Об оценке сходимости метода бидиагонализации[$]

С. А. Горейнов[@]

Получена оценка погрешности матричной билинейной аппроксимации, получаемой методом бидиагонализации в условиях работы в точной арифметике, являющаяся более точной, чем применение к данной задаче аппроксимации известных результатов о сходимости метода Ланцоша.

Пусть $A \in \mathbb{C}^{m\times n}$ — заданная матрица. Рассмотрим задачу отыскания $\min_{U,V} k$, где $U \in \mathbb{C}^{m\times k}$, $V \in \mathbb{C}^{n\times k}$ удовлетворяют условию $\|A-UV^*\|_F \leqslant \varepsilon$. Матричные приближения такого сорта, называемые билинейными — весьма полезный инструмент при аппроксимации больших плотных матриц, имеющих отношение к так называемым асимптотически гладким функциям и нелокальным (интегральным, интегро-дифференциальным) операторам, возникающим в задачах математической физики [11, 3]. Для „хороших“ матриц A ранг билинейной аппроксимации k значительно меньше m и n. В связи с этим были развиты методы, позволяющие получать билинейные аппроксимации таких матриц за линейное по совокупности m и n время [12, 9]. В данной работе обсуждается метод бидиагонализации [4], который применим к более широкому классу матриц, но имеет время работы $O(mn)$. Основной результат — оценка погрешности

[$]Работа выполнена при частичной финансовой поддержке Российского фонда фундаментальных исследований (гранты РФФИ №№05-01-00721 и 06-01-08052) и программы фундаментальных исследований отделения математических наук РАН «Вычислительные и информационные проблемы решения больших задач» по проекту «Матричные методы и технологии для задач со сверхбольшим числом неизвестных».

[@]Институт вычислительной математики РАН

получаемой аппроксимации в условиях работы в точной арифметике, являющаяся более точной, чем применение известных результатов о сходимости метода Ланцоша [7, 8] к нашей задаче.

Вот решение нашей задачи в терминах сингулярного разложения исходной матрицы A.

Теорема 1 (Шмидт, Мирский [10]). *Рассмотрим сингулярное разложение матрицы* $A \in \mathbb{C}^{m\times n}$*:*

$$A = U\Sigma V^*, \qquad U^*U = V^*V = I, \qquad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_p),$$

где $p = \min(m, n)$, $U \in \mathbb{C}^{m\times p}$, $V \in \mathbb{C}^{n\times p}$, *и сингулярные числа* $\{\sigma_k\}$ *упорядочены по невозрастанию. Обозначим символами* U_k, V_k *матрицы, составленные из первых* k *столбцов* U *и* V. *Пусть также* Σ_k *обозначает ведущую подматрицу* Σ *порядка* k. *Тогда*

$$\min_{B:\ \operatorname{rank} B = k} \|A - B\|_F^2 = \|A - U_k\Sigma_k V_k{}^*\|_F^2 = \sigma_{k+1}^2 + \cdots + \sigma_p^2. \tag{1}$$

Таким образом, первые k сингулярных векторов и чисел, где k — минимальный индекс, обеспечивающий неравенство $\sigma_{k+1}^2 + \cdots + \sigma_p^2 \leqslant \varepsilon$, образуют искомую билинейную аппроксимацию A.

Один из известных способов (приближенного) отыскания нескольких старших собственных векторов и чисел эрмитовой матрицы — итерационный алгоритм Ланцоша [2]. Если применить его к матрице A^*A и преобразовать с тем, чтобы в расчетных формулах выражение A^*A не возникало[1], получится алгоритм бидиагонализации, используемый для (приближенного) отыскания нескольких старших сингулярных векторов и чисел A. Выпишем управляющие соотношения обоих методов.

Пусть q_1, $q_2, \ldots, q_k, \ldots$ — векторы процесса Ланцоша, пусть Q_k обозначает матрицу, составленную из первых k этих векторов,

[1]в противном случае имеется нежелательный эффект при реальных расчетах, характеризуемый фразами „возведение обусловленности в квадрат“ или „извлечение квадратного корня из машинного ε“.

и пусть

$$T_k = \begin{pmatrix} \omega_1 & t_1^* & & \\ t_1 & \ddots & \ddots & \\ & \ddots & \ddots & t_{k-1}^* \\ & & t_{k-1} & \omega_k \end{pmatrix}$$

обозначает соответствующую трехдиагональную матрицу. Тогда имеет место соотношение

$$A^*AQ_k = Q_kT_k + q_{k+1}t_ke_k^T, \tag{2}$$

причем $Q_k^*Q_k = I$, $Q_k^*q_{k+1} = 0$.

