CIS 700: “algorithms for Big Data”

Lecture 7: Sketching for Linear Algebra

Slides at http://grigory.us/big-data-class.html

Grigory Yaroslavtsev

http://grigory.us

Penn
University of Pennsylvania

Least Squares Regression

- Solving an overconstrained linear system
- For $d \ll n$ given:
  - matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$
  - vector $\boldsymbol{b} \in \mathbb{R}^{n}$
- Find $\mathbf{x}^* \in \mathbb{R}^d$ that minimizes: $||\boldsymbol{Ax} - \boldsymbol{b}||_2$
- Normal equation: $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}^T\boldsymbol{b}$
- If $\boldsymbol{A}$ has rank d then $\boldsymbol{x}^* = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$
- Takes $O(nd^2)$ time to compute (using naïve matrix multiplication)

Sketching for Least Squares Regression

- Use JL matrix $\boldsymbol{S} \in \mathbb{R}^{r \times n}$ where $r = \Theta\left(\frac{d}{\epsilon^2}\right) \ll n$
- Solve $\min_x \left|\left|\boldsymbol{SAx} - \boldsymbol{Sb}\right|\right|_2$ instead
- Standard JL: time $O(nrd + rd^2) > O(nd^2)$
- Sparse JL: time $O(nd^2/\epsilon + rd^2)$
- Fast JL: time $O(nd \log n + rd^2)$
- Subspace embeddings from JL:
  - JL only gives a guarantee for a fixed vector
  - We need the guarantee for the column space of A

Oblivious Subspace Embeddings

- Subspace embedding for A:

$$||SAx||_2^2 = (1 \pm \epsilon)||Ax||_2^2$$

- SE for $A \equiv$ SE for U where U is the orthonormal basis for the column space of A
- Least Squares Regression: use SE for (A,b)

$$\min_x ||\boldsymbol{Ax} - \boldsymbol{b}||_2 \rightarrow \min_x ||\boldsymbol{SAx} - \boldsymbol{Sb}||_2 = \min_x ||\boldsymbol{S}(\boldsymbol{Ax} - \boldsymbol{b})||_2$$

- Oblivious Subspace Embedding (OSE): matrix S

chosen independently of A, works for any fixed A

- JL transforms can be used as oblivious subspace embeddings

JLT(ϵ, δ, f)

- JLT(ϵ, δ, f): $S \in \mathbb{R}^{k \times n}$ that for any f-element subset $V \subseteq \mathbb{R}^n$ for all $v, v' \in V$ satisfies that:

$$|\langle Sv, Sv' \rangle - \langle v, v' \rangle| \leq \epsilon \left|\left|v\right|\right|_2 \left|\left|v'\right|\right|_2$$

- For unit vectors v, v':

$$|\langle Sv, Sv' \rangle - \langle v, v' \rangle| \leq \epsilon$$

- $\langle Sv, Sv' \rangle =$

$$\frac{1}{2}\left(\left|\left|S(v + v')\right|\right|_2^2 - \left|\left|Sv\right|\right|_2^2 - S\left|\left|v'\right|\right|_2^2\right)$$

$$= \frac{1}{2}\left((1 \pm \epsilon)\left|\left|v + v'\right|\right|_2^2 - (1 \pm \epsilon)\left|\left|v\right|\right|_2^2 - (1 \pm \epsilon)\left|\left|v'\right|\right|_2^2\right)$$

$$= \langle v, v' \rangle \pm O(\epsilon)$$

- Suffices to take regular JL of dimension $d = \Omega(1/\epsilon^2 \log f/\delta)$

OSE construction

- $S = \{y \in \mathbb{R}^n \mid \exists x\colon\ y = Ax, \left||y|\right|_2 = 1\}$
- ϵ-net argument: find a set $N \subseteq S$ such that if

$$\langle \boldsymbol{Sw}, \boldsymbol{Sw}' \rangle = \langle \boldsymbol{w}, \boldsymbol{w}' \rangle \pm \epsilon \qquad \forall \boldsymbol{w}, \boldsymbol{w}' \in N$$

then $\left||\boldsymbol{Sy}|\right|_2^2 = (1 \pm \epsilon)\left||\boldsymbol{y}|\right|_2^2 \ \forall \boldsymbol{y} \in S$

