратной матрицы (докажите).

24.9 Алгоритмы метода окаймления

Пусть $A_n = [a_{i-j}]$ — строго регулярная теплицева матрица порядка $n+1$ и нужно решить систему $Au = f$. Метод окаймления подразумевает последовательное решение систем

$$A_k \begin{bmatrix} u_{0k} \\ \dots \\ u_{kk} \end{bmatrix} = \begin{bmatrix} f_0 \\ \dots \\ f_k \end{bmatrix}, \quad 0 \le k \le n.$$

Чтобы это сделать, достаточно научиться вычислять решения следующих специальных систем:

$$A_k \begin{bmatrix} x_{0k} \\ x_{1k} \\ \dots \\ x_{kk} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}, \quad A_k \begin{bmatrix} y_{0k} \\ \dots \\ y_{k-1\,k} \\ y_{kk} \end{bmatrix} = \begin{bmatrix} 0 \\ \dots \\ 0 \\ 1 \end{bmatrix}.$$

После этого все сводится к рекуррентным вычислениям вида

$$\begin{bmatrix} u_{0k} \\ \dots \\ u_{k-1\,k} \\ u_{kk} \end{bmatrix} = \begin{bmatrix} u_{0\,k-1} \\ \dots \\ u_{k-1\,k-1} \\ 0 \end{bmatrix} + s_k \begin{bmatrix} y_{0k} \\ \dots \\ y_{k-1\,k} \\ y_{kk} \end{bmatrix}, \quad s_k = f_k - \sum_{l=0}^{k-1} a_{k-l} u_{l\,k-1}.$$

Данные вычисления используют лишь величины y_{lk}. Но чтобы их найти, полезно иметь также величины x_{lk}. Достаточно заметить,что уравнения

$$\begin{bmatrix} x_{0k} \\ x_{1k} \\ \dots \\ x_{k-1\,k} \\ x_{kk} \end{bmatrix} = \alpha_k \begin{bmatrix} x_{0\,k-1} \\ x_{1\,k-1} \\ \dots \\ x_{k-1\,k-1} \\ 0 \end{bmatrix} + \beta_k \begin{bmatrix} 0 \\ y_{0\,k-1} \\ \dots \\ y_{k-2\,k-1} \\ y_{k-1\,k-1} \end{bmatrix}, \quad \begin{bmatrix} y_{0k} \\ y_{1k} \\ \dots \\ y_{k-1\,k} \\ y_{kk} \end{bmatrix} = \gamma_k \begin{bmatrix} x_{0\,k-1} \\ x_{1\,k-1} \\ \dots \\ x_{k-1\,k-1} \\ 0 \end{bmatrix} + \delta_k \begin{bmatrix} 0 \\ y_{0\,k-1} \\ \dots \\ y_{k-2\,k-1} \\ y_{k-1\,k-1} \end{bmatrix}$$

однозначно разрешимы относительно коэффициентов $\alpha_k, \beta_k, \gamma_k, \delta_k$. В самом деле, после их умножения слева на A_k, учитывая теплицев вид матрицы, находим

$$\begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \\ 0 \end{bmatrix} = \alpha_k \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \\ \phi_k \end{bmatrix} + \beta_k \begin{bmatrix} \psi_k \\ 0 \\ \dots \\ 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 1 \end{bmatrix} = \gamma_k \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \\ \phi_k \end{bmatrix} + \delta_k \begin{bmatrix} \psi_k \\ 0 \\ \dots \\ 0 \\ 1 \end{bmatrix},$$

$$\phi_k = \sum_{l=0}^{k-1} a_{k-l} x_{l\,k-1}, \quad \psi_k = \sum_{l=0}^{k-1} a_{-l-1} y_{l\,k-1}.$$

Таким образом,

$$\begin{bmatrix} 1 & \psi_k \\ \phi_k & 1 \end{bmatrix} \begin{bmatrix} \alpha_k & \gamma_k \\ \beta_k & \delta_k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Обратимость самой левой матрицы очевидна из следующего равенства:

$$A_k \begin{bmatrix} x_{0\,k-1} & 0 \\ x_{1\,k-1} & y_{0\,k-1} \\ \dots & \dots \\ x_{k-1\,k-1} & y_{k-2\,k-1} \\ 0 & y_{k-1\,k-1} \end{bmatrix} = \begin{bmatrix} 1 & \psi_k \\ 0 & 0 \\ \dots & \dots \\ 0 & 0 \\ \phi_k & 1 \end{bmatrix}.$$

Нужно учесть, что $x_{0\,k-1} = y_{k-1\,k-1} \neq 0$ и ранг не меняется при умножении на обратимую матрицу.

Ясно, что вычисление коэффициентов $\alpha_k, \beta_k, \gamma_k, \delta_k$ и пересчет векторов на k-м шаге требует лишь $O(k)$ операций. Общее число операций для всех k от 1 до n оказывается равным $O(n^2)$.

На основе этих алгоритмов можно получить и более эффективные, *супербыстрые* алгоритмы — с числом операций $O(n \log^2 n)$. Итерационные методы с циркулянтными предобусловливателями *при определенных предположениях* могут дать приближенное решение с затратой $cn \log n$ операций, где c зависит от предписанной точности, но не зависит от n.

Задачи

1. Пусть F_n — матрица Фурье порядка n. Докажите, что матрица $n^{-1/2} F_n$ унитарная.

2. Докажите равенство (24.4.1).

3. Доказать, что два многочлена степени n можно перемножить с затратой $O(n \log_2 n)$ арифметических операций.

4. Даны числа $x_1, \dots, x_n$. Доказать, что коэффициенты многочлена $f(x) = \prod_{i=1}^{n} (x - x_i)$ можно найти с затратой $O(n \log_2^2 n)$ арифметических операций.

5. Пусть $A_n = [a_{i-j}]_{n \times n}$ и $C_n = [c_{i-j}]_{n \times n}$ — теплицева матрица и ее оптимальный циркулянт. Докажите, что

$$c_k = \sum_{i,j:\, i-j=k \,(\mathrm{mod}\, n)} a_{i-j} = \frac{(n-k)\, a_k + k\, a_{n-k}}{n}, \quad 0 \leq k \leq n-1.$$

6. Пусть $A_n = [a_{i-j}]_{n\times n}$ и $C_n = [c_{i-j}]_{n\times n}$ — последовательность теплицевых матриц и построенных для них оптимальных циркулянтов, и пусть $\sum\limits_{k=-\infty}^{\infty} |a_k|^2 < +\infty$. Докажите, что $||A_n - C_n||_F^2 = o(n)$. Докажите также, что при условии

$$||C_n^{-1}||_2 \, ||A_n - C_n||_F = o(\sqrt{n})$$

сингулярные числа матриц $C_n^{-1}A_n$ имеют кластер в точке 1.

