

AMS subject classification: 65F10, 65G99

Формулы оценки ошибок и их анализ для CG, Bi-CG и GMRES

П. Джайн, К. Манглани, М. Венкатапати

Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, 560012, India

E-mails: puneet798@gmail.com (Джайн П.), krish2100@gmail.com (Манглани К.), muruges@iisc.ac.in (Венкатапати М.)

Английская версия этой статьи печатается в журнале “Numerical Analysis and Applications” № 2, Vol. 16, 2023.

Джайн П., Манглани К., Венкатапати М. Формулы оценки ошибок и их анализ для CG, Bi-CG и GMRES // Сиб. журн. вычисл. математики / РАН. Сиб. отд-ние. — Новосибирск, 2023. — Т. 26, № 2. — С. 161–181.

Современные требования к точности в измерениях и инженерных моделях увеличивают число обусловленности задач. В то время как увеличение точности чисел с плавающей запятой привело к стабильным вычислениям, увеличилась неопределенность в вопросе сходимости при использовании невязки в качестве критерия останова. Мы представляем анализ неопределенности сходимости при использовании относительной невязки в качестве критерия останова для итеративного решения линейных систем, а также получаемое увеличение/уменьшение вычислений при заданной допустимой погрешности. Это показывает, что оценка ошибки важна для эффективного или точного решения, даже когда число обусловленности матрицы невелико. Формула оценки ошибки $\mathcal{O}(1)$ для итераций алгоритма CG была предложена более двух десятилетий назад. Недавно формула оценки ошибки $\mathcal{O}(k^2)$ была описана для алгоритма GMRES, который допускает также несимметричные линейные системы, где k — номер итерации. Мы предлагаем небольшую модификацию этой оценки ошибки GMRES для повышения устойчивости. В данной работе мы также предлагаем формулу оценки ошибки $\mathcal{O}(n)$ для A -нормы и l_2 -нормы вектора ошибки в алгоритме Bi-CG. Надежная работа этих оценок в качестве критерия останова увеличивает экономию и точность вычислений по мере увеличения числа обусловленности и размера задач.

DOI: 10.15372/SJNM20230204

Ключевые слова: ошибка, критерии останова, число обусловленности, сопряженные градиенты, Bi-CG, GMRES.

Jain P., Manglani K., Venkatapathi M. Error estimators and their analysis for CG, Bi-CG and GMRES // Siberian J. Num. Math. / Sib. Branch of Russ. Acad. of Sci. — Novosibirsk, 2023. — Vol. 26, № 2. — P. 161–181.

The demands of accuracy in measurements and engineering models today render the condition number of problems larger. While a corresponding increase in the precision of floating point numbers ensured a stable computing, the uncertainty in convergence when using residue as a stopping criterion has increased. We present an analysis of the uncertainty in convergence when using relative residue as a stopping criterion for iterative solution of linear systems, and the resulting over/under computation for a given tolerance in error. This shows that error estimation is significant for an efficient or accurate solution even when the condition number of the matrix is not large. An $\mathcal{O}(1)$ error estimator for iterations of the CG algorithm was proposed more than two decades ago. Recently, an $\mathcal{O}(k^2)$ error estimator was described for the GMRES algorithm which allows for non-symmetric linear systems as well, where k is the iteration number. We suggest a minor modification in this GMRES error estimation for increased stability. In this work, we also propose an $\mathcal{O}(n)$ error estimator for A -norm and l_2 -norm of the error vector in Bi-CG algorithm. The robust performance of these estimates as a stopping criterion results in increased savings and accuracy in computation, as condition number and size of problems increase.

Keywords: error, stopping criteria, condition number, Conjugate Gradients, Bi-CG, GMRES.

1. Введение

Решение системы линейных уравнений вида $Ax = b$ является повсеместным требованием в науке и технике (где A — заданная матрица; x и b — неизвестный и известный векторы соответственно; $x \in \mathbb{R}^n$ и $b \in \mathbb{R}^n$, если $A \in \mathbb{R}^{n \times n}$). Итерационные методы, такие как метод сопряженных градиентов (CG) [7], метод бисопряженных градиентов (Bi-CG) [4] и обобщенный метод минимальной невязки (GMRES) [14], в основном известные как методы подпространства Крылова, обычно используются для решения больших линейных систем. При использовании таких методов итерации следует останавливать, когда норма ошибки $\epsilon_k = x - x_k$ меньше желаемого допуска, где x — неизвестное решение линейной системы, x_k — его итерационная аппроксимация на k -й итерации. Поскольку фактическая ошибка неизвестна, относительная невязка $\frac{\|r_k\|}{\|b\|}$ считается критерием остановки, когда $r_k = b - Ax_k$ — вектор невязки на k -й итерации. Такие критерии остановки могут использоваться, когда система хорошо обусловлена, но они могут быть ненадежными даже при умеренном увеличении относительного числа обусловленности A ($\kappa \gtrsim 10$, где $\kappa = \|A\|_2 \|A^{-1}\|_2$) в зависимости от выбора начального приближения. Хорошо известно, что использование относительной невязки может привести к слишком ранней остановке итераций, когда норма ошибки намного превышает допуск, или к недостаточно ранней остановке, когда для требуемой точности должно выполняться слишком много операций с плавающей запятой [7, 12]. Более того, когда число обусловленности матрицы велико, невязка алгоритмов, таких как Bi-CG, не обязана иметь монотонное поведение и осциллирует [1], тогда как фактическая ошибка может (хотя и медленно) сходиться, и аналогичным образом, колебания ошибок в алгоритме GMRES возможны, когда невязка сходится медленно. Отметим, что l_2 -нормы относительных ошибок и невязок могут значительно отличаться, как показано ниже:

$$\kappa \frac{\|r_k\|}{\|b\|} \geq \frac{\|\epsilon_k\|}{\|x\|} \geq \frac{\|r_k\|}{\|b\|} \frac{1}{\kappa}. \quad (1)$$

Даже при использовании итерационных алгоритмов с предобуславливателями, уменьшающими число обусловленности матрицы, уменьшение числа обусловленности решаемой обратной задачи не гарантируется, как это требуется для более быстрой сходимости (см. пункт 3.3.1). Это наблюдается для матриц больших размерностей или более высоких чисел обусловленности, что во многих случаях делает использование предобуславливателей неэффективным. Более того, большинство методов с использованием предобуславливателей часто не работает в случае несимметричных и неопределенных матриц [13]. Таким образом, уже для $\kappa(A) \gtrsim 10^2$ либо точность, либо эффективность вычислений ухудшаются из-за вышеуказанной дилеммы.

Вследствие требований к точности в измерениях и технике как размер, так и число обусловленности большинства задач должны быть большими; это делает точные критерии остановки и перезапуска очень важными для обеспечения вычислительной эффективности решателей. Это послужило толчком к появлению методов оценивания норм ошибки в итерационных решателях. Аппроксимации ошибок x_k для итерационного метода были предложены более двух десятилетий назад [2, 3]. Также были представлены оценки ошибок для алгоритма CG [11], а для решения несимметричных линейных систем с использованием методов полной ортогонализации (FOM) и GMRES-формулы для оценки ошибок были предложены совсем недавно [10]. Мы предлагаем небольшую модификацию этой формулы для оценки GMRES для повышения ее стабильности и точности.

Мы также получим эффективную формулу оценки ошибки при решении несимметричных линейных систем с использованием Bi-CG.

В следующих пунктах мы представляем анализ, показывающий важность этих оценок ошибок для методов CG, Bi-CG и GMRES. Мы показываем, что использование относительной невязки в качестве оценки истинной относительной ошибки менее надежно, чем использование формулы для оценки ошибки. Мы оцениваем эту “неопределенность” и показываем, что ее *ожидаемое* значение пропорционально числам обусловленности, т.е. прямым или обратным числам обусловленности или их средневзвешенным значениям, в зависимости от используемого алгоритма. Численные результаты, усредненные по большому числу линейных систем, используются для демонстрации теоремы неопределенности сходимости. Демонстрируется уменьшение/увеличение вычислений, вызванное неопределенностью, а выигрыши, благодаря использованию этих формул оценок, получить довольно просто. Наш анализ показывает, что эти формулы оценки ошибок являются надежными и значительно повышают эффективность/точность вычислений.

Пункт 2 кратко знакомит с представленной работой и предыдущими исследованиями по оценкам ошибок методов CG и GMRES. В п. 3 представлены модификация формулы оценки GMRES, предлагаемые оценки ошибки Bi-CG и анализ оценок ошибки. В п. 4 обсуждаются экспериментальные результаты по модифицированным оценкам ошибки GMRES, предлагаемые оценки ошибки Bi-CG и работа оценок ошибок в качестве эффективных критериев останова для этих итерационных методов решения линейных систем.

2. Предшествующие работы

2.1. Оценки ошибки для алгоритма CG

Один из наиболее часто используемых методов решения линейных систем с вещественной симметричной положительно определенной матрицей — алгоритм сопряженных градиентов. Квадратичный алгоритм CG Ланцоша (CGQL) [6, 11, 12] обеспечивает границы A -нормы ошибки, т.е. $\|x - x_k\|_A$ на каждой итерации CG с использованием правил Ланцоша и квадратур с задержкой в d итераций при оценивании. Границы квадрата A -нормы ошибки на итерации k для CG имеют следующий вид [6]:

$$\|\epsilon_k\|_A^2 \approx \|r_0\|_2^2 \left[(T_{k+d}^{-1})_{(1,1)} - (T_k^{-1})_{(1,1)} \right]. \quad (2)$$

Здесь T_k — трехдиагональная матрица, полученная с использованием алгоритма Ланцоша на k -й итерации, элементы которой могут быть вычислены одинаково хорошо из коэффициентов алгоритма CG; $(T_k^{-1})_{(1,1)}$ представляет $(1, 1)$ -элемент матрицы T_k^{-1} . Кроме того, ϵ и r — вектор ошибки и вектор невязки алгоритма CG соответственно.

