

## Глава 13. ЭЛЕМЕНТАРНЫЙ АНАЛИЗ ПОГРЕШНОСТЕЙ

### 13.1. Предварительные замечания

Решение любой вычислительной задачи можно представить формулой

$$y = F(x). \quad (13.1.1)$$

Здесь  $x \in \mathbb{C}^n$  – вектор исходных данных,  $y \in \mathbb{C}^m$  – вектор результатов,  $F$  – оператор, действующий из  $\mathbb{C}^n$  в  $\mathbb{C}^m$ .

В реальной ситуации вектор  $x$  содержит неопределенность, т.е. "на вход" подается не вектор  $x$ , а некоторый другой вектор  $\tilde{x}$ . Точно так же оператор реализуется неточно, т.е. фактически работает некоторый другой оператор  $\tilde{F}$ .

Таким образом, вместо требуемого результата  $y$ , мы получаем "на выходе" некоторый другой результат  $\tilde{y}$ . Пусть  $\tilde{y} = \tilde{F}(\tilde{x})$ . Тогда можно записать, что

$$\Delta y = \tilde{y} - y = (\tilde{y} - \tilde{F}(\tilde{x})) + (\tilde{F}(\tilde{x}) - F(\tilde{x})) + (F(\tilde{x}) - F(x)).$$

Мы представили погрешность результата в виде суммы двух слагаемых. При этом  $\Delta^T y = F(\tilde{x}) - F(x)$  – погрешность, возникающая при точной реализации вычислительного алгоритма из-за погрешности в исходных данных, а  $\Delta^M y = \tilde{F}(\tilde{x}) - F(\tilde{x})$  – погрешность, возникающая из-за неточной реализации алгоритма.

Существуют различные наименования для этих составляющих. Мы будем называть  $\Delta y_T$  *трансформированной погрешностью*, а  $\Delta y_M$  – *погрешностью метода*.

В предыдущих главах мы предполагали, что все исходные данные рассматриваемых задач заданы точно и точно выполняются арифметические операции. Однако в действительности и в исходных данных обычно имеется неопределенность (так как они являются, как правило, либо результатами измерений, либо результатами вычислений), и арифметические операции выполняются либо с округлением, либо с усечением. Это приводит к появлению в результатах неопределенности, без оценки которой пользоваться этими результатами нельзя. Могут появляться и несуществующие на самом деле "решения" задач.

Пример. (см. п.1.1). Решая очевидно несовместную систему

$$\begin{cases} 7x + 9y = 16 \\ 14x + 18y = 32.01 \end{cases}$$

методом Гаусса-Жордана без выбора ведущего элемента с помощью микрокалькулятора МК-56, получаем "ответ":

$$x = -6427.571, \quad y = 38890.667.$$

Проиллюстрируем проблему геометрически.

Как известно, линейное алгебраическое уравнение с двумя переменными задает на плоскости прямую, а система из двух таких уравнений – пару прямых. Если матрица коэффициентов системы не вырождена, то прямые не параллельны и имеют единственную точку пересечения, координаты которой – единственное решение системы (рис. 13.1).

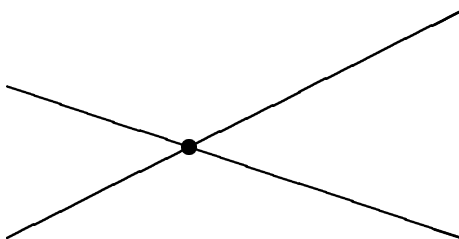


Рис. 13.1

Наличие неопределенности в исходных данных (коэффициентах и свободных членах системы) превращает каждую прямую в целое семейство прямых. В частном случае, когда неопределенность содержится только в свободных членах, прямые каждого семейства параллельны между собой, и каждое уравнение системы порождает "толстую" прямую (полосу, толщина которой растет с ростом неопределенности. "Решением" системы теперь является образующаяся при пересечении "толстых" прямых "толстая" точка, размеры которой характеризуют неопределенность решения (рис. 13.2).

