EE270
Large scale matrix computation, optimization and learning

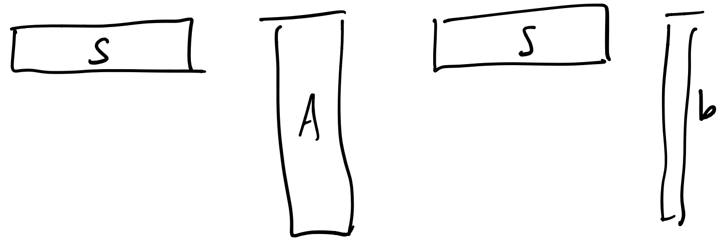
Instructor : Mert Pilanci

Stanford University

Tuesday, Feb 9 2021

Randomized Linear Algebra
Lecture 9: High-dimensional Problems, Least-norm Solutions and Randomized Methods

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

Gaussian Sketch

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

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

- Unbiased $\mathbb{E}\left[\tilde{x}\right] = x_{LS}$
  since $\tilde{x} = x_{LS} + \underbrace{(A^T S^T SA)^{-1} A^T S^T Sb^{\perp}}_{\text{zero mean}}$

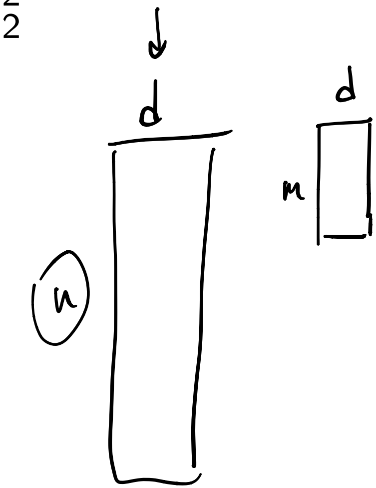
Gaussian Sketch

$f(x) = \|Ax - b\|_2^2$

$f(x) =$

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

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



- Unbiased $\mathbb{E}\left[\tilde{x}\right] = x_{LS}$

since $\tilde{x} = x_{LS} + \underbrace{(A^TS^TSA)^{-1}A^TS^TSb^{\perp}}_{\text{zero mean}}$

- Variance

$$\mathbb{E}\|A(\tilde{x} - x_{LS})\|_2^2 = f(x_{LS})\frac{d}{m-d-1}$$

valid for $m > d + 1$ where $f(x) = \|Ax - b\|_2^2$

Gaussian Sketch

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

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

- Unbiased $\mathbb{E}\left[\tilde{x}\right] = x_{LS}$
  since $\tilde{x} = x_{LS} + \underbrace{(A^T S^T SA)^{-1} A^T S^T S b^{\perp}}_{\text{zero mean}}$

- Variance

$$\mathbb{E}\|A(\tilde{x} - x_{LS})\|_2^2 = f(x_{LS})\frac{d}{m-d-1}$$

  valid for $m > d+1$ where $f(x) = \|Ax - b\|_2^2$
- Function value
  $f(\tilde{x}) = \|A\tilde{x} - b\|_2^2 = \|A(\tilde{x} - x_{LS})\|_2^2 + \|Ax_{LS} - b\|_2^2$
- $\mathbb{E}f(\tilde{x}) - f(x_{LS}) = f(x_{LS})\frac{d}{m-d-1}$

Variance Reduction by Averaging

- Let $S_1, \ldots, S_r$ be $\frac{1}{\sqrt{m}}\times$ i.i.d. Gaussian. $\mathbb{E}[S_i^T S_i] = I$

