Записки научных
семинаров ПОМИ
Том 334, 2006 г.

А. О. Родников, Б. А. Самокиш

ОЦЕНКА ОШИБКИ ОКРУГЛЕНИЯ В ЗАДАЧЕ ИСКЛЮЧЕНИЯ

Задачей исключения в монографии [1] авторы называют задачу вычисления линейной формы $\mu = (g, x^*) = g_1 x_1^* + \ldots + g_n x_n^*$ от решения x^* линейной системы

$$Ax = f. \tag{1}$$

Эта задача достаточно типична для приложений. Например, в линейных задачах аппроксимации, в проекционных методах нас интересуют не коэффициенты разложения по координатной системе, а, например, значения в точках. Термин "задача исключения" употребляется в работах [2]–[6]. В них рассмотрены различные алгоритмы решения этой задачи и далеко идущих ее обобщений, с особым вниманием к точности результата. В работе [7] рассмотрен случай, когда функционал μ заменяется на оператор $\mathbb{R}^n \to \mathbb{R}^k$, $k < n$. Вводятся числа обусловленности и даны методы их приближенной оценки – нестрогие, но экономичные. Впрочем, случай общей скалярной формы g также рассмотрен: определяется число обусловленности – близкое к нашему, но не тождественное ему.

Разумеется, для решения задачи исключения достаточно решить систему (1) подставить решение в линейную форму. Авторы [1] предлагают модифицированную схему исключения, позволяющую непосредственно вычислить μ, избежав вычисления x^*. Ниже предлагается аналогично модифицировать метод ортогонализации строк. Так как ошибка округления при вычислении x^* имеет порядок числа обусловленности $k(A) = \|A\| \, \|A^{-1}\|$ матрицы A, представляется вероятным, что такие более непосредственные методы позволят уменьшить ошибку округления.

Выясним сначала, какова неустранимая погрешность решения задачи исключения. Будем считать, что $\|A\| \le 1$. Рассмотрим возмущенную систему:

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

С точностью до малых второго порядка

$$x + \Delta x = A^{-1}f + A^{-1}\Delta f - A^{-1}\Delta A A^{-1} f,$$

так что

$$\begin{aligned}(x+\Delta x, g+\Delta g) - (x, g) &= (A^{-1}f, g) + (A^{-1}\Delta f, g) + (A^{-1}f, \Delta g) \\ &\quad - (A^{-1}\Delta A A^{-1} f, g) - (A^{-1}f, g) \\ &= (A^{-1}f, \Delta g) + (\Delta f, A^{T^{-1}} g) - (\Delta A A^{-1} f, A^{T^{-1}} g),\end{aligned}$$

$$\|\Delta\mu\| \le \|A^{-1}f\|\,\|\Delta g\| + \|A^{T^{-1}}g\|\,\|\Delta f\| + \|A^{-1}f\|\,\|A^{T^{-1}}g\|\,\|\Delta A\|.$$

Величины $\|A^{-1}f\|$, $\|A^{T^{-1}}g\|$, по-видимому, и следует считать числами обусловленности задачи исключения. Возможна, следовательно, такая ситуация, когда при плохо обусловленной матрице погрешность решения задачи исключения невелика.

Для решения задачи исключения мы применим метод исключения и метод ортогонализации строк.

Метод исключения

Запишем расширенную матрицу системы и дополним ее строкой коэффициентов вычислямой линейной формы

$$\left(\begin{array}{c|c} A & f \\ \hline -g^T & 0 \end{array}\right).$$

Значения форм, задаваемых строками расширенной матрицы $(a_{i1}, \dots, a_{in}, f_i)$, на векторе $(x_1^*, \dots, x_n^*, -1)$ – нули, а дополнительной строки – μ. Проведем процесс исключения в схеме, добавляя к каждой строке предыдущие (с множителем); тогда значение последней формы не изменится, и если в дополнительной строке в позициях $1, \dots, n$ окажутся нули, то в правом нижнем углу окажется значение формы.

Ортогонализация

Мы будем применять к строкам матрицы A так называемый модифицированный процесс Грама–Шмидта [8]. Модификация состоит в том, что когда мы ортогонализуем данный вектор к некоторой последовательности уже построенных ортогональных векторов, то, вычтя из данного вектора проекцию на первый вектор, далее вычитаем проекцию на второй вектор не исходного данного вектора, а полученной на первом шаге разности, на третий вектор проектируем полученную на втором шаге разность и т.д.

Кроме того, проведя ортогонализацию к $i-1$ вектору, в качестве следующего вектора мы выбираем наибольший по норме из оставшихся.

Тогда процесс ортогонализации строк протекает по следующим формулам.

Пусть $a_i^{(1)} = a_i = (a_{i1}, a_{i2}, \ldots, a_{in})$ – строки исходной матрицы. Не умаляя общности, считаем, что $1 \geq \|a_1^{(1)}\| \geq \|a_i^{(1)}\|$. Первый шаг состоит в ортогонализации всех строк к 1-ой:

$$a_i^{(2)} = a_i^{(1)} - k_i^{(1)} a_1^{(1)}, \quad k_i^{(1)} = \frac{(a_i^{(1)},\, a_1^{(1)})}{(a_1^{(1)},\, a_1^{(1)})}, \quad i = 2, \ldots, n.$$

И т.д., на m-ом шаге, считая $\|a_{m-1}^{(m-1)}\| \geq \|a_i^{(m-1)}\|$, $i \geq m-1$, ортогонализуем полученные на предыдущем шаге строки к $a_{m-1}^{(m-1)}$:

$$a_i^{(m)} = a_i^{(m-1)} - k_i^{(m-1)} a_{m-1}^{(m-1)}, \quad k_i^{(m-1)} = \frac{(a_i^{(m-1)},\, a_{m-1}^{(m-1)})}{(a_{m-1}^{(m-1)},\, a_{m-1}^{(m-1)})},$$

$$f_i^{(m)} = f_i^{(m-1)} - k_i^{(m-1)} f_{m-1}^{(m-1)}. \tag{2}$$

Применяя такой процесс к схеме задачи исключения, на каждом шаге выполняем ортогонализацию также и дополнительной строки

$$g^{(m)} = g^{(m-1)} - k_{n+1}^{(m-1)} a_{m-1}^{(m-1)}, \quad k_{n+1}^{(m-1)} = \frac{(g^{(m-1)},\, a_{m-1}^{(m-1)})}{(a_{m-1}^{(m-1)},\, a_{m-1}^{(m-1)})}. \tag{3}$$