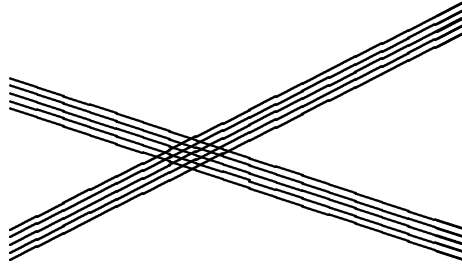


Рис. 13.2

Если матрица коэффициентов системы ортогональна, то прямые перпендикулярны, и размеры "решения" будут того же порядка, что и ширина полос. Если угол между полосами близок к нулю, то размеры решения могут во много раз превышать ширину полос. При параллельных же прямых может возникнуть "решение, не существующее в точной арифметике" (рис. 13.3).

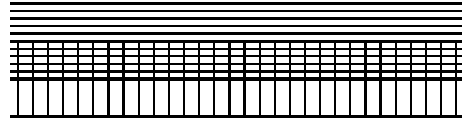


Рис. 13.3

Эти геометрические построения показывают (хотя и не доказывают), что "плохими" являются системы, близкие к вырожденным. Чтобы придать этому выражению точный смысл, нам потребуется ввести некоторые новые понятия.

### 13.2. Норма матрицы

Для квадратичной формы, порожденной эрмитовой матрицей  $A$ , известно неравенство (см. п.9.2)

$$\langle Ax, x \rangle \leq \lambda_{\max} \cdot \|x\|^2, \quad (13.2.1)$$

где  $\lambda_{\max}$  – наибольшее собственное число матрицы  $A$ .

Пусть теперь  $B$  – произвольная матрица размера  $m \times n$ . Из (13.2.1) следует, что

$$\|Bx\|^2 = \langle Bx, Bx \rangle = \langle B^* Bx, x \rangle \leq \sigma_{\max}^2 \cdot \|x\|^2 \quad \text{или} \quad \|Bx\| \leq \sigma_{\max} \cdot \|x\|, \quad (13.2.2)$$

где  $\sigma_{\max}$  – наибольшее сингулярное число матрицы  $B$ . Если вектор  $x$  *ненулевой*, то, разделив на его норму обе части (13.2.2), получим

$$\frac{\|Bx\|}{\|x\|} \leq \sigma_{\max}. \quad (13.2.3)$$

Покажем, что в (13.2.3) равенство достигается. Если  $v$  – правый, а  $u$  – левый сингулярные векторы, соответствующие наибольшему сингулярному числу, то (см. п.10.1):

$$\|Bv\| = \|\sigma_{\max} u\| = \sigma_{\max} = \sigma_{\max} \|v\| = \|B\| \cdot \|v\| \quad \text{и} \quad \frac{\|Bv\|}{\|v\|} = \sigma_{\max}.$$

Геометрическая интерпретация неравенства (13.2.3) очевидна: при умножении ненулевого вектора на матрицу слева евклидова норма этого вектора изменяется. Норма вектора в  $\mathbb{R}^3$  – это длина соответствующего ему направленного отрезка. Поэтому естественно назвать отношение  $\frac{\|Bx\|}{\|x\|}$  "коэффициентом растяжения" вектора этой матрицей. Тогда неравенство (13.2.3) показывает, что коэффициент растяжения для каждой матрицы не может быть больше, чем ее наибольшее сингулярное число.

Определение. Нормой матрицы называется ее наибольшее сингулярное число

$$\|B\| = \max_{x \neq 0} \frac{\|Bx\|}{\|x\|} = \sigma_{\max}. \quad (13.2.4)$$

**Замечание.** Мы уже отмечали ранее, что норма вектора в линейном пространстве может вводиться различными способами. Соответственно появятся различные нормы матрицы. Если в двух конечномерных линейных пространствах  $X$  и  $Y$  введены нормы векторов  $\|\cdot\|_X$  и  $\|\cdot\|_Y$  соответственно, то линейному оператору, порожденному матрицей  $B$ , можно сопоставить число

$$\|B\|_{X \rightarrow Y} = \max_{x \neq \theta} \frac{\|Bx\|_Y}{\|x\|_X}.$$

Это число называют нормой оператора (нормой матрицы).