7. Пусть $A_n = [a_{i-j}]_{n\times n}$ и $C_n = [c_{i-j}]_{n\times n}$ — последовательность теплицевых матриц и построенных для них оптимальных циркулянтов. Докажите, что собственные значения матриц C_n суть значения *частичных сумм Чезаро*:

$$\lambda_k(C_n) = \sigma_n\left(k\frac{2\pi}{n}\right), \quad 0 \le k \le n; \qquad \sigma_n(t) = \sum_{k=0}^{n}\left(1 - \frac{k}{n}\right) a_k \, e^{\mathbf{i}kt}.$$

8. Если $a_{-k} = -a_{n-k}$ при $1 \le k \le n-1$, то теплицева матрица порядка n с элементами a_{i-j} называется *косоциркулянтом*. Докажите, что матрица, обратная к косоциркулянту, будет косоциркулянтом.

9. Пусть $A = [a_{i-j}]$ — теплицева матрица порядка $n+1$, для которой обе системы

$$A\begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix}, \quad A\begin{bmatrix} z_0 \\ z_1 \\ \dots \\ z_n \end{bmatrix} = \begin{bmatrix} 0 \\ a_{-n} \\ \dots \\ a_{-1} \end{bmatrix}$$

имеют решение. Докажите, что матрица A обратима.

10. Докажите, что все элементы матрицы, обратной к строго регулярной теплицевой матрице порядка n, можно вычислить за $O(n^2)$ операций.

11. Пусть $x_1, \dots, x_n, y_1, \dots, y_n$ — попарно различные числа. Докажите, что матрица $A = [1/(x_i - y_j)]$ (известная как *матрица Коши*) обратима. Докажите, что если $P = \operatorname{diag}(x_1, \dots, x_n)$ и $Q = \operatorname{diag}(y_1, \dots, y_n)$, то (Q, P)-ранг матрицы A^{-1} равен 1.

12. Пусть матрица A обратима, а P и Q — произвольные матрицы такого же порядка. Доказать, что $\operatorname{rank}(A - PAQ) = \operatorname{rank}(A^{-1} - QA^{-1}P)$.

Глава 25

25.1 Нелинейные аппроксимации

В задачах численного анализа важнейшими являются два типа нелинейной аппроксимации:

- аппроксимации с разделением переменных:

$$f(x_2, x_2) \approx \sum_{k=1}^{r} \phi_{1k}(x_1)\phi_{2k}(x_2),$$

  или, в случае d переменных,

$$f(x_1, ..., x_d) \approx \sum_{k=1}^{d} \phi_{1k}(x_1)...\phi_{dk}(x_d);$$

- аппроксимации с разложением по избыточной (линейно зависимой) системе "простых" функций

$$f(x) \approx \alpha_1\Phi_1(x) + ... + \alpha_m\Phi_m(x)$$

  при условии, что заданная точность должна быть получена при минимальном числе ненулевых коэффициентов $\alpha_1, ..., \alpha_m$. Такие системы рекурсивно строятся с помощью *вейвлетов* — функций из тех же подпространств, в которых обычно находятся ошибки аппроксимации достаточно гладких функций (в ортогональных дополнениях к полиномам степени не выше некоторой заданной степени).

Задача о разделении переменных в теории матриц сводится к изучению аппроксимаций вида $A \approx A_r = u_1v_1^\top + ... + u_rv_r^\top$. Она решается на основе сингулярного разложения (см. главу 2). Если порядок матрицы A равен n и $r \ll n$, то матрица A_r определяется существенно меньшим числом параметров, чем исходная матрица: $2rn \ll n^2$.

Если матрица A невырожденная, то ее минимальное сингулярное число σ_n положительно, и, как следует из результатов главы 2, при $r \leq n - 1$

получаем $||A - A_r||_2 \geq \sigma_n$. Это означает, что точность $\varepsilon < \sigma_n$ недостижима! Однако, в невырожденных матрицах нередко выделяются совокупности элементов (помимо прямоугольных блоков, это могут быть элементы нижнего (верхнего) треугольника, элементы внутри ленты и т. д.), с высокой точностью приближаемых элементами в тех же позициях некоторой матрицы малого ранга.

25.2 Малый ранг и ленточные матрицы

Яркий пример "кусочно-малоранговой" структуры — матрицы, обратные к ленточным.

Пусть $1 \leq p, q \leq n$. Матрица A порядка n с элементами a_{ij} называется *ленточной* типа (p, q), если $a_{ij} = 0$ при $i - j < -p$ or $i - j > q$, и *неприводимой*, если все элементы на граничных линиях $i - j = -p$ и $i - j = q$ отличны от нуля.

Обозначим через $\mathcal{U}_p$ и $\mathcal{L}_q$ пространства верхних и нижних треугольных матриц порядка p и q соответственно. Матрица A называется *семисепарабельной* типа (p, q), если

$$A = S + \begin{bmatrix} 0 & U \\ 0 & 0 \end{bmatrix}, \quad \mathrm{rank} S = q, \quad U \in \mathcal{U}_{n-q}, \tag{25.2.1}$$

$$A = R + \begin{bmatrix} 0 & 0 \\ L & 0 \end{bmatrix}, \quad \mathrm{rank} R = p, \quad L \in \mathcal{L}_{n-p}. \tag{25.2.2}$$

Теорема 25.2.1 *Невырожденная матрица является неприводимой ленточной типа (p, q) в том и только том случае, когда ее обратная матрица является семисепарабельной типа (p, q).*

Доказательство. [1] Предположим, что A — невырожденная семисепарабельная матрица типа (p, q). Запишем (25.2.1) в блочной форме

$$A = \begin{bmatrix} u_1 v_1 & u_1 v_2 + U \\ u_2 v_1 & u_2 v_2 \end{bmatrix}, \quad S = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \begin{bmatrix} v_1 & v_2 \end{bmatrix},$$

полагая, что число столбцов в блоках u_1, u_2 и число строк в блоках v_1, v_2 равно $q = \mathrm{rank} S$. Отсюда ясно, что блоки u_2 и v_1 обратимы. Исключение блока в позиции $(1, 1)$ описывается матричным равенством

$$\begin{bmatrix} I & -(u_1 v_1)(u_2 v_1)^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} u_1 v_1 & u_1 v_2 + U \\ u_2 v_1 & u_2 v_2 \end{bmatrix} = \begin{bmatrix} 0 & u_1 v_2 + U - (u_1 v_1)(u_2 v_1)^{-1}(u_2 v_2) \\ u_2 v_1 & u_2 v_2 \end{bmatrix} =$$

[1] Данный факт неоднократно переоткрывался и передоказывался; идея излагаемого здесь рассуждения принадлежит Д. К. Фаддееву.

$$= \begin{bmatrix} 0 & U \\ u_2v_1 & u_2v_2 \end{bmatrix}.$$

Следовательно, матрица U невырожденная и при этом $A^{-1} = \begin{bmatrix} * & * \\ U^{-1} & * \end{bmatrix}$. Аналогично, из (25.2.2) получаем: $A^{-1} = \begin{bmatrix} * & L^{-1} \\ * & * \end{bmatrix}$. Таким образом, матрица A является ленточной типа (p, q).