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

- let $\tilde{x} = \frac{1}{r}\sum_{i=1}^r x_i$
- Unbiased $\mathbb{E}\left[\tilde{x}\right] = x_{LS}$
- Variance is reduced by $\frac{1}{r}$
- $\mathbb{E}\|A(\tilde{x} - x_{LS})\|_2^2 = f(x_{LS})\frac{1}{r}\frac{d}{m-d-1}$

$$\mathbb{E}\tilde{x} = \frac{1}{r}\sum \underbrace{\mathbb{E}x_i}_{x_{LS}} = x_{LS}$$

$$\mathrm{Var}\left(\frac{1}{r}\sum_{i=1}^r x_i\right) = \frac{1}{r^2}\cdot\underline{\mathrm{Var}\left(\sum x_i\right)}$$

$$= \frac{1}{r^2}\cdot\sum_{i=1}^r \mathrm{Var}(x_i)$$

$$= \frac{r}{r^2}\cdot\mathrm{Var}(x_1)$$

Variance Reduction by Averaging

$b = Ax_{LS} + b^{\perp}$

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

$= \| A(x - x_{LS}) \|_2^2 + \underbrace{\|b^{\perp}\|_2^2}_{f(x_{LS})}$

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

- Let $S_1, \ldots, S_r$ be $\frac{1}{m} \times$ i.i.d. Gaussian. $\mathbb{E}[S^T S] = I$

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

- let $\tilde{x} = \frac{1}{r} \sum_{i=1}^{r} x_i$
- Unbiased $\mathbb{E}\left[\tilde{x}\right] = x_{LS}$
- Variance is reduced by $\frac{1}{r}$
- $\mathbb{E}\|A(\tilde{x} - x_{LS})\|_2^2 = f(x_{LS}) \frac{1}{r} \frac{d}{m-d-1}$
- $\mathbb{E} f(\tilde{x}) - f(x_{LS}) = f(x_{LS}) \frac{1}{r} \frac{d}{m-d-1}$

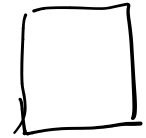
$$f(x_{LS}) \leq \mathbb{E} f(\tilde{x}) \leq f(x_{LS}) \cdot \left(1 + \frac{1}{r} \frac{d}{m-d-1}\right)$$

High-dimensional Least Squares Problems

Right sketching.

$A \cdot S$

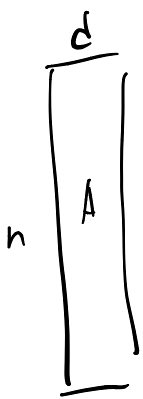
$n = d$



Square equation system

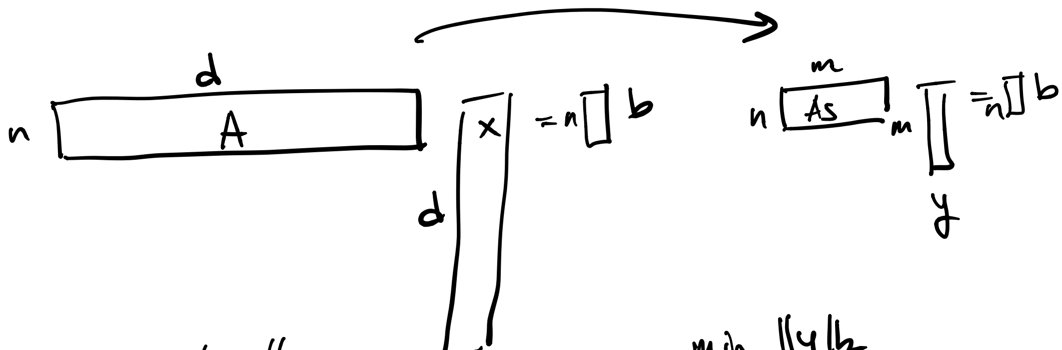
- $A \in \mathbb{R}^{n \times d}$ where $d > n$
- no unique solution

tall LS



$n >> d$

wide LS (high-dim)



