CSE 392: Matrix and Tensor Algorithms for Data Spring 2024

Lecture 6 — Feb. 5th

Instructor: Shashanka Ubaru *Scribe: Yifan Zhang*

1 Randomized Methods via Sampling

Consider computing the inner product between two large vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. One simple way of accelerating the computation is to uniformly sample $m \ll n$ entries $\hat{\boldsymbol{x}}$ of $\boldsymbol{x}$ and $\hat{\boldsymbol{y}}$ of $\boldsymbol{y}$, and deliver $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \approx \frac{n}{m} \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \rangle$. This naive approach can have bad performance if the vectors are sparse, so the sampled entries have a high probability not capturing significant information. There are two approaches to improve this. First, one can use a randomized isometry $\boldsymbol{U}$ to make the entries in $\boldsymbol{Ux}$ and $\boldsymbol{Uy}$ have similar magnitudes, and then do the uniform sampling. Second, one can use importance sampling to capture larger entries in the vectors, say using a sampling probability proportional to the magnitude of each entry. One example of the first approach is the subsampled randomized Hadamard transform (SRHT), which will be discussed in later lectures. In this lecture, the focus is on the second approach. In particular, we will introduce sampling based approximate matrix multiplication and low rank approximation.

2 Sampling Based Approximate Matrix Multiplication

Matrix multiplications are nothing but a bunch of vector inner products. Consider two large matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times p}$. To compute $\boldsymbol{M} = \boldsymbol{AB}$, we subsample c columns of $\boldsymbol{A}$ and get the corresponding rows of $\boldsymbol{B}$. Denote $\boldsymbol{S} \in \mathbb{R}^{n \times c}$ the column sampling operator, $c \ll n$, and $\widehat{\boldsymbol{A}} = \boldsymbol{AS} \in \mathbb{R}^{m \times c}$, $\widehat{\boldsymbol{B}} = \boldsymbol{S}^\top \boldsymbol{B} \in \mathbb{R}^{c \times p}$, then we deliver

$$\boldsymbol{AB} \approx \widehat{\boldsymbol{A}}\widehat{\boldsymbol{B}} = \boldsymbol{ASS}^\top \boldsymbol{B},$$

with some appropriate scaling included in the matrix $\boldsymbol{S}$. We assume the choice of each column sample is i.i.d. Denote p_i the probability of sampling column i in sample. Using the idea of importance sampling, one choice of p_i is the **length-squared sampling** scheme where

$$p_i = \frac{\big\|\boldsymbol{A}_{(:,i)}\big\|_2 \big\|\boldsymbol{B}_{(:,i)}\big\|_2}{\sum_{j=1}^n \big\|\boldsymbol{A}_{(:,j)}\big\|_2 \big\|\boldsymbol{B}_{(:,j)}\big\|_2}.$$

Denote z_i the (random) index of sampled column in sample i, $i = 1, \ldots, c$. Denote $\boldsymbol{a}_{z_i}$ and $\boldsymbol{b}_{z_i}$ the sampled column and row in sample i (viewed as column vectors). The approximated matrix product from the sample is then

$$\boldsymbol{M} = \boldsymbol{AB} \approx \frac{1}{c} \cdot \sum_{i=1}^{c} \frac{1}{p_{z_i}} \boldsymbol{a}_{z_i} \boldsymbol{b}_{z_i}^\top := \widehat{\boldsymbol{M}} \tag{1}$$

The scaling constants are chosen so that the constructed approximation is unbiased. We can prove this and bound the variance of $\widehat{M}$.

Theorem 2.1. *Using the notation defined above, we have*

$$\mathbb{E}\,\widehat{\boldsymbol{M}} = \boldsymbol{M};$$
$$\mathbb{E}\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\right\|_F^2 \leqslant \frac{1}{c}\|\boldsymbol{A}\|_F^2\|\boldsymbol{B}\|_F^2.$$