Суть CGQL заключается в вычислении разницы между $(1, 1)$ -элементами обратных матриц к двум трехдиагональным матрицам, сгенерированным алгоритмом Ланцоша с такими же начальными векторами, что и алгоритм CG. Однако для вычислений были предложены другие эквивалентные формулы для аппроксимации ошибки CG [8], включая LDL^T -факторизацию соответствующей матрицы Якоби, имеющейся в CG [12]. Позднее были предложены оценки l_2 -нормы ошибки алгоритма CG [9], включая расширения для предобусловленных версий [18]. Следует также отметить, что при задержке оценивания d , $d > 1$, с учетом $\|\epsilon_{k+d}\|_A \ll \|\epsilon_k\|_A$ аппроксимация A -нормы ошибки на итерации k также задается с помощью коэффициентов α_j для CG следующим образом:

$$\|\epsilon_k\|_A^2 \approx \sum_{j=k}^{k+d-1} \alpha_j r_j^\top r_j.$$

2.2. Формула оценивания ошибки алгоритма GMRES

Обобщенный алгоритм минимальных невязок GMRES представляет собой итерационный метод решения линейных систем даже для общих несимметричных матриц [14]. Пусть V_k — матрица, столбцы которой являются ортонормированными базисными векторами v_j , $j = 1, \dots, k$, подпространства Крылова $\mathcal{K}_k(A, r_0)$, где r_0 — первоначальная невязка. Итерации GMRES определяются в виде $x_k = x_0 + V_k z_k$, где вектор z_k имеет веса для ортонормированных базисных векторов v_j в матрице V_k . При сохранении обозначений из [10] для оценок нормы ошибки FOM и GMRES на k -й итерации GMRES матрица Хессенберга H_k , получаемая в результате итерационного ортогонального преобразования матрицы системы, может быть разложена по блокам следующим образом:

$$H_k = \begin{pmatrix} H_{k-d} & W_{k-d} \\ Y_{k-d}^\top & \tilde{H}_{k-d} \end{pmatrix}.$$

Здесь H_{k-d} и $\tilde{H}_{k-d}$ — квадратная и верхняя матрицы Хессенберга соответственно, W_{k-d} в общем случае представляет собой полную прямоугольную матрицу.

Пусть $h_{i,j}$ — (i, j) -й элемент матрицы Хессенберга, e_k — k -й столбец единичной матрицы, вектор t_k — последний столбец $(H_k^\top H_k)^{-1}$ с $t_{k,k}$ в качестве его последнего элемента. Тогда переменные γ_{k-d} , w_{k-d} , δ_{k+1} и u_k можно определить следующим образом:

$$\gamma_{k-d} = \frac{h_{k-d+1,k-d}(e_{k-d}, H_{k-d}^{-1}e_1)}{1 - h_{k-d+1,k-d}(e_{k-d}, H_{k-d}^{-1}w_{k-d})},$$

$$w_{k-d} = W_{k-d}\tilde{H}_{k-d}^{-1}e_1, \quad \delta_{k+1} = \frac{h_{k+1,k}^2}{1 + h_{k+1,k}^2 t_{kk}}, \quad u_k = \delta_{k+1} t_k.$$

Оценку ошибки ϵ_{k-d} на $(k-d)$ -й итерации GMRES можно аппроксимировать следующим образом [10]:

$$\frac{\|\epsilon_{k-d}\|^2}{\|r_0\|^2} \approx \gamma_{k-d}^2 \left\| \tilde{H}_{k-d}^{-1} e_1 \right\|^2 + \left\| \gamma_{k-d} H_{k-d}^{-1} w_{k-d} + (e_{k-d}, H_{k-d}^{-1} e_1) u_{k-d} \right\|^2. \quad (3)$$

Обратите внимание, что приведенная выше оценка включает только одно умножение плотной матрицы на вектор и требует $\mathcal{O}(k^2)$ арифметических операций на каждой итерации, а это незначительно для всех $k \ll n$. Для несходящихся задач при $k \rightarrow n$ общая дополнительная стоимость, связанная с оценкой, может составлять почти треть стоимости итераций GMRES, однако это происходит редко в приложениях с большим n , и таким образом, в среднем формула оценки может обеспечить значительные преимущества.

3. Методы

3.1. Предлагаемая модификация формулы оценки ошибки GMRES

Представленную выше оценку ошибки в GMRES также можно переписать в виде

$$\frac{\|\epsilon_{k-d}\|^2}{\|r_0\|^2} = \underbrace{\frac{\|\epsilon_{k-d}\|^2 - \|\epsilon_k\|^2}{\|r_0\|^2}}_{\text{term1}} + \underbrace{\|s_k\|^2}_{\text{term2}} + \underbrace{\zeta_k}_{\text{term3}}, \quad (4)$$

где $s_k = (e_k, H_k^{-1} e_1) u_k$ и

$$\zeta_k = 2h_{k+1,k}(e_k, H_k^{-1}e_1) \left((H_n^{-1}e_{k+1})^k, H_k^{-1}e^1 + s_k \right) - \\ 2h_{k-d+1,k-d}(e_{k-d}, H_{k-d}^{-1}e_1) \left((H_n^{-1}e_{k-d+1})^{k-d} - (H_k^{-1}e_{k-d+1})^{k-d}, H_{k-d}^{-1}e_1 + s_{k-d} \right),$$

а индексы k и $k-d$ были введены для компонент вектора и представляют соответствующие усеченные размеры.

Правая часть (4) была разделена на три слагаемых (term1, term2, term3), где ζ_k предполагается пренебрежимо малым. Кроме того, H_n^{-1} , необходимая для оценивания ζ_k , не может быть вычислена на k -й итерации. Обратите внимание, что для большого d , который представляет собой задержку оценивания, второе слагаемое $\|s_k\|^2$ также становится пренебрежимо малым. Однако на практике d остается небольшим (≈ 10). Следовательно, в таких случаях term2 может внести значительный вклад в значение оценки ошибки. Мы предлагаем включить term2 в оценку ошибки. Операция взятия абсолютного значения в (5) необходима для сохранения положительной оценки l_2 -нормы ошибки с использованием усеченного выражения. Включив s_k в (3) и учитывая абсолютные значения, имеем

$$\frac{\|e_{k-d}\|^2}{\|r_0\|^2} \approx \left| \gamma_{k-d}^2 \left\| \tilde{H}_{k-d}^{-1}e_1 \right\|^2 + \left\| \gamma_{k-d} H_{k-d}^{-1}w_{k-d} + (e_{k-d}, H_{k-d}^{-1}e_1)u_{k-d} \right\|^2 - \|s_k\|^2 \right|. \quad (5)$$

3.2. Предлагаемая формула оценки A -нормы и l_2 -нормы ошибки в алгоритме Bi-CG

Тогда как A -норма ошибки в CG монотонно минимизируется с количеством итераций, A -норма ошибки в Bi-CG может осциллировать и может быть даже не определена, когда A — неопределенная матрица. Рассмотрим абсолютные значения билинейных произведений невязок Bi-CG:

$$\|\epsilon_k\|_A^2 = |\epsilon_k^\top A \epsilon_k| = |r_k^\top (A^\top)^{-1} r_k| = |r_k^\top A^{-1} r_k|. \quad (6)$$

Обратите внимание, что когда $A \in \mathbb{R}^{N \times N}$ и $r \in \mathbb{R}^N$, $r_k^\top (A^\top)^{-1} r_k$ — скалярная величина, которая всегда положительна, тогда A является положительно определенной. Кроме того, квадрат l_2 -нормы ошибки определяется следующим выражением:

$$\|\epsilon_k\|_2^2 = \epsilon_k^\top \epsilon_k = r_k^\top (A^\top)^{-1} A^{-1} r_k. \quad (7)$$

Если $A = A^\top$, то A -норма и l_2 -норма ошибки определяются как $r_k^\top A^{-1} r_k$ и $r_k^\top A^{-2} r_k$ соответственно. В следующих пунктах получены оценки A -нормы и l_2 -нормы ошибки на каждой итерации Bi-CG. Для справки: Bi-CG включен как алгоритм (см. п. Приложение).

3.2.1. Оценивание A -нормы ошибки

Записав последовательную разность между A -нормой ошибки на итерациях k и $k+1$, мы получим следующее соотношение, когда A — несимметричная матрица (см. п. Приложение для алгоритма Bi-CG):

$$r_k^\top (A^\top)^{-1} r_k - r_{k+1}^\top (A^\top)^{-1} r_{k+1} = r_k^\top (A^\top)^{-1} r_k - (r_k - \alpha_k A p_k)^\top (A^\top)^{-1} (r_k - \alpha_k A p_k) \\ = \alpha_k p_k^\top r_k + \alpha_k r_k^\top (A^\top)^{-1} A p_k - \alpha_k^2 p_k^\top A p_k.$$

В приведенном выше уравнении первый и третий члены в правой части могут быть вычислены обычным образом с использованием итераций алгоритма Bi-CG. Второй член включает вычисление A^{-1} . Следовательно, мы уменьшаем его. Поскольку $r_{k+1} = r_k - \alpha_k A p_k$ и $A p_k = \frac{r_k - r_{k+1}}{\alpha_k}$, мы получим следующее соотношение:

$$\begin{aligned} r_k^\top A^{-1} r_k - r_{k+1}^\top A^{-1} r_{k+1} &= \alpha_k r_k^\top p_k + \alpha_k r_k^\top (A^\top)^{-1} \left(\frac{r_k - r_{k+1}}{\alpha_k} \right) - \alpha_k^2 p_k^\top A p_k \\ &\Downarrow \\ r_{k+1}^\top A^{-1} r_{k+1} &= -\alpha_k r_k^\top p_k + r_{k+1}^\top A^{-1} r_k + \alpha_k^2 p_k^\top A p_k. \end{aligned}$$

