S©MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru

Том 10, стр. 504–516 (2013) УДК 519.612.2
MSC 65F05, 65G50

ОБ УСТОЙЧИВОСТИ ОДНОГО АЛГОРИТМА ВСТРЕЧНОЙ ПРОГОНКИ

А.Н. МАЛЫШЕВ

Abstract. A two-sided sweep algorithm is proposed for numerical solution of systems of linear equations with nonsingular tridiagonal $n \times n$-matrices, whose arithmetical cost is about $18n$ operations. We prove the componentwise backward stability of the algorithm.

Keywords: tridiagonal matrix, two-sided sweep algorithm, componentwise backward error.

1. Введение

Рассматривается система линейных уравнений $Tx = f$ с трёхдиагональной матрицей $T \in \mathbb{R}^{n\times n}$ и правой частью $f \in \mathbb{R}^n$. Диагональные элементы T задаются компонентами вектора a_i, $i = 1, \ldots, n$, поддиагональные элементы — компонентами b_i, $i = 2, \ldots, n$, а наддиагональные элементы — компонентами c_i, $i = 1, \ldots, n-1$.

Если матрица T симметричная положительно определённая или строго диагонально-доминантная, то система $Tx = f$ эффективно решается методом исключения Гаусса. В общем случае обычно применяется метод исключения Гаусса с выбором ведущих элементов по столбцам. В русскоязычной литературе по численным методам метод исключения Гаусса в применении к линейным системам с трёхдиагональными матрицами, как правило, называется прогонкой.

Malyshev, Alexander N., On the stability of a two-sided sweep algorithm.

Поступила 22 мая 2013 г., опубликована 2 августа 2013 г.

Метод исключения Гаусса с выбором ведущих элементов для систем с трёхдиагональными матрицами является обратно устойчивым к ошибкам округления по норме. В самом деле, приближённое решение $\tilde{x}$, вычисленное в арифметике машинных чисел с плавающей точкой, является точным решением возмущённой системы линейных уравнений $(T+E)\tilde{x}=f$, так что

$$\|E\|_\infty = O(\epsilon_{\text{machine}})\|T\|_\infty,$$

см., например, [5]. Напомним, что $1+\epsilon_{\text{machine}}$ — наименьшее из машинных чисел больших 1. Оказывается, что последняя компонента $\tilde{x}_n$ удовлетворяет более точным оценкам. А именно, $\tilde{x}_n$ равна последней компоненте точного решения возмущённой системы $(\hat{T}_n+E_n)x=f$, так что справедливы оценки

$$|(\hat{T}_n)_{ij}-T_{ij}| = O(\epsilon_{\text{machine}})|T_{ij}|\ \forall ij \quad \text{и} \quad \|E_n\|_\infty \le O(r_{\min})\left(\|T\|_\infty+1\right), \tag{1}$$

где $r_{\min}$ — наименьшее положительное машинное число с нормализованной мантиссой.

Существуют модификации алгоритма прогонки, называемые встречной прогонкой, для которых оценки вида (1) могут иметь место для любой компоненты $\tilde{x}_k$, $k=1,\dots,n$. Более точно, $\tilde{x}_k$ равна последней компоненте точного решения возмущённой системы $(\hat{T}_k+E_k)x=f$, так что справедливы оценки

$$|(\hat{T}_k)_{ij}-T_{ij}| = O(\epsilon_{\text{machine}})|T_{ij}|\ \forall ij \quad \text{и} \quad \|E_k\|_\infty \le O(r_{\min})\left(\|T\|_\infty+1\right). \tag{2}$$

В настоящей работе описывается один из вариантов встречной прогонки и приводится строгий анализ его устойчивости к ошибкам округления с оценками типа (2). Предыдущие публикации по теме включают работы [1], [6] и [3], где рассматриваются другие варианты встречных прогонок. Новым в настоящей работе являются сам вариант встречной прогонки и обратный анализ ошибок округления при его выполнении в арифметике машинных чисел.

Заметим, что альтернативой встречной прогонке является итерационное уточнение на основе гауссова исключения с выбором ведущего элемента по столбцу. Однако эффективность такого подхода доказана только при существенных ограничениях на матрицу системы линейных уравнений и её решение. Подробные детали об итерационном уточнении можно найти в [5].

2. Базовый алгоритм встречной прогонки

В предположении, что для всех $i=1,\dots,n-1$ ведущие подматрицы $T(1\colon i,1\colon i)$ невырождены, трёхдиагональная матрица T допускает LU-разложение вида $T=LU$, где

$$L=\begin{bmatrix} 1 & & & \\ l_2 & 1 & & \\ & \ddots & \ddots & \\ & & l_n & 1 \end{bmatrix}, \quad U=\begin{bmatrix} d_1 & u_1 & & \\ & d_2 & u_2 & \\ & & \ddots & \ddots \\ & & & d_n \end{bmatrix}$$

— двухдиагональные матрицы, однозначно определяемые формулами

$$\begin{aligned} &u_i = c_i, && i=1,\dots,n-1,\\ &d_1=a_1,\quad l_i=b_i/d_{i-1},\quad d_i=a_i-l_iu_{i-1}, && i=2,\dots,n. \end{aligned}$$

Подобным образом, если ведущие подматрицы $T(i\colon n, i\colon n)$ невырождены при $i = 2, \ldots, n$, то матрица T допускает UL-разложение $T = \underline{U}\,\underline{L}$, где

$$\underline{U} = \begin{bmatrix} 1 & \underline{u}_1 & & \\ & 1 & \underline{u}_2 & \\ & & \ddots & \ddots \\ & & & 1 \end{bmatrix}, \quad \underline{L} = \begin{bmatrix} \underline{d}_1 & & & \\ \underline{l}_2 & \underline{d}_2 & & \\ & \ddots & \ddots & \\ & & \underline{l}_n & \underline{d}_n \end{bmatrix}$$

— двухдиагональные матрицы, однозначно определяемые формулами

$$\begin{aligned} &\underline{l}_i = b_i, && i = n, n-1, \ldots, 2, \\ &\underline{d}_n = a_n, \quad \underline{u}_i = c_i/\underline{d}_{i+1}, \quad \underline{d}_i = a_i - \underline{u}_i\underline{l}_{i+1}, && i = n-1, n-2, \ldots, 1. \end{aligned}$$

Базовый алгоритм встречной прогонки основан на комбинации разложений LU и UL в виде «перекрученного» в позиции k разложения $T = S_k Z_k$, где

$$S_k = \begin{bmatrix} 1 & & & & & & & \\ l_2 & \ddots & & & & & & \\ & \ddots & 1 & & & & & \\ & & l_k & 1 & \underline{u}_k & & & \\ & & & & 1 & \ddots & & \\ & & & & & \ddots & \underline{u}_{n-1} & \\ & & & & & & 1 & \end{bmatrix},$$

$$Z_k = \begin{bmatrix} d_1 & u_1 & & & & & & \\ & \ddots & \ddots & & & & & \\ & & d_{k-1} & u_{k-1} & & & & \\ & & & D_k & & & & \\ & & & \underline{l}_{k+1} & \underline{d}_{k+1} & & & \\ & & & & \ddots & \ddots & & \\ & & & & & \underline{l}_n & \underline{d}_n & \end{bmatrix},$$

откуда компоненты решения системы уравнений $Tx = f$ вычисляются по формулам $x_k = \frac{(S_k^{-1} f)_k}{D_k}$, $k = 1, \ldots, n$.

