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

Distributed low rank approximation

- *We have fast algorithms for low rank approximation, but can they be made to work in a distributed setting?*

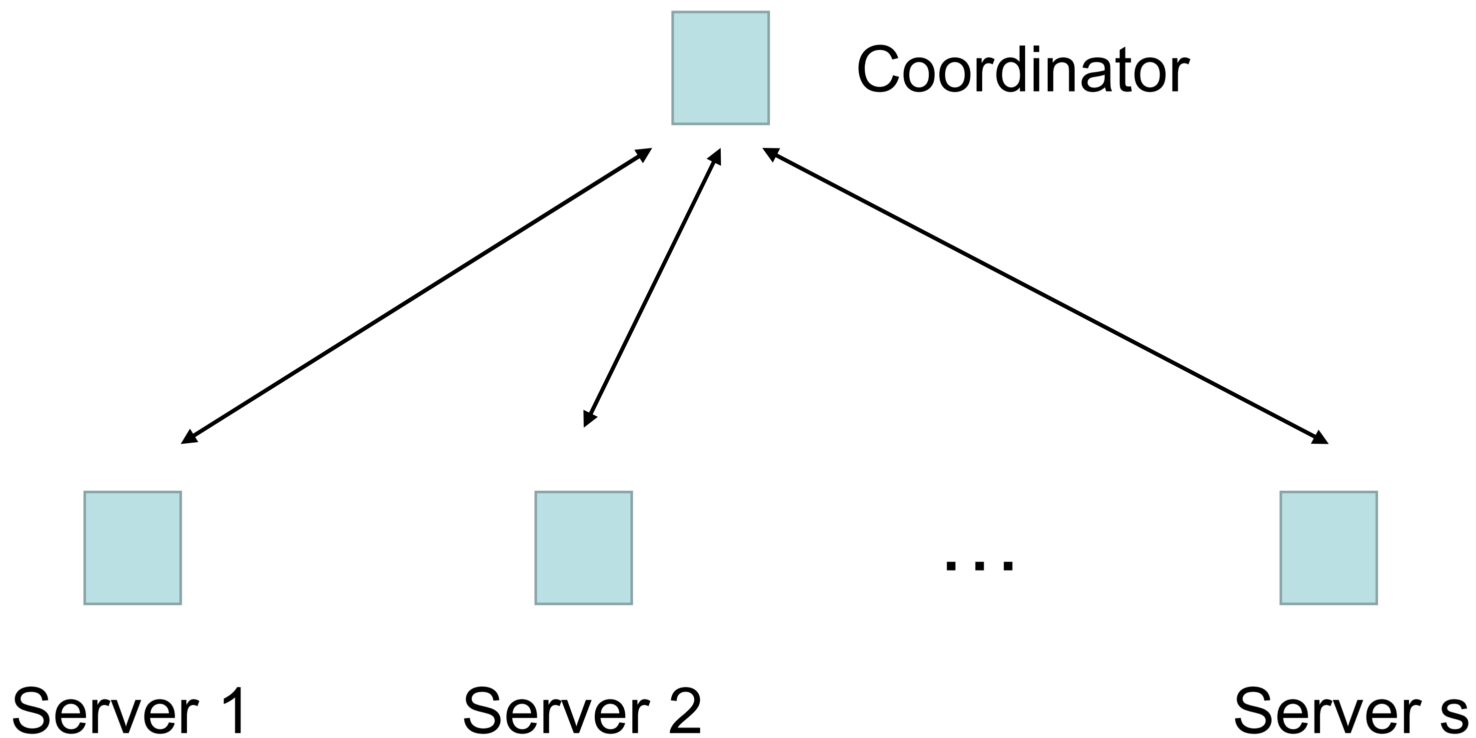
- Matrix A distributed among s servers

- For t = 1, …, s, we get a customer-product matrix from the t-th shop stored in server t. Server t's matrix = A^t

- Customer-product matrix $A = A^1 + A^2 + \ldots + A^s$
  - Model is called the arbitrary partition model

- More general than the row-partition model in which each customer shops in only one shop

The Communication Model



- Each player talks only to a Coordinator via 2-way communication
- Can simulate arbitrary point-to-point communication up to factor of 2 (and an additive O(log s) factor per message)

Communication cost of low rank approximation

- Input: n x d matrix A stored on s servers
  - Server t has n x d matrix A^t
  - $A = A^1 + A^2 + \ldots + A^s$
  - Assume entries of A^t are O(log(nd))-bit integers