Ошибка $\epsilon_k = x - x_k$ задается путем $A^{-1} r_k$. Кроме того, ϵ_k можно записать в виде взвешенной суммы направлений поиска от k до n :

$$\epsilon_k = \sum_{j=k}^n \alpha_j p_j \implies \epsilon_k - \epsilon_{k+d} = \sum_{j=k}^{k+d} \alpha_j p_j \implies \epsilon_k \approx \sum_{j=k}^{k+d} \alpha_j p_j. \quad (8)$$

Если ϵ_{k+d} обозначает ошибку на итерации $k+d$ и $\epsilon_{k+d} \ll \epsilon_k$ для некоторого $d > 0$, то следующими членами ряда можно пренебречь. Вектор ошибки в (8) может дать оценку A -нормы ошибки после задержки в d итераций:

$$\|\epsilon_{k+1}\|_A^2 = r_{k+1}^\top A^{-1} r_{k+1} \approx -\alpha_k r_k^\top p_k + r_{k+1}^\top \sum_{j=k}^{k+d} \alpha_j p_j + \alpha_k^2 p_k^\top A p_k. \quad (9)$$

Когда A является положительно определенной, приведенное выше выражение всегда положительно и, таким образом, обеспечивает нижнюю границу квадрата A -нормы ошибки. Уравнение (9) получается непосредственно из Bi-CG, и, когда $A = A^\top$, эта оценка для A -нормы идентична ранее предложенной оценке CGQL A -нормы (5). Поскольку векторно-матричное произведение $A p_k$ известно в терминах векторов невязок, приведенное выше оценивание занимает всего $2n$ арифметических операций по сравнению с $\mathcal{O}(n^2)$ арифметическими операциями в каждой итерации Bi-CG. Суммирование d членов в приведенном выше выражении означает задержку в d итераций вначале, когда оценки ошибки нет.

3.2.2. Оценивание l_2 -нормы ошибки

Записав последовательную разность между l_2 -нормой ошибки на итерациях k и $k+1$ для Bi-CG, мы получим следующее соотношение, когда A — несимметричная матрица:

$$\begin{aligned} r_k^\top (A^\top)^{-1} A^{-1} r_k - r_{k+1}^\top (A^\top)^{-1} A^{-1} r_{k+1} &= r_k^\top (A^\top)^{-1} A^{-1} r_k - \\ &\quad (r_k - \alpha_k A p_k)^\top (A^\top)^{-1} A^{-1} (r_k - \alpha_k A p_k). \end{aligned} \quad (10)$$

Сделав перестановку в (10), имеем

$$r_k^\top (A^\top)^{-1} A^{-1} r_k - r_{k+1}^\top (A^\top)^{-1} A^{-1} r_{k+1} = 2\alpha_k p_k^\top A^{-1} r_k - \alpha_k^2 \|p_k\|^2. \quad (11)$$

С использованием (8) и $\epsilon_k = A^{-1} r_k$ приведенное выше уравнение можно переписать как

$$\|\epsilon_{k+1}\|_2^2 = r_{k+1}^\top (A^\top)^{-1} A^{-1} r_{k+1} \approx \left(\sum_{j=k}^{k+d} \alpha_j p_j \right)^\top \sum_{j=k}^{k+d} \alpha_j p_j - 2\alpha_k p_k^\top \sum_{j=k}^{k+d} \alpha_j p_j + \alpha_k^2 \|p_k\|^2. \quad (12)$$

Оценивание включает только $2n$ арифметических операций на каждой итерации, и для первых d итераций не имеется оценки ошибки. Эта оценка добавляет незначительные вычислительные затраты по сравнению с $\mathcal{O}(n^2)$ -стоимостью каждой итерации Bi-CG, особенно при увеличении n .

3.3. Анализ формул оценок

3.3.1. Число обусловленности задачи

Число обусловленности задачи позволяет оценить ожидаемую точность результата вычисления [19]. В этом случае относительное число обусловленности для вычисления b при заданных A и x (прямая задача) и число обусловленности для вычисления x из A и b (обратная задача) определяются как

$$\kappa_f(A, x) = \|A\| \frac{\|x\|}{\|Ax\|}, \quad (13)$$

$$\kappa_b(A, b) = \|A^{-1}\| \frac{\|b\|}{\|A^{-1}b\|}. \quad (14)$$

Они помогают нам узнать *относительное* изменение выходных данных при *относительном* изменении входных данных с округлениями, тем самым ограничивая точность выходных данных. Норма $\|\cdot\|$, если не указано иначе, считается l_2 -нормой; $\kappa_f(A, x)$ и $\kappa_b(A, b)$ являются фиксированными для данной задачи $Ax = b$. Число обусловленности матрицы определяется произведением $\kappa_f(A, x)$ и $\kappa_b(A, b)$,

$$\kappa = \|A\| \|A^{-1}\|.$$

3.3.2. Метрики оценивания

Ниже приведена *относительная* ошибка наших оценок относительной ошибки на k -й итерации:

$$\max \left(\frac{\frac{\chi_k}{\|x\|} - \frac{\|\epsilon_k\|}{\|x\|}}{\frac{\|\epsilon_k\|}{\|x\|}}, \frac{\frac{\|\epsilon_k\|}{\|x\|} - \frac{\chi_k}{\|x\|}}{\frac{\chi_k}{\|x\|}} \right), \quad (15)$$

где χ_k — оценка нормы ошибки на k -й итерации, x — решение задачи $Ax = b$. Максимум из двух членов, заключенных в скобки, отражает тот факт, что оценка ошибки может быть больше или меньше истинной относительной ошибки на любой итерации, и это может привести к соответствующей степени уменьшения/увеличения вычислений. Уравнение (15) также можно кратко переписать в виде

$$\frac{\left| \frac{\chi_k}{\|x\|} - \frac{\|\epsilon_k\|}{\|x\|} \right|}{\min \left(\frac{\|\epsilon_k\|}{\|x\|}, \frac{\chi_k}{\|x\|} \right)}. \quad (16)$$

Тогда как (16) является количественной оценкой неопределенности с использованием относительной невязки или формулы оценки ошибки на конкретной итерации, коэффициенты неопределенности $U.R.(\chi)$ и линейные коэффициенты неопределенности $L.U.R.(\chi)$ определяются как средняя неопределенность в оценивании истинной относительной ошибки за $(n - d)$ -итераций:

$$U.R.(\chi) = \frac{1}{n-d} \sum_{k=0}^{n-d-1} \left(\frac{\left| \frac{\chi_k^2}{\|x\|^2} - \frac{\|\epsilon_k\|^2}{\|x\|^2} \right|}{\min \left(\frac{\|\epsilon_k\|^2}{\|x\|^2}, \frac{\chi_k^2}{\|x\|^2} \right)} \right), \quad (17)$$

$$L.U.R.(\chi) = \frac{1}{n-d} \sum_{k=0}^{n-d-1} \left(\frac{\left| \frac{\chi_k}{\|x\|} - \frac{\|\epsilon_k\|}{\|x\|} \right|}{\min \left(\frac{\|\epsilon_k\|}{\|x\|}, \frac{\chi_k}{\|x\|} \right)} \right), \quad (18)$$

где d — задержка в оценивании в терминах итераций. Когда относительная невязка используется для остановки итерационного решения линейной системы при заданном допуске относительной ошибки, неопределенность, соответственно, следующая:

$$U.R.(r) = \frac{1}{n-d} \sum_{k=0}^{n-d-1} \left(\frac{\left| \frac{\|r_k\|^2}{\|b\|^2} - \frac{\|\epsilon_k\|^2}{\|x\|^2} \right|}{\min \left(\frac{\|\epsilon_k\|^2}{\|x\|^2}, \frac{\|r_k\|^2}{\|b\|^2} \right)} \right), \quad (19)$$

$$L.U.R.(r) = \frac{1}{n-d} \sum_{k=0}^{n-d-1} \left(\frac{\left| \frac{\|r_k\|}{\|b\|} - \frac{\|\epsilon_k\|}{\|x\|} \right|}{\min \left(\frac{\|\epsilon_k\|}{\|x\|}, \frac{\|r_k\|}{\|b\|} \right)} \right). \quad (20)$$

Ниже мы представим теорему о верхних границах математических ожиданий $U.R.(r)$ и $L.U.R.(r)$ для сходящихся задач, имеющих фиксированные сингулярные значения и постоянное прямое число обусловленности, мы покажем, что $\mathbb{E}(U.R.(\chi)) \leq \mathbb{E}(U.R.(r))$, и придем к выводу, что использование формулы оценки ошибки в таких случаях в качестве критерия остановки дает в среднем большую экономию и точность вычислений благодаря их надежности.

3.3.3. Теорема о математическом ожидании $U.R.$ и $L.U.R.$

Теорема. Для линейных задач несингулярных действительных матриц A с множеством фиксированных сингулярных значений, фиксированным прямым числом обусловленности $\kappa_f(A, x) \gg 1$ и фиксированным обратным числом обусловленности $\kappa_b(A, b) \gg 1$ имеем

$$\frac{\alpha}{\kappa_b(A, b)^2} + \frac{1-\alpha}{\kappa_f(A, x)^2} \leq \mathbb{E}(U.R.(r)) + o(2) \leq \alpha \kappa_f(A, x)^2 + (1-\alpha) \kappa_b(A, b)^2,$$

и

$$\frac{\alpha}{\kappa_b(A, b)} + \frac{1-\alpha}{\kappa_f(A, x)} \leq \mathbb{E}(L.U.R.(r)) + o(2) \leq \alpha \kappa_f(A, x) + (1-\alpha) \kappa_b(A, b),$$

где n — размерность матрицы A , α — средняя вероятность относительной невязки при предсказании относительной ошибки решения, т. е.

$$\alpha = \frac{1}{n-d} \sum_{k=0}^{n-d-1} P_{\text{cond}} \left(\frac{\|r_k\|}{\|b\|} \geq \frac{\|\epsilon_k\|}{\|x\|} \right),$$

где P_{cond} — условная вероятность с условиями, указанными в начале теоремы. Выборочное пространство для условной вероятности и условного ожидания генерируется путем выборки сингулярной слева матрицы U и сингулярной справа матрицы V из равномерно распределенных ортогональных матриц и единичного вектора $\hat{b}$ из распределения Гаусса.