Мы использовали при определении нормы матрицы *евклидову* норму вектора. Поэтому полное наименование этой нормы – *норма, подчиненная евклидовой норме вектора*.

Установим свойства матричной нормы (любой).

$$\boxed{1. \quad \|B\| \geq 0 \quad - \text{ следует из определения нормы матрицы}}$$

$$\|B\| = 0 \iff \|Bx\| = 0 \quad \text{для любого } x \in \mathbb{C}^n \iff Bx = \theta \quad \text{для любого } x \in \mathbb{C}^n \iff B = \Theta \quad \text{нулевая матрица.}$$

$$\boxed{2. \quad \|B\| = 0 \iff B = \Theta.}$$

$$\|\alpha B\| = \max_{x \neq \theta} \frac{\|\alpha Bx\|}{\|x\|} = |\alpha| \cdot \max_{x \neq \theta} \frac{\|Bx\|}{\|x\|} = |\alpha| \cdot \|B\|.$$

$$\boxed{3. \quad \|\alpha B\| = |\alpha| \cdot \|B\|.}$$

Для любого  $x \in \mathbb{C}^n$

$$\|(A+B)x\| = \|Ax+Bx\| \leq \|Ax\| + \|Bx\| \leq \|A\| \cdot \|x\| + \|B\| \cdot \|x\| = (\|A\| + \|B\|) \cdot \|x\|.$$

Отсюда

$$\|A+B\| = \max_{x \neq \theta} \frac{\|(A+B)x\|}{\|x\|} \leq \max_{x \neq \theta} \frac{(\|A\| + \|B\|) \cdot \|x\|}{\|x\|} = \|A\| + \|B\|.$$

$$\boxed{4. \quad \|A+B\| \leq \|A\| + \|B\|.}$$

$$\|(AB)x\| = \|A(Bx)\| \leq \|A\| \cdot \|Bx\| \leq \|A\| \cdot \|B\| \cdot \|x\|. \quad \text{Поэтому}$$

$$\|AB\| = \max_{x \neq \theta} \frac{\|(AB)x\|}{\|x\|} \leq \max_{x \neq \theta} \frac{\|A\| \cdot \|B\| \cdot \|x\|}{\|x\|} = \|A\| \cdot \|B\|.$$

$$\boxed{5. \quad \|AB\| \leq \|A\| \cdot \|B\|.}$$

Следующие свойства верны для матричной нормы, *подчиненной евклидовой норме вектора*.

В п.10.1 доказано, что сингулярные числа матриц  $B^*$  и  $B$  совпадают. Поэтому

$$\boxed{6. \quad \|B^*\| = \|B\|.}$$

Если  $B$  – *обратимая* матрица, то  $\|B^{-1}\| = \frac{1}{\sigma_{\min}(B)}$ , так как если  $\sigma^2$  – собственное число матриц  $B^*B$  и  $BB^*$ , то собственное число матрицы  $(B^{-1})^*B^{-1} = (BB^*)^{-1} = \frac{1}{\sigma^2}$ . Итак,  $\sigma_{\max}(B^{-1}) = \frac{1}{\sigma_{\min}(B)}$ .

$$\boxed{7. \quad \|B^{-1}\| = \frac{1}{\sigma_{\min}(B)}.$$

Умножение на унитарную матрицу не изменяет норму вектора. Поэтому

$$\boxed{8. \quad \text{Норма унитарной матрицы равна единице.}}$$

### 13.3. Трансформированная погрешность решения системы линейных алгебраических уравнений. Число обусловленности матрицы

Пусть в системе

$$Ax = b \quad (13.3.1)$$

невырожденная  $n \times n$ -матрица  $A$  известна точно и все арифметические операции выполняются без округлений и усечений, а свободный член содержит погрешность, т.е. фактически вместо системы (13.3.1) решается система

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

Вычитая из этого уравнения (13.3.1), получим

$$A \cdot \Delta x = \Delta b \quad \text{или} \quad \Delta x = A^{-1} \cdot \Delta b. \quad (13.3.2)$$

