Randomized Algorithms

Low-Rank Approximation

(Part 2)

Recap

$$A \approx QB$$

$$m\underset{n}{\Big[\qquad\Big]} \approx m\underset{\ell}{\Big[\quad\Big]}\overset{n}{\Big[\qquad\Big]}\ell$$

"Many large data matrices are well-approx. by low-rank matrices"

Idea: Randomized projection onto low-dim. subspace and do direct NLA there.

Stage A: $m \times \ell$ i.i.d. Gaussian random matrix X

"Sketch" $\Rightarrow$ $Y = AX$

ONB $\Rightarrow$ $Y = QR$

$\boxed{\text{hope that } A \approx QQ^*A}$

Stage B: Compute $B = Q^*A$ and factor B with a direct method, e.g.

$\Rightarrow B = \tilde{U}\tilde{\Sigma}\tilde{V}^* \quad \Rightarrow A \approx (Q\tilde{U})\tilde{\Sigma}\tilde{V}^*$

Key Question $\Rightarrow$ Given $\varepsilon > 0$, when is $\|A - QQ^*A\| < \varepsilon$?

$\Rightarrow$ If we choose $Q = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix}$, i.e., the top k dominant singular vectors of A s.t. $\sigma_{k+1} < \varepsilon$, then $\|A - QQ^*A\| = \sigma_{k+1} < \varepsilon$. By Eckart-Young, this is the <u>optimal</u> Q to choose, as it minimizes $\|A - QQ^*A\| = \|A - \underbrace{A_k}_{\text{k-truncated SVD of } A}\|$.

$\Rightarrow$ Remarkably, stage A can provide a Q that is not far from optimal!

<u>Thm 1</u> A $m \times n$ real matrix. Select target rank $k \geq 2$ and oversampling param. $p \geq 2$ s.t. $k + p \leq \frac{1}{2}\min\{m,n\}$. Then, Stage A produces $m \times (k+p)$ ONB s.t.

$$\mathbb{E}\|A - QQ^*A\|_2 \leq \sigma_{k+1}\left[1 + \frac{4\sqrt{k+p}}{p-1}\sqrt{\min\{m,n\}}\right].$$

If $4\log_2 4 \leq p \leq \min\{n,m\}$, then

$$\|A - QQ^*A\|_2 \leq \sigma_{k+1}\left[1 + 9\sqrt{k+p}\sqrt{\min\{m,n\}}\right]$$

with probability $\geq 1 - 3p^{-p}$.

$\Rightarrow$ see H.M.T. Thm 1.1 and Cor. 10.9.

$\Rightarrow$ for $p = 10$, failure probability $= 3 \times 10^{-10}$!

<u>Why is oversampling so powerful?</u>

top k singular vectors

$$Y = \underbrace{U \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} - V_1^* - \\ - V_2^* - \end{bmatrix}}_{A = U\Sigma V^*} \underbrace{\begin{bmatrix} | & & | \\ x_1 & \cdots & x_\ell \\ | & & | \end{bmatrix}}_{X}$$

$$= \underbrace{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix} \begin{bmatrix} - v_1^* - \\ - v_k^* - \end{bmatrix}}_{A_k = \text{k-truncated SVD}} \begin{bmatrix} | & & | \\ x_1 & \cdots & x_\ell \\ | & & | \end{bmatrix}$$

$$+ \underbrace{\begin{bmatrix} | & & | \\ u_{k+1} & \cdots & u_n \\ | & & | \end{bmatrix} \begin{bmatrix} \sigma_{k+1} & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \begin{bmatrix} - v_{k+1}^* - \\ \vdots \\ - v_n^* - \end{bmatrix}}_{E} \begin{bmatrix} | & & | \\ x_1 & \cdots & x_\ell \\ | & & | \end{bmatrix}$$

$$\Rightarrow \quad Y = \underbrace{A_k X}_{W} + EX$$

In Lecture 24, we saw that W is a basis for $\mathrm{span}\{u_1, \dots, u_k\}$, the <u>optimal</u> basis, as long as $V_1^* X$ has rank $= k$.

<u>Intuition</u> If the columns of EX are small relative to those of W, then Y contains an ONB that is not far from optimal.

So we have two questions:

$\Rightarrow$ How big can $\|EXe_j\|$ be?

$\Rightarrow$ How small can $\|A_k X e_j\|$ be?

These are questions about <u>singular values</u> of $A_k X$ and EX.

<u>Two facts about Gaussian Matrices</u>

<u>Prop 1</u> Fix real matrix S and let X be $m \times n$ standard Gaussian. Then, $\mathbb{E}[SX] = \sqrt{n}\|S\|_2 + \|S\|_F \leq 2\sqrt{n}\|S\|_2$

Prop 2 Let X be $m \times n$ standard Gaussian with $n-m \geq 1$ and $n \geq 2$. Then

$$\mathbb{E}\left[1/\sigma_{min}(X)\right] < \frac{e\sqrt{n}}{n-m}$$

and for each $t > 0$,

$$\mathbb{P}\left[\sigma_{min}(X) < 1/t\right] \leq \frac{1}{\sqrt{2\pi(n-m+1)}}\left[\frac{e\sqrt{n}}{n-m+1}\right]^{n-m+1} t^{-(n-m+1)}$$

$\Rightarrow$ Probability of small $\sigma_{min}(X)$ shrinks exponentially as $n-m$ grows!

How big can $\|EX\|$ be?

By prop 1, $\|EX\| \leq 2\sqrt{n}\,\|E\|_2 = 2\sqrt{n}\,\sigma_{k+1}$ ✓

$\Rightarrow$ Small when σ_{k+1} is small

How small can $\|A_k X e_j\|$ be?

$$A_k X = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix} \begin{bmatrix} - & v_1^* & - \\ & \vdots & \\ - & v_k^* & - \end{bmatrix} \begin{bmatrix} | & & | \\ x_1 & \cdots & x_\ell \\ | & & | \end{bmatrix}$$

$$= \begin{bmatrix} | & & | \\ \sigma_1 u_1 & \cdots & \sigma_k u_k \\ | & & | \end{bmatrix} \begin{bmatrix} - & v_1^* & - \\ & \vdots & \\ - & v_k^* & - \end{bmatrix} \begin{bmatrix} | & & | \\ x_1 & \cdots & x_\ell \\ | & & | \end{bmatrix}$$

orthogonal columns

$k \times \ell$ Gaussian matrix

$\Rightarrow$ By Prop 2, probability of $V_1^* X$ having small singular value decreases exponentially as the oversampling param $\ell - k = p$ grows!

$\Rightarrow$ The $k \times \ell$ Gaussian then mixes the columns of $U_k \Sigma_k$ so the columns of $A_k X$ are typically on the order of σ_1 (and no worse than $\mathcal{O}(\sigma_k)$).

Oversampling parameter controls probability that span(Y) is far from span(W). As p increases, the probability that span(Y) is far from optimal span(W) decreases rapidly.