COMPSCI 690RA: Randomized Algorithms and Probabilistic Data Analysis

Prof. Cameron Musco

University of Massachusetts Amherst. Spring 2022.

Lecture 8

Logistics

- I released Problem Set 3 on Monday – it is due 4/15 at 8pm.
- Project progress report due next Friday, 4/8. Submit via email.
- Weekly quiz due next Tuesday at 8pm.

Summary

Last Week: Random sketching and subspace embedding.

- Subspace embedding from the distributional Johnson-Lindenstrauss lemma and an ϵ-net argument.
- Application to fast over-constrained linear regression.
- Proof of distributional JL via the Hanson-Wright inequality.
- You'll see two more applications of subspace embeddings on the problem set, along with problems practicing the use of ϵ-nets and the Hanson-Wright inequality.

Today:

- Subspace embedding via sampling.
- The matrix leverage scores.
- Analysis via matrix concentration bounds.
- Spectral graph sparsifiers.

Quiz Review

Question **4**

Not complete

Points out of 1.00

Flag question

Edit question

Consider applying linear sketching in the streaming setting. There is some underlying matrix $A \in \mathbb{R}^{n \times d}$, initially zero. In each step you receive an update of the form (i, j, v) which modifies A_{ij} by adding v to it. You pick a random sketching matrix $S \in \mathbb{R}^{m \times n}$, and maintain the sketch SA over the updates of A. Storing this sketch requires storing only $m \times d$ entries, as opposed to $n \times d$ for storing all of A

What is the runtime required to update SA to reflect a single entry update to A?

- a. O(mn)
- b. O(m)
- c. O(n)
- d. O(d)
- e. O(nd)

Quiz Review

Which of the following concentration bounds can be apply to show that, for a random $x \in \mathbb{R}^n$ with i.i.d. ± 1 entries, and some fixed $A \in \mathbb{R}^{n \times n}$, that $x^T Ax$ is concentrated around its mean? Select all that apply.

- ☐ a. Chebyshev inequality
- ☐ b. Hanson-Wright Inequality
- ☐ c. Markov bound
- ☐ d. Bernstein bound

Question 6

Not complete

Points out of 1.00

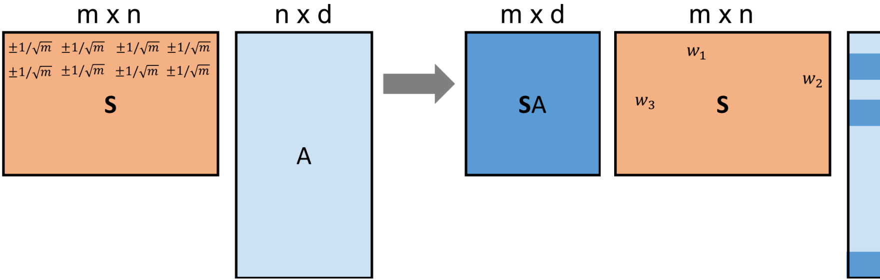
Flag question

Edit question

Subspace Embedding

$\mathbf{S} \in \mathbb{R}^{m \times n}$ is an ϵ-subspace embedding for $A \in \mathbb{R}^{n \times d}$, if for all $x \in \mathbb{R}^d$,

$$(1-\epsilon)\|Ax\| \leq \|\mathbf{S}Ax\|_2 \leq (1+\epsilon)\|Ax\|_2.$$

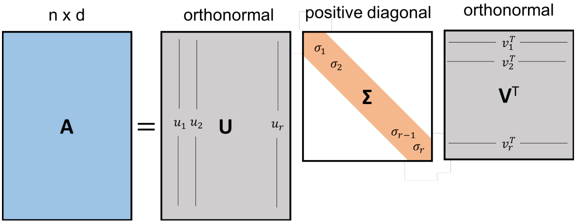


Last Time: If $\mathbf{S}$ is a random sign matrix, and $m = O\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$, then for any A, $\mathbf{S}$ is an ϵ-subspace embedding with probability $\geq 1-\delta$.

