Matrix Approximation via Sampling, Subspace Embedding

Lecturer: Anup Rao Scribe: Rakshith Sharma, Peng Zhang

02/01/2016

1 Solving Linear Systems Using SVD

Two applications of SVD have been covered so far. Today we look at a third application. The problem is that, given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$, find an $\mathbf{x} \in \mathbb{R}^n$ such that $\|\mathbf{Ax} - \mathbf{b}\|$ is minimized. SVD will form the core of the solution and there are no assumptions made on the matrix $\mathbf{A}$. We denote the singular value decomposition of $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{UDV}^T$$

Setting the gradient of $f(x) = \|\mathbf{Ax} - \mathbf{b}\|$ to zero, we get

$$\nabla_j f(\mathbf{x}) = 2 \sum_i (\langle \mathbf{A}^{(i)}, \mathbf{x} \rangle - \mathbf{b}_i) \mathbf{A}_j^{(i)} = 0$$

where $\langle \cdot\ ,\ \cdot \rangle$ is inner product and $\mathbf{A}^{(i)}$ is the i^{th} row of $\mathbf{A}$. Hence,

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b} \tag{1}$$

Observe that this system always has a solution (unique or otherwise).

Solving for x: If $\mathbf{A}^T\mathbf{A}$ is full rank, then the solution is

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \tag{2}$$

If $\mathbf{A}^T\mathbf{A}$ is not full rank, we claim the following is the best solution in the least square sense:

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{\dagger}\mathbf{A}^T\mathbf{b} \tag{3}$$

where $(\mathbf{A}^T\mathbf{A})^{\dagger} = \mathbf{VD}^{-2}\mathbf{V}^T$ is known as the pseudoinverse.

To support this claim, we need to show that this is a solution to (1). Using the SVD of $\mathbf{A}$ and the proposed $\mathbf{x}$,

$$\begin{aligned}(\mathbf{A}^T\mathbf{A})\mathbf{x} &= (\mathbf{VDU}^T\mathbf{UDV}^T)\mathbf{VD}^{-2}\mathbf{V}^T\mathbf{A}^T\mathbf{b} \\ &= \mathbf{VV}^T\mathbf{A}^T\mathbf{b} \\ &= \mathbf{VV}^T\mathbf{VDU}^T\mathbf{b} \\ &= \mathbf{VDU}^T\mathbf{b} = \mathbf{A}^T\mathbf{b}\end{aligned}$$

Hence, the least square solution to $\mathbf{Ax} = \mathbf{b}$ is given by

$$\mathbf{x}_{LS} = (\mathbf{A}^T\mathbf{A})^{\dagger}\mathbf{A}^T\mathbf{b}$$

2 Matrix Approximation via Sampling

In this section, we present a method to estimate the best rank-k approximation to matrices using column selection method.

2.1 Approximating matrix product by sampling

This subsection establishes fast methods to approximate matrix-vector and matrix-matrix product which will later be used in the algorithm to obtain the best rank-k approximation of a given matrix. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{v} \in \mathbb{R}^n$, where $\mathbf{A}^{(i)}$ denotes the i^{th} column of $\mathbf{A}$.

$$\mathbf{Av} = \sum_i^n \mathbf{A}^{(i)} \mathbf{v}_i$$

We estimate this product by sampling just one column from $\mathbf{A}$. Define the random variable X to be

$$X = \frac{\mathbf{A}^{(i)} \mathbf{v}_i}{p_i} \text{ with probability } p_i$$

Then the expected value of X is

$$\mathsf{E}(X) = \sum_i \mathbf{A}^{(i)} \mathbf{v}_i = \mathbf{Av}$$

The Variance of X is

$$\mathsf{Var}(X) = \mathsf{E}(\|X\|^2) - \|\mathsf{E}(X)\|^2 = \sum_i \frac{\|\mathbf{A}^{(i)} \mathbf{v}_i\|^2}{p_i} - \|\mathbf{Av}\|^2$$

