EE270
Large scale matrix computation, optimization and learning

Instructor : Mert Pilanci

Stanford University

Tuesday, Feb 2 2020

Randomized Linear Algebra
Lecture 8: Randomized Least Squares Bias and Variance, Streaming Data

Least Squares Problems and Random Projection

- Given $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^d$
  find the best linear fit $Ax \approx b$ according to

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2 + \lambda x^T x$$

$$\begin{array}{cc} & d \\ n & \begin{bmatrix} A \end{bmatrix} \end{array}$$

rank(A) = d if A is full column rank

- no regularization, i.e., $\lambda = 0$
- If A is full column rank then
- $x_{LS} = (A^T A \; {+\lambda I})^{-1} A^T b = A^{+} b$

$$\approx (A^T S^T S A)^{-1} A^T S^T S b = (SA)^{+}(Sb)$$

$$= \arg\min_{x \in \mathbb{R}^d} \|\tilde{A}x - \tilde{b}\|_2^2$$

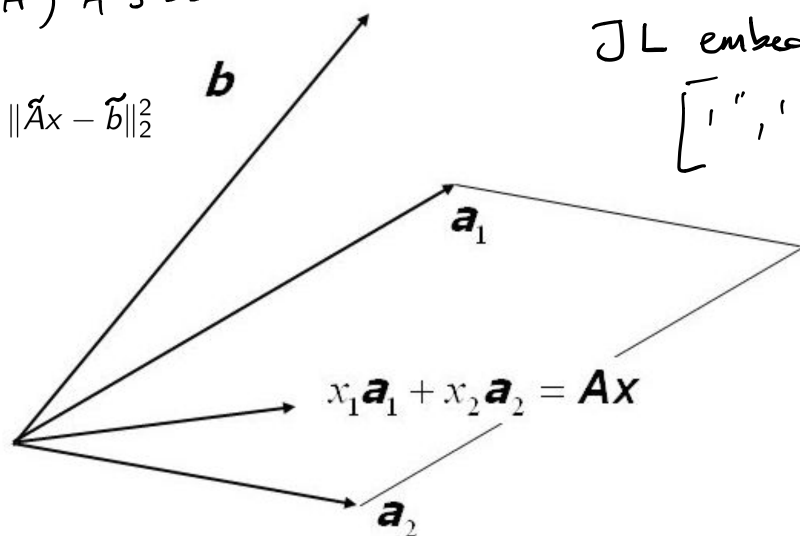
$\tilde{A} = SA$

$\tilde{b} = Sb$

Left sketching

S: sampling matrix (need a distribution)

S: iid distribution or JL embedding



Faster Least Squares Optimization: Random Projection

- **Left-sketching**

  Form SA and Sb where $S \in \mathbb{R}^{m \times n}$ is a random projection matrix

- Solve the smaller problem

$$\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2$$

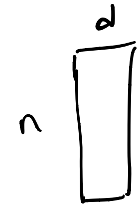
- using any classical method.

  Direct method complexity md^2

Approximation Result

- Suppose that $n \gg d$
- Let $S \in \mathbb{R}^{m \times d}$ be a Johnson-Lindenstrauss Embedding

$$x_{LS} = \arg\min_{x \in \mathbb{R}^d} \underbrace{\|Ax - b\|_2^2}_{f(x)}$$



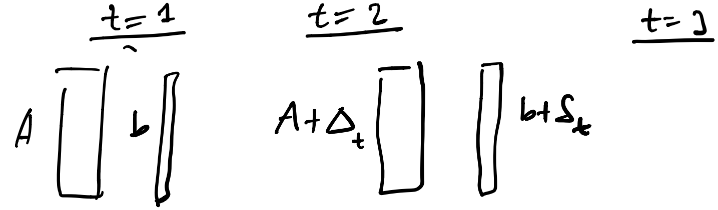
$$\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2$$

$\mathrm{rank}(A) \le d$

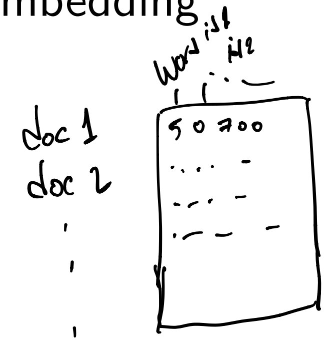
- **Lemma** If $m \geq$ constant $\times \frac{rank(A)}{\epsilon^2}$ then,
- $f(x_{LS}) \leq f(\tilde{x}) \leq (1 + \epsilon^2) f(x_{LS})$
- $\|A(x_{LS} - \tilde{x})\|_2^2 \leq \epsilon^2$ with high probability

error prob. is exponentially small.