Далее, пусть $A = \begin{bmatrix} a & b \\ U & c \end{bmatrix}$. Вспомним формулы Фробениуса:

$$A^{-1} = \begin{bmatrix} U^{-1}ch & U^{-1} - U^{-1}chaU^{-1} \\ h & -haU^{-1} \end{bmatrix}, \quad h = (b - aU^{-1}c)^{-1}.$$

Для получения (25.2.1) остается заметить, что

$$A^{-1} = \begin{bmatrix} U^{-1}ch \\ h \end{bmatrix} \begin{bmatrix} I & aU^{-1} \end{bmatrix} + \begin{bmatrix} 0 & U^{-1} \\ 0 & 0 \end{bmatrix}.$$

Равенство (25.2.2) получается "транспонированием". $\square$

Матрица A называется *квазисепарабельной* типа (p, q), если ранг любой подматрицы в ее верхней треугольной части не выше p, а ранг любой подматрицы в ее нижней треугольной части не выше q.

Теорема 25.2.2 *Невырожденная матрица является квазисепарабельной типа (p, q) в том и только том случае, когда то же верно для ее обратной матрицы.*

Доказательство. Достаточно установить, что для любых согласованных блочных разбиений

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

с квадратными блоками A_{11}, B_{11} имеет место равенство $\mathrm{rank}B_{12} = \mathrm{rank}A_{12}$.

Если блок A_{11} обратим, то это следует из формул Фобениуса. Если нет, рассмотрим матрицы $A(t) = A + tI$ с вещественным параметром t. Для всех достаточно малых $0 < t < \varepsilon$ блок $A_{11}(t)$ обратим (почему?), откуда следует, что $\mathrm{rank}B_{12}(t) = \mathrm{rank}A_{12}(t) \ \ \forall\, 0 < t < \varepsilon$. Очевидно, $A(t) \to A$ при $t \to 0 \quad \Rightarrow \quad \mathrm{rank}B_{12} = \mathrm{rank}A_{12}$. $\square$

Следствие 25.2.1 *Матрица, обратная к невырожденной ленточной матрице типа (p, q) является квазисепарабельной типа (p, q).*

25.3 Многоуровневые матрицы

Пусть A — матрица размеров $m \times n$, $I = \{1, 2, ..., m\}$ и $J = \{1, ..., n\}$. Тогда

$$A(\hat{I}, \hat{J}), \quad \hat{I} \subset I, \quad \hat{J} \subset J,$$

обозначает подматрицу на пересечении строк с номерами из $\hat{I}$ и столбцов с номерами из $\hat{J}$. Любые разбиения на непересекающиеся подмножества

$$I_1 \cup ... \cup I_{m_1} = I, \quad J_1 \cup ... \cup J_{n_1} = J$$

позволяют естественным образом рассматривать A как блочную матрицу с блоками $A(I_k, J_l)$, $1 \le k \le m_1$, $1 \le l \le n_1$. Эти блоки назовем *блоками 1-го уровня*.

Пусть имеются также разбиения

$$\tilde{I}_1 \cup ... \cup \tilde{I}_{m_2} = I, \quad \tilde{J}_1 \cup ... \cup \tilde{J}_{n_2} = J, \quad m_2 > m_1, \quad n_2 > n_1,$$

в которых подмножества не пересекаются и являются частями подмножеств предыдущих разбиений. Блоки $A(\tilde{I}_k, \tilde{J}_l)$ назовем *блоками 2-го уровня*. Таким образом, каждый блок 1-го уровня рассматривается как блочная матрица, блоки которой суть блоки 2-го уровня. При наличии p вложенных разбиений говорят, что матрица имеет p уровней.

Для матриц, имеющих p уровней, можно с большой пользой рассматривать их *многоуровневые расщепления* вида

$$A = A_1 + ... + A_p,$$

где матрица A_k является *разреженной блочной матрицей*, составленной из блоков k-го уровня. Разреженность означает присутствие заведомо нулевых блоков, остальные блоки можно называть *формально ненулевыми* (им *разрешается* быть ненулевыми). Будем предполагать, что если блок матрицы A_k формально ненулевой, то принадлежащие ему блоки матриц A_l при $l > k$ нулевые.

На Рис. 25.1 показаны блоки трех уровней, “светлые” блоки уровней 1 и 2 нулевые и составлены из блоков уровней 2 и 3 соответственно. В данном случае имеем трехуровневое расщепление $A = A_1 + A_2 + A_3$.

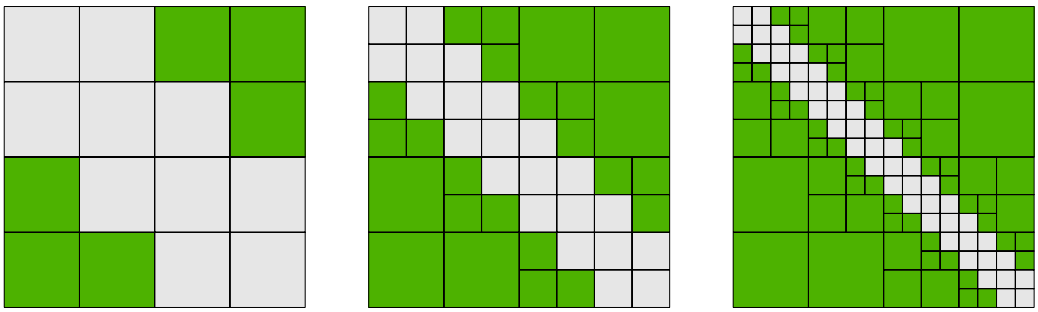


Рисунок 25.1. Матрицы с числом уровней 1, 2, и 3.

При дискретизации характерных для приложений интегральных операторов можно строить расщепления, в которых любой ненулевой блок любого уровня имеет гарантированно малый ранг, а сумма размеров всех ненулевых блоков много меньше общего числа элементов исходной матрицы. Если r — оценка сверху для рангов и τ — сумма размеров ненулевых блоков, то расщепление можно задавать посредством $r\tau$ параметров, а при умножении такой матрицы на вектор выполняется $O(r\tau)$ операций. Попробуем разобраться, почему возможны такие расщепления.