$\min \|x\|_2$ s.t. $Ax = b$ $\longleftrightarrow$ $\min \|y\|_2$ $ASy = b$

High-dimensional Least Squares Problems

$A = U\Sigma V^T \quad A^+ = A^T(AA^T)^{-1}$

$$\min_x \ \tfrac{1}{2}\|x\|_2^2 + \lambda^T(Ax - b) = \cancel{\tfrac{1}{2}\|A^T\lambda\|_2^2} - \tfrac{1}{2}\lambda^T A A^T \lambda - \lambda^T b$$

$$x + A^T\lambda = 0 \Rightarrow x = -A^T\lambda = A^T(AA^T)^{-1} b = A^+ b$$

$$-AA^T\lambda - b = 0$$

$$\lambda = -(AA^T)^{-1} b$$

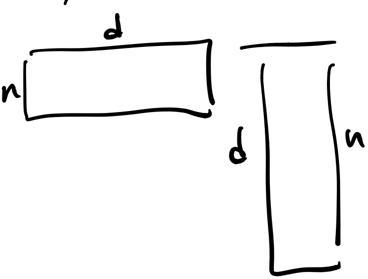
invertible iff A is full row rank

- $A \in \mathbb{R}^{n\times d}$ where $d > n$
- no unique solution
- minimum (ℓ_2) norm solution is unique

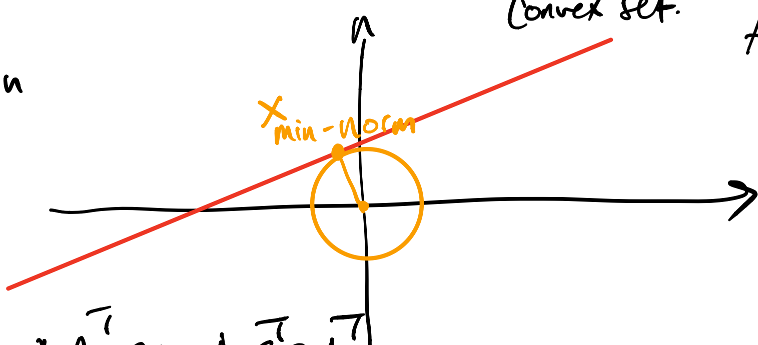
$$x_{\text{min-norm}} = \arg\min_{Ax=b} \tfrac{1}{2}\|x\|_2^2$$

Convex set.

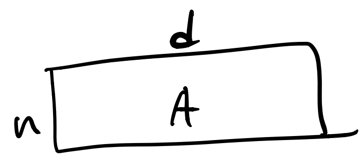
AA^T



$O(n^2 d)$ cost



$A_{11}x_1 + A_{12}x_2 = b$



Idea. $AA^T \approx A S^T S A^T$

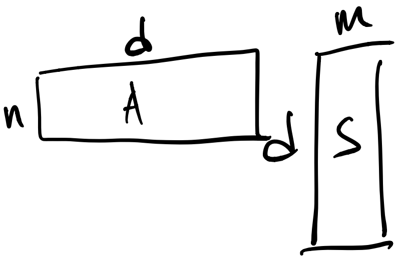
Minimum norm solution and SVD

$$x_{\text{min-norm}} = \arg\min_{Ax=b} \|x\|_2^2$$

Random projection to reduce dimension: Right Sketch

$$x_{\text{min-norm}} = \arg\min_{Ax=b} \|x\|_2^2$$

- We can right multiply A and form AS where $S \in \mathbb{R}^{d\times m}$ and solve



$$\arg\min_{ASz=b} \|z\|_2^2$$

$z = (AS)^{+} b$ $\quad$ m-dimensional

Idea 1) let $\hat{x} = S\,(AS)^{+} b$

Random projection to reduce dimension: Right Sketch

$$x_{\text{min-norm}} = \arg\min_{Ax=b} \|x\|_2^2$$

- We can right multiply A and form AS where $S \in \mathbb{R}^{d\times m}$ and solve

$$\arg\min_{ASz=b} \|z\|_2^2$$

- How do we use $z \in \mathbb{R}^m$?