We choose p_i's to minimize $\mathsf{Var}(X)$. This can be achieved by setting $p_i \propto \|\mathbf{A}^{(i)}\|_2^2$. After normalizing, we get

$$p_i = \frac{\|\mathbf{A}^{(i)}\|_2^2}{\|\mathbf{A}\|_F^2}$$

and

$$\mathsf{Var}(X) = \sum_i \|\mathbf{A}\|_F^2 \mathbf{v}_i^2 - \|\mathbf{Av}\|_2^2 \leq \|\mathbf{A}\|_F^2 \|\mathbf{v}\|_2^2$$

We proceed to reduce the variance by repeating the above process s times, and get s random variables $X_1, \ldots, X_s$. Let $Y = \frac{1}{s} \sum_i X_i$. Then we have

$$\mathsf{E}(Y) = \mathbf{Av} \text{ and } \mathsf{Var}(Y) \leq \frac{1}{s} \|\mathbf{A}\|_F^2 \|\mathbf{v}\|^2$$

We could also have an estimated probability distribution of the form

$$p_i \geqslant c \frac{\|\mathbf{A}^{(i)}\|_2^2}{\|\mathbf{A}^{(i)}\|_F^2} = LS_{col}(\mathbf{A}, c) \tag{4}$$

where $c \leq 1$. (If $\sum p_i < 1$, we could pick a zero vector with the remaining probability). With this choice for p_i's, we get

$$\mathsf{Var}(Y) \leq \frac{1}{cs}\|\mathbf{A}\|_F^2\|\mathbf{v}\|^2$$

We now use this to compute product of two matrices. Given $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, then

$$\mathbf{AB} = [\mathbf{AB}^{(1)}, \mathbf{AB}^{(2)}, \ldots, \mathbf{AB}^{(p)}]$$

Let $p_i = \frac{\|\mathbf{A}^{(i)}\|^2}{\|\mathbf{A}\|_F^2}$ be the probability of picking column i. Denote this distribution as $LS_{col}(\mathbf{A})$. We pick s columns of $\mathbf{A}$, say $j_1, j_2, \ldots, j_s$, according to this distribution.

We approximate $\mathbf{AB}$ by the following random matrix Y

$$Y = \frac{1}{s}\sum_{t=1}^{s}\frac{\mathbf{A}^{(j_t)}\mathbf{B}_{(j_t)}}{p_{j_t}}$$

where the subscripts denote row indices and superscripts denote column indices. We have $\mathsf{E}(Y) = \mathbf{AB}$ and

$$\mathsf{Var}(Y) \leq \frac{1}{cs}\sum_i \|\mathbf{A}\|_F^2\|\mathbf{B}^{(i)}\|_2^2 = \frac{1}{cs}\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2 \tag{5}$$

We now have the framework required to study the algorithm to compute the rank-k approximation of a matrix "quickly".

2.2 Fast algorithm for rank-k approximation

Algorithm:

- Sample s columns of $\mathbf{A}$ according to $LS_{col}(\mathbf{A}, c)$. Let $\mathbf{C}$ be the matrix containing these columns, that is,

$$\mathbf{C} = \frac{1}{\sqrt{cs}}\left(\frac{\mathbf{A}^{(i_1)}}{\sqrt{p_{i_1}}}, \ldots, \frac{\mathbf{A}^{(i_s)}}{\sqrt{p_{i_s}}}\right)$$

- Find the k left singular vectors of $\mathbf{C}$, say $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(k)}$.
- Output $\tilde{\mathbf{A}} := \sum_i \mathbf{u}^{(i)}\mathbf{u}^{(i)T}\mathbf{A}$ as the rank-k approximation of $\mathbf{A}$.

Theorem 2.1 ([KV08]).

$$\mathsf{E}\|\mathbf{A} - \tilde{\mathbf{A}}\|^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + 2\sqrt{\frac{k}{s}}\|\mathbf{A}\|_F^2$$

where $\tilde{\mathbf{A}}$ is a rank-k approximation and in our case equals $\mathbf{UU}^T\mathbf{A}$ (the last term would be $2\sqrt{\frac{k}{cs}}\|\mathbf{A}\|_F^2$ if (4) is used).

We proof Theorem 2.1 through the following lemma.