25.4 Матрицы и функции

Предположим, что $a_{st} = f(x_s, y_t)$, где $x_1, ..., x_n$ и $y_1, ..., y_n$ — точки в пространстве $\mathbb{R}^d$. При $n = 1, 2...$ будем рассматривать различные сетки x_s и y_t, полагая, что все точки принадлежат одному и тому же параллелепипеду $S = [a_1, b_1] \times ... \times [a_d, b_d]$.

Для каждого $k = 1, 2, ...$ рассмотрим равномерные сетки, разбивающие ребра S на 2^{k+1} равных отрезков меньшей длины. Всего имеется $(2^{k+1})^d$ декартовых произведений этих отрезков. Занумеровав их произвольным образом, получаем представление S в виде объединения равных параллелепипедов S_{ik} объема h_k:

$$S = \bigcup_{i=1}^{2^{(k+1)d}} S_{ik}, \qquad h_k = \frac{h_0}{2^{(k+1)d}},$$

где h_0 — объем исходного параллелепипеда.

Основное предположение о сетках: пусть в множество S_{ik} попадают $\mu(S_{ik})$ точек x_s и $\nu(S_{ik})$ точек y_t, тогда

$$\max\{\mu(S_{ik}),\, \nu(S_{ik})\} \;\le\; ch_k n \quad \forall \;\; i,\, k,\, n, \tag{25.4.3}$$

где $c > 0$ не зависит от i, k и n.

Основное предположение о функции: для любого $\varepsilon > 0$ функция $f(x, y)$ при $x \in S_{ik}$ и $y \in S_{jl}$ в том случае, если $S_{ik} \cap S_{jl} = \emptyset$, равномерно приближается с точностью ε суммой $r = r(\varepsilon)$ функций с разделенными переменными:

$$\left| f(x, y) - \sum_{\alpha=1}^{r} \Phi_\alpha(x)\Psi_\alpha(y) \right| \le \varepsilon \quad \forall \;\; x \in S_{ik}, \;\; \forall \;\; y \in S_{jl}. \tag{25.4.4}$$

Пусть при каждом k множество номеров $\{1, ..., n\}$ разбито на непересекающиеся подмножества I_{ik} таким образом, что если $s \in I_{ik}$,то $x_s \in S_{ik}$.

Пусть аналогичным свойством обладает еще одно разбиение, состоящее из подмножеств J_{ik}: если $t \in J_{ik}$, то $y_t \in S_{ik}$.

Лемма 25.4.1 *Если имеет место (25.4.4), то при условии $S_{ik} \cap S_{jl} = \emptyset$ элементы блока $\hat{A} = A(I_{ik}, J_{jl})$ приближаются с точностью ε элементами некоторой матрицы $\hat{A}_r$ ранга не выше r:*

$$|(\hat{A})_{st} - (\hat{A}_r)_{st}| \le \varepsilon \quad s \in I_{ik}, \quad t \in J_{jl}, \quad \mathrm{rank}\hat{A}_r \le r = r(\varepsilon). \qquad (25.4.5)$$

Доказательство. Достаточно взять

$$(\hat{A}_r)_{st} = \sum_{\alpha=1}^{r} \Phi_\alpha(x_s)\Psi_\alpha(y_t). \quad \square$$

Теорема 25.4.1 *Пусть выполнены основные предположения о сетках и функции. Тогда элементы матрицы $A = [f(x_s, y_t)]$ порядка n приближаются с точностью ε элементами многоуровневой матрицы с расщеплением, в котором ранг каждого ненулевого блока не выше $r = r(\varepsilon)$, а сумма размеров всех ненулевых блоков есть $O(n \log_2 n)$.*

Доказательство. Построим многоуровневую матрицу $B \approx A$, имеющую p уровней и расщепление $B = B_1 + ... + B_p$, в котором матрица B_k состоит из блоков $B(I_{ik}, J_{jk})$. [2]

При $1 \le k < p$ полагаем, что если $S_{ik} \cap S_{jk} \ne \emptyset$, то $B(I_{ik}, J_{jk}) = 0$, а соответствующие элементы матрицы B должны входить в блоки уровней с номерами $l > k$. Если $k = 1$, то все остальные блоки B_1 формально ненулевые.

Далее, при $1 < k \le p$, пусть $S_{ik} \subset S_{i'\,k-1}$. Тогда если $S_{i'\,k-1} \cap S_{j'\,k-1} =$ и $S_{jk} \subset S_{j'\,k-1}$, то $B(I_{ik}, J_{jk}) = 0$, так как соответствующие элементы B уже отнесены к блокам уровней с номерами $l < k$. Таким образом, для каждого подмножества I_{ik} формально ненулевых блоков $B(I_{ik}, J_{jk})$ не больше 6^d (почему?). Сумма размеров каждого из них не превышает $2ch_kn$, а всего имеется h_0/h_k подмножеств I_{ik} $\Rightarrow$ сумма размеров всех формально ненулевых блоков матрицы B_k не больше $2 \cdot 6^d\, h_0\, n$. Следовательно, сумма размеров всех ненулевых блоков имеет вид $O(np)$.

Остается выбрать число уровней p так, чтобы размеры блоков уровня p были не больше r:

$$cnh_k \le r.$$

Ясно, что это можно сделать при $p = O(\log_2 n)$. Тогда ранги всех ненулевых блоков не больше r при общей сумме их размеров $O(n \log_2 n)$. $\square$

[2] В случае $d = 1$ при $k = 1$ и $k = 2$ имеем 4 и 16 блоков соответственно (см. Рис. 25.1).

25.5 Асимптотически сепарабельные функции

Во всех практически интересных ситуациях выполняется неравенство

$$r(\varepsilon) \le c \log^{\gamma} \varepsilon^{-1}, \tag{25.5.6}$$

где $c, \gamma > 0$ — некоторые константы. В таких случаях функцию $f(x, y)$ будем называть *асимптотически сепарабельной*.

Пример асимптотически сепарабельной функции:

$$f(x, y) = \frac{1}{|x - y|^{\theta}}, \quad \theta > 0. \tag{25.5.7}$$

Пусть матрица порождена асимптотически сепарабельной функцией с константой γ в неравенстве 25.5.6. Тогда ε-приближение к матрице A задается с помощью $O(n \log_2 n \log^{\gamma} \varepsilon^{-1})$ параметров. Так же выглядит оценка сложности при приближенном (с точностью ε) умножении матрицы A на вектор.

