EE270
Large scale matrix computation, optimization and learning

Instructor : Mert Pilanci

Stanford University

Tuesday, Feb 2 2020

Randomized Linear Algebra
Lecture 7: Least Squares Optimization and Random Projections

Recap: Johnson Lindenstrauss Lemma

- Let $\epsilon \in (0, \frac{1}{2})$. Given any set of points $\{x_1, ..., x_n\}$ in $\mathbb{R}^d$, there exists a map $S : \mathbb{R}^n \to \mathbb{R}^m$ with $m = \frac{9\log(n)}{\epsilon^2 - \epsilon^3}$ such that

$$1 - \epsilon \le \frac{\|Sx_i - Sx_j\|_2^2}{\|x_i - x_j\|_2^2} \le 1 + \epsilon$$

Recap: Johnson Lindenstrauss Lemma

- Let $\epsilon \in (0, \frac{1}{2})$. Given any set of points $\{x_1, ..., x_n\}$ in $\mathbb{R}^d$, there exists a map $S : \mathbb{R}^n \to \mathbb{R}^m$ with $m = \frac{9\log(n)}{\epsilon^2 - \epsilon^3}$ such that

$$1 - \epsilon \leq \frac{\|Sx_i - Sx_j\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \epsilon$$

- Note that the target dimension m is **independent of the original dimension** d, and depends **only on the number of points** n and the accuracy parameter.

Recap: Johnson Lindenstrauss Lemma

- Let $\epsilon \in (0, \frac{1}{2})$. Given any set of points $\{x_1, ..., x_n\}$ in $\mathbb{R}^d$, there exists a map $S : \mathbb{R}^n \to \mathbb{R}^m$ with $m = \frac{9\log(n)}{\epsilon^2 - \epsilon^3}$ such that

$$1 - \epsilon \leq \frac{\|Sx_i - Sx_j\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \epsilon$$

- Note that the target dimension m is **independent of the original dimension** d, and depends **only on the number of points** n and the accuracy parameter.
- more surprises: picking an $m \times d$ random matrix $S = \frac{1}{\sqrt{m}} G$ with $G_{ij} \sim N(0, 1)$ standard normal works with high probability!

True ‘projections’: random subspaces also work

- Pick $S_{(i)}$ uniformly random on the unit sphere
- Pick $S_{(i+1)}$ uniformly random on the unit sphere and $\perp S_{(i)}, ...S_{(1)}$
- S is a projection matrix, which projects onto a uniformly random subspace

$$\mathbb{P}\left\{\left|\|Su\|_2 - \sqrt{\frac{m}{d}}\right| > t\right\} \leq 2e^{\frac{-t^2 d}{2}}$$

- Applying union bound for all points $i, j = 1, ..., d$ gives a similar result
- Random i.i.d. S matrices are easier to generate and approximately orthogonal: $\mathbb{E}S^T S = I$

Computationally cheaper random matrices

- Gaussian $S_{ij} = \frac{1}{\sqrt{m}} N(0, 1)$
- Rademacher

$$S_{ij} = \begin{cases} +\frac{1}{m} & \text{with probability } \frac{1}{\sqrt{m}} \\ -\frac{1}{\sqrt{m}} & \text{with probability } \frac{1}{2} \end{cases} \tag{1}$$

- Bernoulli-Rademacher

$$S_{ij} = \begin{cases} +\frac{\sqrt{3}}{\sqrt{m}} & \text{with probability } \frac{1}{2} \\ 0 & \text{with probability } \frac{2}{3} \\ -\frac{\sqrt{3}}{\sqrt{m}} & \text{with probability } \frac{1}{2} \end{cases} \tag{2}$$

- other sparse matrices (e.g. one non-zero per column)
- Fourier transform based matrices

Random projection for Approximate Matrix Multiplication

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

$$\mathbb{P}\left[\|AB - C\|_F > 3\epsilon\|A\|_F\|B\|_F\right] \leq \delta$$

- 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 $m = \frac{c_3}{\epsilon} \log \frac{1}{\delta}$

Random projection for Approximate Matrix Multiplication

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

$$\mathbb{P}\left[\|AB - C\|_F > 3\epsilon \|A\|_F \|B\|_F\right] \leq \delta$$

- 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

Least Squares Regression

- Predict the value of a continuous target variable y

  $(a_1, b_1), \ldots, (a_n, b_n)$

  $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$

$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix}$ $b = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$