Lemma 2.2. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{C} \in \mathbb{R}^{m\times s}$, *and* $\mathbf{U} \in \mathbb{R}^{m\times k}$ *consisting of the top* k *singular vectors of* $\mathbf{C}$. *Then,*

$$\|\mathbf{A} - \mathbf{U}\mathbf{U}^T\mathbf{A}\|_F^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + 2\sqrt{k}\|\mathbf{A}\mathbf{A}^T - \mathbf{C}\mathbf{C}^T\|_F$$

Proof.

$$\begin{aligned}\|\mathbf{A} - \mathbf{U}\mathbf{U}^T\mathbf{A}\|_F^2 &= tr((\mathbf{A} - \mathbf{U}\mathbf{U}^T\mathbf{A})^T(\mathbf{A} - \mathbf{U}\mathbf{U}^T\mathbf{A}))\\ &= tr(\mathbf{A}^T\mathbf{A} - \mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{A} - \mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{A} + \mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{U}\mathbf{U}^T\mathbf{A})\\ &= tr(\mathbf{A}^T\mathbf{A} - \mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{A})\\ &= \|\mathbf{A}\|_F^2 - \|\mathbf{U}^T\mathbf{A}\|_F^2\end{aligned}$$

The third equality is due to $\mathbf{U}^T\mathbf{U} = I_k$. Then,

$$\begin{aligned}&\|\mathbf{A} - \mathbf{U}\mathbf{U}^T\mathbf{A}\|_F^2 - \|\mathbf{A} - \mathbf{A}_k\|_F^2\\ =&\|\mathbf{A}\|_F^2 - \|\mathbf{U}^T\mathbf{A}\|_F^2 - (\|\mathbf{A}\|_F^2 - \|\mathbf{A}_k\|_F^2)\\ =&\|\mathbf{A}_k\|_F^2 - \|\mathbf{U}^T\mathbf{A}\|_F^2\\ =&\|\mathbf{A}_k\|_F^2 - \|\mathbf{C}_k\|_F^2 + \|\mathbf{C}_k\|_F^2 - \|\mathbf{U}^T\mathbf{A}\|_F^2\\ =&\sum_{i=1}^{k}(\sigma_i^2(\mathbf{A}) - \sigma_i^2(\mathbf{C})) + \sum_{i=1}^{k}(\sigma_i^2(\mathbf{C}) - \|(\mathbf{U}^{(i)})^T\mathbf{A}\|_F^2)\\ \leq&\sqrt{k\sum_i(\sigma_i^2(\mathbf{C}) - \sigma_i^2(\mathbf{A}))^2} + \sqrt{k\sum_i(\sigma_i^2(\mathbf{C}) - \|(\mathbf{U}^{(i)})^T\mathbf{A}\|_F^2)^2}\\ \leq&\sqrt{k\sum_i(\sigma_i^2(\mathbf{C}) - \sigma_i^2(\mathbf{A}))^2} + \sqrt{k\sum_i(u_i^T(CC^T - AA^T)u_i)^2}\\ \leq&\sqrt{k}\|\mathbf{C}\mathbf{C}^T - \mathbf{A}\mathbf{A}^T\|_F + \sqrt{k}\|\mathbf{C}\mathbf{C}^T - \mathbf{A}\mathbf{A}^T\|_F\\ =&2\sqrt{k}\|\mathbf{C}\mathbf{C}^T - \mathbf{A}\mathbf{A}^T\|_F\end{aligned}$$

Here we first used the Cauchy-Schwarz inequality on both summations and then the Hoffman-Wielandt inequality on the first summation.