25.6 Метод крестовой аппроксимации

Согласно теореме 25.4.1, все элементы матрицы A принадлежат различным блокам, допускающим ε-приближение ранга не выше $r = r(\varepsilon)$. Получение такого приближения — задача, решаемая на основе сингулярного разложения матрицы (см. главу 1). Можно ли предложить более эффективный метод для ее решения?

Пусть A обозначает любой из указанных блоков исходной матрицы. Поставим такой вопрос: если известно, что ε-приближение ранга не выше r существует, то можно ли найти $O(\varepsilon)$-приближение, используя лишь элементы каких-то r столбцов и r строк (т. е. *крест*) матрицы A?

Ответ положительный. [3] Приводимая ниже теорема [4] представляет собой матричный аналог принципа максимальных объемов при выборе узлов интерполяции (см. главу 15).

Теорема 25.6.1 *Пусть для матрицы A существует матрица B ранга r такая, что $||B - A||_2 \le \varepsilon$. Тогда, если A имеет блочное разбиение вида*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

[3] Он впервые получен в работе: *С. А. Горейнов, Н. Л. Замарашкин, Е.Е.Тыртышников.* Псевдоскелетные аппроксимации матриц. – ДАН России, 343 (2): 151–152 (1995).

[4] *S. A. Goreinov, E. E. Tyrtyshnikov.* The maximal-volume concept in approximation by low-rank matrices, Contemporary Mathematics, Vol. 280, 47–51 (2001).

где A_{11} — невырожденная подматрица порядка r с максимальным по модулю определителем среди всех подматриц порядка r, то

$$\left| \left(A - \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} A_{11}^{-1} \begin{bmatrix} A_{11} & A_{12} \end{bmatrix} \right)_{ij} \right| \;\leq\; (r+1)\varepsilon, \quad 1 \leq i,j \leq n.$$

Доказательство. Не ограничивая общности, можно считать, что A — квадратная матрица порядка n. Обозначим через

$$\sigma_1(A) \geq ... \geq \sigma_r(A) > \sigma_{r+1}(A) \geq .. \geq \sigma_n(A) \geq 0$$

ее сингулярные числа. Рассмотрим окаймление блока A_{11} с помощью строки $i > r$ и столбца $j > r$:

$$\hat{A} = \begin{bmatrix} & & & a_{1j} \\ & A_{11} & & ... \\ & & & a_{rj} \\ a_{i1} & ... & a_{ir} & a_{ij} \end{bmatrix}.$$

Исключая i-строку, получаем

$$\begin{bmatrix} & & & 0 \\ & I & & ... \\ & & & 0 \\ -\begin{bmatrix} a_{i1} & ... & a_{ir} \end{bmatrix} A_{11}^{-1} & & & 1 \end{bmatrix} \hat{A} = \begin{bmatrix} & & & a_{1j} \\ & A_{11} & & ... \\ & & & a_{rj} \\ 0 & ... & 0 & \varepsilon_{ij} \end{bmatrix},$$

$$\varepsilon_{ij} = a_{ij} - \begin{bmatrix} a_{i1} & ... & a_{ir} \end{bmatrix} A_{11}^{-1} \begin{bmatrix} a_{1j} \\ ... \\ a_{rj} \end{bmatrix}.$$

Наша цель — доказать, что $|\varepsilon_{ij}| \leq (r+1)\varepsilon$. Конечно, это так, если $\varepsilon_{ij} = 0$. Поэтому пусть $\varepsilon_{ij} \neq 0 \;\;\Rightarrow\;\;$ матрица $\hat{A}$ обратима, и нетрудно показать, что

$$(\hat{A}^{-1})_{r+1\,r+1} = \varepsilon_{ij}^{-1} = \det A_{11} / \det \hat{A}.$$

Согласно условию теоремы, ε_{ij}^{-1} — наибольший по модулю элемент $\hat{A}^{-1}$. Значит,

$$\sigma_1(\hat{A}^{-1}) \;\leq\; (r+1)\,|\varepsilon_{ij}^{-1}| \quad\Rightarrow\quad |\varepsilon_{ij}| \;\leq\; (r+1)\sigma_{r+1}(\hat{A}).$$

Из соотношений разделения для сингулярных чисел при исключении столбцов или строк следует, что $\sigma_{r+1}(\hat{A}) \leq \sigma_{r+1}(A) \leq \varepsilon$. $\square$

Пусть A — матрица размеров $m \times n$ и $m, n \gg r$. Теперь мы знаем, что для получения аппроксимации ранга r достаточно иметь крест, составленный из r строк и столбцов матрицы A, т. е. лишь $r(m+n) \ll mn$ ее элементов.

Выбор "правильного" креста в полном соответствии с теоремой 25.6.1 — задача не очень простая. Но часто она может решаться с помощью простого эвристического алгоритма, представляющего собой метод последовательного исключения элементов с выбором ведущего элемента по столбцу.

Эвристика связана с решением о том, какой столбец будет следующим — например, путем просмотра некоторой части элементов "активной" подматрицы и выбора столбца, содержащего максимальный из них по модулю.

Еще одно место, где присутствует эвристика — это критерий остановки: точность аппроксимации проверяется лишь на части элементов. В общем случае такой подход не может дать полной гарантии качества получаемой аппроксимации. Тем не менее, он успешно применяется во многих практических задачах, и, кроме того, для достаточного важного класса матриц, возникающих при решении интегральных уравнений, гарантию все же дать можно.

25.7 Суперфункции

Очевидный способ аппроксимации вектора $u = [u_1, ..., u_n]^\top \approx u_\varepsilon$ такой: если $|u_i| \leq \varepsilon||u||$, то $(u_\varepsilon)_i = 0$, иначе $(u_\varepsilon)_i = u_i$. В этом преобразовании есть смысл, когда среди компонент u_i много относительно малых. Если это не так для вектора u, то можно попытаться перейти к вектору $v = Qu$. Оказывается, существуют семейства матриц Q, при умножении на которые малые компоненты появляются, при этом одна и та же матрица Q "работает" сразу для многих векторов u. Такого типа матрицы определяют так называемые *дискретные вейвлет-преобразования*. С их помощью от плотных матриц A можно переходить к *псевдоразреженным* матрицам $QAQ^\top$.

Классические вейвлет-преобразования связаны с разложениями функций вида

$$f(x) = \sum_k u_k \phi(x - k),$$

где $\phi(x)$ — специально подбираемая *суперфункция* с носителем на отрезке $[0, m]$, и переходом к разложениям по системе "плохих" (ортогональных всем полиномам, степень которых меньше заданного числа p) функций, называемых *вейвлетами*. При построении суперфункции ключевыми являются *уравнение перехода* [5]

$$\phi(x) = \sum_{k=0}^{m} c_k \phi(2x - k), \qquad c_0, c_m \neq 0, \tag{25.7.8}$$

и два свойства:

- *аппроксимация*: полиномы степени меньше p на конечных отрезках представимы в виде линейных комбинаций $\phi(x - k)$;

[5] В западной литературе "dilation equation".

