Lecture 10 — February 19, 2024

Instructor: Shashanka Ubaru *Scribe: Addie Duncan*

1 Sketch and Solve

Recall that in the setting when we have a large number of samples compared to the number of features a useful way to approximate the solution to the linear equation $\mathbf{Ax} = \mathbf{b}$ is to first sketch then solve.

The process involves 3 steps:

1. Generate a sketching matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$.
2. Compute Sketches $\mathbf{SA}$ and $\mathbf{Sb}$.
3. Solve:
$$\tilde{\mathbf{x}} = \min_{\mathbf{x}\in\mathbb{R}^d} ||\mathbf{SAx} - \mathbf{Sb}||_2^2$$
for $\epsilon \le 1/3$

If S is a subspace ϵ-embedding for *span*($[\mathbf{Ab}]$) then we can show that

$$||\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}||_2 \le (1+3\epsilon)||\mathbf{Ax}^* - \mathbf{b}||_2$$

where

$$\mathbf{x}^* = \min_{\mathbf{x}\in\mathbb{R}^d} ||\mathbf{Ax} - \mathbf{b}||_2$$
$$\tilde{\mathbf{x}} = \min_{\mathbf{x}\in\mathbb{R}^d} ||\mathbf{S}(\mathbf{Ax} - \mathbf{b})||_2.$$

A similar result holds for other sketching matrices.

Proposition 1. If S is a Countsketch matrix with $m = O(d^2/\epsilon)$ or SRHT with $m = O(d\log d/\epsilon)$, or Gaussian sketch with $m = O(d/\epsilon)$, then

$$||\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}||_2 \le (1+\epsilon)||\mathbf{Ax}^* - \mathbf{b}||_2. \quad (1)$$

Proof. Using Pythagorean theorem with an orthonormal basis $\mathbf{U}$ of $\mathbf{A}$ we can see that 1 is equivalent to showing $||\tilde{\mathbf{y}} - \mathbf{y}^*||_2^2$ is within $O(\epsilon)$ of $||\mathbf{Uy}^* - \mathbf{b}||_2^2$.

In other words let $\mathbf{U}\tilde{\mathbf{y}} = \mathbf{A}\tilde{\mathbf{x}}$ and $\mathbf{Uy}^* = \mathbf{Ax}^*$. Then by the Pythagorean theorem we get

$$||\mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}||_2^2 = ||\mathbf{Ax}^* - b||_2^2 + ||\mathbf{A}\tilde{\mathbf{x}} - \mathbf{Ax}^*||_2^2$$
$$\implies ||\mathbf{U}\tilde{\mathbf{x}} - \mathbf{b}||_2^2 = ||\mathbf{Ux}^* - b||_2^2 + ||\mathbf{U}\tilde{\mathbf{x}} - \mathbf{Ux}^*||_2^2.$$

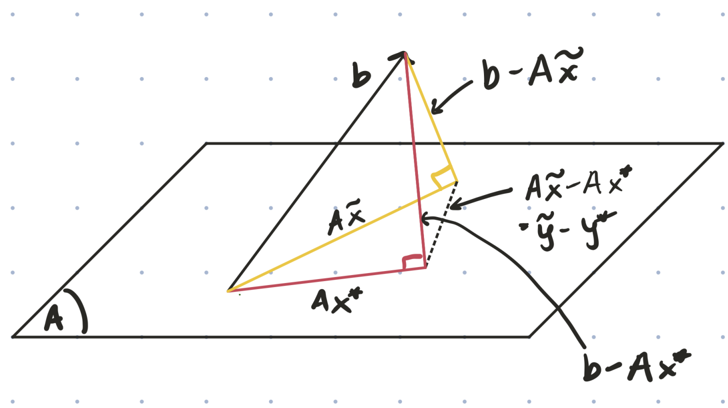

Figure 1: The solution and the sketched solution with respect to the column space of A.

Since $||\mathbf{U}(\tilde{\mathbf{y}} - \mathbf{y}^*)||_2^2 = ||\tilde{\mathbf{y}} - \mathbf{y}^*||_2^2$ we need to show that $||\tilde{\mathbf{y}} - \mathbf{y}^*||_2^2 = O(\epsilon)||\mathbf{U}\mathbf{y}^* - \mathbf{b}||_2^2$.

Figure 1 shows the geometric idea of the argument.