In many applications it is preferable for $\mathbf{S}$ to be a row sampling matrix. The sample can preserve sparsity, structure, etc.

Problem Reformulation

For $A \in \mathbb{R}^{n \times d}$, let $A = U\Sigma V^T$ be its SVD. $U \in \mathbb{R}^{n \times \text{rank}(A)}$, $V \in \mathbb{R}^{d \times \text{rank}(A)}$ are orthonormal, and $\Sigma \in \mathbb{R}^{\text{rank}(A) \times \text{rank}(A)}$ is positive diagonal.)



- For any $x \in \mathbb{R}^d$, let $z = \Sigma V^T x$. Observe that: $\|Ax\|_2 = \|Uz\|_2$ and $\|\mathbf{S}A\|_2 = \|\mathbf{S}Uz\|_2$.
- Thus, to prove that S is an ϵ-subspace embedding for A, it suffices to show that it is an ϵ-subspace embedding for U.
- I.e., it suffices to show that for any $x \in \mathbb{R}^d$,

$$(1-\epsilon)\|Ux\|_2^2 \leq \|\mathbf{S}Ux\|_2^2 \leq (1+\epsilon)\|Ux\|_2^2.$$

Loewner Ordering

Suffices to show that for any $x \in \mathbb{R}^d$,

$$(1-\epsilon)\|x\|_2^2 \leq \|\mathbf{S}Ux\|_2^2 \leq (1+\epsilon)\|x\|_2^2 \implies (1-\epsilon)x^T Ix \leq x^T U^T \mathbf{S}^T \mathbf{S} Ux \leq (1+\epsilon)x^T Ix.$$

This condition is typically denoted by $(1-\epsilon)I \preceq U^T\mathbf{S}^T\mathbf{S}U \preceq (1+\epsilon)I$.

$$M \preceq N \text{ iff } \forall x \in \mathbb{R}^d \; x^T Mx \leq x^T Nx \quad \text{(Loewner Order)}$$

When $(1-\epsilon)N \preceq M \preceq (1+\epsilon)N$, I will write $M \approx_\epsilon N$ as shorthand.

$(1-\epsilon)I \preceq U^T\mathbf{S}^T\mathbf{S}U \preceq (1+\epsilon)I$ is equivilant to all eigenvalues of $U^T\mathbf{S}^T\mathbf{S}U$ lying in $[1-\epsilon, 1+\epsilon]$.

Sampling from U

So Far: We have an orthonormal matrix $U \in \mathbb{R}^{n \times d}$ and we want to sample rows so that $U^T \mathbf{S}^T \mathbf{S} U \approx_\epsilon I$. What are some possible sampling strategies?

Leverage Score Sampling

- $\tau_i = \|U_{i,:}\|_2^2$ is known as the i^{th} leverage score of U.
- Let $p_i = \frac{\tau_i}{\sum_{i=1}^n \tau_i}$.
- Let $\mathbf{S}_{:,j} = e_i^T \cdot \frac{1}{\sqrt{mp_i}}$ with probability p_i.

$$\begin{aligned}
\mathbb{E}[U^T\mathbf{S}^T\mathbf{S}U] = & = \sum_{j=1}^m \mathbb{E}[U^T\mathbf{S}_{:,j}^T\mathbf{S}_{:,j}U] \\
& = \sum_{j=1}^m \sum_{i=1}^n p_i \cdot \left(\frac{1}{\sqrt{mp_i}} U_{i,:}^T\right)\left(\frac{1}{\sqrt{mp_i}} U_{i,:}\right) \\
& = \sum_{j=1}^m \frac{1}{m} U^T U = I.
\end{aligned}$$

Matrix Concentration

We want to show that $U^T \mathbf{S}^T \mathbf{S} U$ is close to $\mathbb{E}[U^T \mathbf{S}^T \mathbf{S} U] = I$. Need to apply a matrix concentration bound.