Тогда в конце получим на месте коэффициентов формы нули, а последний элемент будет равен значению формы:

$$\mu = (g,\ x^*) = -k_{n+1}^{(1)} f_1^{(1)} - k_{n+1}^{(2)} f_2^{(2)} - \ldots - k_{n+1}^{(n)} f_n^{(n)}. \tag{4}$$

Таким образом, в процессе вычислений выполняются неравенства:

$$\|a_i^{(m)}\| \leq \|a_m^{(m)}\|, \quad i \geq m, \tag{5.a}$$

в частности,

$$\|a_1^{(1)}\| \geq \|a_2^{(2)}\| \geq \ldots \geq \|a_n^{(n)}\|. \tag{5.b}$$

Следовательно,

$$\left|k_i^{(m)}\right| = \left|\frac{(a_i^{(m)},\ a_m^{(m)})}{(a_m^{(m)},\ a_m^{(m)})}\right| \leq 1. \tag{5.c}$$

Заметим кроме этого, что для дополнительной строки

$$\left|k_{n+1}^{(m)}\right| = \left|\frac{(g^{(m)},\ a_m^{(m)})}{(a_m^{(m)},\ a_m^{(m)})}\right| \leq \frac{\|g^{(m)}\|}{\|a_m^{(m)}\|}. \tag{6}$$

Теперь мы дадим оценку ошибки округления при решении задачи исключения по описанной схеме ортогонализации строк и убедимся, что она зависит от тех же самых величин $\|A^{-1}f\|$, $\|A^{T^{-1}}g\|$, что и неустранимая погрешность. Аналогичная оценка может быть дана (и получится ещё проще) для метода исключения, разумеется, с выбором главного элемента.

Пусть ε - машинная точность, т.е. такое число, что относительная погрешность округления при умножении и делении, относительная погрешность записи числа в машину при полнозначной мантиссе не превосходят ε.

Нас интересует зависимость погрешности от данных задачи. Зависимость от порядка n мы прослеживать не будем. Поэтому буквой C будем обозначать различные константы, зависящие разве лишь от n.

Теорема.

$$|\delta(g,x)| \leq C\varepsilon + C\varepsilon\|A^{-1}f\| + C\varepsilon\|A^{T^{-1}}g\| + C\varepsilon\|A^{-1}f\|\,\|A^{T^{-1}}g\|,$$

где C - константа, зависящая лишь от порядка матрицы A системы.

Лемма 1. *Если перед началом преобразований матрицы A системы (1) абсолютная погрешность в её элементах оценивалась величиной $C\varepsilon$, то и после преобразований погрешность будет оцениваться величиной $C\varepsilon$.*

Доказательство. Ошибка в коэффициенте ортогонализации $k_m^{(i)}$ составляется из наследственной ошибки $\widetilde{\delta k_m^{(i)}}$ и ошибки округления $\eta_m^{(i)}$:

$$\delta k_m^{(i)} = \widetilde{\delta k_m^{(i)}} + \eta_m^{(i)}.$$

Обозначим $\widetilde{a}_k^{(j)} = a_k^{(j)} + \delta a_k^{(j)}$ реально вычисленное значение.

Тогда с точностью до малых второго порядка

$$\widetilde{\delta k_m^{(i)}} = \frac{(a_m^{(i)} + \delta a_m^{(i)},\ a_i^{(i)} + \delta a_i^{(i)})}{(a_i^{(i)} + \delta a_i^{(i)},\ a_i^{(i)} + \delta a_i^{(i)})} - \frac{(a_m^{(i)},\ a_i^{(i)})}{(a_i^{(i)},\ a_i^{(i)})}$$

$$= \frac{(\delta a_m^{(i)},\ a_i^{(i)})}{(a_i^{(i)},\ a_i^{(i)})} + \frac{(a_m^{(i)},\ a_i^{(i)})}{(a_i^{(i)},\ a_i^{(i)})}\left(1 - 2\frac{(a_i^{(i)},\ \delta a_i^{(i)})}{(a_i^{(i)},\ a_i^{(i)})}\right)$$

$$= \frac{(\delta a_m^{(i)},\ a_i^{(i)})}{(a_i^{(i)},\ a_i^{(i)})} + \frac{1}{(a_i^{(i)},\ a_i^{(i)})}\left(a_m^{(i)},\ \left(E - 2\frac{a_i^{(i)}\ a_i^{(i)^T}}{(a_i^{(i)},\ a_i^{(i)})}\right)\delta a_i^{(i)}\right).$$

Заметим, что $E - 2\frac{a_i^{(i)}\ a_i^{(i)^T}}{(a_i^{(i)},\ a_i^{(i)})}$ – матрица отражения, и её норма равна 1. Оценим $\eta_m^{(i)}$ – ошибку округления. Относительная погрешность числителя равна $C\varepsilon\frac{\|a_m^{(i)}\|\|a_i^{(i)}\|}{|(a_m^{(i)},\ a_i^{(i)})|}$, знаменателя – $C\varepsilon$, всей дроби – $C\varepsilon + C\varepsilon\frac{\|a_m^{(i)}\|\cdot\|a_i^{(i)}\|}{|(a_m^{(i)},\ a_i^{(i)})|}$, и значит абсолютная погрешность $|\eta_m^{(i)}| \le C\varepsilon\frac{\|a_m^{(i)}\|}{\|a_i^{(i)}\|} \le C\varepsilon$.