□

Proof of Theorem 2.1. By Lemma 2.2, we have

$$\mathsf{E}\|\mathbf{A} - \tilde{\mathbf{A}}\|^2 = \mathsf{E}\|\mathbf{A} - \mathbf{U}\mathbf{U}^T\mathbf{A}\|^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + 2\sqrt{k}\mathsf{E}\|\mathbf{A}\mathbf{A}^T - \mathbf{C}\mathbf{C}^T\|_F$$

Since $\mathsf{E}\mathbf{C}\mathbf{C}^T = \mathbf{A}\mathbf{A}^T$, by (5),

$$\mathsf{E}\|\mathbf{A}\mathbf{A}^T - \mathbf{C}\mathbf{C}^T\|_F^2 = \mathsf{Var}(\mathbf{C}\mathbf{C}^T) \leq \frac{1}{s}\|\mathbf{A}\|_F^4.$$

Thus,

$$\mathsf{E}\|\mathbf{A} - \tilde{\mathbf{A}}\|^2 \leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + 2\sqrt{\frac{k}{s}}\|\mathbf{A}\|_F^2.$$

□

3 Fast Approximation Using CUR Decomposition

[KV08] Another way to approximate a matrix is to sample s columns of $\mathbf{A}$ and also sample s rows of $\mathbf{A}$ where $\mathbf{A} \in \mathbb{R}^{m\times n}$. Let $\mathbf{C} \in \mathbb{R}^{m\times s}$ and $\mathbf{R} \in \mathbb{R}^{s\times n}$ be respectively made of columns and rows of $\mathbf{A}$, sampled according to $LS_{col}(\mathbf{A})$ and $LS_{row}(\mathbf{A})$. Then we can compute an $s \times s$ matrix $\mathbf{U}$ such that $\mathbf{CUR} \approx \mathbf{A}$ and

$$\mathsf{E}(\|\mathbf{A} - \mathbf{CUR}\|_F) \leq \|\mathbf{A} - \mathbf{A}_k\|_F + \sqrt{\frac{k}{s}}\|\mathbf{A}\|_F + \left(\frac{4k}{s}\right)^{\frac{1}{4}}\|\mathbf{A}\|_F$$

4 Subspace Embedding

Definition 4.1. *Let $\mathbf{S} \in \mathbb{R}^{r\times n}$ and $V \subset \mathbb{R}^n$ with $\dim(V) = d$. We say $\mathbf{S}$ is a* subspace embedding *for V if $\|\mathbf{Sx}\|_2 = (1 \pm \epsilon)\|\mathbf{x}\|_2$, $\forall\ \mathbf{x} \in V$. We say $\mathbf{S}$ is a* subspace embedding *for matrix $\mathbf{A}$ if it is a subspace embedding for the column space of $\mathbf{A}$.*

Oblivious subspace embedding is defined as follows. $\mathbf{S}$ is a random matrix with distribution $\mathcal{D}$ over $\mathbb{R}^{r\times n}$ such that, for any fixed matrix $\mathbf{A}$, we have $\|\mathbf{SAx}\|_2 = (1 \pm \epsilon)\|\mathbf{Ax}\|_2$ for any $\mathbf{x} \in \mathbb{R}^d$. One idea is that, each entry of $\mathbf{S}$ is an i.i.d. Gaussian random variable $N(0, 1)$. Johnson-Lindenstrauss Lemma shows that, to preserve the pairwise distance, we need $r = O(\log m/\epsilon^2)$ where m is the number of vectors (i.e., $\mathbf{Ax}$). For oblivious subspace embedding, $m = 2^{\Omega(d)}$. Furthermore, $\mathbf{S}$ sampled in this way is a dense matrix, thus computing $\mathbf{SA}$ is expensive. In this lecture, we will show a distribution of $\mathbf{S}$, which has $r = O(d/\epsilon^2)$ and $\mathbf{S}$ is sparse.

An application of subspace embedding is *Linear regression*. The goal of linear regression is $\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2$, where $\mathbf{A} \in \mathbb{R}^{n\times d}$ with $n >> d$. It is equal to $\min_{\mathbf{y}} \|\mathbf{A}_1\mathbf{y}\|_2$, where $\mathbf{A}_1 = (\mathbf{A}\ \mathbf{b})$ and $\mathbf{y} = \begin{pmatrix} \mathbf{x} \\ -1 \end{pmatrix}$. Now, suppose $\mathbf{S}$ is a subspace embedding for $\mathbf{A}_1$, then we have $\|\mathbf{SA}_1\mathbf{y}\|_2 = (1 \pm \epsilon)\|\mathbf{A}_1\mathbf{y}\|_2$ for any $\mathbf{y}$. The goal becomes the following,

$$\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2 = \min_{\mathbf{x}} \|\mathbf{SAx} - \mathbf{Sb}\|_2$$

Since $\mathbf{SA} \in \mathbb{R}^{r\times d}$, it reduces the dimension significantly if r only depends on d.

