CSE 392: Matrix and Tensor Algorithms for Data Spring 2024

Lecture 18 — 03-25

Instructor: Shashanka Ubaru *Scribe: Gabriel H. Brown*

1 Leverage scores and coherence

Recall the definition of row leverage scores of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times r}$. If $\boldsymbol{U}$ is an orthonormal basis for *span*$(\boldsymbol{A})$, then the ith leverage score is given by

$$\ell_i(\boldsymbol{A}) = \sup_{\boldsymbol{x}} \frac{(\boldsymbol{A}_{i:}\boldsymbol{x})^2}{\|\boldsymbol{A}\boldsymbol{x}\|^2} = \|\boldsymbol{U}_{i:}\|^2,\ i \in [n]. \tag{1}$$

One can then sample rows according to probabilities $p_i = \ell_i/r$. It is also possible to approximately compute leverage scores at a reduced complexity.

The coherence of $\boldsymbol{A}$, denoted by $\mu(\boldsymbol{A})$ is the *maximum leverage score*, that is,

$$\mu(\boldsymbol{A}) = \max_{i \in [n]} \ell_i(\boldsymbol{A}). \tag{2}$$

Coherence obeys the following inequalities: $\frac{r}{n} \leq \mu(\boldsymbol{A}) \leq 1$. The first follows from $1 = \sum_{i=1}^n p_i = \frac{1}{r}\sum_{i=1}^n \ell_i \Rightarrow r = \sum_{i=1}^n \ell_i \Rightarrow r \leq n\mu(\boldsymbol{A})$; the second follows from $\ell_i = \|\boldsymbol{U}_{i:}\|^2 = \|e_i \boldsymbol{U}\|^2 \leq 1^2 \cdot 1^2$ by submultiplicativity. We say $\boldsymbol{A}$ is **incoherent** if $\mu(\boldsymbol{A}) \approx \frac{r}{N}$.

One can use leverage scores to sample linear least squares problems, getting approximate solutions at a reduced cost.

Proposition 1. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times r}$ and a fixed vector $\boldsymbol{b} \in \mathbb{R}^n$, let $\boldsymbol{x}^* = \min_{\boldsymbol{x} \in \mathbb{R}^d} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$. Let $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ be a sampling matrix with probabilities $p_i = \ell_i/r$, and $\boldsymbol{S}_{i*} = \boldsymbol{e}_j/\sqrt{mp_j}$ with $\mathbb{P}(j = i) = p_i$. If $m = O(r\log(r/\delta)/\varepsilon)$ and $\tilde{\boldsymbol{x}} = \min_{\boldsymbol{x} \in \mathbb{R}^d} \|\boldsymbol{S}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\|_2$, then, with high probability,

$$\|\boldsymbol{A}\tilde{\boldsymbol{x}} - \boldsymbol{b}\|_2 \leq (1+\varepsilon)\|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\|_2.$$

2 Leverage score sampling for CP-ALS

Our goal is to accelerate CP-ALS by using leverage score sampling on the least squares subproblems that arise for the approximate CP factor matrices of tensor $\mathcal{X}$ (here order 3)

$$\min_{\boldsymbol{A}_1} \|(\boldsymbol{A}_3 \odot \boldsymbol{A}_2)\boldsymbol{A}_1^T - \boldsymbol{X}_{(1)}^T\|_F^2. \tag{3}$$

However, even approximately computing the leverage scores for $(\boldsymbol{A}_3 \odot \boldsymbol{A}_2)$ can be prohibitively expensive. But we can estimate/bound the leverage scores for this Khatri-Rao structured matrix in terms of the leverage scores of the matrices $\boldsymbol{A}_2$ and $\boldsymbol{A}_3$.

Lemma 1 (Cheng, et al.: Theorem 3.2)**.** $\mu(\boldsymbol{A} \odot \boldsymbol{B}) \leq \mu(\boldsymbol{A})\mu(\boldsymbol{B})$

This implies that if two matrices $\boldsymbol{A}, \boldsymbol{B}$ are incoherent, then their Khatri-Rao product $\boldsymbol{A} \odot \boldsymbol{B}$ is also incoherent.

Motivated by Lemma **??**, instead of sampling according to $p_k = \ell_k(\boldsymbol{A}_3 \odot \boldsymbol{A}_2)/r$ we will instead use $p_k = \ell_i(\boldsymbol{A}_3)\ell_j(\boldsymbol{A}_2)/r^2$, which requires only $\mathcal{O}((n_2 + n_3)r)$ work. Specifically, we'll use the following procedure:

- choose $i \sim p_i = \ell_i(\boldsymbol{A}_3)/r$
- choose $j \sim p_j = \ell_j(\boldsymbol{A}_2)/r$
- select row $k = i + (j-1)n_3$.