Тогда

$$|\delta k_m^{(i)}| \le \frac{\|\delta a_m^{(i)}\| + \|\delta a_i^{(i)}\|}{\|a_i^{(i)}\|} + C\varepsilon. \tag{7}$$

Теперь из формулы $a_m^{(2)} = a_m^{(1)} - k_m^{(1)} a_1^{(1)}$, $\delta a_m^{(2)} = \delta a_m^{(1)} - \delta k_m^{(1)} a_1^{(1)} - k_m^{(1)} \delta a_1^{(1)} + \eta_m^{(2)}$ (также с точностью до малых второго порядка) получим $\|\eta_m^{(2)}\| \le C\varepsilon\|a_1^{(1)}\| \le C\varepsilon$, $\|\delta a_m^{(2)}\| \le C\varepsilon + \left(\frac{\|\delta a_m^{(1)}\| + \|\delta a_1^{(1)}\|}{\|a_1^{(1)}\|} + C\varepsilon\right)\|a_1^{(1)}\| + \|\delta a_1^{(1)}\| + C\varepsilon \le C\varepsilon$.

Точно так же для следующего шага из полученной оценки следует $\|\delta a_m^{(3)}\| \le C\varepsilon$, и по индукции для всех i.

Лемма 2. *При всех* $i \geqslant 1$

$$|\delta k_{n+1}^{(i)}| \le \frac{\|\delta g^{(i)}\|}{\|a_i^{(i)}\|} + C\varepsilon\frac{\|g^{(i)}\|}{\|a_i^{(i)}\|^2}.$$

Доказательство. Как и выше, $\delta k_{n+1}^{(i)} = \widetilde{\delta k_{n+1}^{(i)}} + \eta_{n+1}^{(i)}$ – наследственная ошибка плюс текущая ошибка округления, причём

$$\delta k_{n+1}^{(i)} = \frac{(\delta g^{(i)},\ a_i^{(i)})}{(a_i^{(i)},\ a_i^{(i)})} + \frac{1}{(a_i^{(i)},\ a_i^{(i)})}\left(g^{(i)},\ \left(E - 2\frac{a_i^{(i)}\ a_i^{(i)T}}{(a_i^{(i)},\ a_i^{(i)})}\right)\delta a_i^{(i)}\right).$$

Относительная погрешность округления числителя $k_{n+1}^{(i)}$ в выражении (3) есть $C\varepsilon\frac{\|g^{(i)}\|\|a_i^{(i)}\|}{|(g^{(i)},\ a_i^{(i)})|}$, знаменателя – $C\varepsilon$, абсолютная погрешность – $|\eta_{n+1}^{(i)}| \le C\varepsilon\frac{\|g^{(i)}\|}{\|a_i^{(i)}\|}$, и с учётом леммы 1

$$|\delta k_{n+1}^{(i)}| \le \frac{\|\delta g^{(i)}\|}{\|a_i^{(i)}\|} + \frac{\|g^{(i)}\|}{\|a_i^{(i)}\|^2}\|\delta a_i^{(i)}\| + C\varepsilon\frac{\|g^{(i)}\|}{\|a_i^{(i)}\|}$$
$$\le \frac{\|\delta g^{(i)}\|}{\|a_i^{(i)}\|} + C\varepsilon\frac{\|g^{(i)}\|}{\|a_i^{(i)}\|^2}.$$

Лемма 3. *При* $i \ge 2$

$$\|\delta g^{(i)}\| \le C\varepsilon + C\varepsilon\sum_{m=1}^{i-1}\frac{\|g^{(m)}\|}{\|a_m^{(m)}\|}.$$

Доказательство. Из формулы для $g^{(i)}$, $g^{(i)} = g^{(i-1)} - k_{n+1}^{(i-1)}a_{i-1}^{(i-1)}$, найдём

$$\delta g^{(i)} = \delta g^{(i-1)} - \delta k_{n+1}^{(i-1)}a_{i-1}^{(i-1)} - k_{n+1}^{(i-1)}\delta a_{i-1}^{(i-1)} + \eta_i$$

и, используя выражение для $\delta k_{n+1}^{(i)}$ из доказательства предыдущей леммы, запишем

$$\delta g^{(i)} = \left(E - \frac{a\cdot a^T}{(a,\ a)}\right)\delta g^{(i-1)} - \left(g^{(i-1)},\ \left(E - 2\frac{a\cdot a^T}{(a,\ a)}\right)\delta a\right)\frac{a}{(a,\ a)} + \eta_i',$$

где $a = a_{i-1}^{(i-1)}$, $\eta_i' = \eta_{n+1}^{i-1}k_{n+1}^{(i)} + \eta_i$.

Одна матрица есть матрица проектирования, другая – матрица отражения, их нормы не превосходят 1; заметим ещё, что $|k_{n+1}^{(i+1)}| \le \frac{\|g^{(i-1)}\|}{\|a_{i-1}^{(i-1)}\|}$, значит

$$|\eta_i| \le C\varepsilon\|g^{(i-1)}\| + C\varepsilon|k_{n+1}^{(i-1)}|\|a_{i-1}^{(i-1)}\| \le C\varepsilon\|g^{(i-1)}\|.$$

Тогда $\|\delta g^{(i)}\| \le \|\delta g^{(i-1)}\| + C\varepsilon\frac{\|g^{(i-1)}\|}{\|a_{i-1}^{(i-1)}\|}$; суммируя, получаем неравенство леммы.

Лемма 4. *При всех* $i \geqslant 1$

$$|\delta k_{n+1}^{(i)}| \le C\varepsilon \frac{1}{\|a_i^{(i)}\|} + C\varepsilon \frac{1}{\|a_i^{(i)}\|} \sum_{m=1}^{i} \frac{\|g^{(m)}\|}{\|a_m^{(m)}\|}.$$

Доказательство. Требуемое неравенство непосредственно следует из лемм 2 и 3.

