Экономичные критерии останова итераций в методе сопряженных градиентов

И. В. Киреев

Институт вычислительного моделирования СО РАН, Красноярск, Россия
Контактный e-mail: kiv@icm.krasn.ru

Обсуждаются некоторые аспекты численной реализации метода сопряженных градиентов для решения систем линейных алгебраических уравнений с симметричной положительно определенной матрицей при наличии ошибок округления. Рассмотрены как пошаговое поведение некоторых широко распространенных версий алгоритма, так и критерии останова итерационного процесса.

Ключевые слова: метод сопряженных градиентов, критерии останова итераций.

Введение

Идеальный алгоритм решения систем линейных алгебраических уравнений теоретически должен иметь конечное завершение, но при досрочном завершении он должен дать приемлемое приближение к искомому решению. Одним из таких методов является *метод сопряженных градиентов* (MCG).

История MCG началась с работы [1]. Этот стандартный на сегодня вычислительный алгоритм используется для решения больших систем уравнений с симметричными положительно определенными матрицами, для которых прямые методы непрактичны. Алгоритм MCG можно применять не только для решения систем линейных уравнений, но и для исследования спектра матрицы [2].

Журнал "Computing in Science and Engineering" назвал метод сопряженных градиентов одним из 10 лучших алгоритмов XX века [3]. И тем приятнее отметить, что алгоритм ортогонализации степенной последовательности [4], лежащий в основе теоретического обоснования MCG, предложил русский математик А.Н. Крылов. История создания MCG и его развитие (до 1977 г.) отражены в обзоре [5].

1. Постановка задачи

При численном решении системы линейных уравнений каким-либо методом возникает естественный вопрос: как ошибки округления в компьютере влияют на полученное приближенное решение? Существуют два подхода к оценке точности алгоритма [6]: прямой, когда осуществляется непосредственный анализ погрешности (данные точные, но сам алгоритм работает с погрешностью), и обратный, основанный на идее, что реально вычисленное решение рассматривается как точное для той же задачи, но с возмущенными входными данными. При этом само возмущение выбирается так, чтобы его действие оказалось эквивалентным совокупному влиянию всех ошибок округления. Обратный

анализ позволяет оценить совместное влияние ошибок округления и ошибок входных данных на точность результатов.

В 1992 г. впервые был проведен обратный анализ погрешности метода сопряженных градиентов [7]. При этом показано [8], что характер погрешности численного решения линейной системы уравнений MCG при конечной точности машинной арифметики имеет сходство с точными вычислениями по этому же алгоритму, но приложенными к иной матрице, у которой гораздо больше, нежели у первоначальной, собственных значений, распределенных, однако, вблизи спектральных чисел исходной матрицы.

Метод сопряженных градиентов при точных вычислениях приводит к ответу за конечное число шагов, но по сути является итерационным процессом. Его слабым местом является критерий останова — определение номера шага процесса, после которого точность приближения к решению системы линейных уравнений на данном компьютере не может быть существенно улучшена. Практическое применение результатов обратного анализа погрешности для выработки условий прекращения вычислений в MCG опирается на априорные оценки границ спектра матрицы системы, получение которых является очень трудоемкой задачей. Поэтому весьма актуально *построение экономичных критериев останова для MCG*, что является целью данной работы.

2. Численная реализация MCG

Рассмотрим совместную систему линейных алгебраических уравнений

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1}$$

где $\mathbf{x}, \mathbf{b} \in \mathbf{R}^m$ — m-мерное евклидово пространство, $\mathbf{A}$ — квадратная симметричная положительно определенная вещественная матрица порядка m. При сделанных ограничениях решение $\mathbf{x}$ системы (1) доставляет минимум квадратичному функционалу

$$a(\mathbf{y}) = \frac{1}{2}\langle \mathbf{A}\mathbf{y}, \mathbf{y}\rangle - \langle \mathbf{b}, \mathbf{y}\rangle, \tag{2}$$

где $\langle \mathbf{b}, \mathbf{y}\rangle = \sum\limits_{k=1}^{m} b_k{\cdot}y_k$ — скалярное произведение векторов $\mathbf{b}$ и $\mathbf{y}$ из $\mathbf{R}^m$; евклидову норму вектора $\mathbf{y}$ обозначим через $\|\mathbf{y}\| = \sqrt{\langle \mathbf{y}, \mathbf{y}\rangle}$. Для тех же векторов $\mathbf{b}$ и $\mathbf{y}$ их энергетическое скалярное произведение обозначим через $\langle \mathbf{b}, \mathbf{y}\rangle_A = \langle \mathbf{b}, \mathbf{A}\mathbf{y}\rangle$. Тогда $\|\mathbf{y}\|_A = \sqrt{\langle \mathbf{y}, \mathbf{y}\rangle_A}$ — энергетическая норма вектора $\mathbf{y}$. Необходимо заметить [9], что из (1) и (2) следует равенство $\|\mathbf{x}-\mathbf{y}\|_A^2 = 2a(\mathbf{y}) + \|\mathbf{x}\|_A^2$, т. е. величина $a(\mathbf{y})$ характеризует близость вектора $\mathbf{y}$ к решению $\mathbf{x}$ системы (1) в энергетической норме.

Систему (1) численно можно решить методом сопряженных градиентов [9]:

при $n = 0$ *задан* $\mathbf{x}^0$; *вычисляем* $\mathbf{r}^0 = \mathbf{A}\mathbf{x}^0 - \mathbf{b}$ *и полагаем* $\mathbf{s}^1 = \mathbf{r}^0$;

при $n > 0$ *вычисляем* α_n *и* β_n *по одной из формул*

$$\alpha_n = -\frac{\|\mathbf{r}^{n-1}\|^2}{\langle \mathbf{r}^{n-1}, \mathbf{s}^n\rangle_A} = -\frac{\langle \mathbf{r}^{n-1}, \mathbf{s}^n\rangle}{\|\mathbf{s}^n\|_A^2} < 0, \quad \beta_n = -\frac{\langle \mathbf{r}^n, \mathbf{s}^n\rangle_A}{\|\mathbf{s}^n\|_A^2} = \frac{\|\mathbf{r}^n\|^2}{\|\mathbf{r}^{n-1}\|^2} > 0, \tag{3}$$

и находим

$$\mathbf{r}^n = \mathbf{r}^{n-1} + \alpha_n \mathbf{A}\mathbf{s}^n, \quad \mathbf{s}^{n+1} = \mathbf{r}^n + \beta_n \mathbf{s}^n, \quad \mathbf{x}^n = \mathbf{x}^{n-1} + \alpha_n \mathbf{s}^n. \tag{4}$$

