Записки научных
семинаров ПОМИ
Том 248, 1998 г.

А. Ю. Еремин, И. Е. Капорин

ВЛИЯНИЕ НАИБОЛЬШИХ СОБСТВЕННЫХ ЗНАЧЕНИЙ НА ЧИСЛЕННУЮ СХОДИМОСТЬ МЕТОДА СОПРЯЖЕННЫХ ГРАДИЕНТОВ

1. Постановка задачи

Рассмотрим метод сопряженных градиентов, примененный к решению (предобусловленной) системы уравнений очень большого размера

$$Mx = b, \tag{1.1}$$

где матрица M является симметричной положительно определенной,

$$M = M^T > 0.$$

Теоретически, метод сопряженных градиентов эффективно использует наличие изолированных собственных значений вблизи обоих концов спектра собственных значений матрицы M. Таким образом, в этом случае для сходимости метода с заданной точностью требуется выполнить гораздо меньше итераций по сравнению с оценкой, вытекающей из стандартной границы погрешности через максимальное и минимальное собственные значения матрицы M:

$$\frac{\|r_k\|_{M^{-1}}}{\|r_0\|_{M^{-1}}} \leqslant \frac{1}{T_k(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1})}, \tag{1.2}$$

где n – порядок матрицы M, а

$$0 < \lambda_1 \leqslant \cdots \leqslant \lambda_n$$

– собственные значения матрицы M, упорядоченные в порядке неубывания.

Работа выполнена при частичной финансовой поддержке Огранизации научных исследований Нидерландов (the Netherlands Organisation for Scientific Research (NWO)), грант 047 003 017, и ИНТАСа и РФФИ, грант 95-0098.

Однако в случае, когда вычисления проводятся в арифметике с плавающей точкой, наличие изолированных *наибольших* собственных значений матрицы M обычно приводит к существенному увеличению числа итераций метода сопряженных градиентов (по сравнению с его “точным” аналогом). Увеличение числа итераций может быть настолько значительным, что наличие изолированных наибольших собственных значений матрицы M приходится рассматривать скорее как недостаток выбранной стратегии предобусловливания. В настоящей работе приводятся результаты некоторых численных экспериментов, свидетельствующие в пользу такой точки зрения.

2. О процессе Ланцоша и значениях Ритца, связанных с методом сопряженных градиентов

Сравнение точного спектра собственных значений матрицы M со спектром значений Ритца (совпадающих с множеством собственных значений трехдиагональной матрицы T_k, возникающей на k-й итерации метода сопряженных градиентов) позволяет сделать следующие заключения.

Спектр матрицы M обычно можно разделить на следующие три подмножества:

(a) Наименьшие изолированные собственные значения.

(b) Кластеризованные собственные значения (т.е. не отличимые друг от друга и воспринимаемые как непрерывный спектр).

(c) Наибольшие изолированные собственные значения.

При решении задач очень большого размера методом сопряженных градиентов негативное влияние наибольших изолированных собственных значений M можно объяснить следующим образом.

На k-й итерации метода сопряженных градиентов, соответствующий ей спектр Ритца также может быть разбит на аналогичные три подмножества:

(a) Значения Ритца, хорошо аппроксимирующие наименьшие собственные значения матрицы M.

(b) Значения Ритца, плохо аппроксимирующие собственные значения матрицы M или вообще не аппроксимирующие их, лежащие в области кластеризованных собственных значений матрицы M.

(c) Значения Ритца, хорошо аппроксимирующие наибольшие собственные значения матрицы M, но *представленные с избыточными кратностями.*

Как правило, чем больше величина значения Ритца из подмножества (c), тем больше его кратность. Кроме того, эта кратность возрастает примерно как линейная функция от числа итераций k метода сопряженных градиентов. Далее, эта кратность увеличивается при переходе к вычислениям с меньшей точностью представления чисел (напр., при замене типа чисел double precision на тип single precision).