- $N = 1/2$-net:

$$\forall y \in S\ \exists\ w \in N\colon \left||y - w|\right|_2 \leq \frac{1}{2}$$

- $\boldsymbol{y} = \boldsymbol{y}^0 + \boldsymbol{y}^1 + \boldsymbol{y}^2 + \cdots$, where $\left||\boldsymbol{y}^i|\right| \leq \frac{1}{2^i}$ and each $\boldsymbol{y}^i$

is a multiple of a vector in N.

Net argument

- $\boldsymbol{y} = \boldsymbol{y}^0 + \boldsymbol{y}^1 + \boldsymbol{y}^2 + \cdots$, where $\left|\left|\boldsymbol{y}^i\right|\right| \leq \frac{1}{2^i}$ and each $\boldsymbol{y}^i$ is a multiple of a vector in N.
- $\boldsymbol{y} = \boldsymbol{y}^{\mathbf{0}} + (\boldsymbol{y} - \boldsymbol{y}^{\mathbf{0}})$ where $\boldsymbol{y}_{\mathbf{0}} \in \boldsymbol{N}$, $||\boldsymbol{y} - \boldsymbol{y}^{\mathbf{0}}||_{\mathbf{2}} \leq \frac{1}{2}$
- $(\boldsymbol{y} - \boldsymbol{y}^{\mathbf{0}}) = \boldsymbol{y}^{\mathbf{1}} + \left((\boldsymbol{y} - \boldsymbol{y}^{\mathbf{0}}) - \boldsymbol{y}^{\mathbf{1}}\right)$ where $\boldsymbol{y}^{\mathbf{1}} \in N$ and $\left|\left|\left((\boldsymbol{y} - \boldsymbol{y}^{\mathbf{0}}) - \boldsymbol{y}^{\mathbf{1}}\right)\right|\right|_{\mathbf{2}} \leq \frac{\left|\left|\boldsymbol{y} - \boldsymbol{y}^{\mathbf{0}}\right|\right|}{2} \leq 1/4$
- $\left|\left|\boldsymbol{S}\boldsymbol{y}\right|\right|_2^2 = \left|\left|\boldsymbol{S}(\boldsymbol{y}^0 + \boldsymbol{y}^1 + \boldsymbol{y}^2 + \cdots)\right|\right|_2^2$

$$= \sum_{0 \leq i < j < \infty} \left|\left|\boldsymbol{S}\boldsymbol{y}^i\right|\right|_2^2 + 2\langle \boldsymbol{S}\boldsymbol{y}^{\boldsymbol{i}}, \boldsymbol{S}\boldsymbol{y}^{\boldsymbol{j}} \rangle$$

$$\leq \left(\sum_{0 \leq i < j < \infty} \left|\left|\boldsymbol{y}^i\right|\right|_2^2 + 2\langle \boldsymbol{y}^{\boldsymbol{i}}, \boldsymbol{y}^{\boldsymbol{j}} \rangle \right) \pm 2\epsilon\left(\sum_{0 \leq i \leq j < \infty} \left|\left|\boldsymbol{y}^i\right|\right|_2 \left|\left|\boldsymbol{y}^j\right|\right|_2 \right)$$

$$= 1 \pm O(\epsilon)$$

½ -Net construction

- For $0 < \gamma < 1$ there is a γ-net for S of size $\leq \left(1 + \frac{2}{\gamma}\right)^d$
- Choose a maximal set N' of points on S^d such that no two points are within γ of each other
- Balls of radius $\frac{\gamma}{2}$ around the points are disjoint
- Ball of radius $1 + \frac{\gamma}{2}$ around the origin contains all balls
- # points $\leq \left(\frac{1+\frac{\gamma}{2}}{\frac{\gamma}{2}}\right)^d = \left(1 + \frac{2}{\gamma}\right)^d$
- Size of ½-net $\leq 5^d$
- JLT of dimension $\Omega((d + \log\frac{1}{\delta})/\epsilon^2)$ gives OSE