Если все вычисления произведены при отсутствии ошибок округления, то соотношения (3), (4) гарантируют получение решения $\mathbf{x}$ системы (1) не более чем за m шагов. При этом совокупность векторов $\{\mathbf{s}^n\}$ будет $\mathbf{A}$-ортогональной ($\langle \mathbf{s}^n, \mathbf{s}^k \rangle_A = 0,\ n \neq k$), а векторы $\{\mathbf{r}^n\}$ образуют ортогональную, т. е. сопряженную систему ($\langle \mathbf{r}^n, \mathbf{r}^k \rangle = 0,\ n \neq k$), причем $\mathbf{r}^n$ является градиентом функции $a(\mathbf{y})$ из (2) в точке $\mathbf{x}^n$:

$$\left.\frac{\partial a(\mathbf{y})}{\partial \mathbf{y}}\right|_{\mathbf{y}=\mathbf{x}^n} = \mathbf{r}^n = \mathbf{A}\mathbf{x}^n - \mathbf{b}. \tag{5}$$

Последнее обстоятельство отражено и в названии метода — Method of Conjugate Gradients, MCG.

Вариант MCG, в котором на каждом шаге процесса выполнены условия сопряженности $\langle \mathbf{r}^n, \mathbf{r}^{n-1} \rangle = \langle \mathbf{s}^{n+1}, \mathbf{s}^n \rangle_A = 0$, связан с выражениями

$$\alpha_n = -\frac{\|\mathbf{r}^{n-1}\|^2}{\langle \mathbf{r}^{n-1}, \mathbf{s}^n \rangle_A}, \quad \beta_n = -\frac{\langle \mathbf{r}^n, \mathbf{s}^n \rangle_A}{\|\mathbf{s}^n\|_A^2}. \tag{6}$$

На каждом шаге версии (4), (6) MCG необходимо вычислить одно произведение матрицы $\mathbf{A}$ на вектор $\mathbf{s}^n$ и четыре скалярных произведения: $\|\mathbf{r}^n\|^2$, $\langle \mathbf{r}^{n-1}, \mathbf{s}^n \rangle_A$, $\langle \mathbf{r}^n, \mathbf{s}^n \rangle_A$ и $\|\mathbf{s}^n\|_A^2$.

Напомним, что в авторской [1] версии MCG параметры α_n и β_n находятся по формулам

$$\alpha_n = -\frac{\langle \mathbf{r}^{n-1}, \mathbf{s}^n \rangle}{\|\mathbf{s}^n\|_A^2}, \quad \beta_n = -\frac{\langle \mathbf{r}^n, \mathbf{s}^n \rangle_A}{\|\mathbf{s}^n\|_A^2}. \tag{7}$$

Численная реализация алгоритма (4), (7) MCG требует меньших вычислительных затрат по сравнению с предыдущей версией — необходимо вычислить $\mathbf{A}\mathbf{s}^n$, $\|\mathbf{s}^n\|_A^2$, $\langle \mathbf{r}^{n-1}, \mathbf{s}^n \rangle$ и $\langle \mathbf{r}^n, \mathbf{s}^n \rangle_A$.

Предложенная в [10] и алгоритмически реализованная в [2] версия MCG, в которой α_n и β_n определяются соотношениями

$$\alpha_n = -\frac{\|\mathbf{r}^{n-1}\|^2}{\langle \mathbf{r}^{n-1}, \mathbf{s}^n \rangle_A}, \quad \beta_n = \frac{\|\mathbf{r}^n\|^2}{\|\mathbf{r}^{n-1}\|^2}, \tag{8}$$

считается наиболее экономичной, поскольку на очередном шаге алгоритма (4), (8) необходимо вычислять одно произведение матрицы на вектор и только два скалярных произведения.

В монографии [11] обсуждается версия MCG, в которой гарантируется релаксационность процесса, т. е. выполнение неравенства $a(\mathbf{x}^n) < a(\mathbf{x}^{n-1})$ на каждом шаге n, для чего предложено делать два вычисления произведения матрицы на вектор. Однако, если рассматривать MCG как метод генерации направления спуска [12] для численного решения задачи минимизации квадратичной функции $a(\mathbf{y})$, то релаксационность можно получить и за одно вычисление произведения матрицы на вектор. Покажем, как этого можно достичь.

Зафиксируем векторы $\mathbf{y}$, $\mathbf{s}$ и обозначим через α_s решение одномерной задачи безусловной минимизации [12] квадратичной сильно выпуклой функции $a(\mathbf{y} + \alpha\mathbf{s})$ по вещественному аргументу α:

$$\alpha_s = -\frac{\langle \mathbf{A}\mathbf{y} - \mathbf{b}, \mathbf{s} \rangle}{\|\mathbf{s}\|_A^2} = \frac{\langle \mathbf{x} - \mathbf{y}, \mathbf{s} \rangle_A}{\|\mathbf{s}\|_A^2},$$

$$a(\mathbf{y}+\alpha_s\mathbf{s}) = a(\mathbf{y}) - \frac{1}{2}(\alpha_s \cdot \|\mathbf{s}\|_A)^2 = \min_{\alpha\in\mathbf{R}} a(\mathbf{y}+\alpha\mathbf{s}).$$

Очевидно, что если направление спуска $\mathbf{s}$ не ортогонально градиенту функции $a(\mathbf{y})$ в "точке" $\mathbf{y}$, то будет выполнено неравенство $a(\mathbf{y}-\alpha_s\mathbf{s}) < a(\mathbf{y})$, означающее релаксационность перехода от "точки" $\mathbf{y}$ к "точке" $\mathbf{y}-\alpha_s\mathbf{s}$ посредством "спуска по направлению $\mathbf{s}$". Поэтому, если α_n и β_n найти по формулам

$$\alpha_n = -\frac{\langle \mathbf{x}^{n-1}, \mathbf{s}^n\rangle_A - \langle \mathbf{b}, \mathbf{s}^n\rangle}{\|\mathbf{s}^n\|_A^2}, \quad \beta_n = -\frac{\langle \mathbf{r}^n, \mathbf{s}^n\rangle_A}{\|\mathbf{s}^n\|_A^2}, \tag{9}$$

эквивалентным (3) при точных вычислениях, то релаксационность процесса (4), (9) гарантирована с той точностью, с которой вычисляются скалярные произведения и произведение матрицы на вектор. В такой версии MCG на каждом шаге следует выполнить одно произведение матрицы $\mathbf{A}$ на вектор $\mathbf{s}^n$ и четыре скалярных произведения: $\langle \mathbf{x}^{n-1}, \mathbf{s}^n\rangle_A$, $\langle \mathbf{b}^{n-1}, \mathbf{s}^n\rangle$, $\langle \mathbf{r}^n, \mathbf{s}^n\rangle_A$ и $\|\mathbf{s}^n\|_A^2$.