Наконец, количество наибольших собственных значений, которые могут рассматриваться как изолированные, также возрастает с увеличением числа итераций k, так как критерий "изолированности" i-го собственного значения оказывается следующим:

$$\frac{|\lambda_i - \lambda_{i+1}|}{\lambda_i} = O(k^{-2}), \quad \frac{|\lambda_i - \lambda_{i-1}|}{\lambda_i} = O(k^{-2}). \tag{2.1}$$

Суммарное количество всех избыточных кратностей наибольших значений Ритца весьма точно соответствует числу "лишних" итераций метода сопряженных градиентов, выполняемых исключительно ввиду наличия погрешностей арифметики с плавающей точкой [2].

Таким образом, если предобусловливание оказалось "неустойчивым" (т.е. матрица M имеет много изолированных наибольших собственных значений, вызывающих интенсивную потерю ортогональности в рекуррентных соотношениях метода сопряженных градиентов), то весьма значительная часть итераций (50% и более) фактически выполняется лишь для того, чтобы преодолеть неточность вычислений с плавающей точкой.

3. Основные результаты работы

В настоящей работе мы рассмотрим некоторые численные примеры, показывающие что:

(a) при решении плохо обусловленных систем большого размера интенсивность дублирования *наибольших* значений Ритца может быть столь велика, что наличием промежутков между наибольшими собственными значениями матрицы M практически можно пренебречь (так как оно не может быть эффективно использовано);

(b) до тех пор, пока дублирование *наименьших* изолированных собственных значений не влияет на сходимость метода сопряженных градиентов, реальная скорость сходимости может быть достаточно точно описана в терминах оценки погрешности, зависящей только от максимального собственного значения матрицы M и некоторого числа ее наименьших собственных значений (примеры таких оценок можно найти в [1, 2]):

$$\frac{\|r_k\|_{M^{-1}}}{\|r_0\|_{M^{-1}}} \leqslant \min_{0\leqslant j \ll k} \frac{\prod_{i=1}^{j}(\frac{e\lambda_j}{\lambda_i})}{T_{k-j}(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1})}, \quad e = \exp(1); \tag{3.1}$$

(c) как только обнаруживает себя копирование *наименьших* собственных значений матрицы M (что приводит к дополнительному замедлению сходимости метода сопряженных градиентов), реальная скорость сходимости может быть описана в терминах наименьших значений Ритца

$$0 < \mu_1 \leqslant \ldots \leqslant \mu_k$$

(с учетом их дублирования) следующим образом:

$$\frac{\|r_k\|_{M^{-1}}}{\|r_0\|_{M^{-1}}} \leqslant \min_{0\leqslant j \ll k} \frac{\prod_{i=1}^{j}(\frac{e\mu_j}{\mu_i})}{T_{k-j}(\frac{\mu_n+\mu_1}{\mu_n-\mu_1})}, \quad e = \exp(1); \tag{3.2}$$

здесь символ “$\ll$” означает, что берется первый локальный минимум, найденный при $j = 0, 1, \ldots$.

4. Описание тестовых матриц

Для построения тестовых задач использовалось семейство матриц, параметризованных шестью параметрами, которые определяли границы трех сегментов, содержавших спектр матрицы, и число собственных значений в каждом из этих сегментов. Внутри каждого сегмента собственные значения матрицы M были подчинены равномерному распределению, т.е.

$$\begin{aligned}
\lambda_i &= \gamma_0 + (\gamma_1 - \gamma_0)\frac{i-1}{m_L}, & i &= 1, \ldots, m_L,\\
\lambda_i &= \gamma_1 + (\gamma_2 - \gamma_1)\frac{i-m_L-1}{n-m_L-m_R-1}, & i &= m_L+1, \ldots, n-m_R-1,\\
\lambda_i &= \gamma_2 + (\gamma_3 - \gamma_2)\frac{i-n+m_R}{m_R}, & i &= n-m_R+1, \ldots, n,
\end{aligned}$$