Невырожденность ведущих подматриц гарантируется, когда T имеет строгое диагональное преобладание или является симметричной положительно определённой. В других ситуациях лучше использовать LU или UL-разложения с выбором ведущих элементов.

3. Встречная прогонка с выбором ведущих элементов

Один шаг гауссова исключения сверху вниз с выбором ведущего элемента, определяющего LU-разложение, состоит в преобразовании линейных уравнений

$$\begin{bmatrix} d_{i-1} & u_{i-1} & 0 \\ b_i & a_i & c_i \end{bmatrix} \begin{bmatrix} x_{i-1} \\ x_i \\ x_{i+1} \end{bmatrix} = \begin{bmatrix} g_{i-1} \\ f_i \end{bmatrix}, \quad i = 2, \ldots, n,$$

начиная с $d_1 = a_1$, $u_1 = c_1$, $g_1 = f_1$. Если $|b_i| \le |d_{i-1}|$, то первое уравнение остаётся без изменений, а второе заменяется на $d_i x_i + u_i x_{i+1} = g_i$ по формулам

$$l_i = b_i/d_{i-1}, \quad d_i = a_i - l_i u_{i-1}, \quad u_i = c_i, \quad g_i = f_i - l_i g_{i-1}.$$

Если $|b_i| > |d_{i-1}|$, то первое уравнение заменяется на второе, а второе заменяется на $d_i x_i + u_i x_{i+1} = g_i$ по формулам

$$l_i = d_{i-1}/b_i, \quad d_i = u_{i-1} - l_i a_i, \quad u_i = -l_i c_i, \quad g_i = g_{i-1} - l_i f_i.$$

Гауссово исключение снизу вверх с выбором ведущих элементов, определяющее UL-разложение, состоит в преобразовании линейных уравнений

$$\begin{bmatrix} b_i & a_i & c_i \\ 0 & \underline{l}_{i+1} & \underline{d}_{i+1} \end{bmatrix} \begin{bmatrix} x_{i-1} \\ x_i \\ x_{i+1} \end{bmatrix} = \begin{bmatrix} f_i \\ \underline{g}_{i+1} \end{bmatrix}, \quad i = n-1, n-2, \ldots, 1,$$

начиная с $\underline{d}_n = a_n$, $\underline{l}_n = b_n$, $\underline{g}_n = f_n$. Если $|c_i| \le |\underline{d}_{i+1}|$, то второе уравнение остаётся без изменений, а первое заменяется на $\underline{l}_i x_{i-1} + \underline{d}_i x_i = \underline{g}_i$ по формулам

$$\underline{u}_i = c_i/\underline{d}_{i+1}, \quad \underline{d}_i = a_i - \underline{u}_i \underline{l}_{i+1}, \quad \underline{l}_i = b_i, \quad \underline{g}_i = f_i - \underline{u}_i \underline{g}_{i+1}.$$

Если $|c_i| > |\underline{d}_{i+1}|$, то второе уравнение заменяется на первое, а первое заменяется на $\underline{l}_i x_{i-1} + \underline{d}_i x_i = \underline{g}_i$ по формулам

$$\underline{u}_i = \underline{d}_{i+1}/c_i, \quad \underline{d}_i = \underline{l}_{i+1} - \underline{u}_i a_i, \quad \underline{l}_i = -\underline{u}_i b_i, \quad \underline{g}_i = \underline{g}_{i+1} - \underline{u}_i f_i.$$

При $k = 1, \ldots, n-1$ встречные ходы гауссовых исключений дают системы линейных уравнений

$$\begin{bmatrix} d_k & u_k \\ \underline{l}_{k+1} & \underline{d}_{k+1} \end{bmatrix} \begin{bmatrix} x_k \\ x_{k+1} \end{bmatrix} = \begin{bmatrix} g_k \\ \underline{g}_{k+1} \end{bmatrix}, \tag{3}$$

а при $k = n$ используется лишь одно уравнение $d_n x_n = g_n$, откуда $x_n = g_n/d_n$. Неизвестное x_k вычисляется из системы (3) также исключением с выбором ведущего элемента. А именно, если $|u_k| \le |\underline{d}_{k+1}|$, то

$$\underline{\underline{u}}_k = u_k/\underline{d}_{k+1}, \quad D_k = d_k - \underline{\underline{u}}_k \underline{l}_{k+1}, \quad \underline{\underline{g}}_k = g_k - \underline{\underline{u}}_k \underline{g}_{k+1}, \quad x_k = \underline{\underline{g}}_k/D_k.$$

Если же $|u_k| > |\underline{d}_{k+1}|$, то

$$\underline{\underline{u}}_k = \underline{d}_{k+1}/u_k, \quad D_k = \underline{l}_{k+1} - \underline{\underline{u}}_k d_k, \quad \underline{\underline{g}}_k = \underline{g}_{k+1} - \underline{\underline{u}}_k g_k, \quad x_k = \underline{\underline{g}}_k/D_k.$$

Алгоритм 1 [Встречная прогонка с выбором ведущих элементов]

вход: элементы a_i, b_i, c_i и f_i для $i = 1, 2, \ldots, n$
выход: вектор $x = T^{-1}f$
  $d_1 = a_1, \quad g_1 = f_1$
  для $i = 2, 3, \ldots, n$
    если $|b_i| \le |d_{i-1}|$ **, то**
      $l_i = b_i/d_{i-1}, \quad d_i = a_i - l_i u_{i-1}, \quad u_i = c_i, \quad g_i = f_i - l_i g_{i-1}$
    иначе
      $l_i = d_{i-1}/b_i, \quad d_i = u_{i-1} - l_i a_i, \quad u_i = -l_i c_i, \quad g_i = g_{i-1} - l_i f_i$
    конец условного оператора
  конец цикла
  $\underline{d}_n = a_n, \quad \underline{g}_n = f_n$
  для $i = n-1, n-2, \ldots, 1$
    если $|c_i| \le |\underline{d}_{i+1}|$ **, то**
      $\underline{u}_i = c_i/\underline{d}_{i+1}, \quad \underline{d}_i = a_i - \underline{u}_i \underline{l}_{i+1}, \quad \underline{l}_i = b_i, \quad \underline{g}_i = f_i - \underline{u}_i \underline{g}_{i+1}$
    иначе
      $\underline{u}_i = \underline{d}_{i+1}/c_i, \quad \underline{d}_i = \underline{l}_{i+1} - \underline{u}_i a_i, \quad \underline{l}_i = -\underline{u}_i b_i, \quad \underline{g}_i = \underline{g}_{i+1} - \underline{u}_i f_i$
    конец условного оператора
  конец цикла
  $x_1 = \underline{g}_1/\underline{d}_1, \quad x_n = g_n/d_n$
  для $i = 2, 3, \ldots, n-1$
    если $|u_i| \le |\underline{d}_{i+1}|$ **, то**
      $\underline{\underline{u}}_i = u_i/\underline{d}_{i+1}, \quad x_i = (g_i - \underline{\underline{u}}_i \underline{g}_{i+1})/(d_i - \underline{\underline{u}}_i \underline{l}_{i+1})$
    иначе
      $\underline{\underline{u}}_i = \underline{d}_{i+1}/u_i, \quad x_i = (\underline{g}_{i+1} - \underline{\underline{u}}_i g_i)/(\underline{l}_{i+1} - \underline{\underline{u}}_i d_i)$
    конец условного оператора
  конец цикла