Необходимо отметить, что при минимизации выпуклых гладких функций [12] методом сопряженных градиентов возникают соотношения, аналогичные (9), в которых роль матрицы $\mathbf{A}$ играет матрица Гессе целевой функции, а выражение для β_n из (9) обеспечивает $\mathbf{A}$-ортогональность векторов $\mathbf{s}^n$ и $\mathbf{s}^{n+1}$ с той же точностью, с которой вычисляется параметр α_n.

О реальной эффективности приведенных выше версий MCG можно судить по значению энергетической нормы невязки $\|\mathbf{x}-\mathbf{x}^n\|_A$ на каждом шаге процесса. Для того чтобы отслеживать накопление ошибки, присущее только версии алгоритма MCG, была проведена серия численных расчетов для диагональных матриц $\mathbf{A}$; подобный выбор матрицы системы (1) гарантирует, что "кососимметричная составляющая" погрешности вычисления произведения матрицы $\mathbf{A}$ на вектор $\mathbf{s}^n$ будет пренебрежимо мала. На рис. 1 в графическом виде приведены результаты вычислительных экспериментов для систем

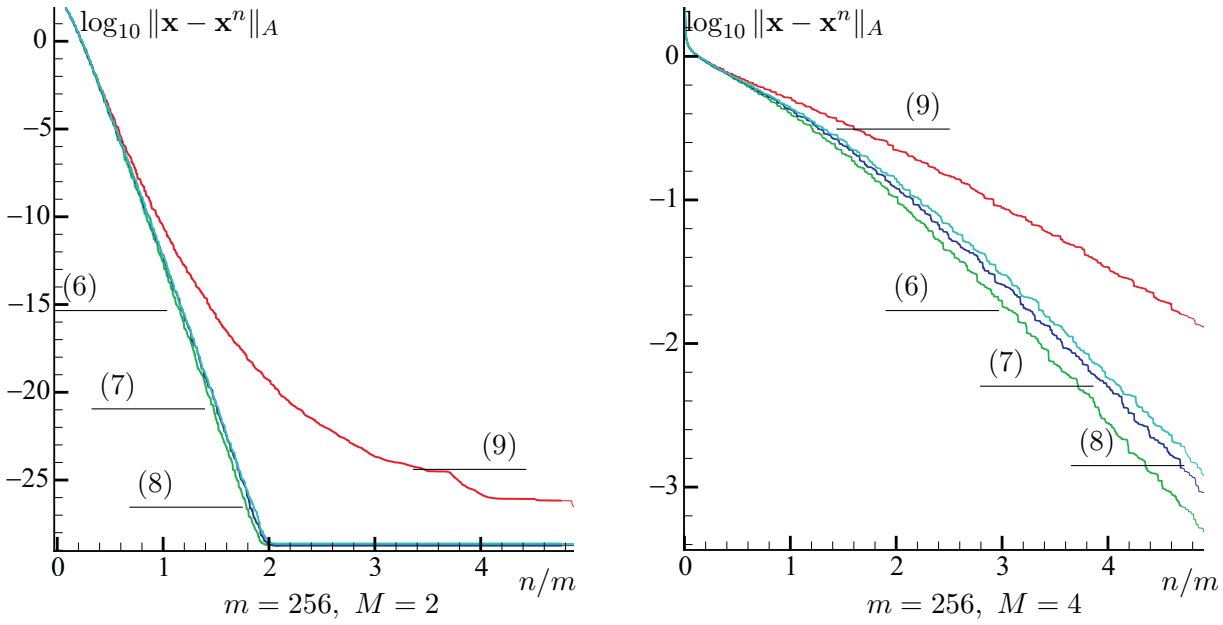

Рис. 1. Изменение по шагам MCG энергетической нормы невязки $\mathbf{x}-\mathbf{x}^n$; (6)–(9) — версии MGG (то же на рис. 2)

с матрицами вида $A_{ii} = i^{-M(i-1)}$, $i = 1, 2, \ldots, m$; здесь M — натуральное число. Число обусловленности такой матрицы $\mathrm{cond}\mathbf{A} = A_{11}/A_{mm} = (m-1)^{M(m-1)}$ очень быстро увеличивается с ростом m и M: если $m = 256$, то при $M = 2$ $\mathrm{cond}\,\mathbf{A} \simeq 6.6 \cdot 10^4$, а при $M = 2$ $\mathrm{cond}\,\mathbf{A} \simeq 4.3 \cdot 10^9$. Вычисления проводились в режиме "double", при котором относительная погрешность представления чисел в компьютере не превышает величину $\varepsilon_c = 2^{-52} \approx 2.22045 \cdot 10^{-16}$, а наименьшее, отличное от нуля, положительное число равно $2^{-1022} \approx 2.22507 \cdot 10^{-308}$ (машинный ноль). Эти характеристики вычислительного процесса можно легко вычислить [13].

Отметим монотонность приведенных зависимостей, характер которых в основном определяется числом обусловленности матрицы системы и величиной ε_c. Это обстоятельство говорит о том, что целесообразно не прерывать итерации после m шагов [2]. Кроме того, "экспериментально" установлено, что наименее эффективной оказывается релаксационная версия MCG (4), (9), в то время как алгоритм (4), (6), обеспечивающий на каждом шаге MCG условия сопряженности $\langle \mathbf{r}^n, \mathbf{r}^{n-1} \rangle = \langle \mathbf{s}^{n+1}, \mathbf{s}^n \rangle_A = 0$, оказался самым точным из рассматриваемых.

3. Критерии останова итераций MCG

Весьма сложным вопросом численной реализации MCG является определение рационального числа итераций, т. е. момента, начиная с которого полученное приближенное решение $\mathbf{x}^n$ не может быть существенно улучшено и итерационный процесс целесообразно прекратить.

Наиболее естественным было бы остановить процесс в тот момент, когда величина $\|\mathbf{x} - \mathbf{x}^n\|_A$ становится соизмеримой с погрешностью ее вычисления. Линейная теория накопления ошибок вычислений [7, 13, 14] позволяет, отправляясь от точностной характеристики вычислений ε_c, получить верхнюю оценку $|\delta\|\mathbf{x} - \mathbf{x}^n\|_A|$ погрешности вычисления $\|\mathbf{x} - \mathbf{x}^n\|_A$. Рисунок 2 иллюстрирует типовой характер зависимости $|\delta\|\mathbf{x} - \mathbf{x}^n\|_A|$ от относительного номера итерации n/m.