Now recall that for a subspace embedding $\mathbf{S}$ we have

$$||\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - I||_2 \leq \frac{1}{2}$$

Then,

$$\begin{aligned}
||\tilde{\mathbf{y}} - \mathbf{y}^*||_2 &\leq ||\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U}(\tilde{\mathbf{y}} - \mathbf{y}^*)||_2 + ||\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U}(\tilde{\mathbf{y}} - \mathbf{y}^*) - (\tilde{\mathbf{y}} - \mathbf{y}^*)||_2 \\
&\leq ||\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U}(\tilde{\mathbf{y}} - \mathbf{y}^*)||_2 + ||\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - I||_2||\tilde{\mathbf{y}} - \mathbf{y}^*||_2 \\
&\leq ||\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U}(\tilde{\mathbf{y}} - \mathbf{y}^*)||_2 + \frac{1}{2}||\tilde{\mathbf{y}} - \mathbf{y}^*||_2 \\
&\leq 2||\mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U}(\tilde{\mathbf{y}} - \mathbf{y}^*)||_2 \\
&\leq 2||\mathbf{U}^T\mathbf{S}^T\mathbf{S}(\mathbf{U}\mathbf{y}^* - \mathbf{b})||_2.
\end{aligned}$$

With high probability, we have that

$$\begin{aligned}
||\mathbf{U}^T\mathbf{S}^T\mathbf{S}(\mathbf{U}\mathbf{y}^* - \mathbf{b})||_2^2 &\leq \frac{9\epsilon}{d^2}||\mathbf{U}||_F^2||\mathbf{U}\mathbf{y}^* - \mathbf{b}||_2^2 \\
&\leq 18\epsilon||\mathbf{U}\mathbf{y}^* - \mathbf{b}||_2^2
\end{aligned}$$

which shows that $||\tilde{\mathbf{y}} - \mathbf{y}^*||_2^2 = O(\epsilon)||\mathbf{U}\mathbf{y}^* - \mathbf{b}||_2^2$ as desired.

■

2 Sampling for Least Squares

Recall leverage scores:

$$\ell_i(\mathbf{A}) := \sup_{\mathbf{x}} \frac{(\mathbf{A}_{i*}\mathbf{x})^2}{||\mathbf{A}\mathbf{x}||^2} = ||\mathbf{U}_{i*}||^2$$

where $\mathbf{U}$ is an orthonormal basis for $span(\mathbf{A}$.

We can use leverage scores to sample rows of $\mathbf{A}$ to approximate a least squares problem. The general idea is to pick m rows of $\mathbf{A}$ with the probability of choosing the i^{th} row is chosen to be $p_i = \ell_i/d$.

For a sampling matrix $\mathbf{S}$ chosen this way we get an ϵ-embedding.

Proposition 2. Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ with $r = \text{rank}(\mathbf{A})$, and $\mathbf{S} \in \mathbb{R}^{m\times n}$ be a sampling amtrix with probabilities $p_i = \ell_i/r$, and $\mathbf{S}_{i*} = \mathbf{e}_j/\sqrt{mp_j}$ with $Pr(j = i) = p_i$. If $m = O(r\log(r/\delta)/\epsilon^2)$, then $\mathbf{S}$ is ϵ-subspace embedding of $span(\mathbf{A})$ with probability $1 - \delta$.

To prove this proposition we will need the matrix Chernoff bound.

Theorem 1 (Matrix Chernoff)**.** *Let $\mathbf{X}_k$ for $k \in [m]$ be i.i.d copies of a symmetric random variable $\mathbf{X} \in \mathbb{R}^{r\times r}$ with $\gamma, \sigma^2 > 0$, $\mathbb{E}[\mathbf{X}] = 0$, $||\mathbf{X}||_2 \le \gamma$, and $||\mathbb{E}[\mathbf{X}^2]||_2 \le \sigma^2$. Then for $\epsilon > 0$,*

$$Pr(||\frac{1}{m}\sum_k \mathbf{X}_k||_2 \ge \epsilon) \le sr\exp(-m\epsilon^2/(\sigma^2 + \gamma^2 + \gamma\epsilon/3)).$$

Proof of Proposition 2. Let $\mathbf{U} \in \mathbb{R}^{n\times r}$ be orthonormal with $span(\mathbf{U}) = span(\mathbf{A})$. Let

$$\mathbf{X}_k = m\mathbf{U}^T[\mathbf{S}_{k*}]^T\mathbf{S}_{k*}\mathbf{U} - I$$

so that

$$\frac{1}{m}\sum_k \mathbf{X}_k = \mathbf{U}^T\mathbf{S}^T\mathbf{S}\mathbf{U} - I, \tag{2}$$

To show that we have an ϵ-embedding we need to bound the spectral norm of (2).

Let

$$\mathbf{X} = \frac{1}{p_j}[\mathbf{U}_{j*}]^T\mathbf{U}_{j*} - I \text{ with } Pr(j = i) = p_i = \ell_i/r = ||\mathbf{U}_{i*}||_2^2/r,$$

then we have the following:

- $\mathbb{E}[\mathbf{X}] = 0$:

$$\begin{aligned}\mathbb{E}[\mathbf{X}] &= \mathbb{E}_j[\frac{1}{p_j}\mathbf{U}_{j*}^T\mathbf{U}_{j*} - I] \\ &= \mathbf{U}^T\mathbf{U} - I \\ &= 0\end{aligned}$$

- $||\mathbf{X}||_2 \le r + 1$:

$$\begin{aligned}||\mathbf{X}||_2 &= ||\frac{1}{p_j}\mathbf{U}_j^T\mathbf{U}_j - I||_2 \\ &\le \max_j \frac{1}{p_j}||\mathbf{U}_j^T\mathbf{U}_j||_2 + ||I||_2 \text{ by triangle inequality}\end{aligned}$$

  Since $p_j = \ell_j/r$,$||\mathbf{U}_j^T\mathbf{U}_j||_2 = \ell_j$ and $||I||_2 = 1$ we conclude that $||\mathbf{X}||_2 \le r + 1$.