Лемма 5. *При* $i \ge 2$

$$|\delta f_i^{(i)}| \le C\varepsilon + C\varepsilon \sum_{m=1}^{i-1} \frac{|f_m^{(m)}|}{\|a_m^{(m)}\|}.$$

Доказательство. Из формулы (2) получим

$$|\delta f_i^{(i)}| \le |\delta f_i^{(1)}| + \sum_{m=1}^{i-1} |\delta k_i^{(m)}|\,|f_m^{(m)}| + \sum_{m=1}^{i-1} |k_i^{(m)}|\,|\delta f_m^{(m)}| + \eta_i^{(i)},$$

где $|\eta_i^{(i)}| \le C\varepsilon \sum_{m=1}^{i-1} |k_i^{(m)}|\,|f_m^{(m)}|$.

Учитывая неравенство (7), отсюда найдём

$$|\delta f_i^{(i)}| \le C\varepsilon + \sum_{m=1}^{i-1} |\delta f_m^{(m)}| + C\varepsilon \sum_{m=1}^{i-1} \frac{|f_m^{(m)}|}{\|a_m^{(m)}\|}.$$

Подставляя эти оценки рекуррентно друг в друга, получаем:

$$|\delta f_i^{(i)}| \le C\varepsilon + C\varepsilon \left(\sum_{m=1}^{i-1} \frac{|f_m^{(m)}|}{\|a_m^{(m)}\|} + \sum_{m=1}^{i-2} \frac{|f_m^{(m)}|}{\|a_m^{(m)}\|} + 2 \sum_{m=1}^{i-3} \frac{|f_m^{(m)}|}{\|a_m^{(m)}\|} + \ldots + 2^{i-3} \frac{|f_1^{(1)}|}{\|a_1^{(1)}\|} \right)$$

$$= C\varepsilon + C\varepsilon \left(2^{i-2} \frac{|f_1^{(1)}|}{\|a_1^{(1)}\|} + 2^{i-3} \frac{|f_2^{(2)}|}{\|a_2^{(2)}\|} + \ldots + \frac{|f_{i-1}^{(i-1)}|}{\|a_{i-1}^{(i-1)}\|} \right) = C\varepsilon + C\varepsilon \sum_{m=1}^{i-1} \frac{|f_m^{(m)}|}{\|a_m^{(m)}\|}.$$

Лемма доказана.

Заметим, однако, что константа C может достигать значительной величины $\sim 2^n$.

Лемма 6.

$$|\delta(g,x)| \le C\varepsilon + C\varepsilon \sum_{i=1}^{n} \frac{|f_i^{(i)}|}{\|a_i^{(i)}\|} + C\varepsilon \sum_{i=1}^{n} \frac{\|g^{(i)}\|}{\|a_i^{(i)}\|} + C\varepsilon \sum_{i=1}^{n} \frac{|f_i^{(i)}|}{\|a_i^{(i)}\|} \sum_{m=1}^{n} \frac{\|g^{(m)}\|}{\|a_m^{(m)}\|}.$$

Доказательство. Из (4) следует, что

$$|\delta(g,x)| \le C\varepsilon + \sum_{i=1}^{n} |\delta k_{n+1}^{(i)}|\,|f_i^{(i)}| + \sum_{i=1}^{n} |k_{n+1}^{(i)}|\,|\delta f_i^{(i)}| + \eta,$$

где η – погрешность округления.

$$|\eta| \le \varepsilon \sum_{i=1}^{n} |k_{n+1}^{(i)}|\,|f_i^{(i)}| \le \varepsilon \sum_{i=1}^{n} \frac{\|g^{(i)}\|}{}\|a_i^{(i)}\||f_i^{(i)}| \le \varepsilon \sum_{i=1}^{n} \frac{\|g^{(i)}\|\,|f_i^{(i)}|}{\|a_i^{(i)}\|^2}$$

в силу условия нормировки строк матрицы схемы. Тогда в силу формулы (6) и лемм 4 и 5 имеем

$$\begin{aligned}
|\delta(g,x)| &\le C\varepsilon + \sum_{i=1}^{n} |\delta k_{n+1}^{(i)}|\,|f_i^{(i)}| + \sum_{i=2}^{n} |k_{n+1}^{(i)}|\,|\delta f_i^{(i)}| + |k_{n+1}^{(1)}|\,|\delta f_1^{(1)}| + |\eta| \\
&\le C\varepsilon + \sum_{i=1}^{n} \left(C\varepsilon \frac{1}{\|a_i^{(i)}\|} + C\varepsilon \frac{1}{\|a_i^{(i)}\|} \sum_{m=1}^{i} \frac{\|g^{(m)}\|}{\|a_m^{(m)}\|} \right) |f_i^{(i)}| \\
&\quad + \sum_{i=2}^{n} \frac{\|g^{(i)}\|}{\|a_i^{(i)}\|} \left(C\varepsilon + C\varepsilon \sum_{m=1}^{i-1} \frac{|f_m^{(m)}|}{\|a_m^{(m)}\|} \right) + C\varepsilon \frac{\|g^{(1)}\|}{\|a_1^{(1)}\|} + |\eta| \\
&= C\varepsilon + C\varepsilon \sum_{i=1}^{n} \frac{|f_i^{(i)}|}{\|a_i^{(i)}\|} + C\varepsilon \sum_{i=1}^{n} \frac{\|g^{(i)}\|}{\|a_i^{(i)}\|} + C\varepsilon \sum_{i=1}^{n} \frac{|f_i^{(i)}|}{\|a_i^{(i)}\|} \sum_{m=1}^{i} \frac{\|g^{(m)}\|}{\|a_m^{(m)}\|} \\
&\quad + C\varepsilon \sum_{i=2}^{n} \frac{\|g^{(i)}\|}{\|a_i^{(i)}\|} \sum_{m=1}^{i-1} \frac{|f_m^{(m)}|}{\|a_m^{(m)}\|} + |\eta| \\
&\le C\varepsilon + C\varepsilon \sum_{i=1}^{n} \frac{|f_i^{(i)}|}{\|a_i^{(i)}\|} + C\varepsilon \sum_{i=1}^{n} \frac{\|g^{(i)}\|}{\|a_i^{(i)}\|} + C\varepsilon \sum_{i=1}^{n} \frac{|f_i^{(i)}|}{\|a_i^{(i)}\|} \sum_{m=1}^{n} \frac{\|g^{(m)}\|}{\|a_m^{(m)}\|}.
\end{aligned}$$