где во всех случаях полагалось $\gamma_2 = 1$. Таким образом,

$$\lambda_1 - \gamma_0, \qquad \lambda_n = \gamma_3,$$

а m_L и m_R представляют собой количество изолированных наименьших и наибольших собственных значений, соответственно. Вводя диагональную матрицу

$$\Lambda = \mathrm{Diag}(\lambda_1, \dots, \lambda_n),$$

мы определяем матрицу M, используемую на итерациях метода сопряженных градиентов, как произведение

$$M = \left(I - 2\frac{vv^T}{v^T v}\right)\lambda\left(I - 2\frac{vv^T}{v^T v}\right)$$

с некоторым ненулевым вектором v. В наших экспериментах использовался вектор

$$v = [1, 1, \dots, 1]^T.$$

Собственные значения матрицы M очевидно равны $\lambda_1, \dots, \lambda_n$, так как $I - 2vv^T/v^T v$ представляет собой элементарную матрицу отражения. В то же время, эта матрица может храниться с использованием лишь $O(n)$ ячеек памяти и может быть умножена на вектор за $O(n)$ арифметических операций. Таким образом, указанное семейство матриц позволяет провести нетривиальные тестовые расчеты по методу сопряженных градиентов для матриц с заранее известным спектром собственных значений при минимальных затратах вычислительных ресурсов.

Для тестовых задач типа (1.1), описанных ниже, использовалась правая часть вида

$$b = Mx_*,$$

где точное решение x_* выбиралось в виде

$$x_* = [n, n/2, n/3, \dots, 1]^T.$$

В качестве начального приближения для метода сопряженных градиентов использовался нулевой вектор.

5. Насколько много избыточных итераций может появиться из-за наличия изолированных наибольших собственных значений

Для того, чтобы нарушились свойства ортогональности "точных" рекуррентных соотношений метода сопряженных градиентов, вовсе не обязательно выбирать, например, $\lambda_n = 10^9$, $\lambda_{n-1} =$

10^6, $\lambda_{n-1} = 10^3$, а остальные собственные значения – между 0 и 1, как это делалось в [2] и других работах. При решении недостаточно хорошо обусловленных систем для сходимости могут потребоваться до нескольких тысяч итераций, и для того, чтобы некоторое собственное значение не оказалось изолированным в смысле (2.1), спектр должен быть весьма плотным слева от λ_n. В противном случае, потеря ортогональности из-за дублирования наибольших значений Ритца возникнет уже после первых нескольких десятков итераций.

Рассмотрим следующие тестовые задачи:

Задача 1.	Задача 2.	Задача 3.
$n = 100000,$	$n = 100000,$	$n = 100000,$
$m_L = 200,$	$m_L = 200,$	$m_L = 200,$
$m_R = 20,$	$m_R = 50,$	$m_R = n - 200,$
$\gamma_0 = 0.001,$	$\gamma_0 = 0.001,$	$\gamma_0 = 0.001,$
$\gamma_1 = 0.5,$	$\gamma_1 = 0.5,$	$\gamma_1 = 1.0,$
$\gamma_2 = 1.0,$	$\gamma_2 = 1.0,$	$\gamma_2 = 1.0,$
$\gamma_3 = 10.0.$	$\gamma_3 = 10.0.$	$\gamma_3 = 10.0.$

Промежутки между соседними наибольшими собственными значениями в задачах 1 и 2 составляют лишь 5% и 2% от λ_n, соответственно. Эти задачи сравнивались с задачей 3, имеющей то же число обусловленности 10^4, но очень плотный спектр слева от λ_n.

После достижения сходимости итераций с точностью $\varepsilon = 10^{-7}$ (так как использовалась арифметика с однократной точностью, то не было смысла брать меньшие значения ε), производилось вычисление всех значений Ритца, т.е. собственных значений трехдиагональной матрицы, построенной к последней итерации. Из результатов работ [1] и [2] следует, что "теоретическое" число итераций может быть оценено как количество (численно) различных значений Ритца. Таким образом, общее количество избыточных ритцевских копий для всех собственных значений принималось равным количеству "избыточных итераций", связанных с вмешательством погрешностей арифметики с плавающей точкой.