Application: Streaming data



- Suppose that $n \gg d$
- Let $S \in \mathbb{R}^{m \times d}$ be a Johnson-Lindenstrauss Embedding



$$x_{LS} = \arg\min_{x \in \mathbb{R}^d} \underbrace{\|Ax - b\|_2^2}_{f(x)}$$

Randomized Linear Algebra

$O(md)$ space + $O(md^2)$ compute per time instant

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2$$

Idea: Update $SA \leftarrow SA + S\Delta$ and $Sb \leftarrow Sb + S\delta$

- A and b are dynamically updated and we need to find x_{LS} at any time

  $A_{t+1} = A_t + \Delta_t$ and $y_{t+1} = y_t + \Delta_t$

  Can we form and update $A_t^T A_t \in \mathbb{R}^{d \times d}$?

$x_{LS} = (A^T A)^{-1} A^T b$

Classical Linear Algebra

$(A+\Delta)^T (A+\Delta) = A^T A + \Delta^T A + \cdots$

$O(nd)$ space and $O(nd^2)$ compute per time instant

Application: Streaming data

- Suppose that $n \gg d$
- Let $S \in \mathbb{R}^{m \times d}$ be a Johnson-Lindenstrauss Embedding

$$x_{LS} = \arg\min_{x \in \mathbb{R}^d} \underbrace{\|Ax - b\|_2^2}_{f(x)}$$

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2$$

- A and b are dynamically updated and we need to find x_{LS} at any time

  $A_{t+1} = A_t + \Delta_t$ and $y_{t+1} = y_t + \Delta_t$

  Can we form and update $A_t^T A_t \in \mathbb{R}^{d \times d}$?
- Linear sketch can be updated on the fly

  $SA_{t+1} = SA_t + S\Delta_t$ and $Sy_{t+1} = Sy_t + S\Delta_t$

Gaussian Sketch

- Let S be $\frac{1}{m}\times$ i.i.d. Gaussian. $\mathbb{E}[S^T S] = I$

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2$$

- Is $\mathbb{E}\left[\tilde{x}\right]$ equal to x_{LS}?

Gaussian Sketch

- Let S be $\frac{1}{m}\times$ i.i.d. Gaussian. $\mathbb{E}[S^T S] = I$

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2$$

- Is $\mathbb{E}[\tilde{x}]$ equal to x_{LS}? Yes! Only Gaussian.
- Assuming $A^T S^T SA$ is invertible, we have

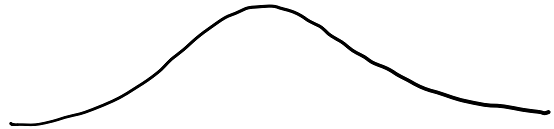
$$\tilde{x} = (A^T S^T SA)^{-1} A^T S^T Sb$$

let $b = Ax_{LS} + b^\perp$ where $b^\perp \perp Range(A)$

$$\tilde{x} = (A^T S^T SA)^{-1} A^T S^T S(Ax_{LS} + b^\perp)$$
$$= x_{LS} + (A^T S^T SA)^{-1} A^T S^T Sb^\perp$$

- $\mathbb{E}(A^T S^T SA)^{-1} A^T S^T Sb^\perp = 0$ since $Sb^\perp$ and SA are uncorrelated zero mean Gaussian.

Gaussian Sketch: Variance



$f(x_{LS}) = \min_x \|Ax-b\|_2^2$

- Let S be i.i.d. Gaussian

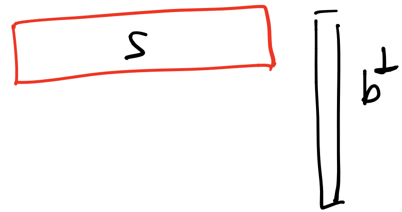
$$\tilde{x} = \arg\min_{x\in\mathbb{R}^d} \underbrace{\|SAx - Sb\|_2^2}_{f(x)} = x_{LS} + (A^T S^T SA)^{-1} A^T S^T Sb^{\perp}$$
$$= x_{LS} + (SA)^{\dagger} Sb^{\perp}$$

- Analyzing the variance $\mathbb{E}\|A\tilde{x} - x_{LS}\|_2^2$
- **Lemma (a)** Conditioned on the matrix SA

$$\tilde{x} \sim N\Big(x_{LS}, \frac{f(x_{LS})}{m}(A^T S^T SA)^{-1}\Big)$$

Gaussian Sketch: Variance

$\mathbb{E}\, S^T S = I$



$\mathbb{E}(S b^\perp) = \mathbb{E}\, s_1^T b^\perp = 0$

- Let S be i.i.d. Gaussian $\times \frac{1}{\sqrt{m}}$

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2 = x_{LS} + (A^T S^T S A)^{-1} A^T S^T S b^\perp$$

$$= x_{LS} + (SA)^\dagger S b^\perp$$

independent

- Analyzing the variance $\mathbb{E}\|A\tilde{x} - x_{LS}\|_2^2$
- **Lemma (a)** Conditioned on the matrix SA

$$\tilde{x} \sim N\left(x_{LS}, \frac{f(x_{LS})}{m} (A^T S^T S A)^{-1}\right)$$

$\mathbb{E}\, Sb^\perp (Sb^\perp)^T$

$= \begin{bmatrix} \mathbb{E}\, s_1^T b^\perp \cdot s_1^T b^\perp & \cdots \\ \underbrace{\mathbb{E}\, s_2^T b^\perp}_{0} \underbrace{\mathbb{E}\, s_1^T b^\perp}_{0} & \end{bmatrix}$