The guarantee for this procedure is as follows.

Theorem 1 (Larsen and Kolda: Theorem 6)**.** *Let* $\boldsymbol{A}_i \in \mathbb{R}^{n_i \times r}$*,* $\boldsymbol{X}_{(1)} \in R^{n_2 n_3 \times n_1}$ *and consider the linear least squares problem*

$$\arg\min_{\boldsymbol{A}_1} \|(\boldsymbol{A}_3 \odot \boldsymbol{A}_2)\boldsymbol{A}_1^T - \boldsymbol{X}_{(1)}^T\|_F^2.$$

with optimal solution $\boldsymbol{A}_1^*$*. Now let* $\tilde{\boldsymbol{A}}_1$ *be the optimal solution to the problem*

$$\arg\min_{\boldsymbol{A}_1} \|(\boldsymbol{S}(\boldsymbol{A}_3 \odot \boldsymbol{A}_2)\boldsymbol{A}_1^T - \boldsymbol{S}\boldsymbol{X}_{(1)}^T\|_F^2.$$

where $\boldsymbol{S} \in \mathbb{R}^{s \times n_2 n_3}$ *is the leverage score sampling matrix which samples according to the procedure described above.*
If $s = r^4 \max\{1700 \log(r/\delta), 1/(\delta\epsilon)\}$*, then*

$$Pr\left[\|(\boldsymbol{A}_3 \odot \boldsymbol{A}_2)\tilde{\boldsymbol{A}}_1^T - \boldsymbol{X}_{(1)}^T\|_F^2 \leq (1+\epsilon)\|(\boldsymbol{A}_3 \odot \boldsymbol{A}_2)(\boldsymbol{A}_1^*)^T - \boldsymbol{X}_{(1)}^T\|_F^2\right] \geq 1 - \delta.$$

Larsen and Kolda also suggest additional practical tips for efficient implementation:

- **hybrid approach**: deterministically include all rows whose leverage scores/probabilities are above some threshold and randomly sample from the remaining rows; using the hybrid strategy, they demonstrate equally good or better decompositions with the same number of total samples
- **unfoldings**: never form $\boldsymbol{X}_{(i)}^T$ explicitly if $\mathcal{X}$ is sparse, instead precompute linear indices for every nonzero for each mode to directly form a sparse unfolding/right hand side; this requires $3\text{nnz}(\mathcal{X})$ extra memory
- **estimate residual**: calculating the residual is necessary to determine when the approximation is sufficiently converged, but computing the residual can take many times longer then updating all three factor matrices; therefore, as a *practical hack* with no theoretical guarantees the authors suggest estimating the residual based on a random sample of tensor elements (using a stratified sampling to correct for problems introduced by sparsity)

We conclude this discussion by comparing the complexities of the two main kernels in CP-ALS: computation of the residual and the formation and solution of (one) linear least squares problem. Here s_{fit} is the user specified number of elements used to sample and estimate the residual and j corresponds the the mode being updates by the least squares subproblem.

operation	sparse/dense	complexity (big $\mathcal{O}$)	sampled complexity
residual	dense	$rn_1n_2n_3$	rs_{fit}
	sparse	$r\text{nnz}(\mathcal{X})$	rs_{fit}
least squares (mode j)	dense	$rn_1n_2n_3$	$3sr + sr^2 + srn_j$
	sparse	$rn_1n_2n_3$	$3sr + sr^2 + r\text{nnz}(\boldsymbol{X}_{(j)})$

3 Troubles with CP decomposition

Finally, we summarize some mathematical troubles that plague the CP decomposition, stressing that these are independent of algorithm.

- **ill-posedness**: from de Silva and Lim there is no guarantee that the best rank k approximation exists; for example no rank 3 $2 \times 2 \times 2$ tensor has a best rank 2 approximation, and a random $m \times n \times p$ tensor has no best rank 2 approximation with probability; in general this rules out the possibility of a theorem like Echart-Young-Mirsky for tensors
- **complexity of determining rank**: given a tensor $\mathcal{T}$, determining its real rank is NP hard (and so are many other tensor problems, see Lim)
- **many local minima**: the best rank k approximation problem is non-convex and non-linear, and so may have many local minima; no known results suggest that the standard algorithms frequently find "good" local minima