Theorem (Matrix Chernoff Bound)

Consider independent symmetric random matrices $\mathbf{X}_1, \ldots, \mathbf{X}_m \in \mathbb{R}^{d \times d}$, *with* $\mathbf{X}_i \succeq 0$, $\lambda_{\max}(\mathbf{X}_i) \leq R$, *and* $\mathbf{X} = \sum_{i=1}^m \mathbf{X}_i$. *Let* $M = \mathbb{E}[\mathbf{X}]$. *Then:*

$$\Pr\left[\lambda_{\min}(\mathbf{X}) \leq (1-\epsilon)\lambda_{\min}(M)\right] \leq d \cdot \left[\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right]^{\lambda_{\min}(M)/R}$$

$$\Pr\left[\lambda_{\max}(\mathbf{X}) \geq (1+\epsilon)\lambda_{\max}(M)\right] \leq d \cdot \left[\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right]^{\lambda_{\min}(M)/R}$$

Matrix Concentration Applied to Leverage Score Sampling

Theorem (Matrix Chernoff Bound)

Consider independent symmetric random matrices $\mathbf{X}_1, \ldots, \mathbf{X}_m \in \mathbb{R}^{d \times d}$, *with* $\mathbf{X}_i \succeq 0$, $\lambda_{\max}(\mathbf{X}_i) \leq R$, *and* $\mathbf{X} = \sum_{i=1}^m \mathbf{X}_i$. *Let* $M = \mathbb{E}[\mathbf{X}]$. *Then:*

$$\Pr\left[\lambda_{\max}(\mathbf{X}) \geq (1+\epsilon)\lambda_{\max}(M)\right] \leq d \cdot \left[\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right]^{\lambda_{\min}(M)/R}$$

- In our setting, $\mathbf{X}_i = U^T \mathbf{S}_{:,j}^T \mathbf{S}_{:,j} U$. $\mathbf{X}_i = \frac{1}{mp_i} U_{i,:}^T U_{i,:}$ with probability p_i.
- $M = \mathbb{E}[\mathbf{X}] =$
- $R =$
- $\Pr[U^T \mathbf{S}^T \mathbf{S} U \succeq (1+\epsilon) I] \leq d \cdot \left[\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right]^{m/d} \lesssim d \cdot e^{-\epsilon^2 \cdot m/d}$
- If we set $m = O\left(\frac{d \log(d/\delta)}{\epsilon^2}\right)$ we have $\Pr[U^T \mathbf{S}^T \mathbf{S} U \succeq (1+\epsilon) I] \leq \delta$.

Subpace Embedding via Sampling

Theorem (Subspace Embedding via Leverage Score Sampling)

For any $A \in \mathbb{R}^{n\times d}$ with left singular vector matrix U, let $\tau_i = \|U_{i,:}\|_2^2$ and $p_i = \frac{\tau_i}{\sum \tau_i}$. Let $\mathbf{S} \in \mathbb{R}^{m\times n}$ have $\mathbf{S}_{:,j}$ independently set to $\frac{1}{\sqrt{mp_i}} \cdot e_i^T$ with probability p_i.

Then, if $m = O\left(\frac{d\log(d/\delta)}{\epsilon^2}\right)$, with probability $\geq 1-\delta$, $\mathbf{S}$ is an ϵ-subspace embedding for A.

Matches oblivious random projection up to the $\log d$ factor.

Leverage Score Intuition

Check-In

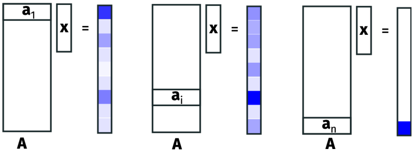
Check-in Question: Would row-norm sampling from A directly rather than its left singular vectors U have worked to give a subspace embedding?