Для указанных трех задач на Рис. 1 приведены графики величины значений Ритца μ_i как функции их номера (напомним,

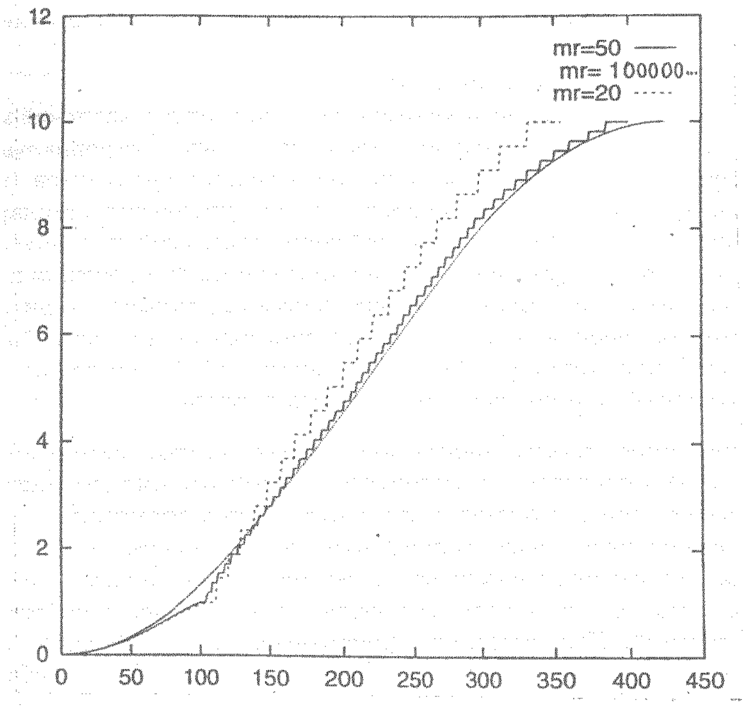

Рис. 1. Величины значений Ритца в зависимости от их номера для задач 1–3.

что значения Ритца упорядочены по возрастанию). Для первых двух задач интенсивное возникновение копий Ритца для наибольших собственных значений очевидным образом приводит к кривым типа “лестницы”. Каждая “ступенька” соответствует одному собственному значению матрицы M, в то время как ее длина пропорциональна количеству избыточных ритцевских копий этого собственного значения. Для задачи 1 сходимость была достигнута за 355 итераций, 219 из которых были избыточными. Для задачи 2 эти числа составляли 401 и 241, соответственно. В то же время, для задачи 3 метод сопряженных градиентов достиг сходимости за 425 итераций, копирования значений Ритца не наблюдалось, и это число итераций было весьма близко к его “точному” значению. Соответствующая кривая на Рис. 1 выглядит вполне гладкой и весьма близка к сдвинутому и отмасштабированному графику функции $1-\cos(x)$. (Эта функция возникает, когда значения Ритца близки к корням сдвинутого многочлена Чебышева.)

Отсюда можно заключить, что:

(a) только отделенность *наименьших* собственных значений может быть достаточно надежно учтена некоторыми нетривиальными оценками сходимости метода сопряженных градиентов (напр., (3.1)) которые используются с целью уточнения стандартной оценки (1.2); (b) основное действие стратегий предобусловливания, предназначенных для использования при решении плохо обусловленных задач большого размера, должно сводиться (помимо улучшения числа обусловленности) не только к "разрежению" наименьших собственных значений M, но также и к "уплотнению" наибольших собственных значений.