4. Некоторые свойства Алгоритма 1

Справедливы очевидные неравенства $|l_i| \le 1$, $|\underline{u}_i| \le 1$ и $|\underline{\underline{u}}_k| \le 1$, откуда следуют оценки $|u_i| \le |c_i|$, $|\underline{l}_i| \le |b_i|$, $|d_i| \le |a_i| + |c_{i-1}|$, $|\underline{d}_i| \le |a_i| + |b_{i+1}|$ и $|D_k| \le |a_k| + |c_{k-1}| + |b_{k+1}|$. В итоге, $|d_i| \le \|T\|_1$, $|\underline{d}_i| \le \|T\|_1$, $|D_k| \le \|T\|_1$.

По индукции легко доказывается, что $g_k = \sum_{i=1}^{k} \left[g_i \prod_{j \le k} (-l_j) \right]$, где каждое из k произведений содержит подходящие множители $-l_j$. Аналогично, $\underline{g}_{k+1} = \sum_{i=k+1}^{n} \left[g_i \prod_{j>k} (-\underline{u}_j) \right]$. Следовательно, имеет место представление

$$\underline{\underline{g}}_k = \sum_{i=1}^{k} \left[g_i \prod_{j \le k} (-l_j) \right] - \underline{\underline{u}}_k \sum_{i=k+1}^{n} \left[g_i \prod_{j>k} (-\underline{u}_j) \right]$$

или

$$\underline{\underline{g}}_k = -\underline{\underline{u}}_k \sum_{i=1}^{k} \left[g_i \prod_{j \le k} (-l_j) \right] + \sum_{i=k+1}^{n} \left[g_i \prod_{j>k} (-\underline{u}_j) \right].$$

В итоге, $|\underline{\underline{g}}_k| \le \|f\|_1$.

Для вычисления компоненты решения x_k Алгоритм 1 использует $k-1$ исключений Гаусса сверху вниз с матрицами преобразований $L_i \in \mathbb{R}^{n \times n}$ и $n-k$

исключений Гаусса снизу вверх с матрицами преобразований $\underline{L}_i \in \mathbb{R}^{n\times n}$. В результате, матрица $\underline{L}_k^{-1}\dots\underline{L}_{n-1}^{-1}L_k^{-1}\dots L_2^{-1}T$ имеет k-ую строку вида

$$[0,\dots,0,D_k,0,\dots,0].$$

Умножение на L_i^{-1} слева преобразует строки $i-1$ и i, так что они умножаются слева на

$$\begin{pmatrix}1 & 0\\ -l_i & 1\end{pmatrix} \text{ или } \begin{pmatrix}0 & 1\\ 1 & -l_i\end{pmatrix}$$

в зависимости от справедливости неравенства $|d_{i-1}| \le |b_i|$. Следовательно, матрица L_i является блочнодиагональной с одним 2×2-блоком

$$\begin{pmatrix}1 & 0\\ l_i & 1\end{pmatrix} \text{ или } \begin{pmatrix}l_i & 1\\ 1 & 0\end{pmatrix} \tag{4}$$

на пресечении строк и столбцов $i-1$ и i. Все остальные диагональные блоки имеют размер 1×1 и равны 1. Аналогично, матрица $\underline{L}_i$ является блочнодиагональной с одним 2×2-блоком

$$\begin{pmatrix}1 & \underline{u}_i\\ 0 & 1\end{pmatrix} \text{ или } \begin{pmatrix}0 & 1\\ 1 & \underline{u}_i\end{pmatrix} \tag{5}$$

на пресечении строк и столбцов i и $i+1$, а все остальные диагональные блоки имеют размер 1×1 и равны 1.

5. Обратный анализ ошибок округления

Чтобы облегчить понимание доказательств, вначале рассмотрим вычисления в арифметике вещественных машинных чисел с плавающей точкой, в которой отсутствует ситуация потери значимости «underflow». Анализ влияния «underflow» на результаты вычислений будет добавлен позже.

Напомним, что операция округления произвольного вещественного числа r, не являющегося машинным, до одного из двух соседних машинных чисел, обозначаемая через $fl(r)$, подчиняется правилу

$$fl(r) = r(1+\epsilon_1) = r/(1+\epsilon_2), \tag{6}$$

где $|\epsilon_1| \le \epsilon_{\text{rel}}$ и $|\epsilon_2| \le \epsilon_{\text{rel}}$, а $\epsilon_{\text{rel}} = \epsilon_{\text{machine}}$. Если проводится округление до ближайшего машинного числа, то $\epsilon_{\text{rel}} = \epsilon_{\text{machine}}/2$. Операции op сложения, вычитания, умножения и деления машинных чисел f_1 и f_2 определяются как округление точного результата операции $fl(f_1\ op\ f_2)$. Одно из приличных описаний машинной арифметики содержится в монографии [5].

В дальнейшем для удобства будем помечать вычисленные в машинной арифметике значения тильдой сверху.