- $\mathbb{E}[\mathbf{X}^2] \leq (r-1)I$:

$$\begin{aligned}
\mathbb{E}[\mathbf{X}^2] &= \mathbb{E}_j[[\frac{1}{p_j}\mathbf{U}_j^T\mathbf{U}_j - i]^2] \\
&= \mathbb{E}_j[\frac{1}{p_j}\mathbf{U}_j^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_j] - 2\mathbb{E}_j[\frac{1}{p_j}\mathbf{U}_j^T\mathbf{U}_j]I \\
&= \mathbb{E}_j[\frac{1}{p_j}\mathbf{U}_j^T\mathbf{U}_j\mathbf{U}_j^T\mathbf{U}_j] - I \\
&\leq \mathbb{E}_j[\frac{||\mathbf{U}_j||^2}{p_j^2}\mathbf{U}_j\mathbf{U}_j^T] - I \\
&\leq \mathbb{E}_j[||\mathbf{U}_j||^2\frac{1}{p_j}\mathbf{U}_j^T\mathbf{U}_j] - I \\
&\leq \mathbb{E}_j[||\mathbf{U}_j||^2\frac{r}{\ell_j}\mathbf{U}_j^T\mathbf{U}_j] - I \\
&\leq r\mathbb{E}_j[\mathbf{U}_j^T\mathbf{U}_j] - I \\
&= rI - I
\end{aligned}$$

Thus $||\mathbb{E}[\mathbf{X}^2]||_2 \leq r-1$.

■

3 Preconditioning for Least Squares

When preconditioning for Least Squares we can use an Iterative Refinement method.

Recall that Iterative Refinement is the process of solving for $x^* = ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2$ where in the j^{th} iteration we set

$$\mathbf{x}^{j+1} = \mathbf{x}^j + \mathbf{A}^T\mathbf{r}$$

where $\mathbf{r} = \mathbf{A}\mathbf{x}^j - \mathbf{b}$.

To precondition Iterative refinement we replace $\mathbf{A}$ with $\mathbf{A}\mathbf{R}^{-1}$ where $\mathbf{R}$ is the preconditioner.

We can use sketching to precondition Least Squares. We set our preconditioner $\mathbf{R}$ to be the "R" in the QR decomposition of $\mathbf{SA}$ where $\mathbf{S}$ is a sketching matrix.

To show why this works take

$$\mathbf{x}^{(j+1)} \leftarrow \mathbf{x}^{(j)} + (\mathbf{R}^T)^{-1}\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{R}^{-1}\mathbf{x}^{(j)}).$$

Then we have

$$\begin{aligned}
\mathbf{A}\mathbf{R}^{-1}(\mathbf{x}^{(j+1)} - \mathbf{x}^*) &= \mathbf{A}\mathbf{R}^{-1}(\mathbf{x}^{(j)} + (\mathbf{R})^{-T}\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{R}^{-1}\mathbf{x}^{(j)}) - \mathbf{x}^*) \\
&= \mathbf{A}\mathbf{R}^{-1}\mathbf{x}^{(j)} - \mathbf{A}\mathbf{R}^{-1}\mathbf{R}^{-T}\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{R}^{-1}\mathbf{x}^{(j)} - \mathbf{x}^*) \\
&= (\mathbf{A}\mathbf{R}^{-1} - \mathbf{A}\mathbf{R}^{-1}\mathbf{R}^{-T}\mathbf{A}^T\mathbf{A}\mathbf{R}^{-1})(\mathbf{x}^{(j)} - \mathbf{x}^*) \\
&= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T - \mathbf{U}\mathbf{\Sigma}^3\mathbf{V}^T)(\mathbf{x}^{(j)} - \mathbf{x}^*) \text{ where } \mathbf{A}\mathbf{R}^{-1} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\
&= \mathbf{U}(\mathbf{\Sigma} - \mathbf{\Sigma}^3)\mathbf{V}^T(\mathbf{x}^{(j)} - \mathbf{x}^*)
\end{aligned}$$

Since $\mathbf{AR}^{-1}$ has singular values in $[1-\epsilon_0, 1+\epsilon_0]$ and the diagonal entries of $\mathbf{\Sigma} - \mathbf{\Sigma}^3$ are at most $\sigma_i(1-(1-\epsilon_0)^2) \leq 3\sigma_i\epsilon_0$ for $\epsilon_0 \leq 1$, we have

$$||\mathbf{AR}^{-1}(\mathbf{x}^{(m+1)} - \mathbf{x}^*)|| \leq 3\epsilon_0 ||\mathbf{AR}^{-1}(\mathbf{x}^{(m)} - \mathbf{x}^*)||.$$

Let $\epsilon_0 = 1/2$ then $O(\log(1/\epsilon))$ iterations suffice to attain ϵ relative error.