По свойствам матричной нормы из (13.3.1) и (13.3.2) имеем

$$\|b\| \leq \|A\| \cdot \|x\|; \quad \|\Delta x\| \leq \|A^{-1}\| \cdot \|\Delta b\|.$$

Перемножив эти неравенства, получим

$$\|b\| \cdot \|\Delta x\| \leq \|A\| \cdot \|A^{-1}\| \cdot \|\Delta b\| \cdot \|x\|, \quad \text{т.е.} \quad \frac{\|\Delta x\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\Delta b\|}{\|b\|}.$$

Введя понятие "относительная погрешность"

$$\delta x = \frac{\|\Delta x\|}{\|x\|}, \quad \delta b = \frac{\|\Delta b\|}{\|b\|},$$

придем к основному неравенству

$$\delta x \leq \text{cond}(A) \cdot \delta b, \quad (13.3.3)$$

где  $\text{cond}(A) = \|A\| \cdot \|A^{-1}\|$  – так называемое *число обусловленности* матрицы  $A$ .

Из (13.3.3) видно, что относительная погрешность результата при точно известной матрице и точном выполнении арифметических операций может превысить относительную погрешность исходных данных не более, чем в  $\text{cond}(A)$  раз.

Покажем, что в (13.3.3) может достигаться равенство. Если нам "очень не повезло" и  $x$ ,  $\Delta x$  – правые сингулярные векторы матрицы  $A$ , причем  $x$  соответствует наибольшему сингулярному числу, а  $\Delta x$  – наименьшему, то

$$\|x\| = 1, \quad \|\Delta x\| = 1, \quad \delta x = 1;$$

$$\|b\| = \|Ax\| = \sigma_{\max}, \quad \|\Delta b\| = \|A\Delta x\| = \sigma_{\min}, \quad \delta b = \frac{\sigma_{\min}}{\sigma_{\max}};$$

$$\frac{\delta x}{\delta b} = \sigma_{\max} \cdot \frac{1}{\sigma_{\min}} = \|A\| \cdot \|A^{-1}\| = \text{cond}(A).$$

Результат, очевидно, не изменится, если  $x$  и  $\Delta x$  не совпадают с сингулярными векторами, а лишь коллинеарны им.

Таким образом, число обусловленности матрицы коэффициентов системы линейных алгебраических уравнений – это наибольшее значение "коэффициента усиления" относительной погрешности в задании свободного члена. При решении "плохо обусловленной" системы, т.е. системы, матрица коэффициентов которой имеет большое число обусловленности, может происходить (даже при точной арифметике!) катастрофическая потеря точности.

Покажем, как можно конструировать плохо обусловленные системы.

Пример. Для матрицы  $A = \begin{bmatrix} 1 & 0 \\ N & 1 \end{bmatrix}$   $N > 0$  имеем

$$A^*A = \begin{bmatrix} 1 + N^2 & N \\ N & 1 \end{bmatrix}, \quad P_{A^*A}(\lambda) = \lambda^2 - (N^2 + 2)\lambda + 1,$$

$$\lambda_{\max}(A^*A) = \frac{N^2 + 2 + N \cdot \sqrt{N^2 + 4}}{2} > N^2 + 1, \quad \lambda_{\min}(A^*A) = \frac{1}{\lambda_{\max}(A^*A)},$$

$$\text{cond}(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} = \left( \frac{\lambda_{\max}(A^*A)}{\lambda_{\min}(A^*A)} \right)^{1/2} = \lambda_{\max}(A^*A) > N^2 + 1(!)$$

Итак, при решении этой системы (всего-то два уравнения) относительная погрешность задания свободного члена может трансформироваться в относительную погрешность результата, усилившись более, чем в  $N^2$  раз!

Анализ трансформированной погрешности усложняется, если ошибки присутствуют и в матрице коэффициентов. В предположении, что относительные погрешности исходных данных *малы*, можно показать, что

$$\delta x \leq \text{cond}(A) \cdot \frac{\delta A + \delta b}{1 - \text{cond}(A) \cdot \delta(A)}$$

(здесь  $\delta A$  – относительная погрешность задания матрицы коэффициентов).

