Тензорные методы и их приложения

Е.Е.Тыртышников

Институт вычислительной математики им. Г.И.Марчука
Российской академии наук

Московский университет им. М.В.Ломоносова

Московский физико-технический институт

tee@inm.ras.ru

Иннополис, 6 июня 2024

Проклятие размерности

Массив $A = [a(i_1, \ldots, i_d)]$ размера $n \times \ldots \times n$ невозможно задавать списком всех элементов даже при небольших d.

Если $d = 83$ и $n = 10$, то число элементов 10^{83} равно числу атомов во Вселенной!

Нужны *модели данных* с приемлемо малым числом параметров и *алгоритмы, работающие только с параметрами*.

Представление матриц малого ранга

Если r - ранг матрицы A порядка n, то

$$A = \sum_{\alpha=1}^{r} u_\alpha v_\alpha^\top = UV^\top$$

$$U = [u_1, \ldots, u_r], \quad V = [v_1, \ldots, v_r]$$

(матрицы размера $n \times r$)

$$2rn \ll n^2$$

Малый ранг повсюду

- Многие матрицы специального вида получают малый ранг после линейного преобразования. Для теплицевой матрицы $A = [a_{i-j}]$ после сдвигов $\operatorname{rank}(PA - AP) = 2$.

- В приложениях, например при решении интегральных уравнений математической физики, многие невырожденные матрицы состоят из блоков малого ранга (мозаично-скелетонный метод, мультипольный метод).

- Невырожденные матрицы могут быть близки к матрицам малого ранга поэлементно (в чебышевской норме):

$$||I_n - R_n||_C \leqslant \varepsilon, \quad \operatorname{rank}(R_n) \leqslant c\frac{\log n}{\varepsilon^2}.$$

Польза приближений малого ранга

- Сжатие данных
  - в случае матриц размера $n \times n$:
    $$r \ll n \quad \Rightarrow \quad 2nr \ll n^2$$
  - в случае d-тензоров размера $n \times \ldots \times n$:
    $$r \ll n^{d-1} \quad \Rightarrow \quad dnr \ll n^d$$
- Выделение наиболее значимой информации и устранение шума
- Быстрые вычисления в малоранговых форматах

Эффективная парадигма вычислений

- Для “больших” объектов выбирается модель их представления на основе “малых” векторов $A = A(p), \quad B = B(q), \quad C = C(s)$.

- Чтобы найти “большой” объект $C = A * B$, мы строим *быстрый алгоритм* приближенного вычисления s по p и q. При этом “большие” объекты A и B не должны возникать явным образом.

- Для широкого класса приложений для параметризации и аппроксимации “больших” объектов можно использовать разложения малого ранга.

Сингулярное разложение матрицы

Можно построить разложение, в котором векторы $u_1, \ldots, u_r$ и $v_1, \ldots, v_r$ ортогональны.

После нормировки

$$u_\alpha v_\alpha^\top = \sigma_\alpha u_\alpha v_\alpha^\top, \quad u_\alpha := u_\alpha / ||u_\alpha||, \quad v_\alpha := v_\alpha / ||v_\alpha||,$$

получаем *сингулярное разложение*:

$$A = \sum_{\alpha=1}^{r} \sigma_\alpha u_\alpha v_\alpha^\top = U \Sigma V^\top, \quad \Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix},$$