Right Sketch

$$x_{\text{min-norm}} = \arg\min_{Ax=b} \underbrace{\|x\|_2^2}_{f(x)}$$

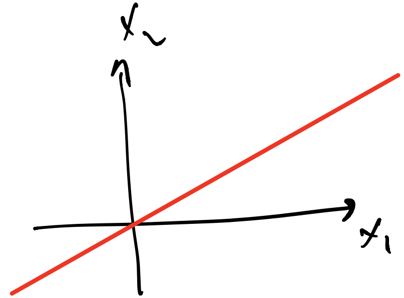
approximation $\tilde{x} = S\tilde{z}$

where $\tilde{z} := \arg\min_{ASz=b} \|z\|_2^2$

Feasible estimate for x: $A\tilde{x} = AS\tilde{z}$

$= b$

Right Sketch



$$\begin{bmatrix}\tilde{x}_1\\ \tilde{x}_2\end{bmatrix} = Sz = \begin{bmatrix}s_1\\ s_2\end{bmatrix} z$$

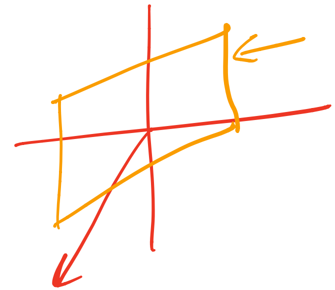
$$x_{\text{min-norm}} = \arg\min_{Ax=b} \underbrace{\|x\|_2^2}_{f(x)}$$

approximation $\quad \tilde{x} = S\tilde{z}$

where $\tilde{z} := \arg\min_{ASz=b} \|z\|_2^2$

- Let S be i.i.d. Gaussian $N(0, \frac{1}{\sqrt{m}})$
- Is $\tilde{x}$ unbiased, i.e., $\mathbb{E}\tilde{x} \overset{?}{=} x_{\text{min-norm}}$

Right Sketch



$$x_{\text{min-norm}} = \arg\min_{Ax=b} \underbrace{\|x\|_2^2}_{f(x)}$$

$$\text{approximation} \quad \tilde{x} = S\tilde{z}$$

$$\text{where } \tilde{z} := \arg\min_{ASz=b} \|z\|_2^2$$

- Let S be i.i.d. Gaussian $N(0, \frac{1}{\sqrt{m}})$
- Is $\tilde{x}$ unbiased, i.e., $\mathbb{E}\tilde{x} \overset{?}{=} x_{\text{min-norm}}$
- Yes, conditioned on SA

$$\tilde{x} \sim N(x_{\text{min-norm}}, V_2 V_2^T b^T (AS^T SA^T)^{-1} b)$$

- $V_2 V_2^T$ is the projection onto the null space of A
- error $\tilde{x} - x_{\text{min-norm}} \in \text{Null}(A)$

$Ax = b$

$A\tilde{x} = b$

$A(x - \tilde{x}) = 0$

$x - \tilde{x} \in \text{Null}(A)$

$\boxed{A}$

$A = U \Sigma V^T$

$= U [\Sigma \;\; 0] [V_1 \;\; V_2]^T$

Right Sketch

$$x_{\text{min-norm}} = \arg\min_{Ax=b} \underbrace{\|x\|_2^2}_{f(x)}$$

$A^+ b = A^T(AA^T)^{-1} b$

$$\text{approximation} \qquad \tilde{x} = S\tilde{z}$$

$$\text{where } \tilde{z} := \arg\min_{ASz=b} \|z\|_2^2$$

$(AS)^+ b = S^T A^T \underbrace{(ASS^TA^T)}_{n\times n \text{ matrix}}{}^{-1} b$

- Let S be i.i.d. Gaussian $N(0, \frac{1}{\sqrt{m}})$
- Is $\tilde{x}$ unbiased, i.e., $\mathbb{E}\tilde{x} \overset{?}{=} x_{\text{min-norm}}$
- Yes, conditioned on SA