- *ортогональность*: функции $\phi(x-k)$ образуют ортогональную систему.

Добавим также условие нормировки $\int \phi(x)dx = 1$, откуда сразу же получается, что

$$\sum_{k=0}^{m} c_k = 2. \tag{25.7.9}$$

Функция $\phi(x)$ полностью определяется коэффициентами c_k. Пусть, например, $m = 3$. Тогда, согласно (25.7.8),

$$\begin{bmatrix}\phi(1)\\ \phi(2)\end{bmatrix} = \begin{bmatrix}c_1 & c_0\\ c_3 & c_2\end{bmatrix}\begin{bmatrix}\phi(1)\\ \phi(2)\end{bmatrix}.$$

Полагая, что $\phi(x)$ является непрерывной, находим $\phi(0) = \phi(3) = 0$. Если значения $\phi(1)$ и $\phi(2)$ найдены, то уравнение (25.7.8) позволяется вычислить $\phi(x)$ в любой точке $x = k/2^l$ при целых k, l.

Как видим, вектор $[\phi(1), \phi(2)]^\top$ является собственным вектором матрицы $\begin{bmatrix}c_1 & c_0\\ c_3 & c_2\end{bmatrix}$ для собственного значения $\lambda = 1$.

25.8 Классические вейвлеты

Опуская детали, расскажем, как получаются классические суперфункции Добеши. Пусть $c_k = 0$ при $k < 0$ или $k > m$. Свойства аппроксимации и ортогональности описываются следующими условиями:

$$\sum_{k=0}^{m}(-1)^k k^l c_k = 0, \quad 0 \le l \le p-1 \ \ \text{(аппроксимация)}, \tag{25.8.10}$$

$$\sum_{k=0}^{m} c_k c_{k-2l} = \delta_{0l} \quad \text{(ортогональность)}. \tag{25.8.11}$$

Из (25.8.11) вытекает, что число m должно быть нечетным.

Вот полный список условий на коэффициенты c_k в случае $m = 3$ и $p = 2$:

$$\begin{aligned} c_0 + c_1 + c_2 + c_3 &= 2,\\ -c_0 + c_1 - c_2 + c_3 &= 0,\\ -c_1 + 2c_2 - 3c_3 &= 0,\\ c_0c_2 + c_1c_3 &= 0. \end{aligned} \tag{25.8.12}$$

Решение данной системы нелинейных уравнений такое:

$$c_0 = \frac{1+\sqrt{3}}{4}, \quad c_1 = \frac{3+\sqrt{3}}{4}, \quad c_2 = \frac{3-\sqrt{3}}{4}, \quad c_3 = \frac{1-\sqrt{3}}{4}. \tag{25.8.13}$$

Функции, ортогональные на конечных отрезках полиномам степени меньше p, должны быть ортогональны линейным комбинациям функций $\phi(x-k)$. Нетрудно проверить, что таким свойством обладают линейные комбинации функций $\psi(x-k)$, если "плохую" суперфункцию $\psi(x)$ выбрать в виде

$$\psi(x)=\sum_{k=0}^{m} d_k\phi(2x-k),\quad d_k(-1)^{m-k}c_{m-k}. \tag{25.8.14}$$

Теперь фиксируем h и рассмотрим линейное пространство V_n периодических функций вида

$$f(x)=\sum_i u_i\phi\left(\frac{x}{h}-i\right),\qquad u_i=u_{i+n}\ \ \forall\ i,$$

и взаимно-однозначное соответствие между функциями из V_n и векторами:

$$f\leftrightarrow u=[u_0,...,u_{n-1}]^\top.$$

Полагая, что n четно, рассмотрим подпространства $V_{n/2}$ и $W_{n/2}$ "хороших" и "плохих" функций:

$$V_{n/2}=\left\{\sum_i v_i\phi\left(\frac{x}{2h}-i\right)\right\},\qquad W_{n/2}=\left\{\sum_i w_i\psi\left(\frac{x}{2h}-i\right)\right\},$$

$$v_i=v_{i+n/2},\quad w_i=w_{i+n/2}\quad\forall\ i.$$

По построению, $V_n=V_{n/2}\oplus W_{n/2}$ — ортогональная сумма подпространств. Функции из $V_{n/2}$ и $W_{n/2}$ определяются векторами $v=[v_0,...,v_{n/2-1}]^\top$ и $w=[w_0,...,w_{n/2}]^\top$. В силу уравнений (25.7.8) и (25.8.14)

$$u=\begin{bmatrix}C_n^\top & D_n^\top\end{bmatrix}\begin{bmatrix}v\\w\end{bmatrix},$$

$$C_n=\frac{1}{2}[c_{j-2i+m}],\quad D_n=\frac{1}{2}[d_{j-2i+m}],\quad 1\le i\le n/2,\ \ 1\le j\le n,$$

$$c_{i\pm n}=c_i,\quad 0\le i\le n-1.$$

В частности, при $m=3$ и $n=8$ имеем

$$C_8=\frac{1}{2}\begin{bmatrix}c_2&c_3&0&0&0&0&c_0&c_1\\c_0&c_1&c_2&c_3&0&0&0&0\\0&0&c_0&c_1&c_2&c_3&0&0\\c_0&c_1&0&0&&0&c_2&c_3\end{bmatrix},\quad D_8=\frac{1}{2}\begin{bmatrix}d_2&d_3&0&0&0&0&d_0&d_1\\d_0&d_1&d_2&d_3&0&0&0&0\\0&0&d_0&d_1&d_2&d_3&0&0\\d_0&d_1&0&0&&0&d_2&d_3\end{bmatrix}.$$

Выбрав *число уровней* s, рассмотрим расщепление

$$V_n=V_{n/2^s}\oplus W_{n/2^s}\oplus...\oplus W_{n/2}.$$

Пусть функции из подпространств $V_{n/2^l}$ и $W_{n/2^l}$ определяются векторами $v^{(s)}$ и $w^{(s)}$ соответственно. Тогда дискретное вейвлет-преобразование Добеши

$$u = Q \begin{bmatrix} v^{(s)} \\ w^{(s)} \\ \dots \\ w^1 \end{bmatrix}$$

определяется следующим образом:

$$Q = Q_0 Q_1 \dots Q_{s-1}, \qquad Q_l = \left[\begin{array}{cc|c} C^\top_{n/2^l} & D^\top_{n/2^l} & \\ \hline & & I_{n-n/2^l} \end{array}\right].$$

Согласно определению коэффициентов c_i и d_i, матрица Q является ортогональной. Поэтому

$$Q^{-1} = Q^\top = Q^\top_{s-1} \dots Q^\top_1 Q^\top_0.$$

Реализация прямого и обратного вейвлет-преобразований требует, как легко подсчитать, всего лишь $O(n)$ арифметических операций — меньше, чем сложность $O(n \log n)$ быстрого преобразовния Фурье! Многие коэффициенты разложения функции из V_n при вейвлетах оказываются малыми именно потому, что вейвлеты определяются как "плохие" функции: если функция является полиномом степени меньше p локально — на носителе вейвлета, то отвечающий ему коэффициент просто равен нулю.