Now, we give the following theorem.

Theorem 4.2 ([Woo14]). *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, there exists an algorithm which generating an $\mathbf{S} \in \mathbb{R}^{r\times n}$ with $r = O\left(d^2/(\epsilon^2 poly(\log d/\epsilon))\right)$. With probability 0.99, $\mathbf{S}$ is $(1 \pm \epsilon)$ subspace embedding for $\mathbf{A}$. Furthermore, $\mathbf{SA}$ can be computed in $O(nnz(\mathbf{A}))$ time.*

$\mathbf{S}$ can be generated in the following way. Each column of $\mathbf{S}$ has exactly one non-zero entry, which is chosen uniformly and independently. We assign this entry to be 1 or -1 with equal probability. More precisely, let $h : [n] \to \{1, \ldots, r\}$ and $\sigma : [n] \to \{-1, 1\}$, then $\mathbf{S}_{ij} = \chi(h(j) = i)\sigma(j)$ with $\chi(h(j) = i)$ being an indicator. (The dimension of $\mathbf{S}$ can be reduced further by applying Johnson-Lindenstrauss Lemma on $\mathbf{SA}$, which gets $\mathbf{S}'\mathbf{SA}$.)

To prove Theorem 4.2, we need the following lemma and theorem.

Lemma 4.3. *If $r \geq \frac{2}{\epsilon^2\delta}$, then for any $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x}\|_2 = 1$, $\mathbf{S}$ generated by the above algorithm satisfies*

$$\mathsf{E}_{\mathbf{S}}\left(\left(\|\mathbf{Sx}\|_2^2 - 1\right)^2\right) \leq \epsilon^2\delta.$$

Theorem 4.4. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{B} \in \mathbb{R}^{n\times d}$. If $r \geq \frac{2}{\epsilon^2\delta}$, then*

$$\mathsf{P}_{\mathbf{S}}\left(\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B}\|_F \geq 2\epsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F\right) \leq \delta.$$

Proof of Theorem 4.2. Let $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d) \in \mathbb{R}^{n\times d}$ be an orthonormal basis of the column space of $\mathbf{A}$. To show $\mathbf{S}$ is a subspace embedding of $\mathbf{A}$, it suffices to show that $\|\mathbf{S}\mathbf{U}\mathbf{x}\|_2 = (1\pm\epsilon)\|\mathbf{x}\|_2, \forall\, \mathbf{x} \in \mathbb{R}^d$. It is equivalent to show that $\mathbf{x}^T(\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U})\mathbf{x} = (1\pm\epsilon)\mathbf{x}^T\mathbf{I}\mathbf{x}$, i.e., $\|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - \mathbf{I}\|_2 \leq \epsilon$. Since $\|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - \mathbf{I}\|_F \geq \|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - \mathbf{I}\|_2$, it suffices to show that $\|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - \mathbf{I}\|_F \leq \epsilon$.

Let $\mathbf{A} = \mathbf{B} = \mathbf{U}$. Since $\mathbf{U}$'s columns are orthonormal vectors, we have that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_d$ and $\|\mathbf{U}\|_F^2 = d$. By Theorem 4.4, we have

$$\mathsf{P}_{\mathbf{S}}\left(\|\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - \mathbf{I}_d\|_F \geq 2\epsilon_1 d\right) \leq \delta.$$

Set $\epsilon_1 = \epsilon/2d$, we have $r = O(d^2/\epsilon^2\delta)$. □

Proof of Theorem 4.4. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be two unit vectors. Then,

$$\langle \mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y}\rangle = \frac{\|\mathbf{S}\mathbf{x}\|_2^2 + \|\mathbf{S}\mathbf{y}\|_2^2 - \|\mathbf{S}(\mathbf{x}-\mathbf{y})\|_2^2}{2}.$$