$$\tilde{x} \sim N(x_{\text{min-norm}}, VV^T b^T (AS^TSA^T)^{-1} b)$$

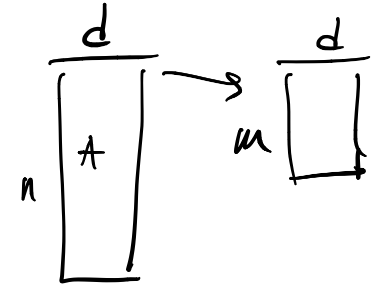
We need at least $m \geq n$

- VV^T is the projection onto the null space of A
- error $\tilde{x} - x_{\text{min-norm}} \in \text{Null}(A)$
- Using $\mathbb{E}(AS^TSA^T)^{-1} = (AA^T)^{-1}\frac{m}{m-n-1}$

$\boxed{m > n+1}$

$$\mathbb{E}\|\tilde{x} - x_{\text{min-norm}}\|_2^2 = \frac{d-n}{m-n-1} f(x_{\text{min-norm}}) = \frac{d-n}{m-n-1}\|x_{\text{min-norm}}\|_2^2$$

Left Sketch vs Right Sketch Summary



- Both are unbiased using Gaussian projections
- A is $n \times d$
- Left sketch $n \geq d$

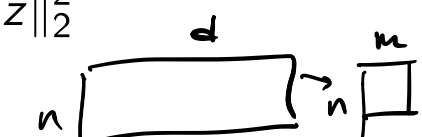
$$\tilde{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_2^2$$

Variance: $\mathbb{E}\|A(\tilde{x} - x_{LS})\|_2^2 = f(x_{LS})\frac{d}{m-d-1}$

dimension of the relevant subspace

- Right sketch $d > n$

$$\tilde{x} = S\tilde{z} \quad \text{where } \tilde{z} := \arg\min_{\underline{ASz=b}} \|z\|_2^2$$



Variance: $\mathbb{E}\|\tilde{x} - x_{\text{min-norm}}\|_2^2 = f(x_{\text{min-norm}})\frac{d-n}{m-n-1}$

dimension of the relevant subspace

Back to Left Sketch: Which sketching matrices are good?

- We need to find conditions to guarantee approximate optimality
- Let $A = U\Sigma V^T$ SVD in compact form

some deterministic options

- $S = U^T$ is $d \times n$
- $S = A^T$

high dim

V_1^T

$S = A$

- For random S matrices $A^T S^T S A$ needs to be invertible

we want it to be close to $A^T A$

Approximate Matrix Multiplication

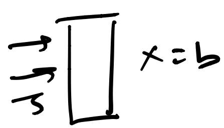
- Let the approximate product of AB be $C = AS^TSB$

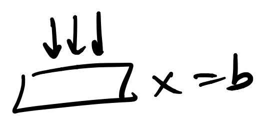
$$\mathbb{P}\left[\|AB - C\|_F > \epsilon \|A\|_F \|B\|_F\right] \le \delta$$

- Follows from JL Moment property
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$random i.i.d. sub-Gaussian, e.g., ± 1, or $N(0,1)$ with $m = \frac{c_1}{\epsilon^2} \log \frac{1}{\delta}$
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$CountSketch matrix (one nonzero per column, which is ± 1 at a uniformly random location) with $m = \frac{c_2}{\epsilon^2 \delta}$
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$Fast JL Transform with $m = \frac{c_3}{\epsilon} \log \frac{1}{\delta}$

Fourier Transform

→ row/colum sampling





Approximate Matrix Multiplication

- Let the approximate product of AB be $C = AS^TSB$

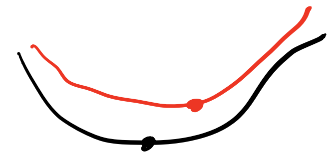
$$\mathbb{P}\left[\|AB - C\|_F > \epsilon\|A\|_F\|B\|_F\right] \leq \delta$$