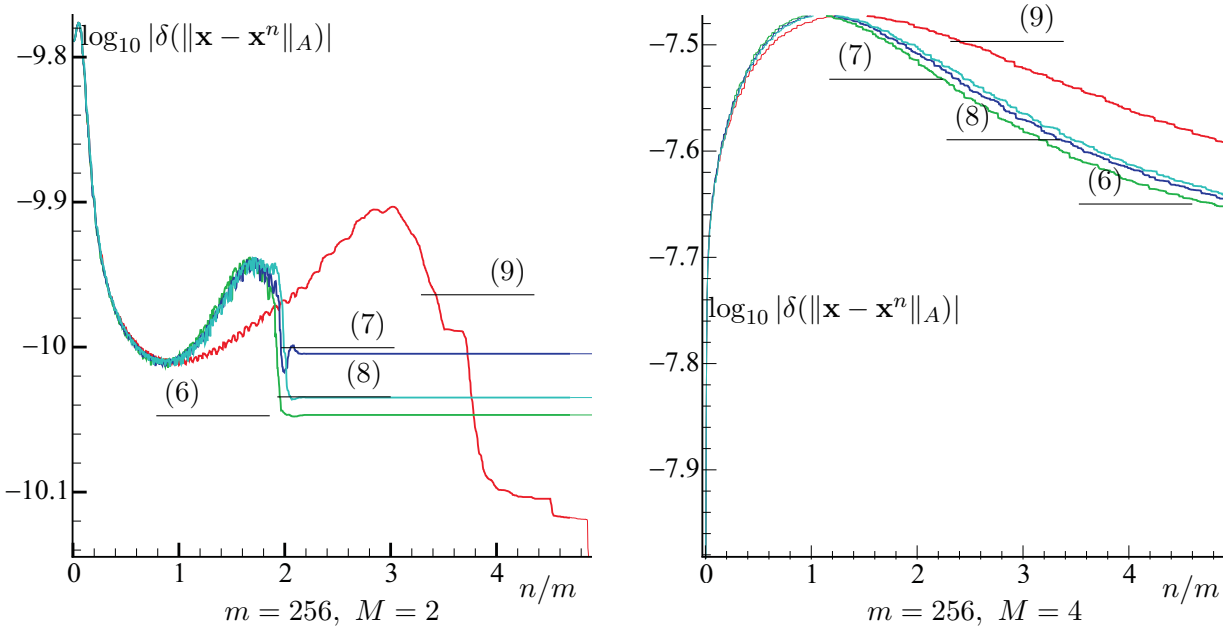

Рис. 2. Верхняя оценка абсолютной погрешности вычисления энергетической нормы невязки

Анализ численных экспериментов и "затратность" вычисления как энергетической нормы, так и верхней (как правило, весьма завышенной) оценки погрешности ее вычисления приводят к выводу, что использовать критерии останова итерационного процесса MCG на базе линейной теории накопления ошибок нецелесообразно.

В работе [10] предложено оценивать накопленную погрешность вычислений, анализируя векторы $\mathbf{r}^n$ и $\mathbf{A}\mathbf{x}^n - \mathbf{b}$. Численные эксперименты показали, что величины $\log_{10}\|\mathbf{r}^n\|$ и $\log_{10}\|\mathbf{A}\mathbf{x}^n - \mathbf{b} - \mathbf{r}^n\|$ для MCG-версий (4), (6)–(4), (8) меняются с ростом n схожим образом (рис. 3, *а*), тогда как соотношения (4), (9) дают несколько иную картину (рис. 3, *б*).