6. Численное тестирование некоторых оценок сходимости метода сопряженных градиентов, принимающих во внимание отделенность наименьших собственных значений

Из сказанного выше следует, что любая оценка сходимости метода сопряженных градиентов, имеющая практическое значение, должна игнорировать отделенность наибольших собственных значений. Действительно, почти весь "теоретический" выигрыш, полученный от учета отделенность наибольших собственных значений, сводится на нет за счет копирования соответствующих значений Ритца в присутствии погрешностей арифметики с плавающей точкой. Отсюда ясно, почему мы отдаем предпочтение оценке (3.1), легко вытекающей из результатов работ [1] и [2]. Однако, как это отмечено в [2] и подтверждено многими численными экспериментами, эта оценка справедлива лишь до тех пор, пока не началось копирование наименьших значений Ритца. В противном случае, остается справедливой только оценка (3.2), предложенная в настоящей работе. Заметим, что эта оценка включает в себя лишь величины $\mu_1, \mu_2, \dots$, и ее можно легко вычислить на любой итерации метода сопряженных градиентов (в отличие от точных крайних собственных значений, которые не известны до тех пор, пока итерации не сошлись с большой точностью). Таким образом, (3.2) дает надежную и реалистическую оценку для величины $\|r_k\|_{M^{-1}}$, точное значение которой не может быть известно при решении прикладных задач. С другой стороны, оценка (3.1) все же полезна для развития практических критериев качества предобусловливания.

Чтобы продемонстрировать, как оценки (1.2), (3.1) и (3.2) соотносятся с реальным поведением погрешности метода сопряженных градиентов, рассмотрим следующую тестовую задачу:

Задача 4.

$$
\begin{aligned}
&n = 100000,\\
&m_L = 200,\\
&m_R = 100,\\
&\gamma_0 = 0,001\\
&\gamma_1 = 0.5,\\
&\gamma_2 = 1.0,\\
&\gamma_3 = 10.0.
\end{aligned}
$$

Правая часть и начальное приближение выбраны такими же, как и ранее. Таким образом, мы опять рассматриваем задачу с отделенными наибольшими собственными значениями. Однако, несмотря на связанное с этим возможное ускорение сходимости, оценки (3.1) (до определенного момента) и (3.2) (для любого числа итераций) дают удовлетворительное описание истории сходимости итераций.

На Рис. 2 показаны графики десятичных логарифмов величины $\|r_k\|_{M^{-1}}$ и правых частей неравенств (1.2), (3.1) и (3.2). До 130-й итерации оценки (1.2) и (3.1) просто совпадают и заметно превосходят реальную погрешность. В то же время, новая оценка (3.2) оказывается гораздо точнее, чем две другие, что можно легко объяснить тем фактом, что на начальных итерациях метода сопряженных градиентов наименьшие значения Ритца значительно превосходят соответствующие наименьшие собственные значения матрицы M.

Между 130-й и 180-й итерациями все три оценки дают примерно одну и ту же оценку погрешности в силу того, что минимальные значения Ритца сошлись к минимальным собственным значениям матрицы M, но число итераций еще не достаточно велико, чтобы выявить преимущество оценок (3.1) и (3.2).

Между 180-й и 310-й итерациями оценки (3.1) и (3.2) также почти совпадают (хотя (3.2) все же немного точнее) и весьма