- Output: Each server outputs the same k-dimensional space W
  - $C = A^1 P_W + A^2 P_W + \ldots + A^s P_W$, where P_W is the projection onto W
  - $|A-C|_F \leq (1+\varepsilon)|A-A_k|_F$
  - Application: k-means clustering

- Resources: Minimize total communication and computation. Also want O(1) rounds and input sparsity time

Work on Distributed Low Rank Approximation

- [FSS]: First protocol for the row-partition model.
  - O(sdk/ε) real numbers of communication
  - Don't analyze bit complexity (can be large)
  - SVD Running time, see also [BKLW]

- [KVW]: O(skd/ε) communication in arbitrary partition model

- [BWZ]: O(skd) + poly(sk/ε) words of communication in arbitrary partition model. Input sparsity time
  - Matching Ω(skd) words of communication lower bound

- Variants: kernel low rank approximation [BLSWX], low rank approximation of an implicit matrix [WZ], sparsity [BWZ]

Outline of Distributed Protocols

- [FSS] protocol
- [KVW] protocol
- [BWZ] protocol

Constructing a Coreset [FSS]

- Let $A = U\,\Sigma V^T$ be its SVD
- Let m = k + k/ϵ
- Let Σ_m agree with Σ on the first m diagonal entries, and be 0 otherwise
- Claim: For all projection matrices Y=I-X onto (d-k)-dimensional subspaces,

$$|AY|_F^2 \leq \left|\Sigma_m V^T Y\right|_F^2 + c \leq (1+\epsilon)|AY|_F^2,$$

where $c = |A - A_m|_F^2$ does not depend on Y

- We can think of S as U_m^T so that $SA = U_m^T U \Sigma V^T = \Sigma_m V^T$ is a sketch
- If $\widetilde{Y}$ is the minimizer of $\left|\Sigma_m V^T Y\right|_F^2$, and Y^* is the minimizer of $|AY|_F^2$, then $\left|A\widetilde{Y}\right|_F^2 \leq \left|\Sigma_m V^T \widetilde{Y}\right|_F^2 + c \leq \left|\Sigma_m V^T Y^*\right|_F^2 + c \leq (1+\epsilon)|AY^*|_F^2 = (1+\epsilon)|A - A_k|_F^2$

Constructing a Coreset

- Claim: For all projection matrices Y=I-X onto (d-k)-dimensional subspaces,

$$|AY|_F^2 \le \left|\Sigma_m V^T Y\right|_F^2 + c \le (1+\epsilon)|AY|_F^2,$$

where $c = |A - A_m|_F^2$ does not depend on Y

- Proof: $|AY|_F^2 = \left|U\Sigma_m V^T Y\right|_F^2 + \left|U(\Sigma - \Sigma_m)V^T Y\right|_F^2$

$$\le \left|\Sigma_m V^T Y\right|_F^2 + |A - A_m|_F^2 = \left|\Sigma_m V^T Y\right|_F^2 + c$$

Also, $\left|\Sigma_m V^T Y\right|_F^2 + |A - A_m|_F^2 - |AY|_F^2$

$$= \left|\Sigma_m V^T\right|_F^2 - \left|\Sigma_m V^T X\right|_F^2 + |A - A_m|_F^2 - |A|_F^2 + |AX|_F^2$$

$$= |AX|_F^2 - \left|\Sigma_m V^T X\right|_F^2$$

$$= \left|(\Sigma - \Sigma_m)V^T X\right|_F^2$$

$$\le \left|(\Sigma - \Sigma_m)V^T\right|_2^2 \cdot |X|_F^2$$

$$\le \sigma_{m+1}^2 k \le \epsilon\, \sigma_{m+1}^2 (m-k) \le \epsilon \sum_{i \in \{k+1,..,m+1\}} \sigma_i^2 \le \epsilon |A - A_k|_F^2 \le \epsilon\, |AY|_F^2$$

Unions of Coresets

- Suppose we have matrices $A^1, \ldots, A^s$ and construct $\Sigma_m^1 V^{T,1}, \Sigma_m^2 V^{T,2}, \ldots, \Sigma_m^s V^{T,s}$ as in the previous slide, together with $c_1, \ldots, c_s$