Введём матрицу

$$K := \begin{pmatrix} 1 & & & & \\ k_2^{(1)} & 1 & & 0 & \\ k_3^{(1)} & k_3^{(2)} & 1 & & \\ \dots & \dots & \dots & \ddots & \\ k_n^{(1)} & k_n^{(2)} & \dots & k_n^{(n-1)} & 1 \end{pmatrix}.$$

Лемма 7.

$$\left(k_{n+1}^{(1)}, \dots, k_{n+1}^{(n)}\right)^T = K^T A^{T^{-1}} g.$$

Доказательство. Из (2) для всякого i, $1 \le i \le n$,

$$a_i^{(i)} = a_i^{(1)} - k_i^{(1)} a_1^{(1)} - \dots - k_i^{(i-1)} a_{i-1}^{(i-1)}.$$

Значит

$$a_i^{(1)} = k_i^{(1)} a_1^{(1)} + \dots + k_i^{(i-1)} a_{i-1}^{(i-1)} + a_i^{(i)}.$$

В матричной записи $A = K\widetilde{A}$, где

$$\widetilde{A} = \left(a_1^{(1)}, a_2^{(2)}, \dots, a_n^{(n)}\right)^T.$$

С другой стороны, $g^T = k_{n+1}\widetilde{A}$, где

$$k_{n+1} = \left(k_{n+1}^{(1)}, k_{n+1}^{(2)}, \dots, k_{n+1}^{(n)}\right).$$

Тогда

$$g^T A^{-1} = k_{n+1}\widetilde{A}A^{-1} = k_{n+1}K^{-1}AA^{-1} = k_{n+1}K^{-1}.$$

Лемма 8.

$$\frac{\|g^{(i)}\|}{\|a_i^{(i)}\|} \le \|K\|\,\|A^{T^{-1}}g\|, \quad i \leqslant 1 \leqslant n.$$

Доказательство. Так как $g^{(i)} = \sum_{j=i}^{n} k_{n+1}^{(j)} a_j^{(j)}$ и векторы $a_j^{(j)}$ взаимно ортогональны, то

$$\|g^{(i)}\|^2 = \sum_{j=i}^{n} \left(k_{n+1}^{(j)}\right)^2 \|a_j^{(j)}\|^2,$$

и в силу неравенств (5)

$$\frac{\|g^{(i)}\|^2}{\|a_i^{(i)}\|^2} \le \sum_{j=i}^{n} \left(k_{n+1}^{(j)}\right)^2.$$

Следовательно,

$$\frac{\|g^{(i)}\|}{\|a_i^{(i)}\|} \le \sqrt{\sum_{j\ge i} {k_{n+1}^{(j)}}^2} \le \left\|\left(k_{n+1}^{(1)},\, k_{n+1}^{(2)},\, \ldots,\, k_{n+1}^{(n)}\right)\right\|_2 \le \|K\|\,\|A^{T^{-1}}g\|$$

по лемме 7.

Лемма 9.

$$\sum_{i=1}^{n} \frac{|f_i^{(i)}|}{\|a_i^{(i)}\|} \le C\|A^{-1}f\|.$$

Доказательство. Обозначим $x = A^{-1}f$. Тогда верно тождество $Ax = f$. В матричной записи

$$\begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \ldots \\ a_n^{(1)} \end{pmatrix} x = \begin{pmatrix} f_1^{(1)} \\ f_2^{(1)} \\ \ldots \\ f_n^{(1)} \end{pmatrix}.$$

После преобразований получим

$$\begin{pmatrix} a_1^{(1)} \\ a_2^{(2)} \\ \ldots \\ a_n^{(n)} \end{pmatrix} x = \begin{pmatrix} f_1^{(1)} \\ f_2^{(2)} \\ \ldots \\ f_n^{(n)} \end{pmatrix}$$

или с учётом нормировки

$$\begin{pmatrix} \frac{a_1^{(1)}}{\|a_1^{(1)}\|} \\ \frac{a_2^{(2)}}{\|a_2^{(2)}\|} \\ \ldots \\ \frac{a_n^{(n)}}{\|a_n^{(n)}\|} \end{pmatrix} x = \begin{pmatrix} \frac{f_1^{(1)}}{\|a_1^{(1)}\|} \\ \frac{f_2^{(2)}}{\|a_2^{(2)}\|} \\ \ldots \\ \frac{f_n^{(n)}}{\|a_n^{(n)}\|} \end{pmatrix}.$$

Обозначим

$$\widetilde{A} := \begin{pmatrix} \frac{a_1^{(1)}}{\|a_1^{(1)}\|} \\ \frac{a_2^{(2)}}{\|a_2^{(2)}\|} \\ \ldots \\ \frac{a_n^{(n)}}{\|a_n^{(n)}\|} \end{pmatrix}, \quad \widetilde{F} := \begin{pmatrix} \frac{f_1^{(1)}}{\|a_1^{(1)}\|} \\ \frac{f_2^{(2)}}{\|a_2^{(2)}\|} \\ \ldots \\ \frac{f_n^{(n)}}{\|a_n^{(n)}\|} \end{pmatrix}.$$

Тогда

$$\sum_{i=1}^{n}\frac{|f_i^{(i)}|}{\|a_i^{(i)}\|}=\|\widetilde{F}\|_1=\|\widetilde{A}x\|_1\le C\|\widetilde{A}\|\,\|x\|_2=C\|A^{-1}f\|,$$

поскольку $\widetilde{A}$ – ортогональная матрица и, следовательно, $\|\widetilde{A}\|=1$.

Доказательство теоремы. Имеем $\|K\|\le n$ в силу (5.c). По леммам 6, 8 и 9 имеем

$$|\delta(g,\,x)|\le C\varepsilon+C\varepsilon\sum_{i=1}^{n}\frac{|f_i^{(i)}|}{\|a_i^{(i)}\|}+C\varepsilon\sum_{i=1}^{n}\frac{\|g^{(i)}\|}{\|a_i^{(i)}\|}+C\varepsilon\sum_{i=1}^{n}\frac{|f_i^{(i)}|}{\|a_i^{(i)}\|}\sum_{m=1}^{n}\frac{\|g^{(m)}\|}{\|a_m^{(m)}\|}$$

$$\le C\varepsilon+C\varepsilon\|A^{-1}f\|+C\varepsilon\|A^{T^{-1}}g\|+C\varepsilon\|A^{-1}f\|\,\|A^{T^{-1}}g\|,$$

где C – константа, зависящая лишь от порядка матрицы A системы (1).