OSE constructions Running Times

nnz(A) = # non-zero entries in A

- OSE from Sparse JL: time $O(nnz(A)d/\epsilon)$
- Fast JL: time $O(nd\log n)$
- [Clarkson, Woodruff'13] possible to construct OSE in time $O(nnz(A))$

Leverage Score Sampling

- **Def (Leverage Score):** For an $n \times k$ matrix Z with orthonormal columns let the leverage score $p_i = \frac{\ell_i^2}{k}$ where $\ell_i^2 = \left|\left|e_i^T Z\right|\right|_2^2 = ||Z_i||_2^2$
- Note: leverage scores form a distribution
- If A doesn't have orthonormal columns we can still pick an orthonormal basis Z for it
- Choice of Z doesn't matter ($Z' = ZR$) where R is orthonormal gives same leverage scores
- All ℓ_i^2 are at most 1

Leverage Score Sampling

- Given: $\beta > 0$ distribution $(q_1, \ldots, q_n)$ with $q_i \geq \beta p_i$
- **Leverage Score Sampling** (Z, s, q):
  - Constructs matrices $\Omega \in \mathbb{R}^{n \times s}$ and $D \in \mathbb{R}^{s \times s}$
  - For each column indep. with replacement pick row i w.p. q_i
  - Set $\Omega_{i,j} = 1$ and $D_{jj} = 1/\sqrt{q_i s}$

LSS as a Subspace Embedding

- **Thm.:** If $Z \in \mathbb{R}^{n \times k}$ has orthonormal columns then for $s > 144k \log\left(\frac{2k}{\delta}\right) / \beta\epsilon^2$ if Ω and D are constructed via LSS(Z, s, q) then for all i w.p. $1 - \delta$:

$$1 - \epsilon \leq \sigma_i^2(D^T \Omega^T Z) \leq 1 + \epsilon$$

- (**Matrix Chernoff**): If $X_1, \ldots X_s$ are i.i.d copies of a symmetric random matrix $X \in \mathbb{R}^{k \times k}$ with $E[X] = 0, \left||X|\right|_2 \leq \gamma$ and $\left||E[X^T X]|\right|_2 \leq s^2$ then for $W = \frac{1}{s} \sum_{i=1}^{s} X_i$ and $\epsilon > 0$:

$$\Pr\left[\left||W|\right|_2 > \epsilon\right] \leq 2k \exp\left(-\frac{s\epsilon^2}{2s^2 + \frac{2\gamma\epsilon}{3}}\right)$$

Proof: LSS as a Subspace Embedding

- $U_i = i$-th sampled row of Z in LSS(Z, s, q)
- $z_j =$ j-th row of Z
- $X_i = I_k - U_i^T U_i / q_i$
- $E[X_i] = I_k - \sum_{j=1}^{n} \frac{q_j z_j^T z_j}{q_j} = I_k - Z^T Z = 0_{k \times k}$
- $\frac{z_j^T z_j}{q_j}$ is a rank-1 matrix with operator norm $\leq \frac{||z_j||_2^2}{q_j} \leq \frac{k}{\beta}$:

$$||X_i||_2 \leq ||I_k||_2 + \left|\left|\frac{U_i^T U_i}{q_i}\right|\right|_2 \leq 1 + k/\beta$$

