EE270
Large scale matrix computation, optimization and learning

Instructor : Mert Pilanci

Stanford University

Tuesday, Feb 2 2020

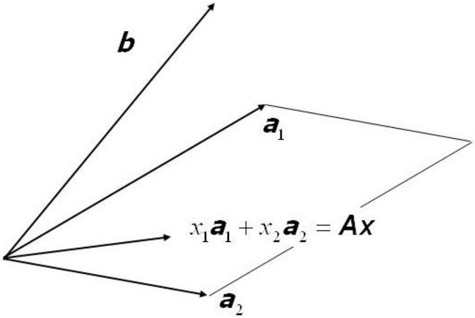
Randomized Linear Algebra
Lecture 8: Randomized Least Squares Bias and Variance, Streaming Data

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

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

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



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

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

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

- **Lemma** If $m \geq \text{constant} \times \frac{rank(A)}{\epsilon^2}$ 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

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 $A_t^T A_t \in \mathbb{R}^{d \times d}$?

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 $\mathbb{E}\left[\tilde{x}\right]$ equal to x_{LS}?
- 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 $b^{\perp} \perp Range(A)$
$$\begin{aligned}\tilde{x} &= (A^T S^T SA)^{-1} A^T S^T S(Ax_{LS} + b^{\perp})\\ &= x_{LS} + (A^T S^T SA)^{-1} A^T S^T Sb^{\perp}\end{aligned}$$
- $\mathbb{E}(A^T S^T SA)^{-1} A^T S^T Sb^{\perp} = 0$ since $Sb^{\perp}$ and SA are uncorrelated zero mean Gaussian.

Gaussian Sketch: Variance

- Let S be i.i.d. Gaussian

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

- 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

- Let S be i.i.d. Gaussian

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Expected Inverse of a Random Matrix

- Where does the formula

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

come from?

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

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

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

Questions?