Рассмотрим исключение Гаусса с матрицей преобразования L_i^{-1}. Предположим, что $|\tilde{d}_{i-1}| \ge |b_i|$. В соответствии с правилом (6)

$$\tilde{l}_i = (1+\epsilon_{i,1})\frac{b_i}{\tilde{d}_{i-1}}, \quad \tilde{d}_i = (1+\epsilon_{i,2})\left[(1+\epsilon_{i,3})a_i - (1+\epsilon_{i,4})\tilde{l}_i\tilde{u}_{i-1}\right],$$
$$\tilde{u}_i = (1+\epsilon_{i,5})c_i,$$

где $|\epsilon_{i,j}| \le \epsilon_{\rm rel}$. Величины $\epsilon_{i,3} = \epsilon_{i,5} = 0$ введены только для удобства в последующих рассуждениях. При $|\tilde d_{i-1}| < |b_i|$ имеют место равенства

$$\tilde l_i = \frac{\tilde d_{i-1}}{(1+\epsilon_{i,1})b_i}, \quad \tilde d_i = (1+\epsilon_{i,2})\left[(1+\epsilon_{i,4})\tilde u_{i-1} - (1+\epsilon_{i,3})\tilde l_i a_i\right],$$
$$\tilde u_i = -(1+\epsilon_{i,5})\tilde l_i c_i,$$

где $\epsilon_{i,4} = 0$, $|\epsilon_{i,j}| \le \epsilon_{\rm rel}$. В обозначениях $\hat a_1 = a_1$,

$$\hat b_i = (1+\epsilon_{i,1})b_i, \quad i = 2,\dots,k, \tag{7}$$
$$\hat a_i = (1+\epsilon_{i,2})(1+\epsilon_{i,3})a_i, \quad i = 2,\dots,k, \tag{8}$$
$$\hat c_i = (1+\epsilon_{i+1,2})(1+\epsilon_{i+1,4})(1+\epsilon_{i,5})c_i, \quad i = 1,\dots,k-1, \tag{9}$$
$$\hat u_{i-1} = (1+\epsilon_{i,2})(1+\epsilon_{i,4})\tilde u_{i-1}, \quad i = 2,\dots,k, \tag{10}$$

получим тождества

$$\begin{bmatrix} 1 & 0 \\ -\tilde l_i & 1 \end{bmatrix} \begin{bmatrix} \tilde d_{i-1} & \hat u_{i-1} & 0 \\ \hat b_i & \hat a_i & \hat c_i \end{bmatrix} = \begin{bmatrix} \tilde d_{i-1} & \hat u_{i-1} & 0 \\ 0 & \tilde d_i & \hat u_i \end{bmatrix} \quad \text{при} \quad |\tilde d_{i-1}| \ge |b_i| \tag{11}$$

и

$$\begin{bmatrix} 0 & 1 \\ 1 & -\tilde l_i \end{bmatrix} \begin{bmatrix} \tilde d_{i-1} & \hat u_{i-1} & 0 \\ \hat b_i & \hat a_i & \hat c_i \end{bmatrix} = \begin{bmatrix} \hat b_i & \hat a_i & \hat c_i \\ 0 & \tilde d_i & \hat u_i \end{bmatrix} \quad \text{при} \quad |\tilde d_{i-1}| < |b_i|. \tag{12}$$

Заметим, что элементы $\tilde l_i$ и $\tilde d_i$ не заменились на элементы с шапочкой. Элементы $\hat c_k$ и $\hat u_k$ будут определены позднее после склейки встречных ходов прогонки.

Ход прогонки снизу вверх удовлетворяет аналогичным соотношениям. А именно, элементы $\underline{\tilde u}_i$ и $\underline{\tilde d}_i$ не заменяются на элементы с шапочкой, а

$$\underline{\hat c}_i = (1+\underline{\epsilon}_{i,1})c_i, \quad i = n-1, n-2, \dots, k+1, \tag{13}$$
$$\underline{\hat a}_i = (1+\underline{\epsilon}_{i,2})(1+\underline{\epsilon}_{i,3})a_i, \quad i = n-1,\dots,k+1, \quad \underline{\hat a}_n = a_n, \tag{14}$$
$$\underline{\hat b}_i = (1+\underline{\epsilon}_{i-1,2})(1+\underline{\epsilon}_{i-1,4})(1+\underline{\epsilon}_{i,5})b_i, \quad i = n, n-1, \dots, k+2, \tag{15}$$
$$\underline{\hat l}_{i+1} = (1+\underline{\epsilon}_{i,2})(1+\underline{\epsilon}_{i,4})\underline{\tilde l}_{i+1}, \quad i = n-1,\dots,k+1, \tag{16}$$
(17)

где $|\underline{\epsilon}_{i,j}| \le \epsilon_{\rm rel}$. Элементы $\underline{\hat b}_{k+1}$ и $\underline{\hat l}_{k+1}$ также будут определены после склейки.

При склейке встречных ходов прогонки на строках k и $k+1$ возникает система линейных уравнений

$$\begin{bmatrix} \tilde d_k & \tilde u_k \\ \tilde l_{k+1} & \underline{\tilde d}_{k+1} \end{bmatrix} \begin{bmatrix} x_k \\ x_{k+1} \end{bmatrix} = \begin{bmatrix} \tilde g_k \\ \underline{\tilde g}_{k+1} \end{bmatrix},$$

с помощью которой вычисляется x_k. При $|\underline{\tilde d}_{k+1}| \ge |\tilde u_k|$ будет

$$\underline{\tilde u}_k = (1+\underline{\epsilon}_{k,1})\frac{\tilde u_k}{\underline{\tilde d}_{k+1}}, \quad \tilde D_k = (1+\underline{\epsilon}_{k,2})^{-1}\left[\tilde d_k - (1+\underline{\epsilon}_{k,3})\underline{\tilde u}_k \underline{\tilde l}_{k+1}\right],$$

где $|\underline{\epsilon}_{k,j}| \le \epsilon_{\rm rel}$. Обозначив $\hat u_k = (1+\underline{\epsilon}_{k,1})\tilde u_k$, $\underline{\hat l}_{k+1} = (1+\underline{\epsilon}_{k,3})\underline{\tilde l}_{k+1}$, $\underline{\epsilon}_{k,4} = 0$, $\hat D_k = (1+\underline{\epsilon}_{k,4})(1+\underline{\epsilon}_{k,2})\tilde D_k$, получим равенство

$$\begin{bmatrix} 1 & -\underline{\tilde u}_k \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \tilde d_k & \hat u_k \\ \underline{\hat l}_{k+1} & \underline{\tilde d}_{k+1} \end{bmatrix} = \begin{bmatrix} \hat D_k & 0 \\ \underline{\hat l}_{k+1} & \underline{\tilde d}_{k+1} \end{bmatrix}. \tag{18}$$

При $|\tilde{\underline{d}}_{k+1}| < |\tilde{u}_k|$ имеем

$$\tilde{\underline{\underline{u}}}_k = \frac{\tilde{\underline{d}}_{k+1}}{(1+\underline{\epsilon}_{k,1})\tilde{u}_k}, \quad \tilde{D}_k = (1+\underline{\epsilon}_{k,2})^{-1}\left[\tilde{\underline{l}}_{k+1} - (1+\underline{\epsilon}_{k,3})^{-1}\tilde{\underline{\underline{u}}}_k\tilde{d}_k\right].$$

Обозначив $\hat{u}_k = (1+\underline{\epsilon}_{k,1})\tilde{u}_k$, $\hat{\underline{l}}_{k+1} = (1+\underline{\epsilon}_{k,3})\tilde{\underline{l}}_{k+1}$, $\underline{\epsilon}_{k,4} = \underline{\epsilon}_{k,3}$, $\hat{D}_k = (1+\underline{\epsilon}_{k,4})(1+\underline{\epsilon}_{k,2})\tilde{D}_k$, получим равенство

$$\begin{bmatrix} 1 & 0 \\ -\tilde{\underline{\underline{u}}}_k & 1 \end{bmatrix} \begin{bmatrix} \tilde{d}_k & \hat{u}_k \\ \hat{\underline{l}}_{k+1} & \tilde{\underline{d}}_{k+1} \end{bmatrix} = \begin{bmatrix} \tilde{d}_k & \hat{u}_k \\ \hat{D}_k & 0 \end{bmatrix}. \tag{19}$$

В обоих случаях, $\hat{c}_k = (1+\underline{\epsilon}_{k,1})(1+\epsilon_{k,5})c_k$ и $\hat{b}_{k+1} = (1+\underline{\epsilon}_{k,3})(1+\underline{\epsilon}_{k,5})b_{k+1}$.