Полученная оценка выражается через те же величины, что и оценка неустранимой погрешности.

В экспериментальных вычислениях рассмотрим следующие задачи.

Задача 1

Задача наилучшего квадратичного приближения:

$$\int_0^1 (f(x)-p_n(x))^2\,dx\to min.$$

Необходимо вычислить значение приближающего полинома в произвольной точке x отрезка $[-1,\ 1]$. В качестве приближаемой функций берётся полином $\sum_{k=0}^{n}x^k$. Схема имеет вид

$$\left(\begin{array}{ccccc|c}
1 & \frac{1}{2} & \frac{1}{3} & \dots & \frac{1}{n+1} & \int_0^1 f(x)dx\\
\frac{1}{2} & \frac{1}{3} & \dots & \dots & \frac{1}{n+2} & \int_0^1 xf(x)dx\\
\frac{1}{3} & \dots & & & & \\
\dots & \dots & \dots & \dots & \dots & \dots\\
\frac{1}{n+1} & \frac{1}{n+2} & \dots & \dots & \frac{1}{2n+1} & \int_0^1 x^nf(x)dx\\
\hline
-1 & -x & -x^2 & \dots & -x^n & 0
\end{array}\right).$$

Задача 2

Задача интерполяции: вычислить значение интерполяционного полинома в произвольной точке x отрезка $[-1,\ 1]$ по известным значениям в узлах интерполяции. В качестве приближаемых функций берутся полиномы Чебышёва, Лежандра, $\sum_{k=0}^{n} c^k x^k$.

Схема имеет вид

$$\left(\begin{array}{ccccc|c} 1 & x_0 & x_0^2 & \dots & x_0^n & B_0 \\ 1 & x_1 & x_1^2 & \dots & x_1^n & B_1 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_n & x_n^2 & \dots & x_n^n & B_n \\ \hline -1 & -x & -x^2 & \dots & -x^n & 0 \end{array}\right),$$

где B_i – значение полинома в i-ом узле.

Задача 3

Задача интерполяции дробно-рациональными функциями, т.е. посредством координатных функций $\varphi_k(x) = \frac{1}{x-a_k}$ на отрезке $[-1, 1]$, $a_k \notin [-1, 1]$.

В качестве приближаемой функций берётся $\sum_{k=0}^{n} \frac{a_k}{x-a_k}$. Схема имеет вид

$$\left(\begin{array}{ccc|c} \varphi_0(x_0) & \dots & \varphi_n(x_0) & f(x_0) \\ \dots & \dots & \dots & \dots \\ \varphi_0(x_n) & \dots & \varphi_n(x_n) & f(x_n) \\ \hline -\varphi_0(x) & \dots & -\varphi_n(x) & 0 \end{array}\right).$$

Результаты

Вычисления проводились на языке Pascal в формате double, так что $\varepsilon \approx 10^{-16}$.

Обозначения:

0 – $\|A^{-1}\|$

1 – $\|A^{-1}f\|$

2 – $\|A^{T^{-1}}g\|$

3 – $k(A) = \|A\|\,\|A^{-1}\|$

4 – Точное значение формы в точке x.

5 – Разница между точным значением формы в точке x и значением, вычисленным подстановкой в форму решения системы, полученного методом Гаусса.

6 – Разница между точным значением формы в точке x и значением, вычисленным подстановкой в форму решения системы, полученного методом отражения.

7 – Разница между точным значением формы в точке x и значением, вычисленным подстановкой в форму решения системы, полученного методом ортогонализации.

8 – Разница между точным значением формы в точке x и значением, вычисленным по схеме исключения методом Гаусса.

9 – Разница между точным значением формы в точке x и значением, вычисленным по схеме исключения методом отражения.

10 – Разница между точным значением формы в точке x и значением, вычисленным по схеме исключения методом ортогонализации.

Задача 1 ($c = 1$).

n	10	25	40
x	$9.75E-01$	$9.75E-01$	$9.75E-01$
0	$4.08E+14$	$4.02E+19$	$1.41E+19$
1	$1.10E+01$	$1.84E+08$	$4.57E+09$
2	$5.37E+07$	$1.54E+10$	$1.24E+10$
3	$1.23E+15$	$1.55E+20$	$6.08E+19$
4	$9.72E+00$	$1.93E+01$	$2.58E+01$
5	$-2.32E-10$	$3.88E-08$	$1.12E-08$
6	$-4.58E-11$	$9.30E-08$	$-6.85E-08$
7	$6.18E-04$	$3.87E+01$	$3.93E+00$
8	$-4.58E-11$	$9.30E-08$	$-6.85E-08$
9	$1.54E-09$	$2.95E-08$	$-6.19E-08$
10	$-2.32E-10$	$3.88E-08$	$1.12E-08$

Задача 2.