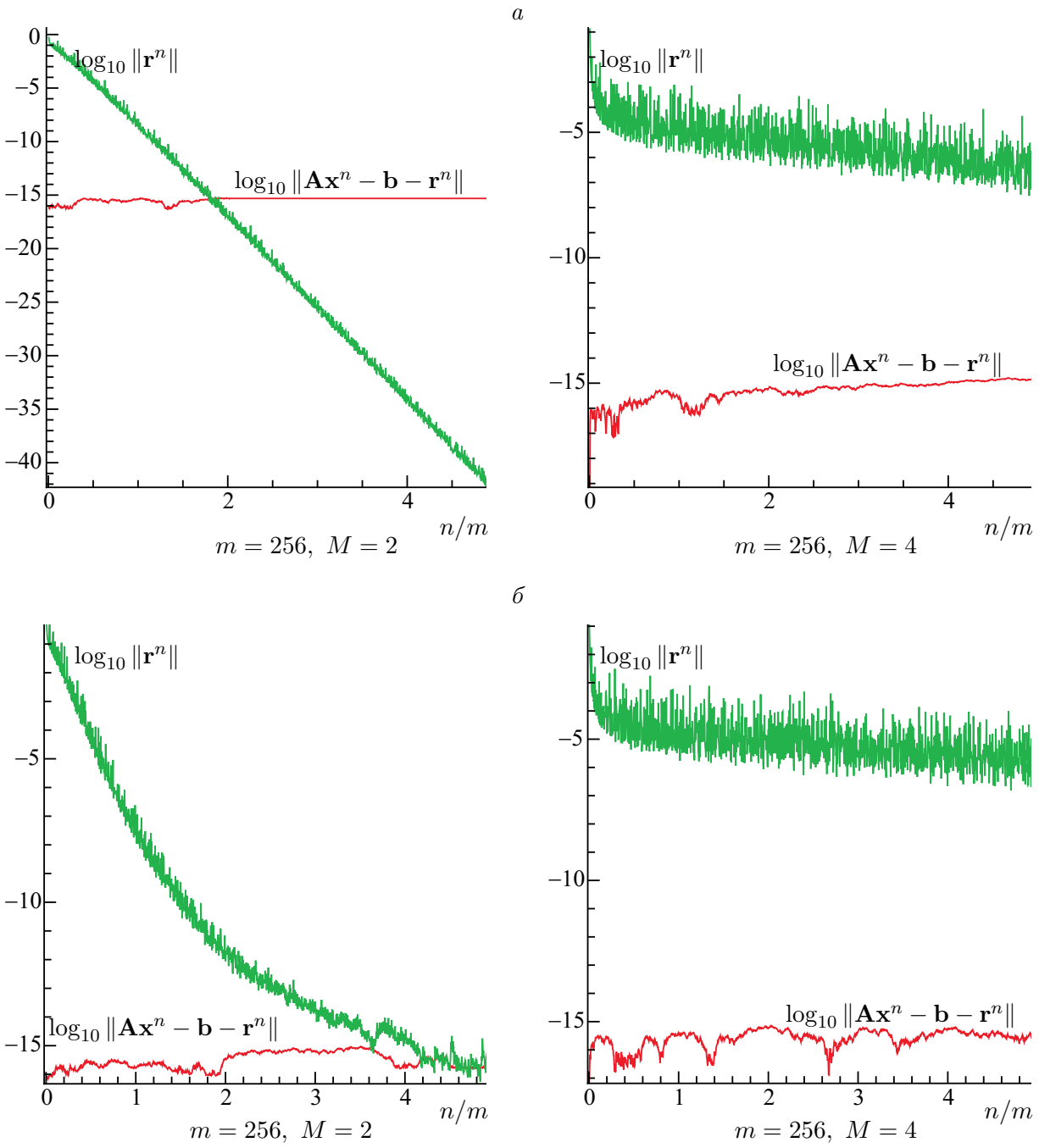

Рис. 3. Изменение $\|\mathbf{r}^n\|$ и $\|\mathbf{A}\mathbf{x}^n - \mathbf{b} - \mathbf{r}^n\|$ по шагам MCG на основе соотношений (4), (7) (*а*) и (4), (9) (*б*)

Общим для рассматриваемых версий MCG является то, что для систем, числа обусловленности которых не очень велики, евклидова норма $\|\mathbf{A}\mathbf{x}^n - \mathbf{b} - \mathbf{r}^n\|$ с ростом числа итераций меняется незначительно, тогда как величина $\|\mathbf{r}^n\|$ уменьшается начиная с некоторого шага. Поэтому в [2] в качестве критерия останова предлагается проверять неравенство

$$\|\mathbf{r}^n\| \le e^{(n/m)^2}\|\mathbf{A}\mathbf{x}^n - \mathbf{b} - \mathbf{r}^n\|, \tag{10}$$

где n — номер текущей итерации, m — порядок исходной системы уравнений (1); при нарушении (10) итерации прекращаются. Экспоненциальный множитель $e^{(n/m)^2}$ призван, наряду с ограничением числа шагов сверху [2] величиной $5m$, бороться с плохой обусловленностью системы уравнений.

Каждая из приведенных выше версий MCG по своей природе есть разновидность метода спуска решения задачи минимизации квадратичного функционала (2), а характерной особенностью таких алгоритмов [16] является высокая производительность первых итераций. Поэтому можно отследить выполнение неравенства (10), рассматривая лишь $\mathbf{s}^{n+1}$-составляющие соответствующих векторов.

Если ввести $\mathbf{e}_s^n = \mathbf{s}^n/\|\mathbf{s}^n\|$, то

$$\langle \mathbf{e}_s^{n+1}, \mathbf{A}\mathbf{x}^n - \mathbf{b} - \mathbf{r}^n\rangle = \left(\langle \mathbf{A}\mathbf{s}^{n+1}, \mathbf{x}^n\rangle - \langle \mathbf{s}^{n+1}, \mathbf{b}\rangle - \langle \mathbf{s}^{n+1}, \mathbf{r}^n\rangle\right)/\|\mathbf{s}^{n+1}\|,$$

и для определения момента окончания итерационного процесса вместо (10) проверяем неравенство

$$|\langle \mathbf{s}^{n+1}, \mathbf{r}^n\rangle| \le e^{(n/m)^2}|\langle \mathbf{A}\mathbf{s}^{n+1}, \mathbf{x}^n\rangle - \langle \mathbf{s}^{n+1}, \mathbf{b}\rangle - \langle \mathbf{s}^{n+1}, \mathbf{r}^n\rangle|. \tag{11}$$

Так как величины $\langle \mathbf{s}^n, \mathbf{r}^{n-1}\rangle$ и $\mathbf{A}\mathbf{s}^n$ определяются на каждом шаге MCG при построении коэффициентов α_n и β_n, то на очередной итерации можно не вычислять произведение матрицы $\mathbf{A}$ на вектор $\mathbf{x}^n$ [17], а достаточно найти лишь скалярные произведения $\langle \mathbf{A}\mathbf{s}^{n+1}, \mathbf{x}^n\rangle$ и $\langle \mathbf{s}^{n+1}, \mathbf{b}\rangle$, что значительно уменьшает вычислительные затраты.

Совместное изменение величин $|\langle \mathbf{e}^{n+1}, \mathbf{b}\rangle|$ и $|\langle \mathbf{e}_s^{n+1}, \mathbf{A}\mathbf{x}^n - \mathbf{b} - \mathbf{r}^n\rangle|$ отражено на рис. 4, *а* для версии (4), (7) и на рис. 4, *б* для версии (4), (9). Сравнение последних графиков с рис. 3 показывает хорошую коррелированность последовательностей $\{|\langle \mathbf{e}^{n+1}, \mathbf{r}^n\rangle|\}$ и $\{\|\mathbf{r}^n\|\}$, $\{|\langle \mathbf{e}_s^{n+1}, \mathbf{A}\mathbf{x}^n - \mathbf{b} - \mathbf{r}^n\rangle|\}$ и $\{\|\mathbf{A}\mathbf{x}^n - \mathbf{b} - \mathbf{r}^n\|\}$ для рассматриваемых версий MCG.

Анализ вычислительных экспериментов [18] показывает, что рассмотренные критерии останова наименее эффективно работают в релаксационной (4), (9) версии MCG, для которой, однако, возможен и иной критерий остановки вычислений.