Proof: LSS as a Subspace Embedding

- $E[X^T X] = I_k - 2E\left[\frac{U_i^T U_i}{q_i}\right] + E\left[\frac{U_i^T U_i U_i^T U_i}{q_i^2}\right]$

$$= \sum_{j=1}^{n} \frac{z_j^T z_j z_j^T z_j}{q_j} - I_k$$

$$\leq \left(\frac{k}{\beta}\right) \sum_{j=1}^{n} z_j^T z_j - I_k$$

$$= \left(\frac{k}{\beta} - 1\right) I_k$$

- $\left|\left|E[X^T X]\right|\right|_2 \leq \left(\frac{k}{\beta} - 1\right)$
- Take $W = \frac{1}{k} \sum_{i=1}^{s} X_i = I_k - Z^T \Omega D D^T \Omega^T Z$
- By Matrix Chernoff for $s = \Theta(k \log \frac{k}{\delta} / (\beta \epsilon^2))$:

$$\Pr\left[\left|\left|I_k - Z^T \Omega D D^T \Omega^T Z\right|\right|_2 > \epsilon\right] \leq \delta$$

LSS as a Subspace Embedding

- $A = Z\Sigma V^T$ (SVD of A)
- $||D^T\Omega^T Ax||_2 =$

$=(1 \pm \epsilon)||\Sigma V^T x||_2$ (all sing. values up to $1 \pm \epsilon$)

$=(1 \pm \epsilon)||Ax||_2$ ($||Zy||_2 = ||y||_2$)

- How to compute q in O($nnz(A)\log n + poly(k)$) time?

Thin Singular Value Decomposition

- $\boldsymbol{A} \in \mathbb{R}^{n\times d}, \boldsymbol{U} \in \mathbb{R}^{n\times d}, \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}, \boldsymbol{V} \in \mathbb{R}^{d\times d}$
- $\boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^T$ (computed in $O(n\,d^2)$) time
- $\boldsymbol{U}$ has orthonormal columns, $\boldsymbol{\Sigma}$ is diagonal, $\boldsymbol{V}$ is unitary ($\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$)
- $\boldsymbol{\Sigma}_{ii} = \boldsymbol{\sigma}_i$ is the $\boldsymbol{i}$-th singular value
- $\boldsymbol{v}_i =$ i-th column of $\boldsymbol{V}$ is the i-th right singular vector:

$$||\boldsymbol{A}\boldsymbol{v}_i||_2 = ||\boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^T\boldsymbol{v}_i||_2 = \left\|\boldsymbol{U}\,\boldsymbol{\Sigma}\,\mathbf{e}_\mathbf{i}^\mathbf{T}\right\|_\mathbf{2} = \boldsymbol{\sigma}_i \left\|\boldsymbol{U}\,\mathbf{e}_\mathbf{i}^\mathbf{T}\right\| = \boldsymbol{\sigma}_i$$

- Moore-Penrose pseudoinverse :

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

- Least squares solution: $\boldsymbol{x}^* = \boldsymbol{A}^+\boldsymbol{b} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^T\boldsymbol{b}$

Approximating Leverage Scores

- **Thm.** A constant-approx. leverage score distribution for $A \in \mathbb{R}^{n \times d}$ can be computed with constant prob. in $O(nnz(A) \log n + poly(d))$ time
- **S** = sparse embedding matrix with $r = O(d^2/\gamma^2)$ rows for constant γ (Count-Sketch matrix)
- One non-zero entry per column of **S** => $\boldsymbol{SA}$ computed in nnz(A) time
- QR-factorization: **QR = SA** where **Q** has orthonormal columns, **S** is upper triangular (takes $O(r\, d^2)$) time using e.g. Gram-Schmidt
- $q_i = \left\|e_i^T \boldsymbol{A}\boldsymbol{R}^{-1}\boldsymbol{G}\right\|_2^2$ where $\boldsymbol{G} \in \mathbb{R}^{k \times t}$ is a matrix of i.i.d $N(0,1/t)$ random variables for $t = O\left(\frac{\log n}{\gamma^2}\right)$
- $\boldsymbol{R}^{-1}\boldsymbol{G}$ in $O(k^2 \log n/\gamma^2)$, $\mathbf{A}(\boldsymbol{R}^{-1}\boldsymbol{G})$ in $O(nnz(\boldsymbol{A}) \log n\,/\gamma^2)$