$$\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0.$$

Приближения малого ранга

Для любой матрицы B ранга $\leqslant k$ имеет место неравенство

$$||A - B|| \geqslant \sqrt{\sum_{\alpha=k+1}^{n} \sigma_\alpha^2}$$

$$\min_B ||A - B|| = ||A - A_k||$$

$$A_k = \sum_{\alpha=1}^{k} \sigma_\alpha u_\alpha v_\alpha^\top$$

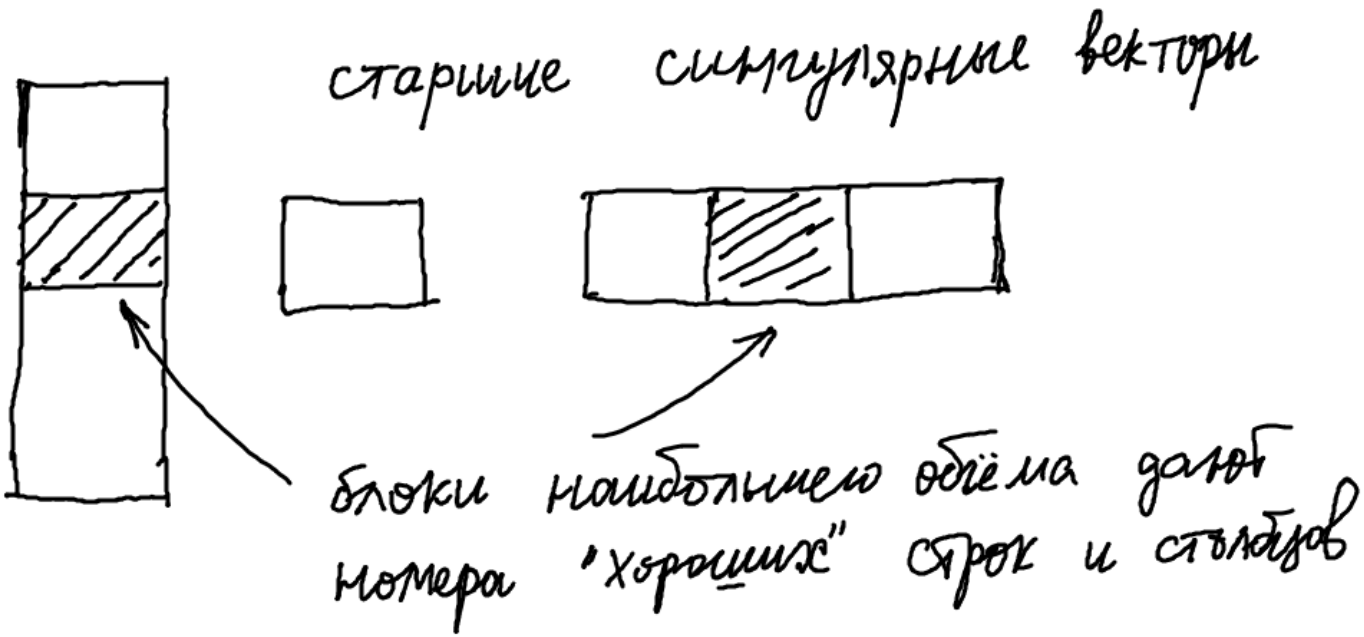
Здесь $||A|| = \sqrt{\sum_{i,j} |a_{ij}|^2}$ (норма Фробениуса).

Крестовая строчно-столбцая аппроксимация

$$A = \underbrace{\begin{bmatrix} & & \\ & \hat{A} & \\ & & \end{bmatrix}}_{C} \Big\} R \quad = C \hat{A}^{-1} R \approx C \hat{G} R$$

$\hat{A}, \hat{G}$ — $r \times r$

Как найти “хороший” крест



Принцип наибольшего объема

Теорема (Горейнов, Тыртышников).

Пусть блок $A_{11} \in \mathbb{C}^{r\times r}$ имеет наибольший объем среди всех $r \times r$-блоков матрицы

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad r = \operatorname{rank}(A + F).$$

Тогда

$$||A - CA_{11}^{-1}R||_C \leqslant (r+1)^2||F||_C,$$

$$C = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}, \quad R = \begin{bmatrix} A_{11} & A_{12} \end{bmatrix}.$$

Как искать “хороший” блок

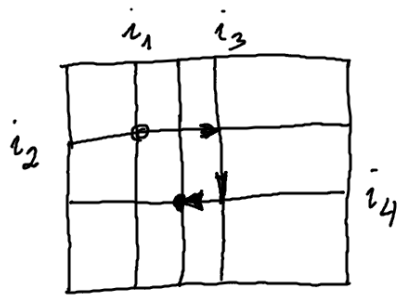
Возьмем обратимый блок $\hat{C} \in \mathbb{R}^{k \times k}$ в матрице $C \in \mathbb{R}^{n \times k}$ и перейдем к матрице

$$C\hat{C}^{-1} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ q_{r+1,1} & \cdots & q_{r+1,r} \\ \cdots & \cdots & \cdots \\ q_{n1} & \cdots & q_{nr} \end{bmatrix}$$

Необходимое условие для того, чтобы блок $\hat{C}$ имел наибольший объем в матрице C: $|q_{ij}| \leqslant 1, \quad r+1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant r$. Если не так, то делаем перестановку двух строк и увеличиваем объем!

D.Knuth, Semi-optimal bases for linear dependencies,
Linear and Mulilinear Algebra, 1985

Поиск больших элементов в матрицах малого ранга



В столбцовой подматрице $C \in \mathbb{R}^{n\times k}$ найдем k "хорощих" строк. Они выделяют строчную подматрицу $R \in \mathbb{R}^{k\times n}$, найдем в ней k "хороших" столбцов. Новые столбцы определяют новую столбцовую подматрицу $C \in \mathbb{R}^{k\times n}$, ищем в ней k "хороших" строк. И так далее.

