[BWZ] Protocol

- Main Problem: communication is $O(skd/\varepsilon) + \text{poly}(sk/\varepsilon)$, but we want $O(skd) + \text{poly}(sk/\varepsilon)$ communication!

- Idea: use projection-cost preserving sketches [CEMMP]

- Let $A$ be an $n \times d$ matrix

- If $S$ is a random $k/\varepsilon^2 \times n$ matrix, then there is a scalar $c \geq 0$ so that for all $k$-dimensional projection matrices $P$:
  $$|A(I - P)|_F^2 \leq |SA(I - P)|_F^2 + c \leq (1 + \varepsilon)|A(I - P)|_F^2$$

- Implication: If $I - \tilde{P}$ is the minimizer of $|SA(I - P)|_F^2$, and $I - P^*$ is the minimizer of $|A(I - P)|_F^2$, then
  $$|A(I - \tilde{P})|_F^2 \leq (1 + \varepsilon)|A - A_k|_F^2$$
  So
  $$|SA - [SA]_k|_F^2 + c \leq (1 + \varepsilon)|A(I - \tilde{P})|_F^2 \leq (1 + O(\varepsilon))|A - A_k^o|_F^2$$
**[BWZ] Protocol**

- Let $S$ be a $k/\varepsilon^2 \times n$ projection-cost preserving sketch
- Let $T$ be a $d \times k/\varepsilon^2$ projection-cost preserving sketch
- Server $t$ sends $SA^tT$ to Coordinator

- Coordinator sends back $SAT = \sum_t SA^tT$ to servers
- Each server computes $k/\varepsilon^2 \times k$ matrix $U$ of top $k$ left singular vectors of $SAT$

Intuitively, $U$ looks like top $k$ left singular vectors of $SA$.

Thus, $U^TSA$ looks like top $k$ scaled right singular vectors of $SA$.

- Server $t$ sends $U^TSA^t$ to Coordinator
- Coordinator returns the space $U^TSA = \sum_t U^TSA^t$ to output

Top $k$ right singular vectors of $SA$ work because $S$ is a projection-cost preserving sketch!
[BWZ] Analysis

- Let $W$ be the row span of $U^TSA$, and $P$ be the projection onto $W$

- Want to show $|A - AP|_F^2 \leq (1 + \epsilon)|A - A_k|_F^2$

- Since $T$ is a projection-cost preserving sketch,

\[ (*) \quad |SA - SAP|_F^2 \leq |SA - UU^TSA|_F^2 \leq (1 + \epsilon)|SA - [SA]_k|_F^2 \]

- Since $S$ is a projection-cost preserving sketch, there is a scalar $c \geq 0$, so that for all $k$-dimensional projection matrices $Q$,

\[ |SA - SAQ|_F^2 + c = (1 \pm \epsilon)|A - AQ|_F^2 \]

- Add $c$ to both sides of (*) to conclude $|A - AP|_F^2 \leq (1 + O(\epsilon))|A - A_k|_F^2$
Conclusions for Distributed Low Rank Approximation

- [BWZ] Optimal $O(sdk) + \text{poly}(sk/\epsilon)$ communication protocol for low rank approximation in arbitrary partition model
  - Handle bit complexity by adding noise (omitted)
  - Input sparsity time
  - 2 rounds, which is optimal [W]

- Communication of other optimization problems?
  - Computing the rank of an $n \times n$ matrix over the reals
  - Linear Programming
  - Graph problems: Matching
  - etc.
Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator Regression
Robust Regression

Method of least absolute deviation ($l_1$-regression)

- Find $x^*$ that minimizes $|Ax-b|_1 = \sum |b_i - \langle A_{i*}, x \rangle|$
- Cost is less sensitive to outliers than least squares
- Can solve via linear programming
Solving $l_1$-regression via Linear Programming

- Minimize $(1,\ldots,1) \cdot (\alpha^+ + \alpha^-)$
- Subject to:
  \[ A x + \alpha^+ - \alpha^- = b \]
  \[ \alpha^+, \alpha^- \geq 0 \]

- Generic linear programming gives poly(nd) time
- Want much faster time using sketching!
Well-Conditioned Bases

- For an $n \times d$ matrix $A$, can choose an $n \times d$ matrix $U$ with orthonormal columns for which $A = UW$, and $|Ux|_2 = |x|_2$ for all $x$

- Can we find a $U$ for which $A = UW$ and $|Ux|_1 \approx |x|_1$ for all $x$?

- Let $A = QW$ where $Q$ has full column rank, and define $|z|_{Q,1} = |Qz|_1$
  - $|z|_{Q,1}$ is a norm

- Let $C = \{z \in \mathbb{R}^d : |z|_{Q,1} \leq 1\}$ be the unit ball of $|.|_{Q,1}$

- $C$ is a convex set which is symmetric about the origin
  - Lowner-John Theorem: can find an ellipsoid $E$ such that: $E \subseteq C \subseteq \sqrt{d}E$, where $E = \{z \in \mathbb{R}^d : z^TFz \leq 1\}$
    - $(z^TFz)^{\frac{1}{5}} \leq |z|_{Q,1} \leq \sqrt{d}(z^TFz)^{\frac{1}{5}}$
    - $F = G^TG$ since $F$ defines an ellipsoid

- Define $U = QG^{-1}$
Well-Conditioned Bases

- Recall $U = QG^{-1}$ where
\[
(z^T F z)^5 \leq |z|_{Q,1} \leq \sqrt{d}(z^T F z)^5 \quad \text{and} \quad F = G^T G
\]