Рассмотрим некоторые свойства числа обусловленности:

1.  $\text{cond}(A^{-1}) = \text{cond}(A)$  (следует из определения).
2.  $\text{cond}(A^*) = \text{cond}(A)$ , так как  $\|A^*\| = \|A\|$ .
3.  $\text{cond}(\alpha A) = \text{cond}(A)$  при  $\alpha \neq 0$ , так как

$$\text{cond}(\alpha A) = \|\alpha A\| \cdot \|(\alpha A)^{-1}\| = |\alpha| \cdot \|A\| \cdot |\alpha^{-1}| \cdot \|A^{-1}\| = \text{cond}(A).$$

**Замечание.** Бытует суеверие (нелепое, как все суеверия), связывающее погрешности решения системы линейных алгебраических уравнений с величиной определителя ее матрицы коэффициентов. Из доказанного выше видно, что произвольно меняя величину этого определителя (умножая матрицу коэффициентов на различные числа), мы не меняем число обусловленности. Приведенный выше пример показывает, что не меняя величину определителя ( $\det \begin{pmatrix} 1 & 0 \\ N & 1 \end{pmatrix} = 1$  при любом  $N$ ), мы можем получить сколь угодно большое число обусловленности. Этот пример показывает также, что большое число обусловленности можно получить и у матрицы небольшого порядка.

4.  $\text{cond}(AB) \leq \text{cond}(A) \cdot \text{cond}(B)$ , так как

$$\text{cond}(AB) = \|AB\| \cdot \|(AB)^{-1}\| \leq \|A\| \cdot \|B\| \cdot \|A^{-1}\| \cdot \|B^{-1}\| = \text{cond}(A) \cdot \text{cond}(B).$$

5.  $\text{cond}(A) \geq 1$ , так как

$$1 = \text{cond}(I) = \text{cond}(A \cdot A^{-1}) \leq \text{cond}(A) \cdot \text{cond}(A^{-1}) = (\text{cond}(A))^2.$$

6. Если  $U$  – унитарная матрица, то

$$\text{cond}(U) = \|U\| \cdot \|U^{-1}\| = \|U\| \cdot \|U^*\| = 1 \cdot 1 = 1;$$

$$\text{cond}(AU) \leq \text{cond}(A) \cdot \text{cond}(U) = \text{cond}(A), \quad A = (AU)U^*, \quad \text{поэтому}$$

$$\text{cond}(A) \leq \text{cond}(AU) \cdot \text{cond}(U^*) = \text{cond}(AU). \quad \text{Отсюда} \quad \text{cond}(AU) = \text{cond}(A).$$

Так же доказывается, что  $\text{cond}(UA) = \text{cond}(A)$ .

**Замечания.** 1. Погрешности исходных данных не могут быть известны по определению. Обычно удается получить лишь некоторую их оценку сверху. Соответственно, и для погрешности результата можно получить только некоторую оценку.

2. Получение точного результата при "зашумленных" исходных данных невозможно. Поэтому трансформированную погрешность часто называют "неустранимой".

### 13.4. Факторизация матриц и число обусловленности

Как уже упоминалось, при решении задач линейной алгебры используются разложения матриц на множители специального вида. В связи с этим рассмотрим вопрос о влиянии такого разложения на трансформированную погрешность.

Итак, пусть система

$$Ax = b \tag{13.4.1}$$

решается с помощью факторизации матрицы  $A$ :

$$A = A_1 \cdot A_2,$$

т.е. вместо системы (13.4.1) решаются последовательно две системы:

$$A_1 y = b, \quad A_2 x = y. \tag{13.4.2}$$

Если вектор  $b$  имеет погрешность  $\Delta b$ , то вектор  $y$  будет получен с погрешностью  $\Delta y = A_1^{-1} \cdot \Delta b$ , а вектор  $x$  – с погрешностью

$$\Delta x = A_2^{-1} \cdot \Delta y = A_2^{-1} \cdot A_1^{-1} \cdot \Delta b = (A_1 A_2)^{-1} \cdot \Delta b = A^{-1} \Delta b. \quad (13.4.3)$$