- Then $\sum_i \quad \left|\Sigma_m^i V^{T,i} Y\right|_F^2 + c_i \ = (1 \pm \epsilon)|AY|_F^2$, where A is the matrix formed by concatenating the rows of $A^1, \ldots, A^s$

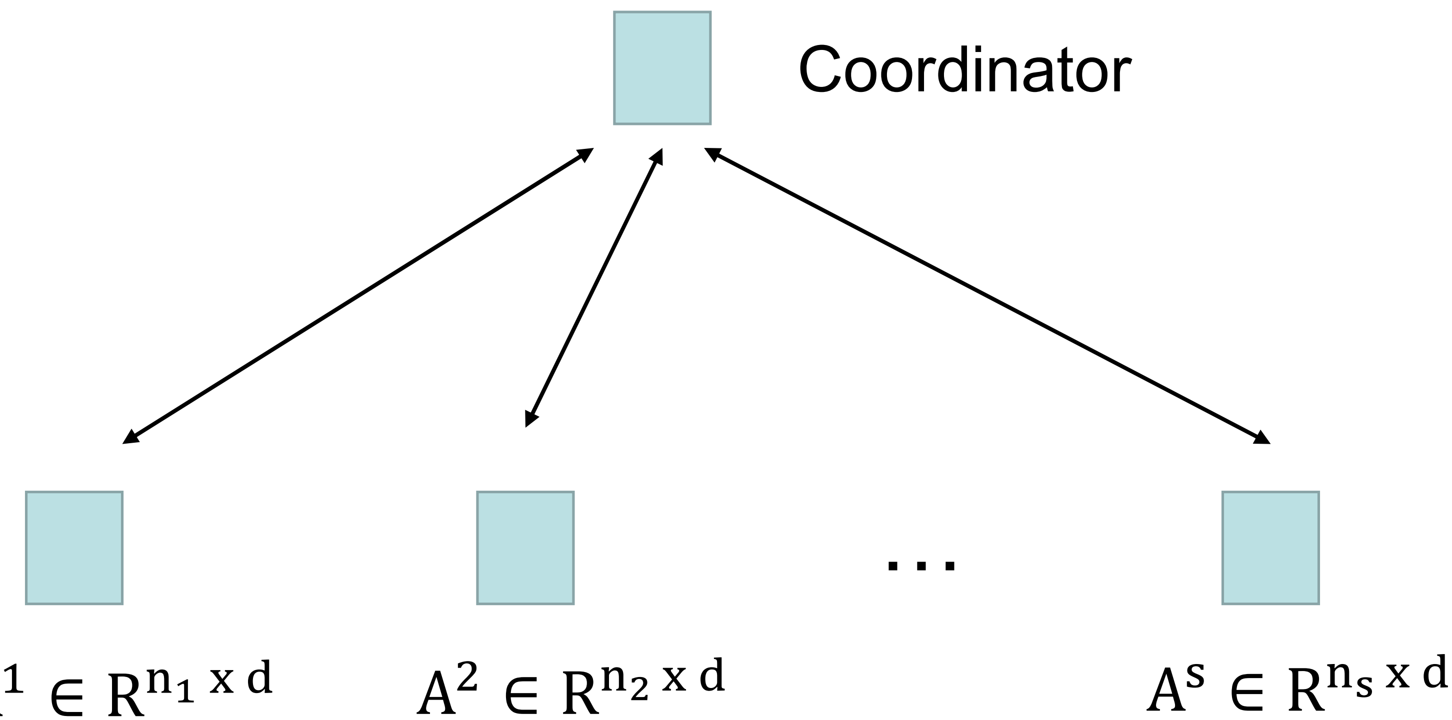
- Let B be the matrix obtained by concatenating the rows of $\Sigma_m^1 V^{T,1}, \Sigma_m^2 V^{T,2}, \ldots, \Sigma_m^s V^{T,s}$

- Suppose we compute $B = U\, \Sigma V^T$ and compute $\Sigma_m V^T$ and $|B - B_m|_F^2$

- Then $\left|\Sigma_m V^T Y\right|_F^2 + c + \sum_i c_i = (1 \pm \epsilon)|BY|_F^2 + \sum_i c_i = (1 \pm O(\epsilon))|AY|_F^2$

- So $\Sigma_m V^T$ and the constant $c + \sum_i c_i$ are a coreset for A

[FSS] Row-Partition Protocol



- Server t sends the top k/ε + k principal components of A^t, scaled by the top k/ε + k singular values Σ^t, together with c^t
- Coordinator returns $c + \sum_i c_i$ and top k/ϵ principal components of $[\Sigma^1 V^1; \Sigma^2 V^2; \ldots; \Sigma^s V^s]$

[FSS] Row-Partition Protocol

[KVW] protocol will handle 2, 3, and 4

Problems:

1. sdk/ε real numbers of communication
2. bit complexity can be large
3. running time for SVDs
4. doesn't work in arbitrary partition model

This is an SVD-based protocol. Maybe our random matrix techniques can improve communication just like they improved computation?