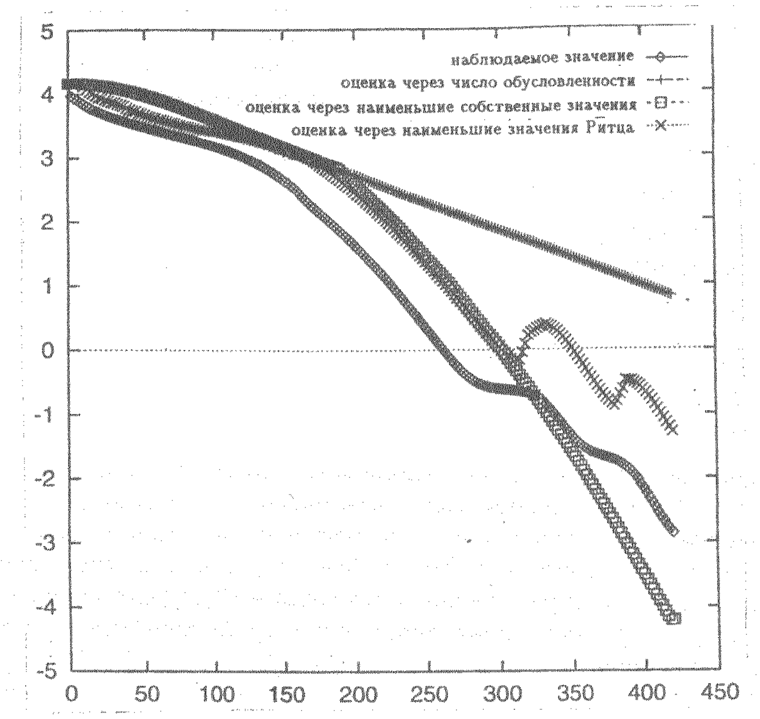

Рис. 2. Оценки итерационной погрешности как функции от номера итерации для задачи 4.

точно повторяют наклон кривой, которая соответствует реальной истории сходимости. Стандартная оценка (1.2) дает здесь весьма завышенную оценку погрешности и не отражает ускорения сходимости итераций, в то время как оценки (3.1) и (3.2) учитывают влияние отделенности наименьших собственных значений.

Однако, ситуация меняется, начиная с 310–320-й итерации, когда происходит копирование наименьшего значения Ритца и наблюдается ощутимое замедление сходимости. Поэтому оценка (3.1) становится слишком оптимистичной и попросту дает ложный результат. Напротив, оценка (3.2) сохраняет свою силу и даже отражает ускорения и замедления процесса сходимости. На этой стадии итерационного процесса оценка (3.2) по-прежнему гораздо точнее, чем (1.2), но их (средние) наклоны становятся ближе друг к другу. Такое поведение реальной погрешности и рассматриваемых оценок весьма типично для ситуаций, когда требуемая относительная точность невязки ε берется близкой к константе машинной точности (около 10^{-7} в нашем случае).

7. Заключительные замечание

Из сказанного выше следует, что стратегии предобусловливания, которые предназначены для использования с итерационными методами типа сопряженных градиентов для решения очень больших и плохо обусловленных систем уравнений, не должны приводить к возникновению предобусловленных матриц со сколько-нибудь "разреженным" спектром наибольших по модулю собственных значений. Хороший пример предобусловливателей, удовлетворяющих указанному требованию, дает блочный метод симметричной последовательной верхней релаксации (Block SSOR).

Формально, если матрица M имеет вид $M = L^{-1}AL^{-T}$, т.е. получена из исходной симметричной положительно определенной матрицы A и ее факторизованной аппроксимации $B = LL^T$, так что

$$B \approx A, \qquad \mathrm{trace}\,(B^{-1}A) \approx 1,$$

то матрица B должна представлять собой "устойчивую" аппроксимацию, т.е. необходимо потребовать, чтобы оценка

$$B^{-1} \leqslant (1+\gamma)A^{-1}$$

выполнялась для некоторого достаточно малого значения $\gamma > 0$.

Литература

1. O. Axelsson and G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient methods.* — Numer. Math. **48** (1986), 499–523.
2. Y. Notay, *On the convergence rate of the conjugate gradients in presence of rounding errors.* — Numer. Math. **65** (1993), 301–317.

Yeremin A. Yu. and Kaporin I. E. The influence of isolated largest eigenvalues on the numerical convergence of the CG method.

This paper considers the dependence of the convergence history of the CG method on largest eigenvalues of a symmetric positive definite matrix. It is demonstrated that, in solving ill-conditioned linear systems, the reproduction of largest eigenvalues can be so intensive that large

eigenvalues cannot be treated as isolated. On the other hand, since the moment at which the smallest isolated eigenvalues start to govern the numerical convergence of the CG method, the character of convergence mainly depends on the smallest Ritz values.

Центр суперкомпьютерных
и массивно параллельных
вычислений при Вычислительном
центре РАН

Поступило 24 апреля 1998 г.