Подведем промежуточный итог. Для вычисления x_k потребуются, кроме компонент правой части f_k, вычисленные значения $\tilde{l}_i$, $i = 2, \ldots, k$, значения $\tilde{\underline{u}}_i$, $i = k+1, \ldots, n-1$, и величины $\tilde{\underline{\underline{u}}}_k$ и $\tilde{D}_k$. Выше было доказано, что существует такая трёхдиагональная матрица $\hat{T}_k$, что матрица

$$\tilde{\underline{L}}_k^{-1} \ldots \tilde{\underline{L}}_{n-1}^{-1} \tilde{L}_k^{-1} \ldots \tilde{L}_2^{-1} \hat{T}_k \tag{20}$$

имеет k-ую строку вида

$$[0, \ldots, 0, \hat{D}_k, 0, \ldots, 0]. \tag{21}$$

Тильда в обозначениях $\tilde{L}_i$, $i = 2, \ldots, k$, и $\tilde{\underline{L}}_i$, $i = k+1, \ldots, n-1$, означает, что эти матрицы содержат $\tilde{l}_i$ и $\tilde{\underline{u}}_i$, а матрица $\tilde{\underline{L}}_k$ содержат $\tilde{\underline{\underline{u}}}_k$.

Трёхдиагональная матрица $\hat{T}_k$ из (20) имеет диагональные элементы вида $(1+\alpha_{i,1})(1+\alpha_{i,2})a_i$, поддиагональные элементы $(1+\beta_{i,1})(1+\beta_{i,2})(1+\beta_{i,3})b_i$ и наддиагональные элементы $(1+\gamma_{i,1})(1+\gamma_{i,2})(1+\gamma_{i,2})c_i$, где $|\alpha_{i,j}| \le \epsilon_{\mathrm{rel}}$, $|\beta_{i,j}| \le \epsilon_{\mathrm{rel}}$, $|\gamma_{i,j}| \le \epsilon_{\mathrm{rel}}$. Таким образом, матрица $\hat{T}_k$ является покомпонентным возмущением матрицы T, удовлетворяющим покомпонентной оценке

$$|\hat{T}_k - T| \le [(1+\epsilon_{\mathrm{rel}})^3 - 1]|T| \approx 3\epsilon_{\mathrm{rel}}|T|. \tag{22}$$

Разберемся с преобразованиями правой части вида $\tilde{\underline{L}}_k^{-1} \ldots \tilde{\underline{L}}_{n-1}^{-1} \tilde{L}_k^{-1} \ldots \tilde{L}_2^{-1} f$. По индукции выводим, что в арифметике машинных чисел

$$\tilde{g}_k = \sum_{i=1}^{k} f_i \left[\prod_j (-\tilde{l}_j) \prod_{j=1}^{2(k-1)} (1+\phi_{i,j}) \right],$$

где $|\phi_{i,j}| \le \epsilon_{\mathrm{rel}}$. Аналогично,

$$\tilde{\underline{g}}_{k+1} = \sum_{i=k+1}^{n} f_i \left[\prod_j (-\tilde{\underline{u}}_j) \prod_{j=1}^{2(n-k-1)} (1+\phi_{i,j}) \right].$$

После склейки встречных ходов прогонки получим представление

$$\tilde{\underline{\underline{g}}}_k = \sum_{i=1}^{n} \left(\prod_j (-\tilde{l}_j) \right) \left(\prod_j (-\tilde{\underline{u}}_j) \right) \left(\prod_j (-\tilde{\underline{\underline{u}}}_j) \right) f_i \prod_{j=1}^{2n-2} (1+\phi_{i,j}).$$

Тем самым доказано, что $\tilde{\underline{\underline{g}}}_k$ равняется k-ой компоненте вектора

$$\tilde{\underline{L}}_k^{-1} \ldots \tilde{\underline{L}}_{n-1}^{-1} \tilde{L}_k^{-1} \ldots \tilde{L}_2^{-1} \check{f}_k, \tag{23}$$

где $(\check{f}_k)_i = f_i \prod_{j=1}^{2n-2}(1+\phi_{i,j})$, $|\phi_{i,j}| \le \epsilon_{\mathrm{rel}}$.

Вследствие представления

$$fl(\frac{\tilde{\underline{\underline{g}}}_k}{\tilde{D}_k}) = (1+\underline{\epsilon}_{k,6})\frac{\tilde{\underline{\underline{g}}}_k}{\tilde{D}_k} = (1+\underline{\epsilon}_{k,6})(1+\underline{\epsilon}_{k,4})(1+\underline{\epsilon}_{k,2})\frac{\tilde{\underline{\underline{g}}}_k}{\hat{D}_k}$$

вычисленное значение x_k — это k-ая компонента точного решения возмущённой системы $\hat{T}_k x = \hat{f}_k$, где $|\hat{T}_k - T| \le [(1+\epsilon_{\rm rel})^3 - 1]|T| \approx 3\epsilon_{\rm rel}|T|$ и $|\hat{f}_k - f| \le [(1+\epsilon_{\rm rel})^{2n+1} - 1]|f| \approx (2n+1)\epsilon_{\rm rel}|f|$.

Сформулированный в предыдущем абзаце результат также имеет место в арифметике машинных чисел компьютеров CRAY. Доказательство этого факта почти дословно повторяет доказательства настоящего параграфа.

6. Учёт эффекта потери точности «underflow»

Наша модель ситуации «underflow» формулируется следующим образом. Если точный результат арифметической операции по модулю меньше минимального положительного машинного числа, то машинный результат операции будет равен нулю. Заметим, что стандарт IEEE допускает модификацию этой модели с помощью денормализованных (или субнормальных) чисел. Будем использовать следующее представление для арифметической операции op над машинными числами f_1 и f_2, см. [4]:

$$\widetilde{f_1\ op\ f_2} = (1+\epsilon_1)(f_1\ op\ f_2) + \epsilon_2, \tag{24}$$

где $|\epsilon_1| \le \epsilon_{\rm rel}$ и $|\epsilon_2| \le \epsilon_{\rm abs}$, причём либо $\epsilon_1 = 0$, либо $\epsilon_2 = 0$. Константа $\epsilon_{\rm abs}$ равна $r_{\min}$, а в арифметике с денормализованными числами она равна $r_{\min} \times \epsilon_{\rm machine}$.