Доказательство. Пусть коэффициент неопределенности на k -й итерации имеет вид

$$\text{U.R.}^{(k)}(r) = \frac{\left| \frac{\|r_k\|^2}{\|b\|^2} - \frac{\|\epsilon_k\|^2}{\|x\|^2} \right|}{\min \left(\frac{\|\epsilon_k\|^2}{\|x\|^2}, \frac{\|r_k\|^2}{\|b\|^2} \right)} \quad \text{и} \quad \alpha_k = P_{\text{cond}} \left(\frac{\|r_k\|}{\|b\|} \geq \frac{\|\epsilon_k\|}{\|x\|} \right).$$

Используя определение в (15) для $\text{U.R.}^{(k)}(r)$, мы можем записать

$$\mathbb{E}(\text{U.R.}^{(k)}(r)) = \alpha_k \mathbb{E}_1 \left(\frac{\frac{\|r_k\|^2}{\|b\|^2} - \frac{\|\epsilon_k\|^2}{\|x\|^2}}{\frac{\|\epsilon_k\|^2}{\|x\|^2}} \right) + (1 - \alpha_k) \mathbb{E}_2 \left(\frac{\frac{\|\epsilon_k\|^2}{\|x\|^2} - \frac{\|r_k\|^2}{\|b\|^2}}{\frac{\|r_k\|^2}{\|b\|^2}} \right),$$

где $\mathbb{E}_1$ означает $\mathbb{E} \left(\bullet \mid \frac{\|r_k\|}{\|b\|} \geq \frac{\|\epsilon_k\|}{\|x\|} \right)$, а $\mathbb{E}_2$ означает $\mathbb{E} \left(\bullet \mid \frac{\|r_k\|}{\|b\|} \leq \frac{\|\epsilon_k\|}{\|x\|} \right)$:

$$\mathbb{E}(\text{U.R.}^{(k)}(r)) = \alpha_k \mathbb{E}_1 \left(\frac{\|r_k\|^2}{\|b\|^2} \frac{\|x\|^2}{\|\epsilon_k\|^2} - 1 \right) + (1 - \alpha_k) \mathbb{E}_2 \left(\frac{\|\epsilon_k\|^2}{\|x\|^2} \frac{\|b\|^2}{\|r_k\|^2} - 1 \right),$$

$$\mathbb{E}(\text{U.R.}^{(k)}(r)) = \alpha_k \mathbb{E}_1 \left(\frac{\|r_k\|^2}{\|\epsilon_k\|^2} \frac{\|x\|^2}{\|b\|^2} \right) + (1 - \alpha_k) \mathbb{E}_2 \left(\frac{\|\epsilon_k\|^2}{\|r_k\|^2} \frac{\|b\|^2}{\|x\|^2} \right) - 1.$$

Пусть $S_k = \frac{\|r_k\|}{\|\epsilon_k\|} \frac{\|x\|}{\|b\|}$, тогда $S_k \in \left\{ \frac{1}{\kappa_b(A, b)}, \kappa_f(A, x) \right\}$ и

$$\mathbb{E}(\text{U.R.}^{(k)}(r)) = \alpha_k \mathbb{E}(S_k^2 | S_k \geq 1) + (1 - \alpha_k) \mathbb{E}\left(\frac{1}{S_k^2} | S_k \leq 1\right) - 1,$$

$$\mathbb{E}(\text{U.R.}^{(k)}(r)) = \alpha_k (\mathbb{E}(S_k^2) - o(1)) + (1 - \alpha_k) \left(\mathbb{E}\left(\frac{1}{S_k^2}\right) - o(1) \right) - 1,$$

$$\mathbb{E}(\text{U.R.}^{(k)}(r)) = \alpha_k (\mathbb{E}(S_k^2)) + (1 - \alpha_k) \left(\mathbb{E}\left(\frac{1}{S_k^2}\right) \right) - o(2),$$

$$\mathbb{E}(S_k^2) = \mathbb{E}\left(\frac{\|r_k\|^2}{\|\epsilon_k\|^2} \frac{\|x\|^2}{\|b\|^2}\right).$$

Ввиду фиксированных условий ожидания $\frac{\|x\|}{\|b\|}$ постоянно, следовательно,

$$\mathbb{E}(S_k^2) = \frac{\|x\|^2}{\|b\|^2} \mathbb{E}\left(\frac{\|r_k\|^2}{\|\epsilon_k\|^2}\right) = \frac{\|x\|^2}{\|b\|^2} \mathbb{E}\left(\frac{\|A\epsilon_k\|^2}{\|\epsilon_k\|^2}\right).$$

Используя определение l_2 -нормы матрицы, имеем

$$\frac{\kappa_f(A, x)^2}{\kappa(A)^2} \leq \mathbb{E}(S_k^2) \leq \kappa_f(A, x)^2.$$

Аналогичным образом

$$\frac{\kappa_b(A, b)^2}{\kappa(A)^2} \leq \mathbb{E}\left(\frac{1}{S_k^2}\right) \leq \kappa_b(A, b)^2.$$

Следовательно, используя приведенные выше границы и α из формулировки теоремы, мы получим

$$\begin{aligned} \frac{\alpha}{\kappa_b(A, b)^2} + \frac{1 - \alpha}{\kappa_f(A, x)^2} - o(2) &\leq \mathbb{E}(\text{U.R.}(r)) \leq \alpha \kappa_f(A, x)^2 + (1 - \alpha) \kappa_b(A, b)^2 - o(2), \\ \frac{\alpha}{\kappa_b(A, b)^2} + \frac{1 - \alpha}{\kappa_f(A, x)^2} &\leq \mathbb{E}(\text{U.R.}(r)) + o(2) \leq \alpha \kappa_f(A, x)^2 + (1 - \alpha) \kappa_b(A, b)^2. \end{aligned}$$

Точно так же мы можем получить границы оценки для $\text{L.U.R.}(r)$:

$$\frac{\alpha}{\kappa_b(A, b)} + \frac{1 - \alpha}{\kappa_f(A, x)} \leq \mathbb{E}(\text{L.U.R.}(r)) + o(2) \leq \alpha \kappa_f(A, x) + (1 - \alpha) \kappa_b(A, b). \quad \square$$

3.3.4. Замечание относительно формул оценки ошибки на основе задержки

По формулам оценок на основе задержки $\chi_k^2 \approx \|\epsilon_k\|^2 - \|\epsilon_{k+d}\|^2$ и

$$\text{U.R.}^{(k)}(\chi) \approx \frac{\|\epsilon_{k+d}\|^2}{\|\epsilon_k\|^2 - \|\epsilon_{k+d}\|^2}.$$

С другой стороны, относительная невязка имеет строгую нижнюю границу $\frac{1}{\kappa(A)} \frac{\|\epsilon\|}{\|x\|}$ и строгую верхнюю границу $\kappa(A) \frac{\|\epsilon\|}{\|x\|}$ для данной матрицы A . Это позволило нам обеспечить условные верхние и нижние границы, когда $\|\epsilon_{k+d}\|^2 < \|\epsilon_k\|^2$. Следовательно, использование формул оценок ошибки на основе задержки является надежным, когда итерационный метод, например GMRES или Bi-CG, имеет уменьшающуюся l_2 -норму ошибки для большинства итераций. Можно утверждать, что значение d можно увеличивать до такой степени, что $\|\epsilon_{k+d}\|^2 < \|\epsilon_k\|^2$ для всех итераций, но на практике необходимо ограничение ($d \ll n$) ввиду требований низкой задержки. Приведенные выше выводы верны без каких-либо ограничений на размерность матриц.

4. Численные результаты

4.1. Модифицированная формула оценки l_2 -нормы ошибки для GMRES

В этом эксперименте мы эмпирически подчеркиваем преимущества включения `term2` из (4) в аппроксимацию ошибки GMRES, как это предлагается в уравнении (5). Мы провели эксперименты с несимметричными положительно определенными матрицами с высоким числом обусловленности (сгенерированными путем собственного разложения с

собственными значениями, экспоненциально разбросанными между 1 и $\kappa(A)$), в которых мы сравнили значения различных формул оценок ошибки и относительную невязку на каждой итерации.

На рисунке 1 мы сравниваем относительную l_2 -норму формулы оценки ошибки, предложенной Меурантом [10], модифицированную l_2 -норму формулы оценки ошибки, предложенной в уравнении (5), относительную невязку и истинную относительную ошибку на каждой итерации для случайно сгенерированной несимметричной матрицы ($\kappa(A) \approx 10^8$) со случайным вектором правой части b и задержкой итераций $d = 10$. Матрица имеет размерность 500, а ее первые 20 самых больших собственных значений имеют величину $\approx 10^7$, остальные находятся в диапазоне $[0.1, 10]$.

Обращаем внимание, что модифицированная формула оценки намного ближе к истинной ошибке, чем исходная формула оценки, поскольку она позволяет избежать нестабильного превышения, наблюдаемого при отсутствии члена смещения. Мы видим, что включение term2 в уравнение (5) помогло избежать нестабильного превышения оценки для некоторых итераций, как показано на рис. 1. Обратите внимание на логарифмический масштаб.