Let X be a random variable and $f(X) = (\mathsf{E}X^2)^{1/2}$. By Minkowski's theorem, $f(X+Y) \leq f(X) + f(Y)$. Thus,

$$\begin{aligned}
&f\left(\langle \mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y}\rangle - \langle \mathbf{x}, \mathbf{y}\rangle\right)\\
=&\frac{1}{2}f\left(\|\mathbf{S}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 + \|\mathbf{S}\mathbf{y}\|_2^2 - \|\mathbf{y}\|_2^2 - \left(\|\mathbf{S}(\mathbf{x}-\mathbf{y})\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2\right)\right)\\
\leq&\frac{1}{2}\left(f\left(\|\mathbf{S}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right) + f\left(\|\mathbf{S}\mathbf{y}\|_2^2 - \|\mathbf{y}\|_2^2\right) + f\left(\|\mathbf{S}(\mathbf{x}-\mathbf{y})\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2\right)\right)\\
\leq&\frac{1}{2}\left(2\sqrt{\epsilon^2\delta} + 2\sqrt{\epsilon^2\delta}\right)\\
=&2\epsilon\sqrt{\delta}
\end{aligned}$$

The last inequality is due to Lemma 4.3, and the fact that $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$ and $\|\mathbf{x}-\mathbf{y}\|_2 \leq \sqrt{2}$.

It is easy to see that $(\mathbf{A}^T\mathbf{B})_{ij} = \langle \mathbf{A}^{(i)}, \mathbf{B}^{(j)}\rangle$. Let X_{ij} denote $\langle \mathbf{S}\mathbf{A}^{(i)}, \mathbf{S}\mathbf{B}^{(j)}\rangle - \langle \mathbf{A}^{(i)}, \mathbf{B}^{(j)}\rangle$. Then,

$$\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B}\|_F^2 = \sum_{ij} X_{ij}^2.$$

By the above calculation,

$$f\left(\frac{X_{ij}}{\|\mathbf{A}^{(i)}\|_2\|\mathbf{B}^{(j)}\|_2}\right) \leq 2\epsilon\sqrt{\delta}.$$

Thus, $f(X_{ij}) \leq 2\epsilon\sqrt{\delta}\|\mathbf{A}^{(i)}\|_2\|\mathbf{B}^{(j)}\|_2$, that is, $\mathbb{E}X_{ij}^2 \leq 4\epsilon^2\delta\|\mathbf{A}^{(i)}\|_2^2\|\mathbf{B}^{(j)}\|_2^2$. Then, we have

$$\mathsf{E}\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B}\|_F^2 \leq 4\epsilon^2\delta\|\mathbf{A}^{(i)}\|_F^2\|\mathbf{B}^{(j)}\|_F^2.$$

By Chebyshev inequality,

$$\mathsf{P}_{\mathbf{S}}\left(\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B}-\mathbf{A}^T\mathbf{B}\|_F \geq 2\epsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F\right) \leq \frac{\mathsf{E}\|\mathbf{A}^T\mathbf{S}\mathbf{S}^T\mathbf{B}-\mathbf{A}^T\mathbf{B}\|_F^2}{4\epsilon^2\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2} \leq \delta.$$

□

Proof of Lemma 4.3. Expand

$$\left(\|\mathbf{S}\mathbf{x}\|_2^2-1\right)^2 = \|\mathbf{S}\mathbf{x}\|_2^4 - 2\|\mathbf{S}\mathbf{x}\|_2^2 + 1$$

We will then bound the expectation of $\|\mathbf{S}\mathbf{x}\|_2^4$ and $\|\mathbf{S}\mathbf{x}\|_2^2$ separately.

$$\mathsf{E}\|\mathbf{S}\mathbf{x}\|_2^2 = \mathsf{E}\mathbf{x}^T\mathbf{S}^T\mathbf{S}\mathbf{x} = \mathbf{x}^T\left(\mathsf{E}\mathbf{S}^T\mathbf{S}\right)\mathbf{x}$$

Consider $\mathsf{E}(\mathbf{S}^T\mathbf{S})_{ij} = \mathsf{E}\langle\mathbf{S}^{(i)},\mathbf{S}^{(j)}\rangle$. According to the distribution of $\mathbf{S}$, each column of $\mathbf{S}$ has exactly one nonzero entry, which is picked as -1 or 1 uniformly and independently. Thus, $\mathsf{E}(\mathbf{S}^T\mathbf{S})_{ii} = 1$ and $\mathsf{E}(\mathbf{S}^T\mathbf{S})_{ij} = 0$ for $i \neq j$, i.e., $\mathsf{E}\mathbf{S}^T\mathbf{S} = \mathbf{I}_n$. It means that $\mathsf{E}\|\mathbf{S}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2 = 1$.