Variational Characterization of Leverage Scores

For a matrix $A \in \mathbb{R}^{n\times d}$ with SVD $A = U\Sigma V^T$, the i^{th} leverage score is given by $\tau_i(A) = \|U_{i,:}\|_2^2$. Consider the maximization problem:

$$\tau_i(A) = \max_{x\in\mathbb{R}^d} \frac{[Ax](i)^2}{\|Ax\|_2^2}.$$

How much can a vector in A's column span 'spike' at position i.



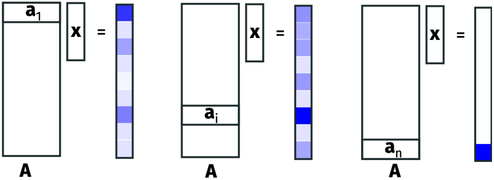
Can rewrite this problem as:

$$\max_{z:\|z\|_2=1} \frac{[Uz](i)^2}{\|Uz\|_2^2} = [Uz](i)^2.$$

What z maximizes this value?

Variational Characterization of Leverage Scores

$$\tau_i(A) = \max_{x \in \mathbb{R}^d} \frac{[Ax](i)^2}{\|Ax\|_2^2}.$$

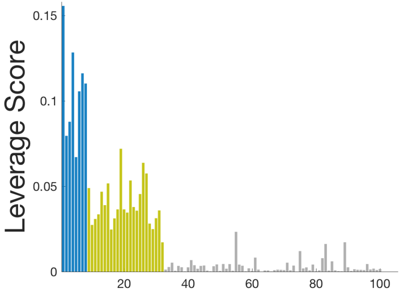


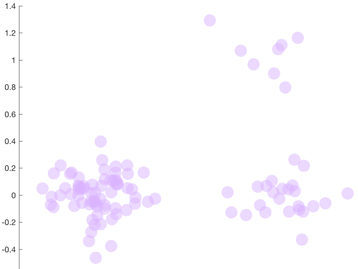
- Remember that we want $\|\mathbf{S}Ax\|_2^2 \approx \|Ax\|_2^2$ for all $x \in \mathbb{R}^d$.
- The leverage scores ensure that we sample all Ax with high enough probability to well approximate $\|Ax\|_2^2$.
- In fact, could prove the subspace embedding theorem by showing that for a fixed $x \in \mathbb{R}^d$, $\|\mathbf{S}Ax\|_2^2 \approx \|Ax\|_2^2$, and then applying a net argument + union bound. Athough you would lose a factor d over the optimal bound.

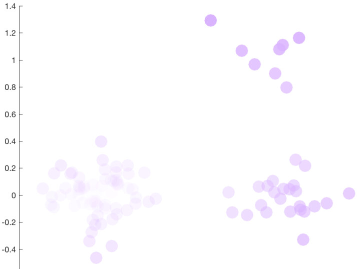
Leverage Score Intuition

- When a_i is not spanned by the other rows of A, $\tau_i(A) = 1$.
- $\tau_i(A)$ is small when many rows are similar to a_i.

Leverage Score Intuition



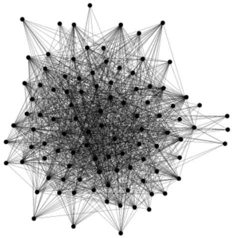


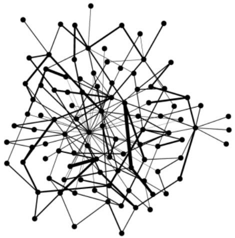


Spectral Graph Sparsification

Graph Sparsification

Given a graph $G = (V, E)$, find a (weighted) subgraph G' with many fewer edges that approximates various properties of G.[1]





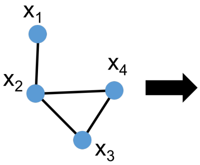
Cut Sparsifier: (Karger) For any set of nodes S,

$$CUT'(S, V \setminus S) \approx_{\epsilon} CUT(S, V \setminus S).$$

[1] Image taken from Nick Harvey's notes https://www.cs.ubc.ca/~nickhar/W15/Lecture11Notes.pdf.