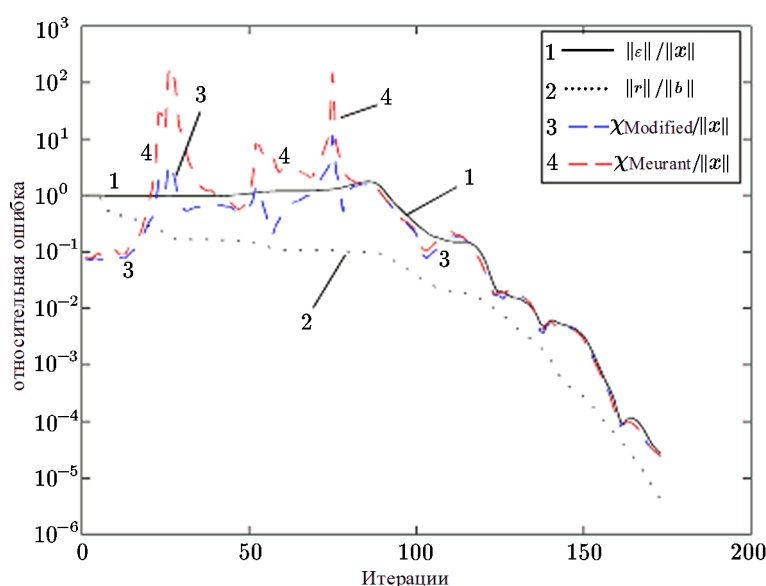


Рис. 1. График сходимости для сравнения различных формул оценок относительной ошибки и истинной относительной ошибки. L.U.R.(невязка) = 12.1; L.U.R.(Меурант) = 10.1; L.U.R.(модифицированная) = 1.2

Также замечено, что формулы оценки ошибки GMRES могут вести себя хаотично, когда численная точность вычислительной системы исчерпана, как показано на рис. 2. Исчерпание численной точности может привести к тому, что матрица Хессенберга H_k , сформированная во время итерации Арнольди, может стать почти сингулярной, что может вызвать большие ошибки в H_k^{-1} и ее функциях. Обратите внимание, что $\|s_k\|$ также является функцией H_k^{-1} и может использоваться как сигнал для предсказания этого исчерпания численной точности (см. рис. 2), что делает эту оценку ошибки надежным критерием останова. Заметим, что $\|s_k\|$ и две формулы оценки ошибки резко увеличиваются, тогда как фактическая ошибка и невязка на самом деле малы. Слагаемое $\|s_k\|$ из (4) может также использоваться для обнаружения исчерпания численной точности.

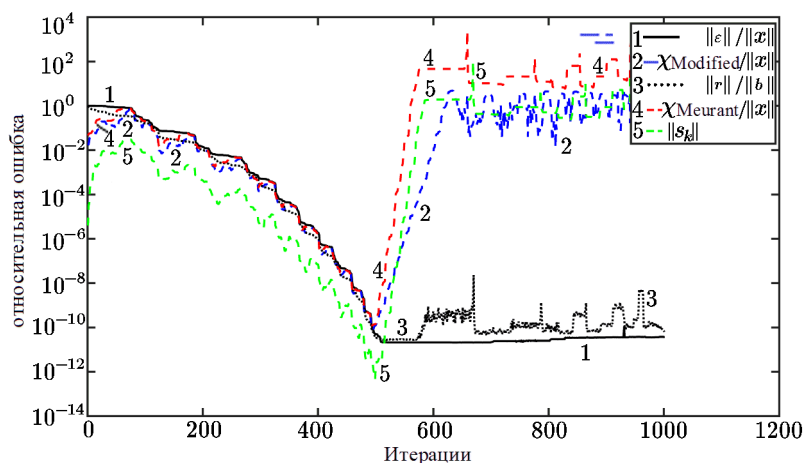


Рис. 2. Поведение оценок относительной ошибки в GMRES (для симметричной положительно определенной матрицы с $\kappa(A) = 10^6$) при исчерпании численной точности

4.2. Формулы A -нормы и l_2 -нормы оценок ошибки для Bi-CG

В этом эксперименте мы хотим эмпирически оценить поведение основанных на задержке формул оценок ошибки Bi-CG, полученных в п. 3.2, при сравнении с относительной невязкой в задачах с высоким числом обусловленности. Мы рассмотрели несимметричные матрицы из множества данных Matrix Market¹ (“sherman5”) со случайным вектором правой части b . Размерность матрицы 3312×3312 ; число обусловленности матрицы 3.9×10^5 ; $d = 10$.

На рис. 3 дано сравнение оцененной относительной ошибки l_2 -нормы, истинной относительной l_2 -нормы ошибки и относительной невязки. На рис. 4 мы сравниваем оцененную относительную A -норму ошибки, относительную невязку и истинную относительную A -норму ошибки на каждой итерации. Мы видим, что предложенные формулы оценки ошибки Bi-CG являются стабильными на всех итерациях в обоих случаях, когда формула оценки ошибки надежна как индикатор фактической ошибки для останова.

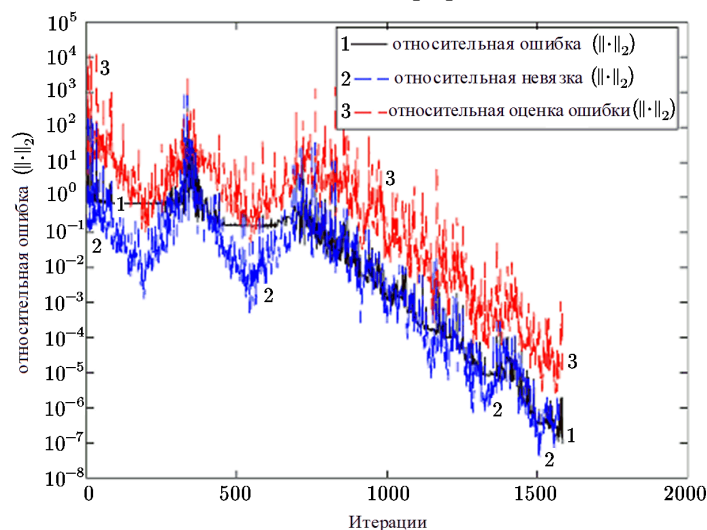


Рис. 3. Оценивание относительной l_2 -нормы ошибки с использованием формулы оценки ошибки Bi-CG и относительной невязки

¹<https://math.nist.gov/MatrixMarket/>

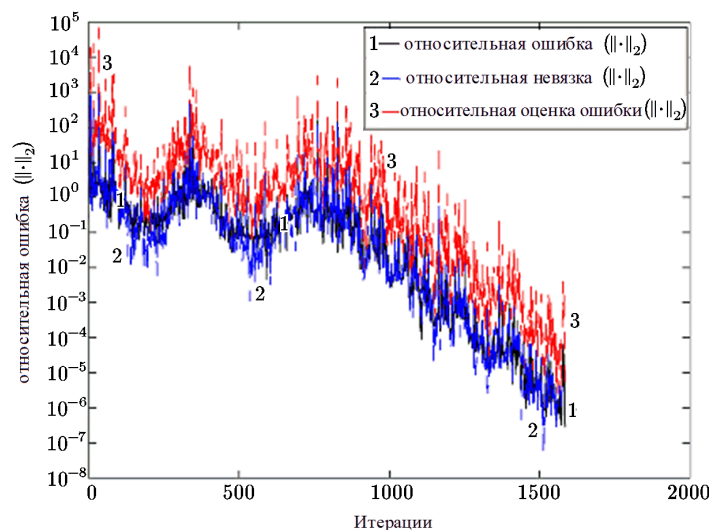


Рис. 4. Оценивание относительной A -нормы ошибки с использованием оценки ошибки Bi-CG и относительной невязки; абсолютные значения рассматриваются для $r^T A^{-1} r$ и ее аппроксимации

4.3. Экспериментальные результаты по эффективности формул оценок ошибки на рандомизированном множестве матриц

Цель экспериментов² — оценивание прогностической эффективности формул оценки ошибки на большом рандомизированном множестве матриц. Это исследование полезно для понимания изменчивости среднего поведения оценок относительно различных аспектов задачи. В предыдущих пунктах мы определили коэффициент неопределенности как среднюю ошибку любого из предикторов истинной относительной ошибки, который можно использовать для остановки.

В этом эксперименте мы определяем коэффициент неопределенности формулы оценки и невязку для заданных прямого и обратного чисел обусловленности для каждой задачи рандомизированного множества. Было взято рандомизированное множество данных, состоящее из 10000 задач с высокими числами обусловленности размерности 100, для которых метрики оценки $L.U.R.(r)$ и $L.U.R.(\chi)$ определяются для задержки оценивания $d = 10$. Мы рассматриваем два вида несимметричных матриц: положительно определенные и общие матрицы в этом множестве данных. Векторы правой части b являются случайными векторами из единичного нормального распределения этой же размерности.

Рисунки 5 и 6 показывают поведение коэффициента линейной неопределенности $L.U.R.$, определяемого путем (18) и (20), относительно числа обусловленности задачи (как прямого, так и обратного) для алгоритмов Bi-CG и GMRES соответственно. Здесь каждый кружок — коэффициент неопределенности, соответствующий относительной невязке для конкретной задачи, т. е. для конкретной матрицы A и вектора правой части b с x_0 , взятым из стандартного гауссова распределения. Аналогичным образом, каждый крестик — это коэффициент неопределенности, соответствующий оценке относительной ошибки.

Мы видим большое увеличение $L.U.R.(r)$ относительной невязки в задачах с высоким числом обусловленности, в то время как $L.U.R.(\chi)$ для формулы оценки относительной ошибки остается ограниченным, что показывает надежность формулы оценок ошибок Bi-CG и GMRES. Также интересно отметить, что коэффициент неопределенности за-

²Код имеется в свободном доступе на <https://github.com/CSPL-IISc/Error-Linsolve>.

висит как от прямого, так и от обратного числа обусловленности задачи. Для Bi-CG L.U.R.(r) увеличивается при прямом числе обусловленности задачи, тогда как L.U.R.(r) увеличивается при обратном числе обусловленности в случае алгоритма GMRES. Обратите внимание, что значения α (см. теорему), которые определяют увеличивающуюся (или уменьшающуюся) зависимость от прямого (или обратного) числа обусловленности, могут варьироваться в зависимости от алгоритма. Поскольку одно и то же множество данных использовалось для тестирования метрик как для алгоритмов Bi-CG, так и для алгоритмов GMRES, масштабы по оси x на двух рисунках различны (обратите внимание, что $\kappa_f(A, x) * \kappa_b(A, b) = \kappa$). Тогда как Bi-CG имеет гораздо большие коэффициенты неопределенности вследствие осциллирующего поведения сходимости, ожидаемые значения уменьшения/увеличения вычислений для требуемого допуска ошибки сопоставимы как для Bi-CG, так и для GMRES, что будет обсуждаться в следующем пункте.