$b^{\perp T} \underbrace{\mathbb{E}\, s_1 s_1^T}_{I \cdot \frac{1}{m}} b^\perp = \frac{b^{\perp T} b^\perp}{m}$

- $Sb^\perp \sim N\left(0, \frac{\|b^\perp\|_2^2}{m} I\right)$ $\quad \frac{\|b^\perp\|_2^2}{m}$
- $\mathbb{E}(\tilde{x} - x_{LS})(\tilde{x} - x_{LS})^T = (SA)^\dagger((SA)^\dagger)^T = (A^T S^T S A)^{-1} \frac{\|b^\perp\|_2^2}{m}$

$= \mathbb{E}(A^T S^T S A)^{-1} A^T S^T S b^\perp \left((A^T S^T S A)^{-1} A^T S^T S b^\perp\right)^T$

Gaussian Sketch: Variance

$(A^T S^T S A)\cdot(\ )(\ )$

- Let S be i.i.d. Gaussian

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2$$

- Analyzing the variance $\mathbb{E}\|A\tilde{x} - x_{LS}\|_2^2$
- **Lemma (a)** Conditioned on the matrix SA

$$\tilde{x} \sim N\left(x_{LS}, \frac{f(x_{LS})}{m}(A^T S^T SA)^{-1}\right)$$

$$A(\tilde{x} - x_{LS}) \sim N\left(0, \frac{f(x_{LS})}{m} A(A^T S^T SA)^{-1} A^T\right)$$

$x \sim N(0, \Sigma)$

$Ax \sim N(0, A \Sigma A^T)$

$\mathbb{E} \cdot Ax \cdot x^T A^T = A \Sigma A^T)$

Gaussian Sketch: Variance

- Let S be i.i.d. Gaussian

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2$$

- Analyzing the variance $\mathbb{E}\|A\tilde{x} - x_{LS}\|_2^2$
- **Lemma (a)** Conditioned on the matrix SA

$$\tilde{x} \sim N\left(x_{LS}, \frac{f(x_{LS})}{m}(A^T S^T SA)^{-1}\right)$$

$$A(\tilde{x} - x_{LS}) \sim N\left(0, \frac{f(x_{LS})}{m} A(A^T S^T SA)^{-1} A^T\right)$$

$$\approx A(A^TA)^{-1} \cdot \frac{m}{m-d-1} A^T \approx I \cdot \frac{m}{m-d-1}$$

Lemma (b) (removing conditioning) for $m > d + 1$

$$\mathbb{E}\left[(A^T S^T SA)^{-1}\right] = (A^T A)^{-1} \frac{m}{m - d - 1}$$

Only for the Gaussian distribution,

Inverse matrices are biased but they are predictable

$$\mathbb{E}\,(A^T S^T S A)^{-1} \neq (A^T A)^{-1}$$

for $m > d+1$

Gaussian Sketch: Variance

$x \sim N(0, \Sigma)$

$\mathbb{E}\|x\|_2^2 = \mathbb{E}\,\mathrm{tr}\, xx^T = \mathbb{E}\,\mathrm{tr}\, x^Tx$
$= \mathrm{tr}\, \underbrace{\mathbb{E}\, xx^T}$
$= \mathrm{tr}\, \Sigma$

- Let S be i.i.d. Gaussian

$$\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2$$

- Analyzing the variance $\mathbb{E}\|A\tilde{x} - x_{LS}\|_2^2$

$\mathrm{tr}\, A(A^TA)^{-1}A^T$ $d \times d$
$= \mathrm{tr}\, \underbrace{(A^TA)^{-1}A^T A}_{I} = d$

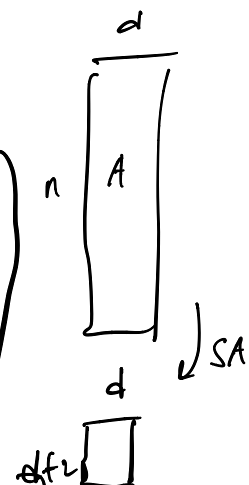
- **Lemma (a)** Conditioned on the matrix SA

tower prop. of cond. expectn

$\mathbb{E}_{SA}\left[\mathbb{E}\left[\|C\|_2^2 \mid SA\right]\right]$
$= \mathbb{E}\|C\|_2^2$

$$\tilde{x} \sim N\Big(x_{LS}, \frac{f(x_{LS})}{m}(A^TS^TSA)^{-1}\Big)$$

$$A(\tilde{x} - x_{LS}) \sim N\Big(0, \frac{f(x_{LS})}{m}A(A^TS^TSA)^{-1}A\Big)$$