Similarly, we bound the expectation of $\|\mathbf{S}\mathbf{x}\|_2^4$.

$$\begin{aligned}\mathsf{E}\|\mathbf{S}\mathbf{x}\|_2^4 &= \mathsf{E}\left(\sum_i\left(\sum_j \mathbf{S}_{ij}\mathbf{x}_j\right)^2\right)^2 \\ &= \sum_{i,i'}\sum_{j_1,j_2,j_1',j_2'} \mathsf{E}\mathbf{S}_{ij_1}\mathbf{S}_{ij_2}\mathbf{S}_{i'j_1'}\mathbf{S}_{i'j_2'}\mathbf{x}_{j_1}\mathbf{x}_{j_2}\mathbf{x}_{j_1'}\mathbf{x}_{j_2'}\end{aligned}$$

Since columns of $\mathbf{S}$ are sampled independently and $\mathsf{E}\mathbf{S}_{ij} = 0$, we have that $\mathsf{E}\mathbf{S}_{ij}\mathbf{S}_{i'j'} = 0$ for all $j \neq j'$. Moreover, $\mathsf{E}\mathbf{S}_{ij}^2 = \mathsf{E}\mathbf{S}_{ij}^4 = \frac{1}{r}$. Thus,

$$\begin{aligned}\mathsf{E}\|\mathbf{S}\mathbf{x}\|_2^4 &= \sum_i\sum_j \mathsf{E}\mathbf{S}_{ij}^4\mathbf{x}_j^4 + \sum_{i,i'}\sum_{j_1\neq j_1'}\mathsf{E}\mathbf{S}_{ij_1}^2\mathbf{S}_{i'j_1'}^2\mathbf{x}_{j_1}^2\mathbf{x}_{j_1'}^2 + 2\sum_i\sum_{j_1\neq j_2}\mathsf{E}\mathbf{S}_{ij_1}^2\mathbf{S}_{ij_2}^2\mathbf{x}_{j_1}^2\mathbf{x}_{j_2}^2 \\ &= \sum_i\sum_j\frac{1}{r}\mathbf{x}_j^4 + \sum_{i,i'}\sum_{j_1\neq j_1'}\frac{1}{r^2}\mathbf{x}_{j_1}^2\mathbf{x}_{j_1'}^2 + 2\sum_i\sum_{j_1\neq j_2}\frac{1}{r^2}\mathbf{x}_{j_1}^2\mathbf{x}_{j_2}^2 \\ &= \sum_j\mathbf{x}_j^4 + \sum_{j_1\neq j_1'}\mathbf{x}_{j_1}^2\mathbf{x}_{j_1'}^2 + 2\sum_{j_1\neq j_2}\frac{1}{r}\mathbf{x}_{j_1}^2\mathbf{x}_{j_2}^2 \\ &\leq 1+\frac{2}{r}\end{aligned}$$

Putting all together, we have

$$\mathsf{E}\left(\|\mathbf{S}\mathbf{x}\|_2^2-1\right)^2 = \mathsf{E}\|\mathbf{S}\mathbf{x}\|_2^4 - 2\mathsf{E}\|\mathbf{S}\mathbf{x}\|_2^2 + 1 \leq \frac{2}{r}.$$

Since $r \geq \frac{2}{\epsilon^2\delta}$, we have $\mathsf{E}\left(\|\mathbf{S}\mathbf{x}\|_2^2-1\right)^2 \leq \epsilon^2\delta$. □

References

[KV08] Ravindran Kannan and Santosh Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4(34):157–288, 2008.

[Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(12):1–157, 2014.