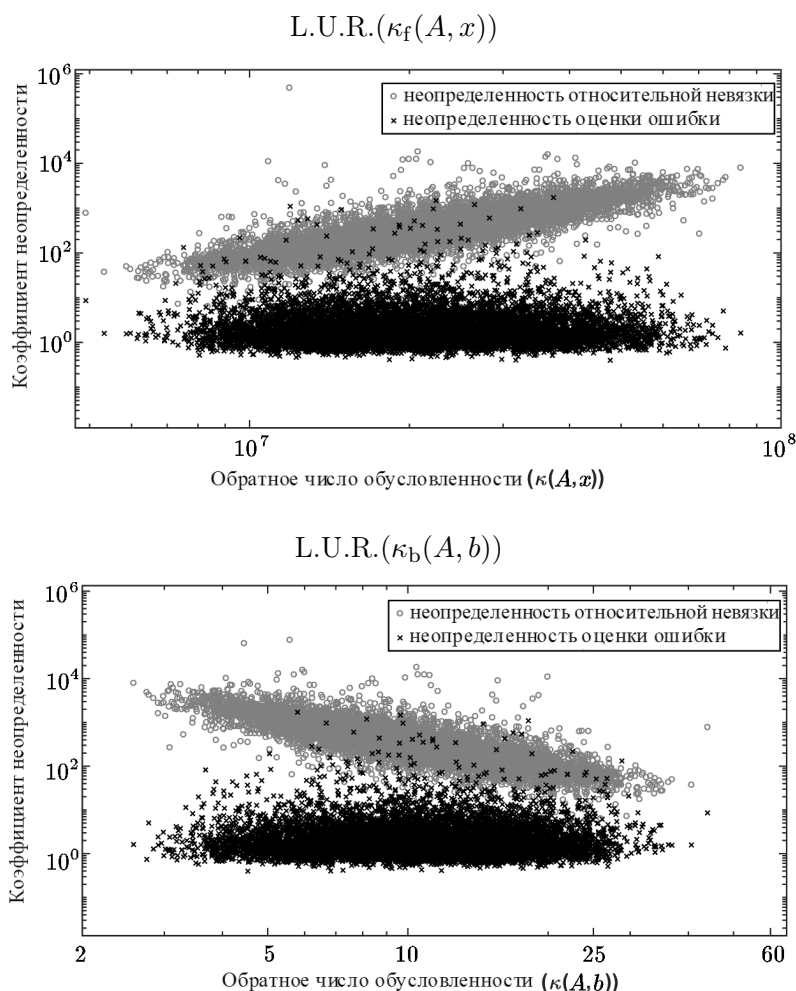


Рис. 5. Неопределенность L.U.R. при использовании формулы оценки l_2 -нормы относительной ошибки или относительной невязки для остановки алгоритма Bi-CG, когда задан допуск относительной ошибки. Прямое и обратное число обусловленности задач варьировалось: среднее L.U.R.(невязка) = 288: среднее L.U.R.(оценка) = 5.9

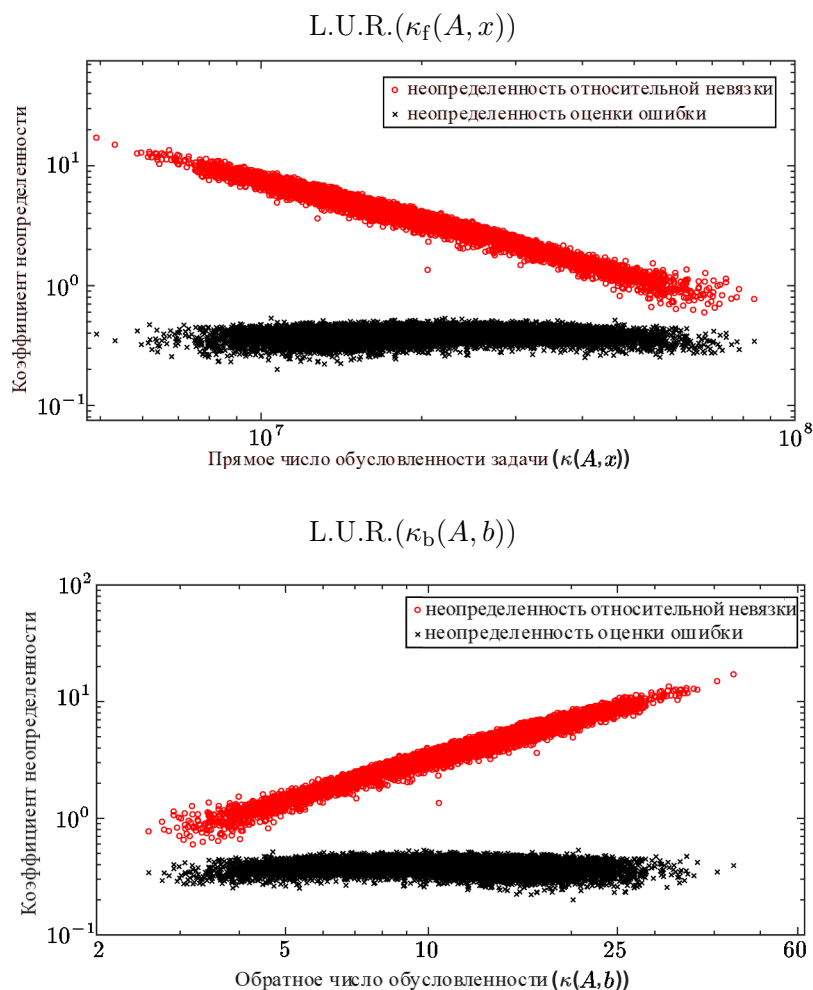


Рис. 6. Неопределенность L.U.R. при использовании формулы оценки l_2 -нормы относительной ошибки или относительной невязки для останова алгоритма GMRES, когда задан допуск относительной ошибки. Прямое и обратное число обусловленности задач варьировалось; среднее $L.U.R.(невязка) = 2.49$; среднее $L.U.R.(оценка) = 0.286$. Обратите внимание, что модифицированные и исходные формулы оценки GMRES почти идентичны на этих диаграммах разброса, где в задачах у нас есть лишь небольшое число измерений $n = 100$

Для дальнейшей демонстрации сделанных выше выводов даже для матриц, имеющих практические применения, из множеств данных Matrix Market и Matrix Sparse Collection были взяты несимметричные матрицы, имеющие различные числа обусловленности и различную размерность, и протестированы с соответствующими случайными нормализованными правыми векторами. Рисунок 7 показывает связь между коэффициентом неопределенности и числами обусловленности (прямым и обратным) соответствующих задач. Мы видим значительное увеличение коэффициента неопределенности относительной невязки с увеличением числа обусловленности задачи, в то время как коэффициент неопределенности формулы оценки ошибки (как для Bi-CG, так и для GMRES) остается почти неизменным при увеличении числа обусловленности задачи. Это означает, что формулы оценки ошибки являются надежными критериями останова. Хотя мы в основном оцениваем метрики оценки для алгоритмов Bi-CG и GMRES, следует учитывать, что эффективность оценки CG неявно описывается оценкой Bi-CG в случае симметричных положительно определенных матриц.

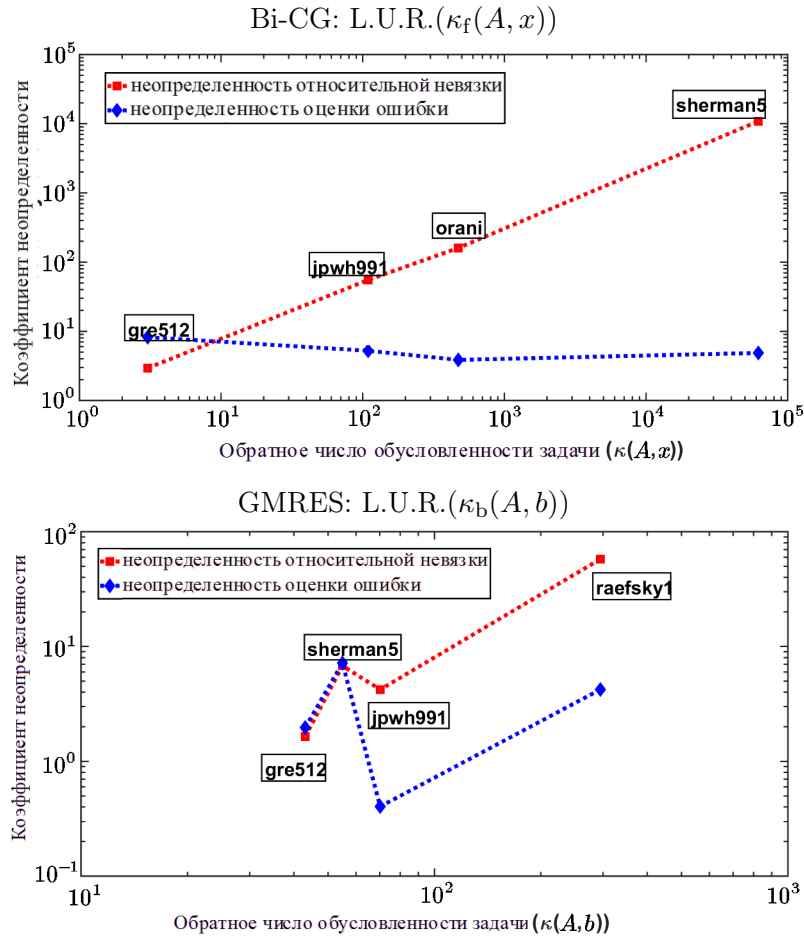


Рис. 7. Неопределенность L.U.R. при использовании формулы оценки относительной l_2 -нормы ошибки или относительной невязки для останова алгоритма GMRES, когда задан допуск относительной ошибки. Матрицы (с указанием размерности), взятые из множества данных Matrix-Market для этой оценки, были следующими: “gre15” (512), “jpwh991” (991), “orani” (2529), “sherman5” (3312) и “raefsky1” (3242)

4.4. Уменьшение/увеличение вычислений с формулами оценки Bi-CG и GMRES

При использовании итерационных алгоритмов желательно останавливать вычисления при допусках ошибок в зависимости от требований определенных приложений. Поскольку имеется большая неопределенность при остановке с использованием относительной невязки, как показано на рис. 5, 6 и 7, уменьшения/увеличения вычислений можно ожидать для любого желаемого допуска к ошибкам в решении. Можно остановить итерации слишком рано (уменьшение вычислений) или выполнить слишком много итераций (увеличение вычислений) ввиду этой неопределенности ошибки решения, что приведет к потере эффективности или точности. В этом эксперименте цель состоит в том, чтобы оценить степень увеличения/уменьшения вычислений вследствие неопределенности ошибки при использовании относительной невязки в качестве индикатора для останова.