- Linear regression $f(a) = x^T a + x_0$

x b

Least Squares Regression

$\mathbb{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

- Predict the value of a continuous target variable y

  $(a_1, b_1), \ldots, (a_n, b_n)$

  $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$

  $\min_x \| Ax - b + x_0 \mathbb{1} \|_2^2$
- Linear regression $f(a) = x^T a + x_0$
- Performance measure: minimum sum of squares

$$\min_{x, x_0} \frac{1}{n} \sum_{i=1}^{n} (b_i - x^T a_i - x_0)^2$$

Least Squares Regression

- Predict the value of a continuous target variable y
  $(a_1, b_1), ..., (a_n, b_n)$
  $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$
- Linear regression $f(a) = x^T a + x_0$
- Performance measure: minimum sum of squares

$$\min_{x,x_0} \frac{1}{n} \sum_{i=1}^{n} (b_i - x^T a_i - x_0)^2$$

$\|x\|_1$

- we can add a regularization term $\lambda||x||_2^2$

$$\min_{x,x_0} \frac{1}{n} \sum_{i=1}^{n} (b_i - x^T a_i - x_0)^2 + \underbrace{\lambda||x||_2^2}_{\text{Ridge or Tikhonov}}$$

Least Squares Regression

- Loss function:
  $L(x, x_0) = \frac{1}{n}\sum_{i=1}^{n}(b_i - x^T a_i - x_0)^2 + \lambda ||x||_2^2$
- $\frac{\partial}{\partial x_0} L(x, x_0) = 0$
  optimal $x_0^* = \frac{1}{n}\sum_{i=1}^{n}(y_i - x^T a_i) = \bar{b} - x^T \bar{a}$
  where $\bar{a} = \sum_{i=1}^{n} a_i$ and $\bar{b} = \sum_{i=1}^{n} b_i$
- plugging x_0^* in $L(x, x_0)$
  $L(x, x_0^*) = \frac{1}{n}\sum_{i=1}^{n}(b_i - \bar{b} - x^T(a_i - \bar{a}))^2 + \lambda ||x||_2^2$

Least Squares Regression

- Loss function:
  $L(x, x_0) = \frac{1}{n}\sum_{i=1}^{n}(b_i - x^T a_i - x_0)^2 + \lambda||x||_2^2$
- $\frac{\partial}{\partial x_0} L(x, x_0) =$
  optimal $x_0^* = \frac{1}{n}\sum_{i=1}^{n}(y_i - x^T a_i) = \bar{b} - x^T \bar{a}$
  where $\bar{a} = \sum_{i=1}^{n} a_i$ and $\bar{b} = \sum_{i=1}^{n} b_i$
- plugging x_0^* in $L(x, x_0)$
  $L(x, x_0^*) = \frac{1}{n}\sum_{i=1}^{n}(b_i - \bar{b} - x^T(a_i - \bar{a}))^2 + \lambda||x||_2^2$
  define centered data $\tilde{a}_i = a_i - \bar{a}$ and $\tilde{b}_i = b_i - \bar{b}$

$$\min_x \; ||\tilde{A}x - \tilde{b}||_2^2 + n\lambda||x||_2^2$$

Least Squares Regression

- Loss function:
  $L(x, x_0) = \frac{1}{n}\sum_{i=1}^{n}(b_i - x^T a_i - x_0)^2 + \lambda||x||_2^2$
- $\frac{\partial}{\partial x_0} L(x, x_0) =$
  optimal $x_0^* = \frac{1}{n}\sum_{i=1}^{n}(y_i - x^T a_i) = \bar{b} - x^T \bar{a}$
  where $\bar{a} = \sum_{i=1}^{n} a_i$ and $\bar{b} = \sum_{i=1}^{n} b_i$
- plugging x_0^* in $L(x, x_0)$
  $L(x, x_0^*) = \frac{1}{n}\sum_{i=1}^{n}(b_i - \bar{b} - x^T(a_i - \bar{a}))^2 + \lambda||x||_2^2$
  define centered data $\tilde{a}_i = a_i - \bar{a}$ and $\tilde{b}_i = b_i - \bar{b}$

$$\min_x \; ||\tilde{A}x - \tilde{b}||_2^2 + n\lambda||x||_2^2$$

$\frac{\partial}{\partial x} L(x, x_0^*) = 2\tilde{A}^T(\tilde{A}x^* - \tilde{b}) + 2n\lambda x^* = 0$
optimal solution $\boxed{x^* = (\tilde{A}^T\tilde{A} + n\lambda I)^{-1}\tilde{A}^T\tilde{b}}$

optional

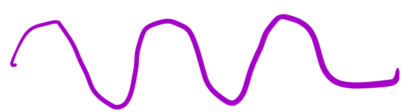
Autoregressive Models

$$b[n] = a[n+1] \approx \sum_k x_k a[n-k]$$

- AR(2) model : two non-zero filter coefficients

$$a[n+1] = -x_0 a[n] - x_1 a[n-1]$$

  and error term $e_n = 0$
- Example: Sine wave $a[n] = \sin(\alpha n)$ satisfies AR(2) model