Тождества (11) и (12) сейчас заменяются на

$$\begin{bmatrix} 1 & 0 \\ -\tilde{l}_i & 1 \end{bmatrix} \begin{bmatrix} \tilde{d}_{i-1} & \hat{u}_{i-1} & 0 \\ \hat{b}_i & \hat{a}_i & \hat{c}_i \end{bmatrix} = \begin{bmatrix} \tilde{d}_{i-1} & \hat{u}_{i-1} & 0 \\ \delta_{i,1} & \tilde{d}_i + \delta_{i,2} & \hat{u}_i + \delta_{i,3} \end{bmatrix} \text{ при } |\tilde{d}_{i-1}| \ge |b_i| \tag{25}$$

и

$$\begin{bmatrix} 0 & 1 \\ 1 & -\tilde{l}_i \end{bmatrix} \begin{bmatrix} \tilde{d}_{i-1} & \hat{u}_{i-1} & 0 \\ \hat{b}_i & \hat{a}_i & \hat{c}_i \end{bmatrix} = \begin{bmatrix} \hat{b}_i & \hat{a}_i & \hat{c}_i \\ \delta_{i,1} & \tilde{d}_i + \delta_{i,2} & \hat{u}_i + \delta_{i,3} \end{bmatrix} \text{ при } |\tilde{d}_{i-1}| < |b_i|. \tag{26}$$

При этом справедливы оценки $|\delta_{i,1}| \le \max(|\tilde{d}_{i-1}|, |b_i|)\epsilon_{\rm abs} \le \epsilon_{\rm abs}(1+\epsilon_{\rm rel})\|T\|_1$, $|\delta_{i,2}| \le \epsilon_{\rm abs}$ и $|\delta_{i,3}| \le \epsilon_{\rm abs}$.

Через Δ_i обозначим $n \times n$-матрицу, i-ая строка которой равна строке

$$[0, \dots, 0, \delta_{i,1}, \delta_{i,2}, \delta_{i,3}, 0, \dots, 0]$$

с $\delta_{i,2}$ в i-ой позиции, а все остальные строки равны нулевым. Во время хода прогонки снизу вверх возникают аналогичные матрицы $\underline{\Delta}_i$, а на склейке — матрица $\underline{\underline{\Delta}}_k$.

При $|\tilde{\underline{d}}_{k+1}| \ge |\tilde{u}_k|$ будет

$$\tilde{\underline{\underline{u}}}_k = \frac{\hat{u}_k}{\tilde{\underline{d}}_{k+1}} + \mu_{k,1}, \quad \tilde{D}_k = (1+\underline{\epsilon}_{k,2})^{-1}\left[\tilde{d}_k - \tilde{\underline{\underline{u}}}_k \hat{\underline{l}}_{k+1}\right] + \mu_{k,2},$$

где $|\mu_{k,j}| \le \epsilon_{\rm abs}$. Обозначив $\underline{\underline{\mu}}_{k,1} = \mu_{k,1}\tilde{\underline{d}}_{k+1}$, $\hat{D}_k = (1+\underline{\epsilon}_{k,2})\tilde{D}_k$, $\underline{\underline{\delta}}_{k,2} = (1+\underline{\epsilon}_{k,2})\mu_{k,2}$, получим равенство

$$\begin{bmatrix} 1 & -\tilde{\underline{\underline{u}}}_k \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \tilde{d}_k & \hat{u}_k \\ \hat{\underline{l}}_{k+1} & \tilde{\underline{d}}_{k+1} \end{bmatrix} = \begin{bmatrix} \hat{D}_k + \underline{\underline{\delta}}_{k,2} & \underline{\underline{\delta}}_{k,1} \\ \hat{\underline{l}}_{k+1} & \tilde{\underline{d}}_{k+1} \end{bmatrix}. \tag{27}$$

При $|\tilde{\underline{d}}_{k+1}| < |\tilde{u}_k|$ имеем

$$\tilde{\underline{\underline{u}}}_k = \frac{\tilde{\underline{d}}_{k+1}}{\hat{u}_k} + \mu_{k,1}, \quad \tilde{D}_k = (1+\underline{\epsilon}_{k,2})^{-1}\left[\tilde{\underline{l}}_{k+1} - (1+\underline{\epsilon}_{k,3})^{-1}\tilde{\underline{\underline{u}}}_k\tilde{d}_k\right] + \mu_{k,2}.$$

Обозначив $\underline{\underline{\delta}}_{k,1} = \mu_{k,1}\tilde{u}_k$, $\hat{D}_k = (1+\underline{\epsilon}_{k,3})(1+\underline{\epsilon}_{k,2})\tilde{D}_k$, $\underline{\underline{\delta}}_{k,2} = (1+\underline{\epsilon}_{k,3})(1+\underline{\epsilon}_{k,2})\mu_{k,2}$, получим равенство

$$\begin{bmatrix} 1 & 0 \\ -\tilde{\underline{\underline{u}}}_k & 1 \end{bmatrix}\begin{bmatrix} \tilde{d}_k & \hat{u}_k \\ \hat{l}_{k+1} & \tilde{\underline{d}}_{k+1} \end{bmatrix} = \begin{bmatrix} \tilde{d}_k & \hat{u}_k \\ \hat{D}_k + \underline{\underline{\delta}}_{k,2} & \underline{\underline{\delta}}_{k,1} \end{bmatrix}. \tag{28}$$

В обоих случаях, $|\underline{\underline{\delta}}_{k,1}| \le \max(|\tilde{\underline{d}}_{k+1}|, |\tilde{u}_k|)\epsilon_{\text{abs}} \le \epsilon_{\text{abs}}(1+\epsilon_{\text{rel}})\|T\|_1$, $|\underline{\underline{\delta}}_{k,2}| \le \epsilon_{\text{abs}}(1+\epsilon_{\text{rel}})$.