Чтобы перейти к алгоритму бидиагонализации, введем разложение Холецкого $T_k = R_k^*R_k$, где матрица

$$R_k = \begin{pmatrix} \rho_1 & \tau_1^* & & \\ & \ddots & \ddots & \\ & & \ddots & \tau_{k-1}^* \\ & & & \rho_k \end{pmatrix}$$

верхняя треугольная (бидиагональная), и обозначим $P_k = AQ_kR_k^{-1}$. В силу (2) матрица P_k имеет ортонормированные столбцы. Подставляя P_k в (2) и умножая полученное равенство на R_k^{-1}, получаем соотношения алгоритма бидиагонализации:

$$\begin{cases} AQ_k = P_kR_k, \\ A^*P_k = Q_kR_k{}^* + q_{k+1}\tau_ke_k{}^*, \end{cases} \tag{3}$$

где $\tau_k = t_k\rho_k^{-1}$.

Соотношения (3) „наращиваются“ точно так же, как (2): скалярные величины τ_k вместе с q_{k+1} можно вычислять нормализацией вектора $A^*p_k - q_k\rho_k{}^*$, а ρ_k вместе с p_k получаются нормализацией вектора $Aq_k - p_{k-1}\tau_{k-1}^*$.

Значки комплексного матричного сопряжения, стоящие кое-где в наших формулах при скалярных величинах t_k, τ_k, ρ_k, означают, что формулы верны и для блочных вариантов рассматриваемых методов [13, 4].

Сходимость метода бидиагонализации, используемого для нахождения билинейного приближения A, контролируется на практике весьма просто.

Лемма 1 ([5]).

$$\|A - P_k R_k Q_k{}^*\|_F^2 = \|A\|_F^2 - \|R_k\|_F^2. \tag{4}$$

Доказательство. В силу первой формулы (3)

$$A - P_k R_k Q_k{}^* = A - AQ_k Q_k{}^* = AQ_{kO} Q_{kO}^*,$$

где Q_{kO} — дополнение Q_k до квадратной унитарной матрицы Q порядка n. Дополнив таким же образом P_k, запишем

$$A\,[\,Q_k \quad Q_{kO}\,] = [\,P_k \quad P_{kO}\,] \begin{bmatrix} R_k & S \\ 0 & T \end{bmatrix},$$

где $S = P_k^* A Q_{kO}$ и $T = P_{kO}^* A Q_{kO}$. Поэтому, в силу унитарной инвариантности и аддитивности по подматрицам фробениусовой нормы,

$$\begin{aligned}
\|A - P_k R_k Q_k{}^*\|_F^2 &= \|AQ_{kO} Q_{kO}^*\|_F^2 = \operatorname{tr}(AQ_{kO} Q_{kO}^* Q_{kO} Q_{kO}^* A^*) \\
&= \|AQ_{kO}\|_F^2 = \|P_k S + P_{kO} T\|_F^2 \\
&= \operatorname{tr}(P_k S + P_{kO} T)^* (P_k S + P_{kO} T) \\
&= \|S\|_F^2 + \|T\|_F^2 = \left\| \begin{bmatrix} R_k & S \\ 0 & T \end{bmatrix} \right\|_F^2 - \|R_k\|_F^2. \quad \square
\end{aligned}$$

Следствие 1.

$$\|A - P_k R_k Q_k{}^*\|_F^2 = \operatorname{tr}(A^* A) - \operatorname{tr} T_k. \tag{4$'$}$$

Если известны оценки чисел Ритца (собственных чисел матрицы T_k), то, ввиду (4$'$), легко получить и теоретическую оценку сходимости. С этой целью приведем здесь известный результат.