Полином $\sum_{k=0}^{n} c^k x^k$, $c = 0.95$.
Равноотстоящие узлы на отрезке $[-1, 1]$.

n	10	25	40
x	$9.75E-01$	$9.75E-01$	$9.75E-01$
0	$3.65+03$	$3.55E+10$	$7.56E+16$
1	$8.62E+00$	$1.47E+01$	$4.22E+04$
2	$2.16E+01$	$2.49E+05$	$5.12E+08$
3	$4.01E+04$	$9.24E+11$	$3.10E+18$
4	$7.72E+00$	$1.17E+01$	$1.30E+01$
5	$-2.66E-15$	$-3.70E-11$	$-7.01E-08$
6	$0.00E+00$	$-8.65E-12$	$1.84E-08$
7	$-3.55E-15$	$7.59E-11$	$-1.18E+00$
8	$8.88E-16$	$-8.65E-12$	$1.84E-08$
9	$-8.88E-16$	$-2.44E-11$	$8.05E-09$
10	$0.00E+00$	$-3.70E-11$	$-7.01E-08$

Полином $\sum_{k=0}^{n} c^k x^k$, $c = 0.95$.
Чебышёвские узлы на отрезке $[-1, 1]$

n	10	25	40
x	$9.75E-01$	$9.75E-01$	$9.75E-01$
0	$8.41E+02$	$1.96E+08$	$6.85E+13$
1	$8.62E+00$	$1.47E+01$	$3.13E+02$
2	$1.80E+00$	$2.44E+00$	$2.86E+00$
3	$9.25E+03$	$5.10E+09$	$2.81E+15$
4	$7.72E+00$	$1.17E+01$	$1.30E+01$
5	$-1.78E-15$	$0.00E+00$	$3.55E-15$
6	$-2.66E-15$	$-8.88E-15$	$3.55E-15$
7	$-4.44E-15$	$-5.33E-15$	$1.90E-03$
8	$-3.55E-15$	$-1.07E-14$	$3.55E-15$
9	$-2.66E-15$	$-3.55E-15$	$8.88E-15$
10	$-1.78E-15$	$-3.55E-15$	$1.78E-15$

Полином Лежандра. Равноотстоящие узлы на отрезке $[-1,\ 1]$

n	10	25	40
x	$9.75E-01$	$9.75E-01$	$9.75E-01$
0	$3.65E+03$	$3.55E+10$	$7.56E+16$
1	$2.08E+02$	$6.87E+07$	$5.25E+13$
2	$2.16E+01$	$2.49E+05$	$5.12E+08$
3	$4.01E+04$	$9.24E+11$	$3.10E+18$
4	$2.76E-01$	$-8.21E-02$	$-2.15E-04$
5	$2.04E-14$	$-3.89E-05$	$1.00E+04$
6	$9.77E-15$	$1.38E-04$	$2.59E+05$
7	$-9.03E-11$	$-3.31E+02$	$3.41E+12$
8	$5.88E-15$	$1.38E-04$	$2.59E+05$
9	$-1.32E-14$	$1.43E-05$	$2.21E+04$
10	$2.68E-14$	$-3.89E-05$	$1.00E+04$

Полином Лежандра. Чебышёвские узлы на отрезке $[-1,\ 1]$

n	10	25	40
x	$9.75E-01$	$9.75E-01$	$9.75E-01$
0	$8.41E+02$	$1.96E+08$	$6.85E+13$
1	$2.08E+02$	$6.87E+07$	$2.96E+13$
2	$1.79E+00$	$2.43E+00$	$2.86E+00$
3	$9.25E+03$	$5.10E+09$	$2.81E+15$
4	$2.76E-01$	$-8.21E-02$	$-2.15E-04$
5	$4.33E-15$	$1.86E-09$	$8.45E-04$
6	$-5.99E-15$	$6.88E-09$	$-2.57E-04$
7	$-1.35E-11$	$-2.31E+00$	$-5.58E+11$
8	$-9.55E-15$	$7.33E-09$	$-1.25E-04$
9	$-1.39E-15$	$2.99E-09$	$4.48E-04$
10	$-8.33E-16$	$1.76E-09$	$8.99E-04$

Полином Чебышёва. Равноотстоящие узлы на отрезке $[-1,\ 1]$

n	10	25	40
x	$9.75E-01$	$9.75E-01$	$9.75E-01$
0	$3.65E+03$	$3.55E+10$	$7.56E+16$
1	$5.77E+02$	$3.18E+08$	$3.15E+14$
2	$2.16E+01$	$2.49E+05$	$5.12E+08$
3	$4.01E+04$	$9.24E+11$	$3.10E+18$
4	$-2.26E-01$	$4.42E-01$	$-6.35E-01$
5	$7.46E-14$	$-1.74E-04$	$6.36E+04$
6	$3.04E-14$	$6.26E-04$	$1.58E+06$
7	$-2.42E-10$	$-1.50E+03$	$2.05E+13$
8	$1.22E-14$	$6.26E-04$	$1.58E+06$
9	$-3.78E-14$	$6.13E-05$	$1.35E+05$
10	$8.06E-14$	$-1.74E-04$	$6.36E+04$

Полином Чебышёва. Чебышёвские узлы на отрезке $[-1,\ 1]$.

n	10	25	40
x	$9.75E-01$	$9.75E-01$	$9.75E-01$
0	$8.41E+02$	$1.96E+08$	$6.85E+13$
1	$5.77E+02$	$3.18E+08$	$1.76E+14$
2	$1.80E+00$	$2.44E+00$	$2.86E+00$
3	$9.25E+03$	$5.10E+09$	$2.81E+15$
4	$-2.26E-01$	$4.42E-01$	$-6.35E-01$
5	$2.22E-14$	$-5.07E-09$	$6.05E-03$
6	$-6.30E-15$	$3.33E-08$	$1.92E-03$
7	$-3.63E-11$	$-1.06E+01$	$-3.29E+12$
8	$-2.59E-14$	$3.43E-08$	$-7.11E-04$
9	$-2.80E-15$	$1.39E-08$	$2.66E-03$
10	$-3.61E-16$	$8.44E-09$	$5.41E-03$

Задача 3.