Approximating Leverage Scores

- $q_i = \left|\left|e_i^T \boldsymbol{A}\boldsymbol{R}^{-\mathbf{1}}\boldsymbol{G}\right|\right|_2^2 \geq (1 - \gamma)\left|\left|e_i^T \boldsymbol{A}\boldsymbol{R}^{-\mathbf{1}}\right|\right|_2^2$
- Singular values of $\boldsymbol{A}\boldsymbol{R}^{-\mathbf{1}} \in [1 - \gamma, 1 + \gamma]$

$$\left|\left|\boldsymbol{A}\boldsymbol{R}^{-\mathbf{1}}\boldsymbol{x}\right|\right|_2^2 = (1 \pm \gamma)\left|\left|\boldsymbol{S}\boldsymbol{A}\boldsymbol{R}^{-\mathbf{1}}\boldsymbol{x}\right|\right|_2^2$$
$$= (1 \pm \gamma)\left|\left|\mathbf{Q}\,\mathbf{x}\right|\right|_2^2$$
$$= (1 \pm \gamma)\left|\left|\mathbf{x}\right|\right|_2^2$$

- $\boldsymbol{U} = \boldsymbol{A}\boldsymbol{R}^{-\mathbf{1}}\boldsymbol{T} =$ o.n.b. for the column space of A
- Singular values of $\boldsymbol{T}$ are $\in [1 - 2\gamma, 1 + 2\gamma]$, otherwise $\left|\left|\boldsymbol{A}\boldsymbol{R}^{-\mathbf{1}}\boldsymbol{T}\boldsymbol{v}\right|\right|_2^2 \leq (1 - 2\gamma)(1 + \gamma) < 1$ but $\left|\left|\boldsymbol{A}\boldsymbol{R}^{-\mathbf{1}}\boldsymbol{T}\boldsymbol{v}\right|\right|_2^2 = \left|\left|\boldsymbol{U}\boldsymbol{v}\right|\right|_2^2 = 1$
- $\left|\left|e_i^T \boldsymbol{A}\boldsymbol{R}^{-\mathbf{1}}\right|\right|_2^2 = \left|\left|e_i^T \boldsymbol{U}\boldsymbol{T}^{-\mathbf{1}}\right|\right|_2^2 \geq (1 - 2\gamma)\left|\left|e_i^T \boldsymbol{U}\right|\right|_2^2 = (1 - 2\gamma)p_i$
- Thus, $q_i \geq (1 - \gamma)(1 - 2\gamma)p_i$

Least Squares Regression

- Dimension $\tilde{O}\left(d^2/\epsilon^2\right)$ can be reduced to $O\left(\frac{d}{\epsilon^2}\right)$ by using sketch matrix $\boldsymbol{S}'' = \boldsymbol{S}'\boldsymbol{S}$ where $\boldsymbol{S}'$ is a dense OSS
- Instead of using leverage scores we could just use $\boldsymbol{S}''$ as OSS and solve LSR in $O(nnz(A) + poly(d/\epsilon))$ time
- Skylark: https://github.com/xdata-skylark/libskylark

L_1-regression

- L_2-regression is too sensitive to outliers
- $\min_x ||\boldsymbol{A}x - \boldsymbol{b}||_1 = \sum_{i=1}^{n} |\boldsymbol{b_i} - \langle \boldsymbol{A_{i,*}}, \boldsymbol{x} \rangle|$
- No closed-form solution
- Best running time by LP in $poly(n, d)$ time
- Maximum Likelihood Estimators for noisy data:
  - L_2 = MLE if noise is Gaussian
  - L_1 = MLE if noise is Laplacian
- L_1 subspace embedding:

$$\forall x: \; ||SAx||_1 = (1 \pm \epsilon)||Ax||_1$$

- Next time: approximate L_1 −regression in O(n $poly(d)$) time.