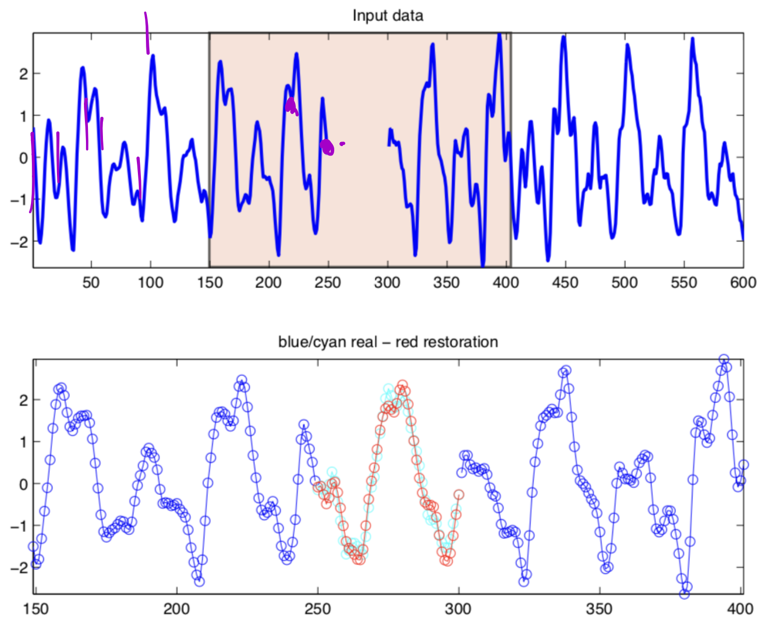
Autoregressive models

- We can predict future values using

$$b[n] = \underbrace{\sum_k a[n-k]x_k}$$

1000

$\min \|Ax-b\|_2$



d

n

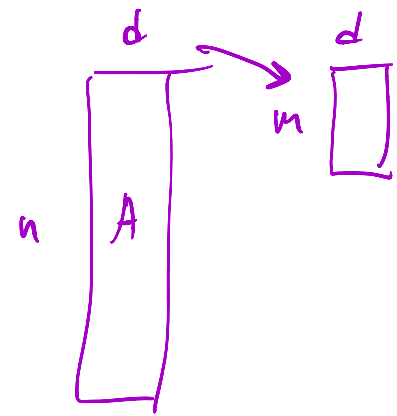
tall.

slide credit: P. Smaragdis

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

$(A^T A)^{-1}$ $O(d^3 \text{ time})$

$\underset{d \times n}{A^T} \underset{n \times d}{A}$: $O(nd^2)$ ← larger n $nd^2 > d^3$ $n > d$

Geometry

$$\begin{bmatrix} a_1^T a_1 & a_1^T a_2 \\ & \end{bmatrix}$$

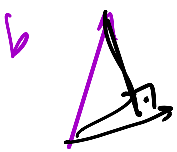
$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2$$

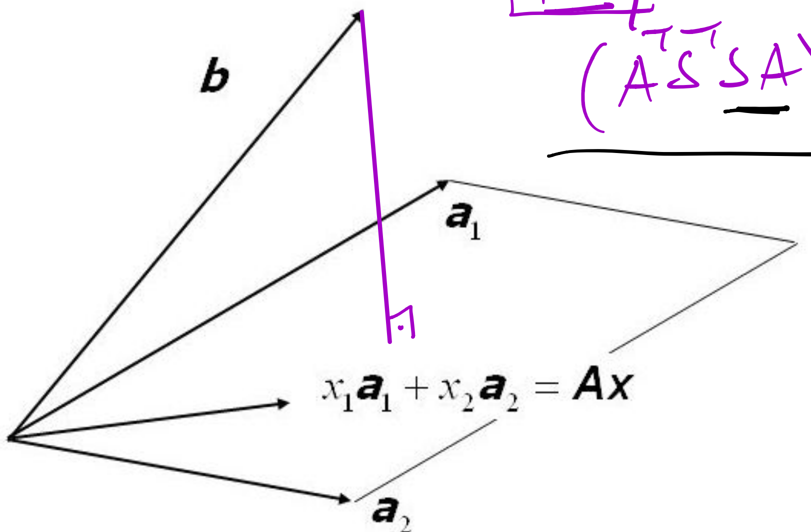
Normal eq.