Рассмотрим последовательность $\{a_n\}$ числовых значений

$$a_n = 2a(\mathbf{x}^n), \tag{12}$$

члены которой, согласно (9), можно вычислить по рекуррентной формуле

$$a_n = a_{n-1} - \left(\frac{\langle \mathbf{x}^{n-1}, \mathbf{s}^n\rangle_A - \langle \mathbf{b}, \mathbf{s}^n\rangle}{\|\mathbf{s}^n\|_A}\right)^2, \quad a_0 = \langle \mathbf{r}^0, \mathbf{x}^0\rangle - \langle \mathbf{b}, \mathbf{x}^0\rangle. \tag{13}$$

С другой стороны, исходя из определения (2) квадратичного функционала $a(\mathbf{y})$, число a_n можно выразить через $\mathbf{r}^n$ и $\mathbf{x}^n$:

$$a_n = \langle \mathbf{r}^n, \mathbf{x}^n\rangle - \langle \mathbf{b}, \mathbf{x}^n\rangle. \tag{14}$$

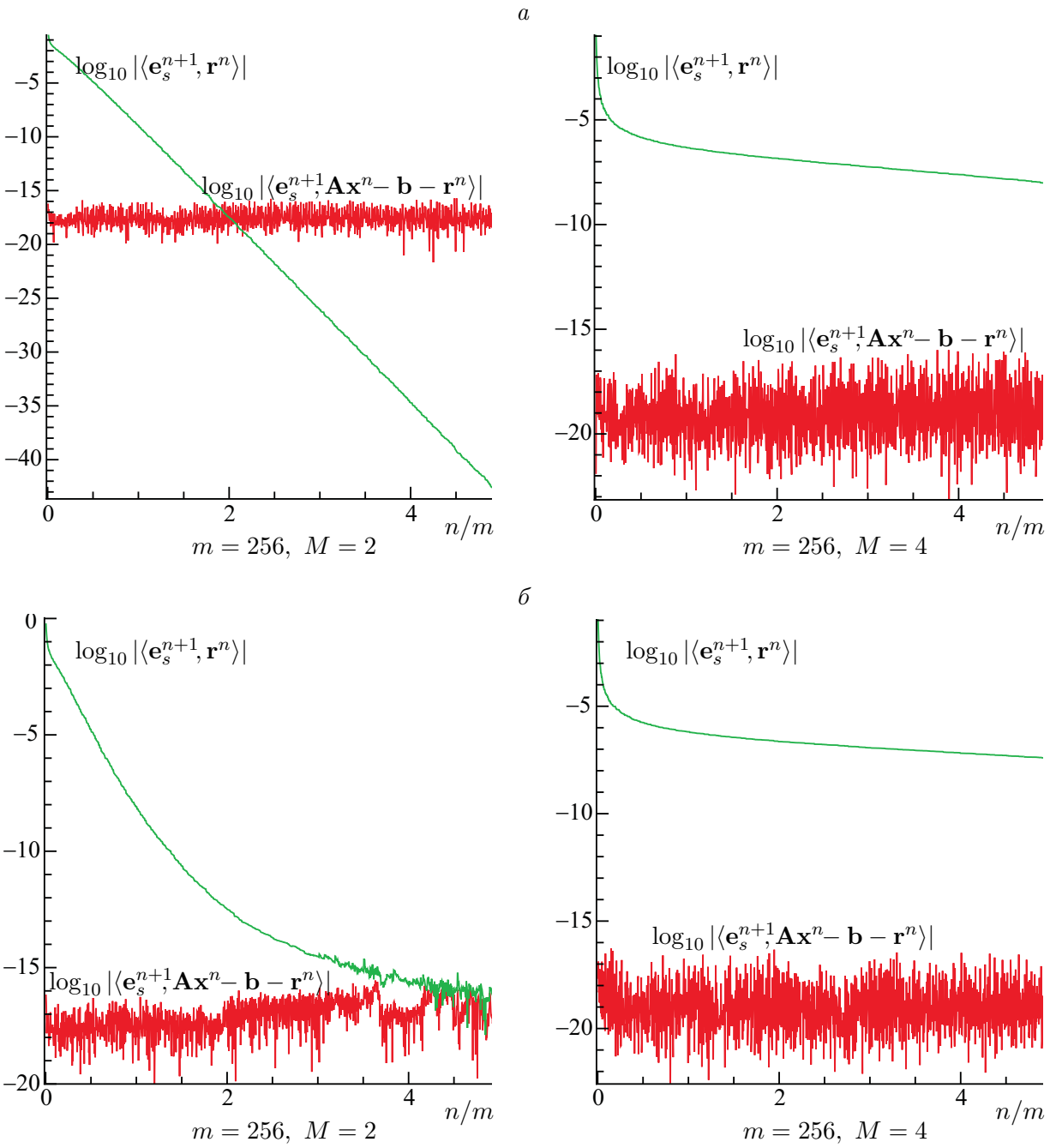

Рис. 4. Изменение модулей проекций векторов $\mathbf{r}^n$ и $\mathbf{A}\mathbf{x}^n - \mathbf{b} - \mathbf{r}^n$ на направление спуска $\mathbf{s}^{n+1}$ по шагам MCG на основе соотношений (4), (7) (*а*) и (4), (9) (*б*)

Замечание к формуле (2) наводит на мысль, что разность между вычисленными по формулам (13) и (14) значениями a_n может служить оценкой накопленной в процессе вычислений погрешности определения $\mathbf{x}^n$, и требование, чтобы на каждой итерации величина $|a_n - a_{n-1}|$ не превышала этой погрешности, является вполне естественным.

Введем

$$b_n = \langle \mathbf{b}, \mathbf{x}^n\rangle, \quad c_n = \langle \mathbf{r}^n, \mathbf{x}^n\rangle. \tag{15}$$

Тогда из (4) следует

$$b_n = \langle \mathbf{b}, \mathbf{x}^n\rangle = \langle \mathbf{b}, \mathbf{x}^{n-1} + \alpha_n \mathbf{s}^n\rangle = b_{n-1} + \alpha_n \langle \mathbf{b}, \mathbf{s}^n\rangle, \quad b_0 = \langle \mathbf{b}, \mathbf{x}^0\rangle.$$

Аналогично для чисел c_n получаем

$$c_n = \langle \mathbf{r}^n, \mathbf{x}^n \rangle = \langle \mathbf{r}^n, \mathbf{x}^{n-1} + \alpha_n \mathbf{s}^n \rangle = \langle \mathbf{r}^{n-1} + \alpha_n \mathbf{A}\mathbf{s}^n, \mathbf{x}^{n-1} \rangle + \alpha_n \langle \mathbf{r}^n, \mathbf{s}^n \rangle =$$
$$= c_{n-1} + \alpha_n \langle \mathbf{x}^{n-1}, \mathbf{s}^n \rangle_A + \alpha_n \langle \mathbf{r}^n, \mathbf{s}^n \rangle, \quad c_0 = \langle \mathbf{r}^0, \mathbf{x}^0 \rangle.$$

Теперь можем сформулировать для MCG в форме (4), (9) критерий прекращения вычислений: считая $\mathbf{x}^0$ заданным,

при $n = 0$, *вычисляем* $\mathbf{r}^0 = \mathbf{A}\mathbf{x}^0 - \mathbf{b}$ *и полагаем* $\mathbf{s}^1 = \mathbf{r}^0$;

$$a_0 = \langle \mathbf{r}^0, \mathbf{x}^0 \rangle - \langle \mathbf{b}, \mathbf{x}^0 \rangle, \quad b_0 = \langle \mathbf{b}, \mathbf{x}^0 \rangle, \quad c_0 = \langle \mathbf{r}^0, \mathbf{x}^0 \rangle;$$