The Graph Laplacian

For a graph with adjacency matrix $A \in \{0,1\}^{n\times n}$ and diagonal degree matrix $D \in \mathbb{R}^{n\times n}$, $L = D - A$ is the graph Laplacian.



$$\underset{\mathbf{D}}{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}} - \underset{\mathbf{A}}{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}} = \underset{\mathbf{L}}{\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}}$$

L can be written as $L = \sum_{(u,v)\in E} L_{u,v}$ where $L_{u,v}$ is an 'edge Laplacian'

$$\underset{\mathbf{L}}{\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}} = \underset{\mathbf{L_{1,2}}}{\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}} + \underset{\mathbf{L_{2,4}}}{\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}} + \cdots$$

Laplacian Smoothness

Observation 1: For any $z \in \mathbb{R}^d$,

$$z^T L z = \sum_{(u,v)\in E} z^T L_{u,v} z = \sum_{(u,v)\in E} (z(i) - z(j))^2.$$

$$\begin{bmatrix} v(1) & v(2) & v(3) & v(4) \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v(1) \\ v(2) \\ v(3) \\ v(4) \end{bmatrix}$$

- $z^T L z$ measures how smoothly z varies across the graph.
- If $z \in \{-1, 1\}^n$ is a cut indicator vector with $z(i) = 1$ for $i \in S$ and $z(i) = -1$ otherwise, then $z^T L z = 4 \cdot CUT(S, V \setminus S)$.
- So G' with (weighted) Laplacian $L' \approx_\epsilon L$ will be a cut sparsifier, with $CUT'(S, V \setminus S) \approx_\epsilon CUT(S, V \setminus S)$ for all S.
- Such a G' is called an ϵ-spectral sparsifier of G.

Laplacian Factorization

Observation 2: $L_{u,v} = b_{u,v} b_{u,v}^T$. So $L = \sum_{(u,v) \in E} b_{u,v} b_{u,v}^T$.

$$\underset{\mathbf{L_{2,4}}}{\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}} = \underset{\mathbf{b_{2,4}}}{\begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}} \underset{\mathbf{b_{2,4}^T}}{\begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix}}$$

1	-1	0	0
0	1	0	-1
0	0	1	-1
-1	0	1	0
1	0	-1	0
0	1	-1	0
1	0	0	-1
0	0	1	-1

vertex-edge incidence matrix B

That is, letting $B \in \mathbb{R}^{m \times n}$ have rows $\{b_{u,v}^T : (u,v) \in E\}$, $L = B^T B$.

- So if a sampling matrix $\mathbf{S}$ is a subspace embedding for B, then $B^T \mathbf{S}^T \mathbf{S} B \approx_\epsilon B^T B \approx_\epsilon L$. I.e., $\mathbf{S}B$ is the weighted vertex-edge incidence matrix of an ϵ-spectral sparsifier of G.
- By our results on subspace embedding, every graph G has an ϵ-spectral sparsifier with just $O(n \log n / \epsilon^2)$ edges.

Some History

- The concept of spectral sparsification was first introduced by Spielman and Teng '04 in their seminal work on fast system solvers for graph Laplacians. In this work, sparsifiers are used as preconditioners (like in Problem Set 3).
- Spielman and Srivastava '08 showed how to construct sparsifiers with $O(n \log n/\epsilon^2)$ edges via effective resistance (leverage score) sampling.
- Batson, Spielman, and Srivastava '08 showed how to achieve $O(n/\epsilon^2)$ edges with a deterministic algorithm.
- Marcus, Spielman, and Srivastava '13 built on this work to give optimal bipartite expanders with any degree and to resolve the famous Kadison-Singer problem in functional analysis.