Определим относительные итерации как I , которые либо сохраняются, либо теряются для желаемого допуска к ошибкам, т. е.
$$I = \frac{|\text{Iter}_{\|r\|/\|b\|} - \text{Iter}_{\|x\|/\|x\|}|}{\min(\text{Iter}_{\|r\|/\|b\|}, \text{Iter}_{\|x\|/\|x\|})}.$$

Для любого допуска, такого как 10^{-4} , I означает число итераций, которые были увеличены/уменьшены при использовании относительной невязки в качестве критерия остановки по сравнению с использованием формулы оценки ошибки в качестве критерия остановки. $\text{Iter}_{\|r\|/\|b\|}$ — число итераций, необходимое для того, чтобы относительная невязка $\|r\|/\|b\|$ достигла желаемого допуска, и у нас есть аналогичная переменная для итераций, необходимых для оценки ошибки.

На рис. 8 показан график I при допуске к ошибкам 10^{-1} , 10^{-2} , 10^{-3} и 10^{-4} при увеличении $\text{L.U.R.}(r)$, усредненном по матричным задачам, использованным в п. 4.3, для обоих алгоритмов. Следует отметить, что значение $I = 1$ означает, что имеется 100 % разница в итерациях, необходимых для того, чтобы $\|r\|/\|b\|$ и $\|\chi\|/\|x\|$ достигли желаемого допуска. Мы видим увеличение I с коэффициентом неопределенности относительной невязки, что показывает важность формулы оценки ошибки при использовании итерационных алгоритмов.

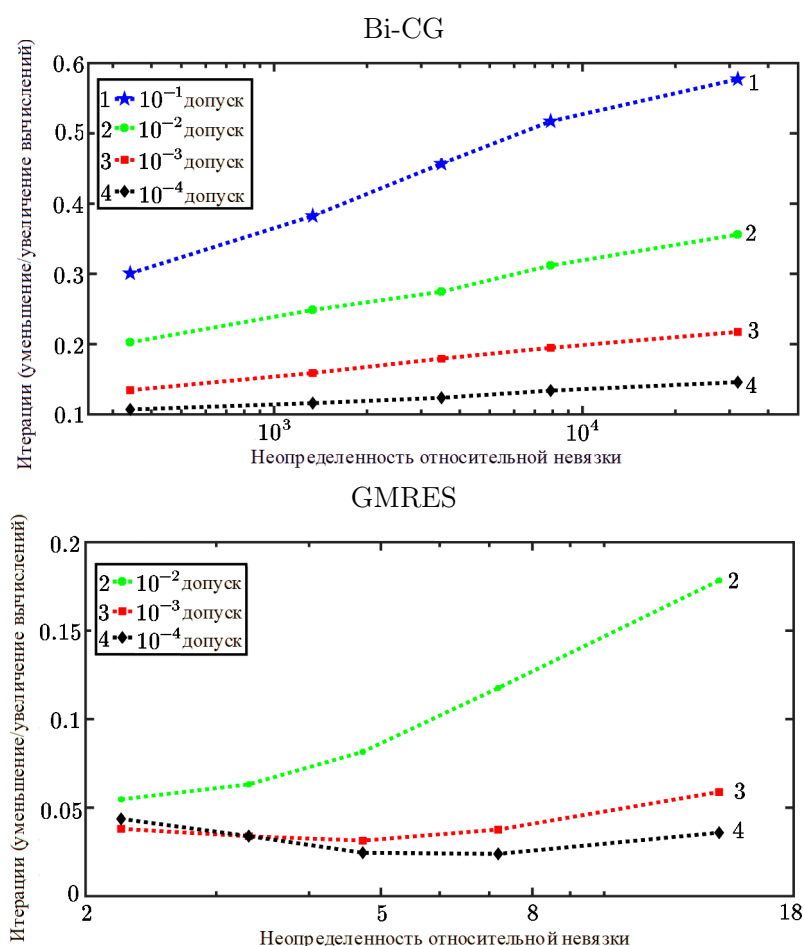


Рис. 8. Доля уменьшенных/увеличенных итераций при остановке итераций для различных допусков по мере увеличения неопределенности относительной невязки L.U.R. Обратите внимание, что неопределенность Bi-CG может быть больше из-за осциллирующей сходимости, но ожидаемые увеличенные и уменьшенные вычисления при использовании невязки для остановки сопоставимы для Bi-CG и GMRES. Прямое и обратное числа обусловленности, соответственно, увеличивались для более высокой неопределенности

5. Выводы

Важность формулы оценки ошибки для эффективной остановки (или перезапуска) очевидна для задач с числами обусловленности $\kappa \gtrsim 100$ и иллюстрируется на численных примерах, а также анализом ожидаемой неопределенности сходимости при использовании нормы вектора невязки для остановки при любом заданном допуске к ошибкам. Результаты, показывающие уменьшение коэффициента неопределенности при использовании формулы оценки ошибки, были представлены как для алгоритмов Bi-CG, так и для GMRES. Также было показано, что предложенная формула оценки для Bi-CG сходится к ранее предложенной формуле оценки для алгоритма CG в случае симметричных положительно определенных матриц. Кроме того, был представлен простой анализ изменений вычислений за счет использования формулы оценки, а также численные результаты большого числа задач умеренной размерности. Тогда как Bi-CG имеет гораздо больший коэффициент неопределенности ввиду своего осциллирующего поведения (и это можно улучшить с использованием расширенных версий алгоритма [4, 5, 16, 17, 20]), ожидаемое уменьшение/увеличение вычислений для требуемого допуска ошибки сопоставимо для Bi-CG и GMRES. На основании результатов, обсуждавшихся в предыдущих пунктах, мы можем заключить, что оценку A -нормы или l_2 -нормы ошибки следует включить в программную реализацию итерационных решателей, а не использовать только относительную норму вектора невязки в качестве критерия остановки.

Заявление авторов

Для данного исследования не было получено финансирования. Авторы не имеют финансовых или нефинансовых интересов.

Приложения

А. Bi-CG и его связь с несимметричным алгоритмом Ланцоша

Алгоритм бисопряженных градиентов Bi-CG является расширением алгоритма сопряженных градиентов CG, который используется для решения системы линейных уравнений и для несимметричной матрицы.

Алгоритм. Алгоритм бисопряженных градиентов

```

1: procedure
   input  $A, A^\top, b, x_0, y_0$ 
2:    $r_0 = b - Ax_0$ 
3:    $\tilde{r}_0 = b - A^\top y_0$ 
4:    $p_0 = r_0$ 
5:    $q_0 = \tilde{r}_0$ 
6:   for  $k = 1, \dots$ , до сближения do
7:      $\alpha_{k-1} = \frac{\tilde{r}_{k-1}^\top r_{k-1}}{q_{k-1}^\top A p_{k-1}}$ 
8:      $x_k = x_{k-1} + \alpha_{k-1} p_{k-1}$ 
9:      $y_k = y_{k-1} + \alpha_{k-1} q_{k-1}$ 
10:     $r_k = r_{k-1} - \alpha_{k-1} A p_{k-1}$ 
11:     $\tilde{r}_k = \tilde{r}_{k-1} - \alpha_{k-1} A^\top q_{k-1}$ 

```

```

12:       $\beta_{k-1} = \frac{\tilde{r}_k^\top r_k}{\tilde{r}_{k-1}^\top r_{k-1}}$ 
13:       $p_k = r_k + \beta_{k-1} p_{k-1}$ 
14:       $q_k = \tilde{r}_k + \beta_{k-1} q_{k-1}$ 
15:  end for
16: end procedure

```

Bi-CG также может быть получен из несимметричного алгоритма Ланцоша. Пусть v_1 и $\tilde{v}_1$ — данные начальные векторы для несимметричного алгоритма Ланцоша (такие что $\|v_1\| = 1$ и $(v_1, \tilde{v}_1) = 1$), двух-/трехчленные рекуррентные соотношения, которые помогают в формировании двух биортогональных подпространств, имеют следующий вид: для $k=1, 2, \dots$

$$z_k = Av_k - w_k v_k - \eta_{k-1} v_{k-1}, \quad \tilde{z}_k = A^\top \tilde{v}_k - w_k \tilde{v}_k - \tilde{\eta}_{k-1} \tilde{v}_{k-1}.$$

Коэффициент w_k вычисляется следующим образом: $w_k = (\tilde{v}_k, Av_k)$. Другие коэффициенты, η_k и $\tilde{\eta}_k$, выбираются (при $(\tilde{z}_k, v_k) = 0$) такими, что $\eta_k \tilde{\eta}_k = (\tilde{z}_k, z_k)$, а новые векторы на шаге $k+1$ задаются следующим образом:

$$v_{k+1} = \frac{z_k}{\eta_k}, \quad \tilde{v}_{k+1} = \frac{\tilde{z}_k}{\tilde{\eta}_k}.$$