при $n > 0$ *вычисляем* $\langle \mathbf{x}^{n-1}, \mathbf{s}^n \rangle_A$, $\langle \mathbf{b}, \mathbf{s}^n \rangle$, $\|\mathbf{s}^n\|_A^2$ *и* $\langle \mathbf{r}^n, \mathbf{s}^n \rangle_A$ *и определяем*

$$a_n = a_{n-1} - \left(\frac{\langle \mathbf{s}^n, \mathbf{x}^{n-1} \rangle_A - \langle \mathbf{s}^n, \mathbf{b} \rangle}{\|\mathbf{s}^n\|_A} \right)^2,$$
$$b_n = b_{n-1} + \alpha_n \langle \mathbf{b}, \mathbf{s}^n \rangle,$$
$$c_n = c_{n-1} + \alpha_n \left(\langle \mathbf{x}^{n-1}, \mathbf{s}^n \rangle_A + \langle \mathbf{r}^n, \mathbf{s}^n \rangle \right),$$
$$\varepsilon_n = |c_n - b_n - a_n|;$$

тогда, если на шаге n *MCG (4), (9) окажется, что*

$$\left(\frac{\langle \mathbf{s}^n, \mathbf{x}^{n-1} \rangle_A - \langle \mathbf{s}^n, \mathbf{b} \rangle}{\|\mathbf{s}^n\|_A} \right)^2 \le e^{(n/m)^2} \varepsilon_n, \tag{16}$$

то вычисления прекращаем.

Характер совместного изменения величин $|2a(\mathbf{x}^n) - 2a(\mathbf{x}^{n-1})|$ и $|2a(\mathbf{x}^n) - a_n|$ отражен на рис. 5. Из сравнения этих данных с рис. 2 можно сделать вывод, что точка пересечения графиков величин $|2a(\mathbf{x}^n) - 2a(\mathbf{x}^{n-1})|$ и $|2a(\mathbf{x}^n) - a_n|$ соответствует тому значению n, при котором начинает устойчиво возрастать верхняя оценка $|\delta \|\mathbf{x} - \mathbf{x}^n\|_A|$ погрешности вычислений энергетической нормы невязки $\|\mathbf{x} - \mathbf{x}^n\|_A$.

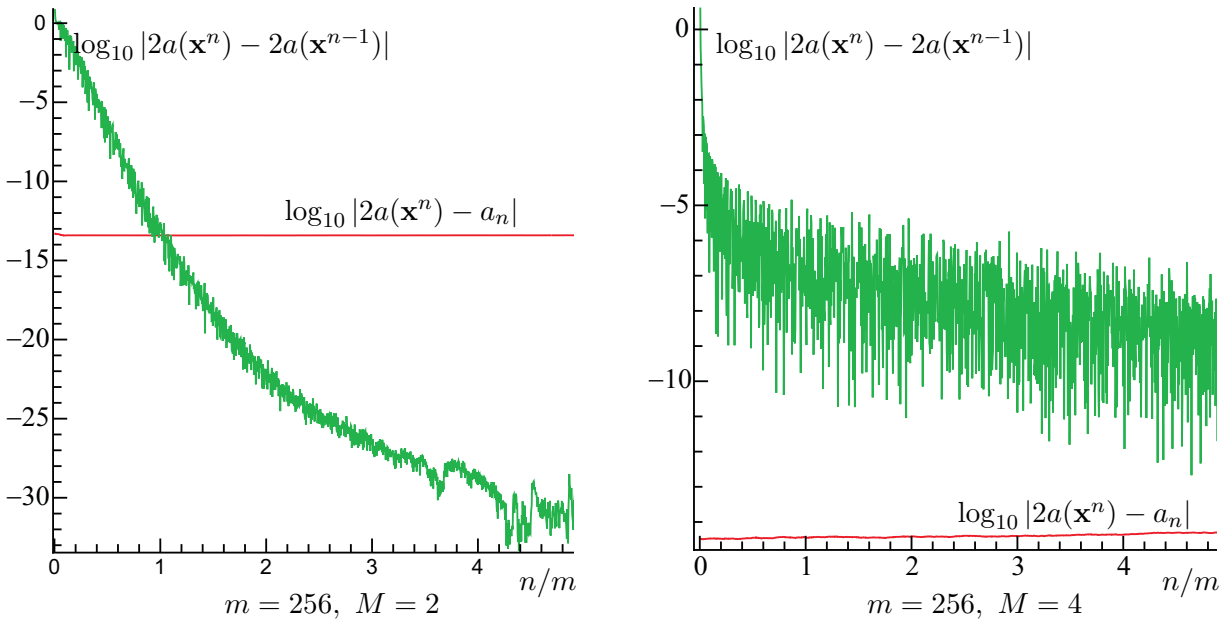

Рис. 5. Пошаговое для MCG на основе соотношений (4), (9) поведение величин $|2a(\mathbf{x}^n) - a_n|$ и $|2a(\mathbf{x}^n) - 2a(\mathbf{x}^{n-1})|$ для различных чисел обусловленности

Заключение

В работе для четырех наиболее распространенных программных реализаций метода сопряженных градиентов решения систем линейных алгебраических уравнений численно исследовано пошаговое поведение энергетических норм невязок с точным решением при наличии погрешностей округления. Численные эксперименты показали, что все рассмотренные версии метода приближают искомое решение примерно с одинаковой погрешностью при числе итераций, не большем числа неизвестных. Однако картина существенно меняется для итераций с номерами, превышающими размерность системы в 3–5 раз. В этом диапазоне шагов наиболее точное приближение дает версия метода, в которой на каждом шаге обеспечиваются ортогональность в энергетической метрике направлений спуска и евклидова перпендикулярность векторов невязок, вычисляемых во время процесса. Наиболее медленным оказался алгоритм, в котором параметры итерационного процесса выбирались из условия его релаксационности. Отношение энергетических норм невязок в этих двух случаях достигало трех десятичных порядков.

Проведен обзор существующих программно независимых условий прекращения вычислений, при выполнении которых предельно возможная точность численного решения считается достигнутой. Для релаксационной версии алгоритма метода сопряженных градиентов предложен новый критерии останова итерационного процесса.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (грант № 14-01-00130).