- $|Ux|_1 = |QG^{-1}x|_1 = |Qz|_1 = |z|_{Q,1}$ where $z = G^{-1}x$

- $z^T F z = (x^T (G^{-1})^T G^T G (G^{-1})x) = x^T x = |x|^2_2$

- So $|x|_2 \leq |Ux|_1 \leq \sqrt{d}|x|_2$

- So $\frac{|x|_1}{\sqrt{d}} \leq |x|_2 \leq |Ux|_1 \leq \sqrt{d}|x|_2 \leq \sqrt{d}|x|_1$
Net for $\ell_1$ – Ball

- Consider the unit $\ell_1$-ball $B = \{x \in \mathbb{R}^d : |x|_1 = 1\}$
- Subset $N$ is a $\gamma$-net if for all $x \in B$, there is a $y \in N$, such that $|x - y|_1 \leq \gamma$
- Greedy construction of $N$
  - While there is a point $x \in B$ of distance larger than $\gamma$ from every point in $N$, include $x$ in $N$
- The $\ell_1$-ball of radius $\gamma/2$ around every point in $N$ is contained in the $\ell_1$-ball of radius $1 + \gamma/2$ around $0^d$
- Further, all such ball are disjoint
- Ratio of volume of $d$-dimensional similar polytopes of radius $1 + \gamma/2$ to radius $\gamma/2$ is $(1 + \gamma/2)^d / (\gamma/2)^d$, so $|N| \leq (1 + \gamma/2)^d / (\gamma/2)^d$
Net for $\ell_1$ — Subspace

- Let $A = UW$ for a well-conditioned basis $U$
  - $|x|_1 \leq |Ux|_1 \leq d|x|_1$ for all $x$

- Let $N$ be a $(\gamma/d)$-net for the unit $\ell_1$-ball $B$

- Let $M = \{Ux \mid x \in N\}$, so $|M| \leq (1 + \gamma/(2d))^d/(\gamma/(2d))^d$

- Claim: For every $x$ in $B$, there is a $y$ in $M$ for which $|Ux - y|_1 \leq \gamma$

- Proof: Let $x'$ in $N$ be such that $|x - x'|_1 \leq \gamma/d$
  Then $|Ux - Ux'|_1 \leq d|x - x'|_1 \leq \gamma$, using the well-conditioned basis property. Set $y = Ux'$

- $|M| \leq \left(\frac{d}{\gamma}\right)^{O(d)}$
Rough Algorithm Overview

\[ \min_{x \in \mathbb{R}^d} |Ax-b|_1 = \min_{x \in \mathbb{R}^d} |Ux - b'|_1 \]

Sample \( \text{poly}(d/\varepsilon) \) rows of \( U \circ b' \) proportional to their \( l_1 \)-norm.

Find \( x' \) such that \( |Ax'-b|_1 \leq \text{poly}(d) \cdot \min_{x \in \mathbb{R}^d} |x|_1 \)

Let \( b' = b - Ax' \) be the residual.

\[ |x|_1/\text{poly}(d) \leq |Ux|_1 \leq \text{poly}(d) \cdot |x|_1 \]

Find a basis \( A = UW \) so that for all \( x \in \mathbb{R}^d \), \( |x|_1/\text{poly}(d) \leq |Ux|_1 \leq \text{poly}(d) \cdot |x|_1 \)

Now generic linear programming is efficient.

Takes \( \text{nnz}(A) \) time.

Solve \( l_1 \)-regression on the sample, obtaining vector \( x \), and output \( x \)
Will focus on showing how to quickly compute

1. A poly(d)-approximation

2. A well-conditioned basis
Sketching Theorem

Theorem

- There is a probability space over \( (d \log d) \times n \) matrices \( R \) such that for any \( n \times d \) matrix \( A \), with probability at least \( 99/100 \) we have for all \( x \):
  \[
  |Ax|_1 \leq |RAx|_1 \leq d \log d \cdot |Ax|_1
  \]

Embedding

- is linear
- is independent of \( A \)
- preserves lengths of an infinite number of vectors
Application of Sketching Theorem

Computing a $d(\log d)$-approximation

- Compute $RA$ and $Rb$

- Solve $x' = \arg\min_x |RAx - Rb|_1$

- Main theorem applied to $A \circ b$ implies $x'$ is a $d \log d$ – approximation

- $RA$, $Rb$ have $d \log d$ rows, so can solve $l_1$-regression efficiently
Application of Sketching Theorem

Computing a well-conditioned basis

1. Compute RA
2. Compute W so that RAW is orthonormal (in the $l_2$-sense)
3. Output $U = AW$

$U = AW$ is well-conditioned because

$$|AWx|_1 \leq |RAWx|_1 \leq (d \log d)^{1/2} |RAWx|_2 = (d \log d)^{1/2} |x|_2 \leq (d \log d)^{1/2} |x|_1$$

and

$$|AWx|_1 \geq |RAWx|_1/(d \log d) \geq |RAWx|_2/(d \log d) = |x|_2/(d \log d) \geq |x|_1/(d^{3/2} \log d)$$
Theorem:

There is a probability space over \((d \log d) \times n\) matrices \(R\) such that for any \(n \times d\) matrix \(A\), with probability at least 99/100 we have for all \(x\):

\[
|Ax|_1 \leq |RAx|_1 \leq d \log d \cdot |Ax|_1
\]

A dense \(R\) that works:

The entries of \(R\) are i.i.d. Cauchy random variables, scaled by \(1/(d \log d)\)