Lemma (b) (removing conditioning) for $m > d + 1$

$$\mathbb{E}\left[(A^TS^TSA)^{-1}\right] = (A^TA)^{-1}\frac{m}{m-d-1}$$

- $\boxed{\mathbb{E}\|\underbrace{A(\tilde{x} - x_{LS})}_{C}\|_2^2} = \mathbb{E}_{SA}\left[\mathbb{E}\left[\frac{f(x_{LS})}{m}\mathrm{tr}A(A^TS^TSA)^{-1}A^T \mid SA\right]\right]$
- $\mathbb{E}\|A(\tilde{x} - x_{LS})\|_2^2 = \frac{f(x_{LS})}{m-d-1}\mathrm{tr}\underbrace{A(A^TA)^{-1}A^T} = f(x_{LS})\frac{d}{m-d-1}$

Expected Inverse of a Random Matrix

$\mathbb{E}\, A^T S^T S A = A^T A$

$\mathbb{E}\, (A^T S^T S A)^{-1} = A^T A \cdot \gamma$

$S =$ —— $A =$ | $\frac{1}{g_1^2 + g_2^2 \cdots}$

- Where does the formula

$$\mathbb{E}\left[(A^T S^T S A)^{-1}\right] = (A^T A)^{-1} \underbrace{\frac{m}{m-d-1}}_{\gamma}$$

come from?

inverse chi-square dist.

as $m \to \infty$ we have $\gamma \to 1$

set $m = \alpha \cdot d$ $\gamma = \frac{\alpha d}{(\alpha-1)d-1} \approx \frac{\alpha}{\alpha-1}$

Randomized algorithm to estimate the inverse!

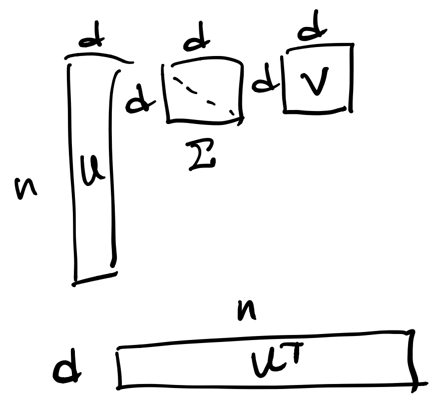
Which sketching matrices are good?

deterministic — slow & exact
random — fast & approx.

- We need to find conditions to guarantee approximate optimality
- Let $A = U\Sigma V^T$ SVD in compact form (drop zero singular vals)

A^TA : $O(nd^2)$
SA is faster

some deterministic options



- $S = U^T$ is $d \times n$ (optimal but slow since we need SVD)
- $\underline{S = A^T}$ (optimal but slow) $\leftarrow O(nd)$

$\min_x \| A^TAx - A^Tb \|_2$

$((A^TA)(A^TA))^{-1} A^TA \cdot A^Tb = (A^TA)^{-1}A^Tb$

$A = U\Sigma V^T$
$A^TA = V\Sigma \cancel{U^TU}\Sigma V^T$
$= \underline{V\Sigma^2V^T}$

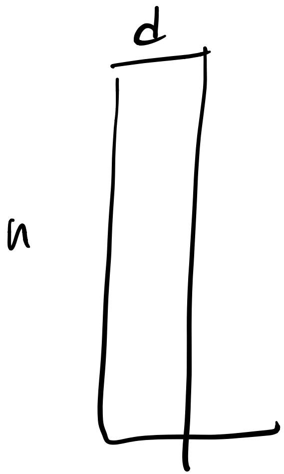
- For random S matrices A^TS^TSA needs to be invertible
  we want it to be close to A^TA

$x_{LS} = (A^TA)^{-1}A^Tb = A^+b$
$=$

$\min_x \|SAx - Sb\|_2^2$ pick $S = U^T$

$= \min_x \| \cancel{U^T \cdot U}\Sigma V^T - U^Tb \|_2^2 = (\Sigma V^T)^+ \cdot U^Tb = \underbrace{(V\Sigma^2V^T)^{-1}}_{A^TA} \underbrace{V\Sigma U^Tb}_{A^Tb}$

$= x_{LS}$ no error!



d

$A^T A : O(nd^2)$

$\approx A^T S^T \underbrace{SA}_{m \times d} : O(md^2)$

$m \approx d$

Questions?