Данный алгоритм с высокой вероятностью приводит к блоку почти максимального объема, в котором имеются почти максимальные элементы. На его основе получен новый эвристический алгоритм глобальной оптимизации. В ряде задач (например, докинга) он оказался на порядки эффективнее традиционных методов.

Представления многомерных массивов

- **Каноническое разложение:**

$$a_{i_1,\ldots,i_d} = \sum_{\alpha_1=1}^{r} u^{(1)}_{i_1,\alpha} \ldots u^{(d)}_{i_d,\alpha}$$

Много теоретических и вычислительных проблем. Массив может быть пределом массивов строго меньшего ранга.

- **Тензорный поезд:**

$$a_{i_1,\ldots,i_d} = \sum_{\alpha_1=1}^{r_1} \ldots \sum_{\alpha_{d-1}=1}^{r_{d-1}} g^{(1)}_{i_1\alpha_1} g^{(2)}_{\alpha_1 i_2 \alpha_2} \ldots g^{(d-1)}_{\alpha_{d-2} i_{d-1} \alpha_{d-1}} g^{(d)}_{\alpha_{d-1} i_d}$$

Теоретических проблем нет. Все вычисления сводятся к работе с матрицами, ассоциированными с массивом.

Модель тензорного поезда

$$a(i_1, i_2, i_3, i_4, i_5) =$$

$$\sum_{\alpha_1,\alpha_2,\alpha_3,\alpha_4} g_1(i_1, \alpha_1)\, g_2(\alpha_1, i_2, \alpha_2)\, g_3(\alpha_2, i_3, \alpha_3)\, g_4(\alpha_3, i_4, \alpha_4)\, g_5(\alpha_4, i_5)$$

$$= \underbrace{A_1^{(i_1)}}_{1\times r_1} \underbrace{A_2^{(i_2)}}_{r_1\times r_2} \underbrace{A_3^{(i_3)}}_{r_2\times r_3} \underbrace{A_4^{(i_4)}}_{r_3\times r_4} \underbrace{A_5^{(i_5)}}_{r_4\times 1}$$

Восстановление тензора и оптимизация в формате тензорного поезда

Тензорный поезд:

$$a_{i_1,\ldots,i_d} = \sum_{\alpha_1=1}^{r_1} \cdots \sum_{\alpha_{d-1}=1}^{r_{d-1}} g^{(1)}_{i_1\alpha_1} g^{(2)}_{\alpha_1 i_2 \alpha_2} \cdots g^{(d-1)}_{\alpha_{d-2} i_{d-1} \alpha_{d-1}} g^{(d)}_{\alpha_{d-1} i_d}$$

$$= G^{(1)}_{i_1} G^{(2)}_{i_2} \cdots G^{(d-1)}_{i_{d-1}} G^{(d)}_{i_d}$$

Алгоритмы строятся на основе структурированных малоранговых представлений для ассоциированных матриц развертки

$$A_k = [a^k_{i_1\ldots i_k; i_{k+1}\ldots i_d}]_{(n_1\ldots n_k)\times(n_{k+1}\ldots n_d)}$$

$$a^k_{i_1\ldots i_k; i_{k+1}\ldots i_d} = a_{i_1,\ldots,i_d}$$

Быстрое суммирование элементов астрономически большого вектора

$$i = \overline{i_1 i_2 \dots i_d} \qquad d = 83$$

$$a(i) = a(i_1, \dots, i_d) = \sum_{\alpha_1, \dots, \alpha_{d-1}} g_1(i_1, \alpha_1) g_2(\alpha_1, i_2, \alpha_2) \dots g_d(\alpha_{d-1}, i_d)$$

$$\sum_{i_1, \dots, i_d} a(i_1, \dots, i_d) = \sum_{\alpha_1, \dots, \alpha_{d-1}} \hat{g}_1(\alpha_1) \hat{g}_2(\alpha_1, \alpha_2) \dots \hat{g}_d(\alpha_{d-1})$$

$$\hat{g}_k = \sum_{i_k} g_k$$

Тензорный поезд и квадратуры

$$I(d) = \int \sin(x_1 + x_2 + \ldots + x_d)\, dx_1 dx_2 \ldots dx_d =$$

$$\mathrm{Im} \int_{[0,1]^d} e^{i(x_1+x_2+\ldots+x_d)}\, dx_1 dx_2 \ldots dx_d = \mathrm{Im}((\frac{e^i - 1}{i})^d).$$

$n = 11$ узлов по одной оси $\Rightarrow$ всего n^d значений! Но вычисляется лишь малая часть.