25.9 Обобщенные вейвлеты

Классические вейвлеты ассоциируются с равномерными сетками. Однако, схожие свойства можно получить с помощью более общей и одновременно более простой конструкции, позволяющей с легкостью вводить нужные для приложений дополнительные требования (например, на границе области определения функций) и строить вейвлеты на неравномерных сетках.

Рассмотрим линейно независимую систему функций $\Phi = [\phi_1, \dots, \phi_n]$ и в их линейной оболочке выберем систему k линейно независмых "хороших" функций $\hat{\Phi} = [\hat{\phi}_1, \dots, \hat{\phi}_k]$ и систему $n-k$ линейно независимых "плохих" функций $\tilde{\Phi} = [\tilde{\phi}_1, \dots, \tilde{\phi}_{n-k}]$, ортогональных полиномам степени меньше p. Эти "плохие" функции и будут рассматриваться как обобщенные вейвлеты.

Предположим, что $\Phi_1 = [\phi_1, \dots, \phi_k]$ и $\Phi_2 = [\phi_{k+1}, \dots, \phi_n]$, а нумерация такова, что в качестве кандидатов на роль "плохих" функций можно рассмотреть Φ_2. Потребуем, чтобы

$$[\hat{\Phi}, \Phi_2] = [\Phi_1, \Phi_2] \begin{bmatrix} D & 0 \\ R & I \end{bmatrix},$$

где матрица D диагональная, а R — разреженная матрица с числом ненулевых элементов $O(n)$. Чтобы обеспечить условия ортогональности, "плохие" функции "подправляются" таким образом:

$$[\hat{\Phi}, \tilde{\Phi}] = [\hat{\Phi}, \Phi_2] \begin{bmatrix} I & -S \\ 0 & I \end{bmatrix},$$

где S — разреженная матрица с числом ненулевых элементов $O(n)$. В итоге

$$[\hat{\Phi}, \tilde{\Phi}] = [\Phi_1, \Phi_2]M, \quad M = \begin{bmatrix} D & 0 \\ R & I \end{bmatrix} \begin{bmatrix} I & -S \\ 0 & I \end{bmatrix}.$$

Далее, пусть функции $\psi_1, ..., \psi_n$ образут дуальную систему: $(\phi_i, \psi_j) = \delta_{ij}$. Пусть $\Psi_1[\psi_1, ..., \psi_k]$ и $\Psi_2 = [\psi_{k+1}, ..., \psi_n]$, и положим

$$[\hat{\Psi}, \tilde{\Psi}] = [\Psi_1, \Psi_2]M^{-\top}.$$

Тогда системы функций $[\hat{\Phi}, \tilde{\Phi}]$ и $[\hat{\Psi}, \tilde{\Psi}]$ также будут дуальными.

Пусть векторы $\hat{u}$ и $\tilde{u}$ содержат коэффициенты разложения по системам $\hat{\Psi}$ и $\tilde{\Psi}$, а векторы $u^{(1)}$ и $u^{(2)}$ — по системам Ψ_1 и Ψ_2. Тогда

$$[\hat{\Psi}, \tilde{\Psi}] \begin{bmatrix} \hat{u} \\ \tilde{u} \end{bmatrix} = [\Psi_1, \Psi_2]M^{-\top} \begin{bmatrix} \hat{u} \\ \tilde{u} \end{bmatrix}.$$

Следовательно,

$$\begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix} = M^{-\top} \begin{bmatrix} \hat{u} \\ \tilde{u} \end{bmatrix}, \qquad \begin{bmatrix} \hat{u} \\ \tilde{u} \end{bmatrix} = M^{\top} \begin{bmatrix} u^{(1)} \\ u^{(2)} \end{bmatrix}.$$

При умножении на матрицу $M^{\top}$ следует ожидать, что некоторые компоненты вектора $\tilde{u}$ будут настолько малыми, что ими можно будет пренебречь. Это заведомо верно для коэффициента при функции $\tilde{\psi}_i$, на носителе которой исходная функция достаточно хорошо приближается полиномом степени не выше p.

Как и в случае классического вейвлет-преобразования, можно выбрать некоторое число уровней s и повторить аналогичное преобразование s раз, применяя его к векторам, имеющим меньшее число элементов и отвечающих "хорошим" дуальным функциям. На практике вся совокупность основных и дуальных функций строится таким образом, что в ней присутствуют функции с носителями разной протяженности — это приводят к малым коэффициентам на разных уровнях. Обычно функции ассоциируются с последовательностью сеток, как и в многосеточном методе. В случае неравномерных сеток в качестве "хороших" функций удобно использовать B-сплайны.

Описанная здесь схема одного шага рекурсии, включающая выбор кандидатов на роль вейвлетов и последующую их ортогонализацию, предложена Свелденсом и называется *схемой лифтинга*. Вообще говоря, схема применима и в случае неструктурированных сеток на плоскости или в трехмерном пространстве (например, для треугольных сеток).

Задачи

1. Покажите, что теорема 25.2.1 теряет силу без условия неприводимости ленточной матрицы.

2. Дана "трехмерная матрица" с элементами $a_{ijk} \in \mathbb{R}$, $1 \le i, j, k \le 2$. Докажите, что существует такое ее представление с разделением индексных переменных вида

$$a_{ijk} = \sum_{s=1}^{r} u_{is} v_{js} w_{ks}, \quad u_{is}, v_{js}, w_{ks} \in \mathbb{R}, \qquad (*)$$

   для которого $r = 3$.

3. Дана "трехмерная матрица" с элементами a_{ijk}, $1 \le i, j, k \le n$. Докажите, что существует разложение вида $(*)$, в котором $r = n^2$.

4. Докажите, что функция (25.5.7) является асимптотически сепарабельной.

5. Докажите, что при нечетном m функция $\psi(x)$ вида (25.8.14) ортогональна $\phi(x)$ вида (25.7.8).

6. Докажите, что ортогональность функций $\phi(x-k)$, удовлетворяющих уравнению перехода (25.7.8), равносильна условию (25.8.11).