По окончании Алгоритма 1 получаем матрицу

$$\begin{aligned}\hat{Z}_k = {} & \tilde{\underline{L}}_k^{-1}\ldots\tilde{\underline{L}}_{n-1}^{-1}\tilde{L}_k^{-1}\ldots\tilde{L}_2^{-1}\hat{T}_k + \sum_{i=2}^{k}\tilde{\underline{L}}_k^{-1}\ldots\tilde{\underline{L}}_{n-1}^{-1}\tilde{L}_k^{-1}\ldots\tilde{L}_{i+1}^{-1}\Delta_i \\ & + \sum_{i=n-1}^{k+1}\tilde{\underline{L}}_k^{-1}\ldots\tilde{\underline{L}}_{i-1}^{-1}\underline{\Delta}_i + \underline{\underline{\Delta}}_k,\end{aligned}$$

у которой k-ая строка имеет только один ненулевой элемент $\hat{D}_k$ в позиции k. Эта матрица получается точным преобразованием $\tilde{\underline{L}}_k^{-1}\ldots\tilde{\underline{L}}_{n-1}^{-1}\tilde{L}_k^{-1}\ldots\tilde{L}_2^{-1}$ матрицы

$$\begin{aligned}\hat{T}_k + E_k &= \tilde{L}_2\ldots\tilde{L}_k\tilde{\underline{L}}_{n-1}\ldots\tilde{\underline{L}}_k\hat{Z}_k \\ &= \hat{T}_k + \sum_{i=2}^{k}\tilde{L}_2\ldots\tilde{L}_i\Delta_i + \sum_{i=n-1}^{k+1}\tilde{\underline{L}}_{n-1}\ldots\tilde{\underline{L}}_{i+1}\underline{\Delta}_i \\ &\quad + \tilde{L}_2\ldots\tilde{L}_k\tilde{\underline{L}}_{n-1}\ldots\tilde{\underline{L}}_k\underline{\underline{\Delta}}_k.\end{aligned}$$

Примечательным фактом является то, что матрица $\tilde{L}_2\ldots\tilde{L}_i\Delta_i$ равна матрице Δ_i, сдвинутой вверх циклически на несколько позиций. Аналогично, матрица $\tilde{\underline{L}}_{n-1}\ldots\tilde{\underline{L}}_{i+1}\underline{\Delta}_i$ равна матрице $\underline{\Delta}_i$, сдвинутой вниз циклически на несколько позиций. Наконец, $\tilde{L}_2\ldots\tilde{L}_k\tilde{\underline{L}}_{n-1}\ldots\tilde{\underline{L}}_k\underline{\underline{\Delta}}_k$ равна $\underline{\underline{\Delta}}_k$, сдвинутой вверх или вниз циклически на несколько позиций.

На основании вышеуказанных свойств

$$\|E_k\|_\infty \le \epsilon_{\text{abs}}\left[(1+\epsilon_{\text{rel}})\|T\|_1 + 2\right](n-1). \tag{29}$$

Влияние ошибок округления на вычисление $\tilde{\underline{\underline{g}}}_k$ эквивалентно возмущению компонент правой части вида

$$f_i\prod_{j=1}^{2n-2}(1+\phi_{i,j}) + \psi_i,$$

где $|\phi_{i,j}| \le \epsilon_{\text{rel}}$, $|\psi_i| \le (2n-1)(1+\epsilon_{\text{rel}})^{n-1}\epsilon_{\text{abs}}$. Учёт ошибок округления для операции $fl(\frac{\tilde{\underline{\underline{g}}}_k}{\hat{D}_k})$ модифицирует возмущения компонент правой части к виду

$$f_i\prod_{j=1}^{2n+1}(1+\phi_{i,j}) + \psi_i,$$

где $|\phi_{i,j}| \le \epsilon_{\rm rel}$, $|\psi_i| \le (2n-1)(1+\epsilon_{\rm rel})^{n+2}\epsilon_{\rm abs} + (1+\epsilon_{\rm rel})\|T\|_1\epsilon_{\rm abs}$. Добавка $(1+\epsilon_{\rm rel})\|T\|_1\epsilon_{\rm abs}$ возникает при «underflow» во время выполнения операции $fl(\frac{\tilde{\tilde{g}}_k}{\tilde{D}_k})$.

7. Сводка оценок

Подведем итог рассуждений предыдущих параграфов. Результат вычисления $\tilde{x}_k$ по Алгоритму 1 в арифметике чисел с плавающей точкой и с учётом «underflow» равен k-ой компоненте возмущённой системы уравнений

$$(\hat{T}_k + E_k)x = \hat{f}_k + h_k, \tag{30}$$

где

$$\begin{aligned}
|\hat{T}_k - T| &\le [(1+\epsilon_{\rm rel})^3 - 1]|T| \approx 3\epsilon_{\rm rel}|T|,\\
\|E_k\|_\infty &\le (n-1)\left[(1+\epsilon_{\rm rel})\|T\|_1 + 2\right]\epsilon_{\rm abs} \approx (n-1)\left(\|T\|_1 + 2\right)\epsilon_{\rm abs},\\
|\hat{f}_k - f| &\le [(1+\epsilon_{\rm rel})^{2n+1} - 1]|f| \approx (2n+1)\epsilon_{\rm rel}|f|,\\
\|h_k\|_\infty &\le (2n-1)(1+\epsilon_{\rm rel})^{n+2}\epsilon_{\rm abs} + (1+\epsilon_{\rm rel})\|T\|_1\epsilon_{\rm abs} \approx (2n-1+\|T\|_1)\epsilon_{\rm abs}.
\end{aligned}$$

Заметим, что в процессе вычислений не возникнет ошибок переполнения, если величины $\|T\|_1$, $\|f\|_1$ и $\|x\|_\infty$ не превышают $r_{\max}/2$, где $r_{\max}$ — наибольшее машинное число.

Обозначим $\delta T = \hat{T}_k - T$, $\delta f = \hat{f}_k - f$ и $\tilde{x} = (\hat{T}_k + E_k)^{-1}(\hat{f}_k + h_k)$. Отбросив малые квадратичные члены в тождестве

$$\tilde{x} - x = -(\hat{T}_k + E_k)^{-1}(\delta T + E_k)x + (\hat{T}_k + E_k)^{-1}(\delta f + h_k),$$

приходим к равенству

$$\tilde{x} - x = -T^{-1}\delta T x - T^{-1}E_k x + T^{-1}\delta f + T^{-1}h_k.$$

Справедливы неравенства

$$\begin{aligned}
|T^{-1}\delta T x| &\le |T^{-1}||\delta T||x| \le 3\epsilon_{\rm rel}|T^{-1}||T||x|,\\
|T^{-1}\delta f| &\le (2n+1)\epsilon_{\rm rel}|T^{-1}||f| \le (2n+1)\epsilon_{\rm rel}|T^{-1}||T||x|,\\
\|T^{-1}E_k x\|_\infty &\le \|T^{-1}\|_\infty\|E_k\|_\infty\|x\|_\infty \le (n-1)\epsilon_{\rm abs}\|T^{-1}\|_\infty(\|T\|_1+2)\|x\|_\infty,\\
\|T^{-1}h_k\|_\infty &\le \epsilon_{\rm abs}(2n-1+\|T\|_1)\|T^{-1}\|_\infty.
\end{aligned}$$

Определим диагональную матрицу D с диагональными элементами $D_i = |x_i|$ при $x_i \ne 0$, и произвольными положительными D_i при $x_0 = 0$. В результате,

$$\begin{aligned}
|\tilde{x}_k - x_k| \le{}& \epsilon_{\rm rel}(2n+4)\|D^{-1}|T^{-1}||T|D\|_\infty|x_k| + \\
&\epsilon_{\rm abs}\|T^{-1}\|_\infty\left[(n-1)(2+\|T\|_1)\|x\|_\infty + (2n-1+\|T\|_1)\right],
\end{aligned} \tag{31}$$

Для примера 1 следующего параграфа $\|D^{-1}|T^{-1}||T|D\|_\infty = O(1)$.