Связь между биортогональными подпространствами и матрицей A можно представить в виде несимметричной трехдиагональной матричной формы (при условии $\tilde{V}_k^\top AV_k = T_k$):

$$T_k = \begin{bmatrix} \omega_1 & \eta_1 & 0 & \cdots & 0 \\ \tilde{\eta}_1 & \omega_2 & \eta_2 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \tilde{\eta}_{k-2} & \omega_{k-1} & \eta_{k-1} \\ 0 & \cdots & 0 & \tilde{\eta}_{k-1} & \omega_k \end{bmatrix}.$$

Б. Связь формулы оценки ошибки Bi-CG и CGQL

Разница между двумя последовательными A -нормами ошибки (и A -мерой для матриц, не являющихся симметричными положительно определенными (SPD) матрицами) в случае алгоритма сопряженных градиентов на итерациях k и $k+1$ может быть определена следующим образом:

$$\|\epsilon_k\|_A^2 - \|\epsilon_{k+1}\|_A^2 = \alpha_k r_k^\top r_k.$$

Следовательно, при задержке в d итераций мы можем легко вычислить A -норму ошибки:

$$\|\epsilon_k\|_A^2 - \|\epsilon_{k+d}\|_A^2 = \sum_{j=k}^{k+d} \alpha_j r_j^\top r_j, \quad \|\epsilon_k\|_A^2 \approx \sum_{j=k}^{k+d} \alpha_j r_j^\top r_j.$$

Для оценки ошибки A -нормы в алгоритме Bi-CG мы получили следующие результаты:

$$\|\epsilon_{k+1}\|_A^2 = r_{k+1}^\top A^{-1} r_{k+1} = -\alpha_k r_k^\top p_k + r_{k+1}^\top A^{-1} r_k + \alpha_k^2 p_k^\top A p_k.$$

С использованием $r_{k+1} = r_k - \alpha_k A p_k$ приведенный выше результат можно записать в виде

$$r_k^\top A^{-1} r_k - r_{k+1}^\top A^{-1} r_{k+1} = \alpha_k r_k^\top p_k + \alpha_k r_k^\top (A^\top)^{-1} A p_k - \alpha_k^2 p_k^\top A p_k.$$

Для симметричной матрицы $A = A^\top$ приведенное выше уравнение можно переписать как

$$r_k^\top A^{-1} r_k - r_{k+1}^\top A^{-1} r_{k+1} = \alpha_k r_k^\top p_k + \alpha_k r_k^\top p_k - \alpha_k^2 p_k^\top A p_k.$$

Далее, используя $p_i^\top r_j = 0$ для $i \neq j$ в CG, а также $r_{k+1} = r_k - \alpha_k A p_k$, и подставляя $A p_k = \frac{r_k - r_{k+1}}{\alpha_k}$, приведенное выше выражение сводится к

$$\begin{aligned} r_k^\top A^{-1} r_k - r_{k+1}^\top A^{-1} r_{k+1} &= \alpha_k r_k^\top p_k + \alpha_k r_k^\top p_k - \alpha_k p_k^\top r_k + \alpha_k p_k^\top r_{k+1} \\ &\Downarrow \\ r_k^\top A^{-1} r_k - r_{k+1}^\top A^{-1} r_{k+1} &= \alpha_k r_k^\top p_k \\ &\Downarrow \\ r_k^\top A^{-1} r_k - r_{k+1}^\top A^{-1} r_{k+1} &= \alpha_k r_k^\top r_k + \beta_{k-1} r_k^\top p_{k-1} \\ &\Downarrow \\ r_k^\top A^{-1} r_k - r_{k+1}^\top A^{-1} r_{k+1} &= \alpha_k r_k^\top r_k. \end{aligned} \tag{21}$$

Уравнение (21) эквивалентно оценке A -нормы в алгоритме CGQL [6], и, таким образом, предлагаемая формула оценки A -нормы для Bi-CG эквивалентна формуле оценки CGQL, когда $A = A^\top$.

В. Связь между несимметричной трехдиагональной матрицей Ланцоша (T_k) и векторами невязки (r и $\tilde{r}$) алгоритма Bi-CG

Прямое соотношение между T_k , r_k и $\tilde{r}_k$ может быть задано следующим образом:

$$(T_n^{-1})_{(1,1)} = (T_k^{-1})_{(1,1)} + \frac{\tilde{r}_k^\top A^{-1} r_k}{\|r_0\|^2}, \tag{22}$$

$$(T_n^{-1})_{(1,1)} = (T_{k+1}^{-1})_{(1,1)} + \frac{\tilde{r}_{k+1}^\top A^{-1} r_{k+1}}{\|r_0\|^2}. \tag{23}$$

Вычитая (22) и (23), мы получим

$$(T_{k+1}^{-1})_{(1,1)} = (T_k^{-1})_{(1,1)} + \frac{\tilde{r}_k^\top A^{-1} r_k - \tilde{r}_{k+1}^\top A^{-1} r_{k+1}}{\|r_0\|^2}.$$

Поскольку $\tilde{r}_k^\top A^{-1} r_k - \tilde{r}_{k+1}^\top A^{-1} r_{k+1} = \alpha_k \tilde{r}_k^\top r_k$ для несимметричной матрицы в Bi-CG, то

$$(T_{k+1}^{-1})_{(1,1)} = (T_k^{-1})_{(1,1)} + \frac{\alpha_k \tilde{r}_k^\top r_k}{\|r_0\|^2}.$$

Таким образом, зная соотношение между двумя последовательными обратными первыми элементами T_k , мы можем связать его с полученной нами формулой оценки для A -нормы ошибки и алгоритмом CGQL. Однако связь между несимметричным алгоритмом Ланцоша и Bi-CG с использованием методов, основанных на квадратурах, будет включать комплексные квадратуры Гаусса [15]. Следовательно, мы применили эквивалентный, но прямой подход к оцениванию с явным использованием соотношений Bi-CG.

Литература

1. **Bjorck A.** Numerical Methods in Matrix Computations. 1st ed. — Springer Cham, 2015. — DOI: 10.1007/978-3-319-05089-8_1.
2. **Brezinski C.** Error estimates for the solution of linear systems // SIAM J. Sci. Comput. — 1999. — Vol. 21, № 2. — P. 764–781.

3. **Brezinski C., Rodriguez G., Seatzu S.** Error estimates for linear systems with applications to regularization // Numerical Algorithms. — 2008. — Vol. 49, № 1-4. — P. 85–104.
4. **Fletcher R.** Conjugate gradient methods for indefinite systems // Numerical Analysis. — Springer, 1976. — P. 73–89. — (Lect. Notes in Mathematics; 506).
5. **Freund R.W., Nachtigal N.M.** QMR: a quasi-minimal residual method for non-Hermitian linear systems // Numerische Mathematik. — 1991. — Vol. 60, № 1. — P. 315–339.
6. **Golub G.H., Meurant G.** Matrices, moments and quadrature II; how to compute the norm of the error in iterative methods // BIT Numerical Mathematics. — 1997. — Vol. 37, № 3. — P. 687–705.
7. **Hestenes M.R., Stiefel E.** Methods of conjugate gradients for solving linear systems // J. Research NBS. — 1952. — Vol. 49, № 6. — P. 409–436.
8. **Meurant G.** The computation of bounds for the norm of the error in the conjugate gradient algorithm // Numerical Algorithms. — 1997. — Vol. 16, № 1. — P. 77–87.
9. **Meurant G.** Estimates of the l_2 norm of the error in the conjugate gradient algorithm // Numerical Algorithms. — 2005. — Vol. 40, № 2. — P. 157–169.
10. **Meurant G.** Estimates of the norm of the error in solving linear systems with FOM and GMRES // SIAM J. Sci. Comput. — 2011. — Vol. 33, № 5. — P. 2686–2705.
11. **Meurant G., Strakos Z.** The Lanczos and conjugate gradient algorithms in finite precision arithmetic // Acta Numer. — 2006. — Vol. 15. — P. 471–542.
12. **Meurant G., Tichy P.** On computing quadrature-based bounds for the A-norm of the error in conjugate gradients // Numerical Algorithms. — 2013. — Vol. 62, № 2. — P. 163–191.
13. **Saad Y.** Preconditioning techniques for nonsymmetric and indefinite linear systems // J. Comput. Appl. Mathematics. — 1988. — Vol. 24, № 1-2. — P. 89–105.
14. **Saad Y., Schultz M.H.** GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems // SIAM J. Sci. and Stat. Comput. — 1986. — Vol. 7, № 3. — P. 856–869.
15. **Saylor P.E., Smolarski D.C.** Why Gaussian quadrature in the complex plane? // Numerical Algorithms. — 2001. — Vol. 26, № 3. — P. 251–280.
16. **Sleijpen G.L., Fokkema D.R.** BiCGstab(l) for linear equations involving unsymmetric matrices with complex spectrum // Electronic Transactions on Numerical Analysis. — 1993. — Vol. 1. — P. 11–32.
17. **Sleijpen G.L., Van der Vorst H.A., Fokkema D.R.** BiCGstab(l) and other hybrid BI-CG methods // Numerical Algorithms. — 1994. — Vol. 7, № 1. — P. 75–109.
18. **Strakos Z., Tichy P.** Error estimation in preconditioned conjugate gradients // BIT Numerical Mathematics. — 2005. — Vol. 45, № 4. — P. 789–817.
19. **Trefethen L.N., Bau III D.** Numerical Linear Algebra. — SIAM, 1997.
20. **Van der Vorst H.A.** Bi-CGSTAB: A fast and smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems // SIAM J. Sci. and Stat. Comput. — 1992. — Vol. 13, № 2. — P. 631–644.

*Поступила в редакцию 07 февраля 2022 г.
После исправления 10 сентября 2022 г.
Принята к печати 30 января 2023 г.*