$$A^T A x = A^T b$$

$\boxed{\text{Idea}}$: Randomized LS. solver

$$(A^T S^T S A) x = A^T S^T S b$$





Singular Value Decomposition

$Ax = b$

$x \neq A^{-1} b$

- Every $A \in \mathbb{R}^{n \times n}$ has a singular value decomposition

$$A = U \Sigma V^T$$

where $U \in \ltimes \times \diagdown$ has orthonormal columns

Σ is diagonal with non-increasing non negative entries

V^T has orthonormal rows

- Pseudoinverse $A^\dagger = V \Sigma^{-1} U^T$
- Least Square solution
$x_{LS} = \underbrace{(A^T A)^{-1} A^T}_{A^\dagger} b = A^\dagger b = V \Sigma^{-1} U^T b$

Classical Methods for Least Squares

- **Direct methods** exact
- not efficient → Cholesky decomposition: Form A^TA and decompose $A^TA = R^TR$ where R is upper triangular. Solve normal equations $(A^TA)^{-1} = (R^TR)^{-1}A^Tb$
- QR decomposition: $A = QR$, solve $Rx = Q^Tb$ $QRx = b$ $Rx = Q^Tb$
- Singular Value Decomposition: $x_{LS} = V\Sigma^{-1}U^Tb$

Direct methods have typically $O(nd^2)$ complexity

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

- **Indirect methods**
- Gradient descent with momentum $\boxed{\text{(Chebyshev iteration) we'll see those}}$
- Conjugate Gradient
- Other iterative methods

Indirect methods have typically $O(\sqrt{\kappa}nd)$ complexity, where κ is the condition number Ax: $O(nd)$ $\kappa = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$

Faster Least Squares Optimization: Random Projection

$\mathbb{E}[(A^T S^T S A)x] = \mathbb{E}[A^T S^T S b]$ approx. normal eq.

$A^T A x = A^T b$ if S - JL

- **Left-sketching**

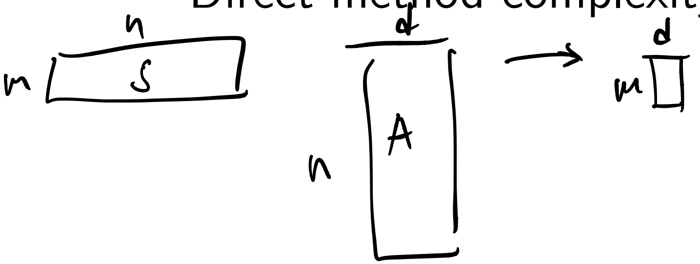
  Form SA and Sb where $S \in \mathbb{R}^{m \times n}$ is a random projection matrix

- Solve the smaller problem

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

- using any classical method.

  Direct method complexity md^2



Faster Least Squares Optimization: Random Projection

Right-sketching

$$\min_{y \in \mathbb{R}^{m}} \|ASy - b\|_2^2$$

- **Left-sketching**
  Form SA and Sb where $S \in \mathbb{R}^{m \times n}$ is a random projection matrix
- Solve the smaller problem

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

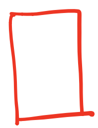
- using any classical method.
  Direct method complexity md^2

Approximation Result

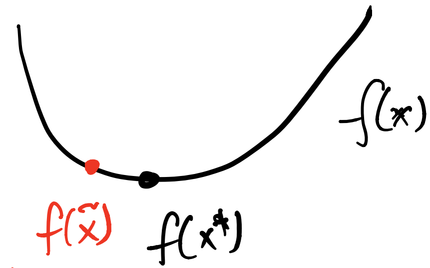
i.i.d subGaussian
Count Sketch
Fast JL

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

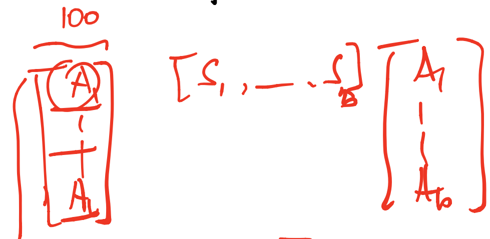
const. d



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



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



- If $m \geq \overbrace{\text{constant}}^{1} \times \overbrace{\frac{rank(A)}{\epsilon^2}}$ 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

$1 + \epsilon^2$ approximation

$\sum S_i \cdot A_i$

$m = \frac{1}{0.01} \cdot rank(A) \Rightarrow 1.01$

Gaussian Sketch

- Let S be $\frac{1}{\sqrt{m}}\times$ 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! $\mathbb{E}\,\tilde{x} = x_{LS}$

Questions?