Proof. To show the estimator is unbiased, we note that each rank one term in (1) has expectation $\boldsymbol{M}$. We are left to bound the variance. To this end, we have

$$\mathbb{E}\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\right\|_F^2 = \mathbb{E}\left\|\frac{1}{c}\sum_{i=1}^{c}\left(p_{z_i}^{-1}\boldsymbol{a}_{z_i}\boldsymbol{b}_{z_i}^\top - \boldsymbol{M}\right)\right\|_F^2 = \frac{1}{c}\,\mathbb{E}\left\|p_{z_1}^{-1}\boldsymbol{a}_{z_1}\boldsymbol{b}_{z_1}^\top - \boldsymbol{M}\right\|_F^2 \leqslant \frac{1}{c}\,\mathbb{E}\left\|p_{z_1}^{-1}\boldsymbol{a}_{z_1}\boldsymbol{b}_{z_1}^\top\right\|_F^2,$$

where we used the independence between the samples. It is then easy to check that

$$\mathbb{E}\left\|p_{z_1}^{-1}\boldsymbol{a}_{z_1}\boldsymbol{b}_{z_1}^\top\right\|_F^2 = \|\boldsymbol{A}\|_F^2\|\boldsymbol{B}\|_F^2.$$

The proof is then complete. ■

3 Sampling Based Low Rank Approximation

In matrix decomposition, CUR decomposition is faster than SVD or even QR. Given a matrix $\boldsymbol{A}$, it chooses c columns of the matrix, $\boldsymbol{C}$, and c rows of the matrix, $\boldsymbol{R}$. Denote $\boldsymbol{T}$ the intersection of the c columns and c rows and denote $\boldsymbol{U} = \boldsymbol{T}^{-1}$. The CUR decomposition is then

$$\boldsymbol{A} = \boldsymbol{CUR}.$$

If $c = \operatorname{rank}(\boldsymbol{A})$, then this decomposition is exact for any choice of columns and rows. Besides it is faster to compute, one other important advantage of CUR is the factor matrices $\boldsymbol{C}$ and $\boldsymbol{R}$ are submatrices of $\boldsymbol{A}$, making the decomposition more interpretable and also sparse in the case $\boldsymbol{A}$ is sparse. If $c < \operatorname{rank}(\boldsymbol{A})$, however, as opposed to Eckart-Young for SVD, there is no guarantee of goodness of fit. It is then an open question on how to choose the subset of columns and rows, so that we have a good chance of getting a good low rank approximation.

In order to find the correct rows and columns, we introduce the notion of leverage scores.

Definition 3.1 (leverage score)**.** *Given a linear subspace $L \subseteq \mathbb{R}^n$. The ith leverage score is $\ell_i(L) = \sup_{\boldsymbol{x}\in L} x_i^2/\|\boldsymbol{x}\|^2$, $i = 1, \ldots, m$.*

Clearly, $\ell_i \in [0,1]$ and $\sum_i \ell_i(L) = \dim(L)$. The leverage scores measures how nicely behaved a sampling method is on L. If the leverage scores are close to $\dim(L)/m$, then for any vector in L, its entries have similar magnitudes. On the other hand, if $\ell_1(L) = 1$, then the first entry can dominate over all other entries, and sampling method can fail if it does not capture this dominating entry.

When the subspace is given by the range of a matrix $\boldsymbol{A}$, the leverage scores can be obtained by

$$\ell_i(\boldsymbol{A}) := \left\|\boldsymbol{U}_{(i,:)}\right\|_2^2,$$

where $\boldsymbol{U}$ is any orthonormal basis of the column space of $\boldsymbol{A}$. Following the motivation of importance sampling, we put more importance on entries that have a large leverage score, and put

$$p_i = \frac{\ell_i(\boldsymbol{A})}{\operatorname{rank}(\boldsymbol{A})}.$$

Similarly, one can also define the leverage scores for the row space of $\boldsymbol{A}$ and define a sampler over the rows.

When we choose the $\boldsymbol{C}$ factor using a leverage score sample, we have the following guarantee on the goodness of fit.

Theorem 3.2. *Let $\boldsymbol{A}_k$ be the best rank k approximation to $\boldsymbol{A}$. Let $\boldsymbol{C}$ be a leverage score sample of c columns of $\boldsymbol{A}$. Denote $\boldsymbol{P}_C$ the orthogonal projection to the column space of $\boldsymbol{C}$. Then with probability at least 0.9,*

$$\|\boldsymbol{A} - \boldsymbol{P}_C \boldsymbol{A}\|_F \leqslant (1+\varepsilon)\|\boldsymbol{A} - \boldsymbol{A}_k\|_F,$$

provided that $c = \Omega\left(\frac{k}{\varepsilon^2}\log\left(\frac{k}{\varepsilon}\right)\right)$.