8. Примеры

Пример 1. Идея первого примера заимствована из работы [2]. Пример демонстрирует, что метод исключения Гаусса с выбором ведущего элемента по столбцу не является обратно устойчивым в смысле покомпонентно относительных возмущений.

Точное решение системы линейных уравнений

$$\begin{pmatrix} \epsilon & \epsilon^{-2} & 0 \\ \epsilon^2 & 0 & -1 \\ 0 & 1 & -\epsilon^3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

равно $x = \begin{bmatrix} \epsilon^{-1} & \epsilon^4 & \epsilon \end{bmatrix}^T / (1+\epsilon^2)$. Положим $\epsilon = \sqrt{\epsilon_{\text{machine}}}/2$. Тогда в арифметике машинных чисел Алгоритм 1 выдаёт решение $\begin{bmatrix} \epsilon^{-1} & \epsilon^4 & \epsilon \end{bmatrix}^T$, что является машинным округлением точного решения, а метод исключения Гаусса с выбором ведущего элемента по столбцу выдаёт $\tilde{x}_1 = \epsilon^{-1}$, $x_2 = 0$ и $\tilde{x}_3 = \epsilon$. Минимальное покомпонентное возмущение матрицы, соответствующее такому приближённому решению, оценивается с помощью параметра Оэтли-Прагера, см., например, [5],

$$\omega = \max_i \frac{|r_i|}{(|T||\tilde{x}|)_i}, \tag{32}$$

где r_i — i-ая компонента невязки $r = f - T\tilde{x}$, а $(|T||\tilde{x}|)_i$ — i-ая компонента вектора $|T||\tilde{x}|$. Для $\tilde{x} = \begin{bmatrix} \epsilon^{-1} & 0 & \epsilon \end{bmatrix}^T$ имеем $\omega = 1$. Это означает, что если $\tilde{x}$ является точным решением возмущённой системы $(T + \delta T)\tilde{x} = f$, где $|\delta T| \le \theta |T|$, то $\theta \ge \omega = 1$. Заметим, что для решения, полученного Алгоритмом 1, $\omega = \epsilon^2/(1+\epsilon^2)$.

Соотношение

$$M = \begin{pmatrix} \frac{1}{|x_1|} & & \\ & \frac{1}{|x_2|} & \\ & & \frac{1}{|x_3|} \end{pmatrix} |T^{-1}||T| \begin{pmatrix} |x_1| & & \\ & |x_2| & \\ & & |x_3| \end{pmatrix} = \begin{pmatrix} 1 & \frac{2\epsilon^2}{1+\epsilon^2} & \frac{2\epsilon^2}{1+\epsilon^2} \\ \frac{2}{1+\epsilon^2} & 1 & \frac{2}{1+\epsilon^2} \\ \frac{2}{1+\epsilon^2} & \frac{2\epsilon^2}{1+\epsilon^2} & 1 \end{pmatrix}$$

позволяет доказать, что решение $\tilde{x}$, вычисленное Алгоритмом 1, удовлетворяет оценке $|\tilde{x}_k - x_k|/|x_k| \le \omega \|M\|_\infty \approx 5\epsilon^2$.

Пример 2. Второй пример взят из работы [3]. Вычисления проводились с помощью программы Матлаб, для которой $\epsilon_{\text{machine}} \approx 2.22 \cdot 10^{-16}$, $\epsilon_{\text{rel}} \approx 1.11 \cdot 10^{-16}$, а $\epsilon_{\text{abs}} = \epsilon_{\text{machine}} r_{\min} \approx 4.94 \cdot 10^{-324}$.

Точное решение трёхдиагональной системы линейных уравнений $Tx = f$ с главной диагональю $a = [-1, 1, 1, \ldots, 1, 1, -1]$, поддиагональю $b = [-1, -1, \ldots, -1]$, наддиагональю $c = [2, 2, \ldots, 2]$ и правой частью $f = [1, 0, 0, \ldots, 0]$ имеет компоненты $x_i = \frac{(-1)^i}{3}$.

Обратная к матрице T имеет элементы $(T^{-1})_{ij} = -\frac{2^{j-i}}{3}$ при $i \le j$ и $(T^{-1})_{ij} = -\frac{(-1)^{i-j}}{3}$ при $i > j$. Следовательно, $\text{cond}_\infty(T) = \frac{4}{3}(2^n - 1)$, $\text{cond}_{\text{Skeel}}(T) = \| |T^{-1}||T| \|_\infty = 2^n - \frac{5}{3}$.

При $n = 60$ результат Алгоритма 1 — точное решение, округленное до ближайших машинных чисел, то есть $\tilde{x} = fl(x)$. Для этого решения параметр, определённый формулой (32), равен $\omega = \epsilon_{\text{rel}}/4$. Метод исключения Гаусса с выбором ведущего элемента по столбцу выдаёт $\tilde{x}_1 = -11$ и $\tilde{x}_{60} = fl(1/3)$, а $\omega = \frac{3}{4}\epsilon_{\text{rel}}$.

Пример 2, в частности, демонстрирует, что обратная устойчивость в смысле покомпонентных относительных возмущений не всегда достаточна для обеспечения точности решения.

Список литературы

[1] I. Babuška, *Numerical stability in problems of linear algebra*, SIAM Journal on Numerical Analysis, **9**:1 (1972), 53–77. MR0386252

[2] I. Bar-On and M. Leoncini, *Stable solution of tridiagonal systems*, Numerical Algorithms, **18** (1998), 361–388. MR1669938

[3] I. Bar-On and M. Leoncini, *Reliable solution of tridiagonal systems of linear equations*, SIAM Journal on Numerical Analysis, **38**:4 (2001), 1134–1153.

[4] Дж. Деммель, *Вычислительная линейная алгебра. Теория и приложения.* Пер. с англ. —М.: Мир, 2001.

[5] N. Higham, *Accuracy and stability of numerical algorithms*. SIAM, Philadephia, PA, 2002. MR1927606

[6] С. И. Фадеев, *Алгоритм универсальной прогонки*, Методы сплайн-функций, Вычислительные системы, **75** (1979), 68–79. MR0599172

Александр Николаевич Малышев
Институт математики им. С. Л. Соболева СО РАН,
пр. академика Коптюга 4,
630090, Новосибирск, Россия
E-mail address: malyshev@math.nsc.ru

Alexander Malyshev
University of Bergen, Department of Mathematics,
Postbox 7800, 5020 Bergen, Norway