Теорема 2 ([13], Theorem 2.6.1, p. 37). *Пусть* λ_1, $\lambda_2, \ldots, \lambda_n$ *— собственные значения матрицы* A^*A, *упорядоченные по невозрастанию. Пусть* μ_1, $\mu_2, \ldots, \mu_{ps}$ *— так же упорядоченные собственные значения матрицы* T_s *процесса Ланцоша с размером блока* p. *Пусть блочный вектор* $v_1 \in \mathbb{C}^{n\times p}$ *содержит первые* p *собственных векторов* A^*A. *Предположим, что* $\lambda_p > \lambda_{p+1}$ *и что* $\cos\theta := \sigma_{\min}(q_1^* v_1) > 0$. *Тогда для собственных значений матрицы* T_s *имеют место неравенства*

$$\lambda_k \geqslant \mu_k \geqslant \lambda_k - {\varepsilon_k}^2, \qquad k = 1, \ldots, p,$$

$$\varepsilon_k^2 = (\lambda_k - \lambda_n)\,\frac{\operatorname{tg}^2\theta}{\mathcal{T}_{s-1}^2\left(\frac{1+\gamma_k}{1-\gamma_k}\right)}, \qquad \gamma_k = \frac{\lambda_k - \lambda_{p+1}}{\lambda_k - \lambda_n}. \tag{5}$$

Символом $\mathcal{T}_{s-1}$ *здесь обозначен многочлен Чебышева первого рода степени* $s-1$.

Однако непосредственные оценки следов T_k приводят к более точному результату; переходим к его изложению. Наряду с ланцошевскими векторами q_k введем и ланцошевские многочлены $\mathfrak{p}_k$: по индукции легко следует, что $q_k = \mathfrak{p}_{k-1}(A^*A)q_1$. Далее, обозначим матрицу собственных векторов A^*A через V, матрицу собственных значений через Λ, и введем вектор $\varphi = V^* q_1$. Тогда

$$Q_k = V\,[\,\varphi \quad \mathfrak{p}_1(\Lambda)\varphi \quad \ldots \quad \mathfrak{p}_{k-1}(\Lambda)\varphi\,],$$

откуда следует, что

$$\sum_{\mu=1}^{n} |\varphi_\mu|^2\, \mathfrak{p}_j(\lambda_\mu)\, \mathfrak{p}_k(\lambda_\mu) = q_j^* q_k = \delta_{kj}.$$

Определение 1 ([1]). *Аддитивное и однородное отображение* $\mathfrak{S}$ *пространства многочленов* $\{p(x)\}$, *рассматриваемых на отрезке* K, *на множество вещественных чисел назовем позитивным функционалом, если всякий раз из* $p(x) \geqslant 0$, $\forall x \in K$, *и* $p(x) \not\equiv 0$ *следует, что* $\mathfrak{S}(p) > 0$.

Поскольку $|\varphi_1|^2 + \cdots + |\varphi_n|^2 = 1$, ясно, что функционал $\mathfrak{S}(p) \equiv \sum_{\mu=1}^n |\varphi_\mu|^2 p(\lambda_\mu)$ позитивен на отрезке $[\lambda_n, \lambda_1]$ для многочленов степени меньшей $2n-2$.

Теорема 3 ([1], с.80). *Функция* $\rho_k(z)$, *определяемая по последовательности многочленов* $\{\mathfrak{p}_k\}$, *ортогональных относительно позитивного функционала* $\mathfrak{S}$, *следующим образом:* $\rho_k(z)^{-1} = \sum_{j=1}^k |\mathfrak{p}_j(z)|^2$, *удовлетворяет соотношению*

$$\rho_k(z) = \min_{\substack{\deg P_k \leqslant k \\ P_k(z)=1}} \mathfrak{S}(|P_k|^2).$$

Следствие 2. $\rho_j(\lambda) \leqslant \sum_{\mu=1}^n |\varphi_\mu|^2 P_j^2(\lambda_\mu)$ *для всякого многочлена* P_j *степени не выше* j *с вещественными коэффициентами, принимающего значение* 1 *в точке* λ.

Таким образом,

$$\begin{aligned} \operatorname{tr} T_k &= \sum_{\mu=1}^k \sum_{j=1}^n |\varphi_j|^2 \lambda_j \, |\mathfrak{p}_{\mu-1}(\lambda_j)|^2 = \sum_{j=1}^n \frac{|\varphi_j|^2 \lambda_j}{\rho_{k-1}(\lambda_j)} \\ &\geqslant \sum_{j=1}^k \frac{|\varphi_j|^2 \lambda_j}{\rho_{k-1}(\lambda_j)} \geqslant \sum_{j=1}^k \frac{|\varphi_j|^2 \lambda_j}{\sum_{\mu=1}^n |\varphi_\mu|^2 P_{k,j}^2(\lambda_\mu)}, \end{aligned}$$

где $P_{k,j}(\lambda)$, $j = 1, \ldots, k$ — многочлен степени не выше $k-1$, принимающий в точке λ_j значение 1.