Список литературы / References

[1] **Hestenes, M.R. and Stiefel, E.** Methods of conjugate gradients for solving linear systems // J. of Res. of National Bureau of Standards. 1952. Vol. 49, No. 5. P. 409–435.

[2] **Wilkinson, J.H. and Reinsch, C.** Handbook for Automatic Computation. Vol. II. Linear Algebra. N.Y.: Springer-Verlag, 1971. 450 p.

[3] **Cipra, B.A.** The best of the 20th century: Editors name top 10 algorithms // Comput. Sci. Eng. 2000. Vol. 332, No. 4. P. 291–293.

[4] **Крылов А.Н.** О численном решении уравнения, которым в технических вопросах определяются частоты малых колебаний материальных систем // Изв. АН СССР. Отделение математических и естественных наук. 1931. № 4. С. 491–539.
Krylov, A.N. On the numerical solution of the equation by which in technical questions frequencies of small oscillations of material systems are determined // Izv. AN SSSR (News of Academy of Sciences of the USSR). Otdel. Matem. i Estest. Nauk. 1931. No. 4. P. 491–539. (in Russ.)

[5] **Golub, G.H. and O'Leary, D.P.** Some history of the conjugate gradient and Lanczos algorithms: 1948–1976 // SIAM Rev. 1989. Vol. 31, No. 1. P. 50–102.

[6] **Воеводин В.В.** Вычислительные основы линейной алгебры. М.: Наука, 1977. 304 с.
Voevodin, V.V. Numerical Foundations of Linear Algebra. Moscow: Nauka, 1977. 304 p. (in Russ.)

[7] **Greenbaum, A.** Estimating the attainable accuracy of recursively computed residual methods. Techn. Rep. TR-95-1515. Department of Comput. Sci., Cornell Univ., Ithaca. N.Y. (May 1995).

[8] **Greenbaum, A., Strakos, Z.** Predicting the Behavior of Finite Precision Lanczos and Conjugate Gradient Computations. Techn. Rep. TR-538, Department of Comput. Sci., Cornell Univ., Ithaca. N.Y. (January 1991).

[9] **Воеводин В.В.** Линейная алгебра. М.: Наука, 1980. 400 с.
Voevodin, V.V. Linear Algebra. Moscow: Nauka, 1980. 400 p. (in Russ.)

[10] **Engeli, M., Ginsburg, Th., Rutishauser, H., Stiefel, E.** Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems. Mitteilungen aus dem Institut für angewandte Mathematik der ETH Zürich, No. 8. Basel, Stuttgart: Birkhaüser Verlag, 1959. 107 p.

[11] **Воеводин В.В.** Численные методы алгебры. Теория и алгорифмы. М.: Наука, 1966. 248 с.
Voevodin, V.V. Computational Methods of Algebra. Theory and Algorithms. Moscow: Nauka, 1966. 248 p. (in Russ.)

[12] **Карманов В.Г.** Математическое программирование: Учеб. пособие. М.: Физматлит, 2004. 264 с.
Karmanov, V.G. Mathematical Programming: Textbook. Moscow: Fizmatlit, 2004. 264 p. (in Russ.)

[13] **Малышев А.Н.** Введение в вычислительную линейную алгебру. Новосибирск: Наука, 1991. 229 с.
Malyshev, A.N. Introduction to Computational Linear Algebra. Novosibirsk: Nauka, 1991. 229 p. (in Russ.)

[14] **Годунов С.К.** Решение систем линейных уравнений. Новосибирск: Наука, 1980. 178 с.
Godunov, S.K. Solving Systems of Linear Equations. Novosibirsk: Nauka, 1980. 178 p. (in Russ.)

[15] **Meurant, G. and Strakoš, Z.** The Lanczos and conjugate gradient algorithms in finite precision arithmetic // Acta Numerica. 2006. Vol. 15. P. 471–542.

[16] **Фаддеев Д.К., Фаддеева В.Н.** Вычислительные методы линейной алгебры. М.: Физматгиз, 1963. 229 с.
Faddeev, D.K., Faddeeva, V.N. Computational Methods in Linear Algebra. Moscow: Fizmatgiz, 1963. 229 p. (in Russ.)

[17] **Чернышёва А.А., Киреев И.В.** Модификация критерия Уилкинсона остановки итераций в методе сопряженных градиентов // Вест. КрасГУ. 2005. № 4. С. 173–177.
Chernysheva, A.A., Kireev, I.V. Modification Wilkinson's criteria stopping iterations in the conjugate gradient method // Bulletin of Krasnoyarsk State Univ. 2005. No. 4. P. 173–177. (in Russ.)

[18] **Киреев И.В.** Численная реализация метода сопряженных градиентов. Красноярск, 2013. (Препр. ИВМ СО РАН; № 13-1. 26 с.)
Kireev, I.V. The Computational Implementation of the Conjugate Gradient Method. Krasnoyarsk, 2013. (Preprint. ICM SB RAS; № 13-1. 26 p.) (in Russ.)

Поступила в редакцию 1 октября 2014 г.,
с доработки — 9 февраля 2015 г.

Inexpensive stopping criteria in the conjugate gradient method

Kireev, Igor' V.

Institute of Computational Modelling SB RAS, Krasnoyarsk, 660036, Russia

Corresponding author: Kireev, Igor' V., e-mail: kiv@icm.krasn.ru

In the paper, some aspects of the numerical implementation of the conjugate gradient method (CGM) for systems of linear algebraic equations with symmetric positive definite matrix in the presence of round-off errors are discussed. With exact calculations, CGM provides an exact solution in a finite number of iteration steps. But in fact CGM is an iterative process and the weak point in an iterative process is in a stopping criterion. It is required to determine the number of the iteration step, after which the accuracy of an approximation to a solution of a system of linear equations may not be considerably improved with a particular computer. Hence, the construction of inexpensive stopping criteria for CGM being the aim of this paper is an urgent problem. For four popular versions of CGM, the step-by-step behavior as well as stopping criteria for an iterative process are considered. Numerical results show that the most accurate approximation is achieved by the CGM-version where descent directions and residual vectors are orthogonal in the energy and Euclidean metrics, respectively, at each iteration step. A practical stopping criteria for CGM is proposed as a formula that enables one to determine the number of the CGM iteration step, starting with which the progress is no longer being made. The application of the constructed criteria to the solution of specific systems of linear algebraic equations with ill-conditioned matrices is demonstrated.

Keywords: conjugate gradient method, stopping criteria.

Acknowledgements. This work has been supported by the Russian Foundation for Basic Research (RFBR) grant No. 14-01-00130.

Received 1 October 2014
Received in revised form 9 February 2015