- Follows from JL Moment property
- $S \in \mathbb{R}^{m\times n} \sim \frac{1}{\sqrt{m}}\times$random i.i.d. sub-Gaussian, e.g., ± 1, or $N(0,1)$ with $m = \frac{c_1}{\epsilon^2}\log\frac{1}{\delta}$
- $S \in \mathbb{R}^{m\times n} \sim \frac{1}{\sqrt{m}}\times$CountSketch matrix (one nonzero per column, which is ± 1 at a uniformly random location) with $m = \frac{c_2}{\epsilon^2\delta}$
- $S \in \mathbb{R}^{m\times n} \sim \frac{1}{\sqrt{m}}\times$Fast JL Transform with $m = \frac{c_3}{\epsilon}\log\frac{1}{\delta}$
- Sparse JL and Fast JL are more efficient
- advantages: doesn't require any knowledge about matrices A and B (**oblivious**)
- optimal sampling probabilities depend on the column/row norms of A and B

Basic Inequality Method

used in statistics
ML.

min $f(x)$ min $\tilde{f}(x)$



second ineq.

first ineq.

- We minimize $\tilde{x} = \arg\min \|S(Ax - b)\|_2^2$
- x_{LS} minimizes $\|Ax - b\|_2^2$
- How far is $\tilde{x}$ from x_{LS}?
- **Step 1**. Establish two optimality (in)equalities for these variables
- $\|Ax_{LS} - b\|_2^2 \leq \|Ax' - b)\|_2^2$ for any x', i.e., $A^T(Ax_{LS} - b) = 0$
- $\|S(A\tilde{x} - b)\|_2^2 \leq \|S(Ax_{LS} - b)\|_2^2$

Basic Inequality Method

$\underbrace{\| S A(\Delta + x_{LS}) - Sb \|_2^2} \le \| S A x_{LS} - S b^T \|_2^2$

$b = A x_{LS} + b^\perp$

$\underbrace{\| SA\Delta + \cancel{SAx_{LS}} - \cancel{SAx_{LS}} - Sb^\perp \|_2^2} \le \| \cancel{SAx_{LS}} - SAx_{LS} - Sb^\perp \|_2^2$

$= \| Sb^\perp \|_2^2$

- We minimize $\tilde{x} = \arg\min \|S(Ax - b)\|_2^2$
- x_{LS} minimizes $\|Ax - b\|_2^2$
- How far is $\tilde{x}$ from x_{LS}?

$\| SA\Delta \|_2^2 + \cancel{\| Sb^\perp \|_2^2} - 2 b^{\perp T} S^T S A\Delta \le \cancel{\| Sb^\perp \|_2^2}$

- **Step 1**. Establish two optimality (in)equalities for these variables
- $\|Ax_{LS} - b\|_2^2 \le \|Ax' - b)\|_2^2$ for any x', i.e., $A^T(Ax_{LS} - b) = 0$
- $\|S(A\tilde{x} - b)\|_2^2 \le \|S(Ax_{LS} - b)\|_2^2$
- **Step 2**. Define error $\Delta = \tilde{x} - x_{LS}$ and re-write these inequalities in terms of Δ

$\tilde{x} = \Delta + x_{LS}$

- $\|SA\Delta\|_2^2 \le 2b^{\perp T}(S^T S - I)A\Delta$
- **Step 3**. Argue $S^T S \approx I$

$\Delta^T A^T \underbrace{S^T S}_{\approx I} A\Delta = \| SA\Delta \|_2^2 \le 2 b^{\perp} S^T S A \Delta = 2 b^{\perp} (\underbrace{S^T S - I}_{\approx I}) A \Delta$ since $b^{\perp T} \cdot A = 0$

Leverage Scores

Questions?