Казалось бы, погрешность результата не зависит от способа факторизации. Однако в наших выкладках мы упустили из виду, что при решении первой системы к трансформированной погрешности прибавится погрешность метода  $\Delta y$ , вызванная неточностью машинной арифметики. При решении второй системы эта погрешность играет роль погрешности в исходных данных и порождает дополнительную трансформированную погрешность  $A_2^{-1} \cdot \Delta y$ . Поэтому большое значение имеют числа обусловленности матриц-сомножителей, и их произведение естественно считать критерием качества факторизации.

Очевидно, что это произведение не меньше числа обусловленности факторизуемой матрицы, т.е. никакая факторизация не может улучшить "плохую" матрицу. Однако неудачная факторизация может "испортить" даже хорошо обусловленную матрицу.

Пример. Произведем  $LU$ -разложение унитарной матрицы (ее число обусловленности равно единице)

$$A = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} = LU = \begin{bmatrix} 1 & 0 \\ \operatorname{tg}(\varphi) & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ 0 & \frac{1}{\cos(\varphi)} \end{bmatrix}.$$

При  $0 < \varphi < \frac{\pi}{2}$  положим в примере из п. 13.3  $N = \operatorname{tg}(\varphi)$ . Тогда  $\operatorname{cond}(L) > 1 + \operatorname{tg}^2(\varphi) = \frac{1}{\cos^2(\varphi)}$ . При значениях  $\varphi$ , близких к  $\frac{\pi}{2}$ , получаются очень плохо обусловленные матрицы. Этот пример показывает, в частности, что за исключением специальных случаев не следует строить  $LU$ -разложение без выбора ведущего элемента.

Хорошими свойствами обладает  $QR$ -разложение:

$$\operatorname{cond}(Q) = 1; \quad \operatorname{cond}(R) = \operatorname{cond}(Q^* A) \geq \operatorname{cond}(A); \quad \operatorname{cond}(A) = \operatorname{cond}(QR) \geq \operatorname{cond}(R).$$

Итак,  $\operatorname{cond}(R) = \operatorname{cond}(A)$ .

Так же доказывается, что в сингулярном разложении  $A = U \Sigma V^*$  имеет место равенство  $\operatorname{cond}(\Sigma) = \operatorname{cond}(A)$ .

Приведем еще пример, когда "хорошей" оказывается модификация  $LU$ -разложения. Можно показать, что процесс  $LDU$ -разложения положительно определенной эрмитовой матрицы проходит без перестановок строк и столбцов, причем это разложение имеет вид  $A = U^* D U$ . Тогда при  $x \neq \theta$

$$\langle D x, x \rangle = \langle (U^*)^{-1} A U^{-1} x, x \rangle = \langle A \cdot (U^{-1} x), U^{-1} x \rangle > 0$$

и, в частности,  $d_{jj} = \langle D e^{(j)}, e^{(j)} \rangle > 0, \quad j = 1, \dots, n$ .

Обозначив  $D^{1/2} = \operatorname{diag} [\sqrt{d_{11}}, \dots, \sqrt{d_{nn}}]$ , получим

$$A = U^* D^{1/2} D^{1/2} U = H^* H, \quad (13.4.4)$$

где  $H = D^{1/2} U$  – верхняя треугольная матрица.

Имеем

$$\operatorname{cond}(H) = \frac{\sigma_{\max}(H)}{\sigma_{\min}(H)} = \left( \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \right)^{1/2} = (\operatorname{cond}(A))^{1/2}.$$

Формула (13.4.4) называется *разложением Холецкого*<sup>1</sup> *положительно определенной матрицы*, а основанный на ней метод решения линейных систем – *методом Холецкого* или *методом квадратного корня*.

<sup>1</sup>Андре-Луи ХОЛЕЦКИЙ (1875-1918) – французский военный геодезист. Изобретенный им алгоритм широко применялся при решении задач геодезии, но опубликован был лишь после смерти автора, в 1924 г.