d	$I(d)$	Rel.error	Time
100	-3.926795e-03	2.915654e-13	0.77
1000	-2.637513e-19	3.482065e-11	11.60
2000	2.628834e-37	8.905594e-12	33.05
4000	9.400335e-74	2.284085e-10	105.49

Тензоризация векторов и матриц

Любой вектор размера $N = n_1 \dots n_d$ можно рассматривать как d-мерный массив, а любую $N \times N$-матрицу

$$a(i, j) = a(i_1 \dots i_d,\ j_1 \dots j_d)$$

как $2d$-мерный массив или же, склеив пары индексов, как d-мерный массив размера $n_1^2 \times \dots \times n_d^2$. вида

$$a(i_1 j_1, \dots, i_d j_d)$$

Тензоризация с последующим построением тензорного поезда может радикально сократить число параметров модели!

Польза даже для одномерных интегралов!

При вычислении интеграла

$$\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}$$

сначала переходим к ограниченной области, а затем применяем формулу прямоугольников.

Чтобы получить машинную точность, нужно взять порядка 2^{77} узлов. Вектор значений функции в этих узлах рассматривается как тензор размера $2 \times 2 \times \ldots \times 2$.

TT-ранги: $\leqslant 12$. Время: меньше 1 секунды на ноутбуке.

Каноническое тензорное разложение

$$a_{i_1 \ldots i_d} = \sum_{\alpha=1}^{r} u^{(1)}_{i_1,\alpha} \ldots u^{(d)}_{i_d,\alpha} \quad \Leftrightarrow \quad a = \sum_{\alpha=1}^{r} u^{(1)}_{\alpha} \otimes \ldots \otimes u^{(d)}_{\alpha}$$

Минимальное r называется *рангом* тензора $A = [a_{i_1 \ldots i_d}]$.

Проблема незамкнутости:

$$T = a \otimes a \otimes b + a \otimes b \otimes a + b \otimes a \otimes a$$

$$(a + \varepsilon b) \otimes (a + \varepsilon b) \otimes (a + \varepsilon b) = a \otimes a \otimes a + \varepsilon T + O(\varepsilon^2)$$

$$\Rightarrow \quad T = \varepsilon^{-1}(a + \varepsilon b)^{\otimes 3} - \varepsilon^{-1} a^{\otimes 3} + O(\varepsilon)$$

Если векторы a и b линейно независимы, то $\operatorname{rank} T = 3$.

Теорема о ранге обобщенного лапласиана

$$T_d = T_1^{(d)} + \ldots + T_d^{(d)}$$

$$T_i^{(d)} = a_1 \otimes \ldots \otimes a_{i-1} \otimes b_i \otimes a_{i+1} \otimes \ldots \otimes a_d$$

Если a_i, b_i линейно независимы, то $\operatorname{rank}(T_d) = d$.

- Tyrtyshnikov E., Tensor decompositions and rank increment conjecture, RJNAMM, 25 (4), 239–246 (2020).

По индукции доказывается более общее утверждение:

$$\operatorname{rank}(T_d + \gamma\, a_1 \otimes \ldots \otimes a_d) = d \quad \forall\, \gamma \in \mathbb{C}.$$

Вопросы о незамкнутости ранговых множеств

Замыкание $\bar{X}_k$ *рангового множества* $X_k := \{T : \operatorname{rank}(T) \leqslant k\}$ *является алгебраическим многообразием.*

- Множества X_1 и $X_{R_{\max}}$ замкнуты.
- Множества X_k при $R_{\text{gen}} \leqslant k < R_{\max}$ незамкнуты.
- Множество X_2 незамкнуто.
- Верно ли, что при $2 \leqslant k < R_{\max}$ множество X_k незамкнуто?
- Каков максимальный ранг R_k предела тензоров ранга k?

$R_{\max}$ и R_{gen} — максимальный и главный ранги в пространстве тензоров заданного размера. Любой тензор является пределом тензоров ранга R_{gen}.

Теорема об увеличении ранга

Для любого $1 < k < R_{\text{gen}}$ *существует алгебраическое многообразие* $W \subsetneq \bar{X}_k$ *такое, что для любого тензора из множества* $X_k \setminus W$ *найдется тензор ранга* 1, *прибавление которого увеличивает ранг данного тензора.*

- Tyrtyshnikov E., Tensor decompositions and rank increment conjecture, RJNAMM, 25 (4), 239–246 (2020).

Гипотеза об увеличении ранга: утверждение верно для *любого* тензора, ранг которого меньше максимального ранга.

Существует последовательнось $n \times n \times n$*-тензоров ранга* n *с предельным тензором ранга* $2n - 1$.

Спасибо за внимание!