Полюсы $-\frac{1}{i+1}-1$. Чебышёвские узлы на отрезке $[-1,\ 1]$.

n	10	25	40
x	$9.75E-01$	$9.75E-01$	$9.75E-01$
0	$7.36E+15$	$1.41E+17$	$2.02E+16$
1	$6.99E+02$	$1.38E+07$	$1.08E+09$
2	$1.75E+00$	$3.30E+01$	$1.51E+01$
3	$1.86E+17$	$1.31E+19$	$3.74E+18$
4	$-6.18E+00$	$-1.40E+01$	$-2.17E+01$
5	$1.78E-15$	$-5.86E-14$	$2.84E-14$
6	$0.00E+00$	$-4.97E-14$	$-1.74E-13$
7	$-1.01E-01$	$-7.89E+00$	$-5.08E+01$
8	$-8.88E-16$	$-8.88E-14$	$-1.78E-13$
9	$3.55E-15$	$1.77E-15$	$1.07E-14$
10	$8.88E-16$	$3.55E-14$	$4.98E-14$

Полюсы $-\frac{1}{i+1}-1$. Равноотстоящие узлы на отрезке $[-1,\ 1]$.

n	10	25	40
x	$9.75E-01$	$9.75E-01$	$9.75E-01$
0	$7.56E+16$	$8.74E+16$	$4.60E+16$
1	$3.72E+03$	$1.62E+06$	$2.13E+08$
2	$1.30E+01$	$4.56E+01$	$6.78E+01$
3	$1.74E+18$	$5.81E+18$	$5.28E+18$
4	$-6.18E+00$	$-1.40E+01$	$-2.17E+01$
5	$1.78E-15$	$1.76E-13$	$-3.55E-14$
6	$-4.44E-15$	$2.84E-14$	$8.88E-13$
7	$1.65E+00$	$-2.96E+01$	$-5.77E+01$
8	$-3.55E-15$	$2.31E-14$	$1.02E-12$
9	$8.88E-16$	$0.00E+00$	$7.11E-15$
10	$-2.66E-15$	$1.85E-13$	$1.31E-13$

Полюсы $-\frac{1}{i+1}-1$. Чебышёвские узлы на отрезке $[-1,\ 1]$.

n	10	25	40
x	$6.00E-01$	$6.00E-01$	$6.00E-01$
0	$7.36E+15$	$1.41E+17$	$2.02E+16$
1	$6.99E+02$	$1.38E+07$	$1.08E+09$
2	$1.62E+00$	$3.19E+01$	$1.39E+01$
3	$1.86E+17$	$1.31E+19$	$3.74E+18$
4	$-7.43E+00$	$-1.69E+01$	$-2.65E+01$
5	$5.33E-15$	$-4.62E-14$	$1.42E-14$
6	$2.66E-15$	$-4.97E-14$	$-1.63E-13$
7	$-1.23E-01$	$-9.17E+00$	$-6.20E+01$
8	$1.78E-15$	$-6.75E-14$	$-2.34E-13$
9	$8.88E-16$	$3.55E-15$	$-1.78E-14$
10	$1.78E-15$	$6.75E-14$	$3.20E-14$

Полюсы $-\frac{1}{i+1}-1$. Равноотстоящие узлы на отрезке $[-1,\ 1]$.

n	10	25	40
x	$6.00E-01$	$6.00E-01$	$6.00E-01$
0	$7.56E+16$	$8.74E+16$	$4.60E+16$
1	$3.72E+03$	$1.62E+06$	$2.13E+08$
2	$6.64E+00$	$3.49E+01$	$5.87E+01$
3	$1.74E+18$	$5.81E+18$	$5.28E+18$
4	$-7.43E+00$	$-1.70E+01$	$-2.65E+01$
5	$1.07E-14$	$-3.20E-14$	$-2.67E-13$
6	$-1.07E-14$	$2.13E-14$	$4.12E-13$
7	$2.02E+00$	$-3.59E+01$	$-6.94E+01$
8	$-1.24E-14$	$2.13E-14$	$4.44E-13$
9	$-8.88E-16$	$3.55E-15$	$0.00E+00$
10	$2.66E-15$	$-2.13E-14$	$7.11E-15$

Выводы

Полученные результаты подтверждают, что, действительно, числами обусловленности являются $\|A^{-1}f\|$, $\|A^{T^{-1}}g\|$, а не число обусловленности матрицы системы. Чрезвычайно любопытно, что способ решения системы и подстановки найденного решения в линейную форму даёт примерно ту же погрешность, за исключением метода ортогонализации. Более того, создаётся впечатление, что в некоторых случаях, чтобы не сказать в большинстве, действительно, числом обусловленности является не произведение норм $\|A^{-1}f\|$, $\|A^{T^{-1}}g\|$, а максимум из них. Особенно выделяется третья задача, где погрешность парадоксально мала.

Литература

1. Д. К. Фаддеев, В. Н. Фаддеева, *Вычислительные методы линейной алгебры.* Издание 3-е, стереотипное. Изд. Лань, СПб., 2002.
2. А. А. Абрамов, *Об одном методе исключения в линейных задачах.* — Ж. вычисл. матем. матем. физ. **20** (1980), 226–230.
3. А. А. Абрамов, В. О. Белаш, *Об одном методе исключения для линейных задач.* — Ж. вычисл. матем. матем. физ. **35** (1995), 499–510.
4. А. А. Абрамов, *О сходимости одного метода исключения.* — Ж. вычисл. матем. матем. физ. **38** (1998), 351–364.
5. А. А. Абрамов, В. И. Ульянова, Л. Ф. Юхно, *О реализации одного метода исключения в линейных задачах с неточно заданными исходными данными.* — Ж. вычисл. матем. матем. физ. **44** (2004), 640–649.
6. Л. Ф. Юхно, *Об одной модификации метода сопряженных градиентов.* — Ж. вычисл. матем. матем. физ. **46** (2006), 7-11.
7. Y. Cao, L. Petzold, *A subspace error estimate for linear systems.* — SIAM J. Matrix Anal. Appl. **24** (2003), 787–801.
8. Дж. Голуб, Ч. Ван Лоун, *Матричные вычисления.* Мир, М., 1999.

Rodnikov A. O., Samokish B. A. An estimate of the round-off error in the elimination problem.

The paper demonstrates that in computing a linear form (g, x) of the solution of a system of linear equations $Ax = f$, the round-off error depends on the quantities $\|A^{-1}f\|$ and $\|A^{T^{-1}}g\|$ rather than on the condition number of the coefficient matrix A. Estimates of the inherent and round-off errors in solving the above problem by the orthogonalization method are provided. Numerical results confirming theoretical conclusions are presented.

С.-Петербургский
государственный университет

Поступило 14 сентября 2006 г.