7. Докажите, что свойство аппроксимации при построении суперфункции равносильно условию (25.8.10).

Литература

1. А.А.Амосов, Ю.А.Дубинский, Н.В.Копченова, Вычислительные методы для инженеров, М., Высшая школа, 1994.

2. К.И.Бабенко, Основы численного анализа, Наука, 1986.

3. Н.С.Бахвалов, Численные методы, Наука, 1975.

4. Н.С.Бахвалов, Н.П.Жидков, Г.М.Кобельков, Численные методы, Наука, 1987.

5. С.М.Белоцерковский, И.К.Лифанов, Численные методы в сингулярных интегральных уравнениях, М., Наука, 1985.

6. И.С.Березин, Н.П.Жидков, Методы вычислений, Физматгиз, 1962.

7. В.В.Воеводин, Вычислительные основы линейной алгебры, Наука, 1977.

8. В.В.Воеводин, Ю.А.Кузнецов, Матрицы и вычисления, М., Наука, 1984.

9. В.В.Воеводин, Е.Е.Тыртышников, Вычислительные процессы с теплицевыми матрицами, Наука, 1987.

10. А.О.Гельфанд, Исчисление конечных разностей, 1952.

11. Ф.Гилл, У.Мюррей, М.Райт, Практическая оптимизауия, М.,Мир, 1985.

12. С.К.Годунов, Решение систем линейных уравнений, Новосибирск, Наука, 1980.

13. С.К.Годунов, Современные аспекты линейной алгебры, Новосибирск, Научная книга, 1997.

14. Дж. Голуб, Ч. Ван Лоун, Матричные вычисления, М., Мир, 1999.

15. И.Ц.Гохберг, И.А.Фельдман, Уравнения в свертках и проекционные методы их решения, М., Наука, 1971.

16. Дж.Деммель, Вычислительная линейная алгебра, М., Мир, 2001.

17. И.Добеши, Десять лекций по вейвлетам, Москва-Ижевск, НИЦ "Регулярная и хаотическая динамика", 2004.

18. Е.Г.Дьяконов, Минимизация вычислительной работы, Наука, 1989.

19. А.Джордж, Дж.Лю, Численное решение больших разреженных систем уравнений, М., Мир, 1984.

20. Ю.С.Завьялов, Б.И.Квасов, В.Л.Мирошниченко, Методы сплайн–функций, Наука, 1980.

21. Х.Д.Икрамов, Численное решение матричных уравнений, М., Наука, 1984.

22. Х.Д.Икрамов, Несимметричная проблема собственных значений, Наука, 1991.

23. Х.Д.Икрамов, Численные методы для симметричных линейных систем, Наука, 1988.

24. В.П.Ильин, Ю.И.Кузнецов, Алгебраические основы численного анализа, Новосибирск, Наука, 1986.

25. В.Г.Карманов, Математическое программирование, Наука, 1975.

26. Д.Каханер, К.Моулер, С.Нэш, Численные методы и программное обеспечение, М., Мир, 1998.

27. В.И.Лебедев, Функциональный анализ и вычислительная математика, М., Физматлит, 2000.

28. О.В.Локуциевский, М.Б.Гавриков, Начала численного анализа, М., ТОО "Янус", 1995.

29. Ч.Лоусон, Р.Хенсон, Численное решение задач метода наименьших квадратов, М., Наука, 1986.

30. Г.И.Марчук, Методы вычислительной математики, Наука, 1989.

31. Г.И.Марчук, В.И.Агошков, Введение в проекционно-сеточные методы, М., Наука, 1981.

32. И.П.Мысовских, Интерполяционные кубатурные формулы, М., Наука, 1981.

33. М.А.Ольшанский, Лекции и упражнения по многосеточным методам, М., Физматлит, 2005.

34. Дж.Ортега, В.Рейнболдт, Итерационные методы решения нелинейных систем уравнений со многими неизвестными, М., Мир, 1975.

35. А.М.Островский, Решение уравнений и систем уравнений, Издательство иностранной литературы, 1963.

36. В.С.Рябенький, Введение в вычислительную математику, М., Наука, 1994.

37. А.А.Самарский, А.В.Гулин, Численные методы, Наука, 1989.

38. А.Г.Сухарев, А.В.Тимохов, В.В.Федоров, Курс методов оптимизации, Наука, 1986.

39. Г. Стрэнг, Линейная алгебра и ее применения, М., Мир, 1980.

40. Г.Стрэнг, Дж.Фикс, Теория метода конечных элементов, М., Мир, 1977.

41. Е. Е. Тыртышников, Теплицевы матрицы, некоторые их аналоги и приложения, Отдел вычислительной математики АН СССР, М., 1989.

42. Е. Е. Тыртышников, Краткий курс численного анализа, М., ВИНИТИ, 1994.

43. Дж.Уилкинсон, Алгебраическая проблема собственных значений, Наука, 1970.

44. Д.К.Фаддеев, В.Н.Фаддеева, Вычислительные методы линейной алгебры, Физматгиз, 1963.

45. Р.П.Федоренко, Введение в вычислительную физику, М., Издательство МФТИ, 1994.

46. Дж.Форсайт, М.Малькольм, К.Моулер, Машинные методы математических вычислений, Мир, 1980.

47. Л.Хейгеман, Д.Янг, Прикладные итерационные методы, М., Мир, 1986.

48. Р.Хорн, Ч.Джонсон, Матричный анализ, Мир, 1989.

49. Е.Чижонков, Численные методы, М., МГУ, 2006.

50. К.Чу, Введение в вейвлеты, М., Мир, 2001.

51. В.В.Шайдуров, Многосеточные методы конечных элементов, М., Мир, 1989.

52. R.Bhatia, *Matrix Analysis*, Springer–Verlag, New York, 1996.

53. A.Böttcher, B.Silbermann, Introduction to large truncated Toeplitz matrices, Springer, New York, 1999.

54. A.Greenbaum, Iterative methods for solving linear systems, SIAM, Philadelphia, 1997.

55. G.Heinig, K.Rost, Algebraic methods for Toeplitz-like matrices and operators, Akademie-Verlag, Berlin, 1984.

56. N.J.Higham, Accuracy and stability of numerical algorithms, SIAM, Philadelphia, 1996.

57. R.Kress, Linear integral equations, Springer-Verlag, 1989.

58. U.Rüde, Mathematical and computational techniques for multilevel adaptive methods, SIAM, Philadelphia, 1993.

59. Y.Saad, Iterative methods for sparse linear systems, PWS Publishing Co., Boston, 1996.

60. G.W.Stewart, J.Sun, Matrix Perturbation Thery, Academic Press, 1990.