Далее, преобразуем знаменатель в последней сумме:

$$\sum_{\mu=1}^n |\varphi_\mu|^2 P_{k,j}^2(\lambda_\mu) \leqslant |\varphi_j|^2 + \sum_{\mu=1,\,\mu\neq j}^n |\varphi_\mu|^2 \big(\max_{\lambda\in K_j} |P_{k,j}(\lambda)|\big)^2,$$

где множество $K_j = [\lambda_n, \lambda_{j+1}] \cup [\lambda_{j-1}, \lambda_1]$.

В связи с последним максимумом выберем в качестве $P_{k,j}$ многочлен степени не выше k, равный 1 в точке λ_j и наименее уклоняющийся от нуля на указанном объединении отрезков, а значение максимума для него обозначим $\mathcal{Z}_{k,j,n}$.

Эти многочлены были исследованы Золотаревым, носят его имя, и допускают явное представление через тета-функции Якоби [6, Lemma 2.1, Problem *BZ*13].

Наконец, имеем

$$\operatorname{tr} T_k \geqslant \sum_{j=1}^{k} \lambda_j \frac{|\varphi_j|^2}{|\varphi_j|^2 + (1 - |\varphi_j|^2)\mathcal{Z}_{k,j,n}^2}.$$

Обозначая через $\varepsilon_k^\star = \sum_{j=k+1}^{n} \lambda_j$ — квадрат погрешности наилучшей k-ранговой аппроксимации A во фробениусовой норме, имеем

$$\|A - P_k R_k Q_k{}^*\|_F^2 \leqslant \varepsilon_k^\star + \sum_{j=1}^{k} \lambda_j \frac{\mathcal{Z}_{k,j,n}^2}{|\varphi_j|^2}. \qquad (6)$$

Мы пришли к следующей теореме.

Теорема 4. *Пусть все проекции стартового вектора процесса бидиагонализации* q_1 *на правые сингулярные векторы матрицы* A *ненулевые и все сингулярные числа* A *различны. Тогда погрешность* k*-ранговой аппроксимации матрицы* A, *порождаемой процессом бидиагонализации, оценивается по формуле* (6).

Список литературы

[1] Ахиезер Н. И. *Классическая проблема моментов и некоторые вопросы анализа, связанные с нею.* М.: Физматлит, 1961.

[2] Парлетт Б. *Симметричная проблема собственных значений.* М.: Мир, 1983.

[3] Тыртышников Е. Е. Тензорные аппроксимации матриц, порожденных асимптотически гладкими функциями. // *Матем. сб.* 2003. Т. 194. №6. С. 147–160.

[4] Golub G., Luk F., Overton M. A block Lanczos method for computing the singular values and corresponding singular vectors of a matrix. // *ACM Transactions on Math. Soft.* 1981. V. 7. №2. P. 149–169.

[5] Goreinov S. A., Tyrtyshnikov E. E., Yeremin A. Yu. *Matrix-free iterative solution strategies for large dense linear systems.* Numer. Linear Algebra Appl. 1997. V.4. №4. P.273–294.

[6] Lebedev V. I. Zolotarev polynomials and extremum problems. // *Russian J. Numer. Anal. Math. Modelling.* 1994. V.9. №3. P. 231–263.

[7] Paige C. C. *The computation of eigenvalues and eigenvectors of very large sparse matrices.* Ph.D.thesis. The University of London. 1971.

[8] Saad Y. *Numerical methods for large eigenvalue problems.* Manchester University Press, 1992.

[9] Oseledets I. V., Savostianov D. V., Tyrtyshnikov E. E. Tucker dimensionality reduction of three-dimensional arrays in linear time. // *SIAM J. Matr. Anal.* 2008. In press.

[10] G. W. Stewart, J. Sun. *Matrix Perturbation Theory.* Academic Press, 1990.

[11] E. E. Tyrtyshnikov. Mosaic-skeleton approximations. // *Calcolo.* 1996. V.33. №1–2. P.47–57.

[12] E. E. Tyrtyshnikov. Incomplete cross approximation in the mosaic-skeleton method. // *Computing.* 2000. V.4. P.367–380.

[13] R. R. Underwood. *An iterative block Lanczos method for the solution of large sparse symmetric eigenproblems.* Ph.D.thesis. Stanford University. 1975.