[KVW] Arbitrary Partition Model Protocol

- Inspired by the sketching algorithms presented earlier
- Let S be one of the k/ε x n random matrices discussed
  - S can be generated pseudorandomly from small seed
  - Coordinator sends small seed for S to all servers
- Server t computes SA^t and sends it to Coordinator
- Coordinator sends $\Sigma_{t=1}^{s} SA^t = SA$ to all servers
- There is a good k-dimensional subspace inside of SA. If we knew it, t-th server could output projection of A^t onto it

[KVW] Arbitrary Partition Model Protocol

Problems:

- Can't output projection of A^t onto SA since the rank is too large

- Could communicate this projection to the coordinator who could find a k-dimensional space, but communication depends on n

[KVW] Arbitrary Partition Model Protocol

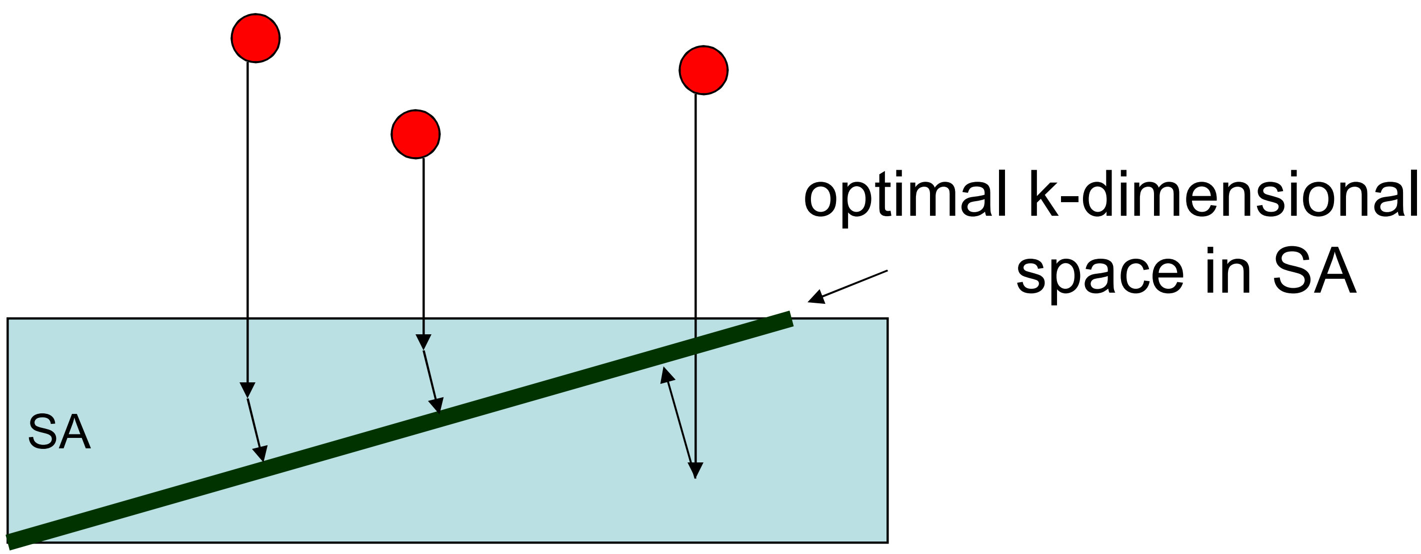
Fix:

- Instead of projecting A onto SA, recall we can solve $\min_{\text{rank}-k\ X} \left|A(SA)^T XSA - A\right|_F^2$
- Let T_1, T_2 be affine embeddings, solve $\min_{\text{rank}-k\ X} \left|T_1 A(SA)^T XSAT_2 - T_1 AT_2\right|_F^2$ (optimization problem is small and has a closed form solution)
- Everyone can then compute XSA and then output k directions

[KVW] protocol

- Phase 1:

- Learn the row space of SA

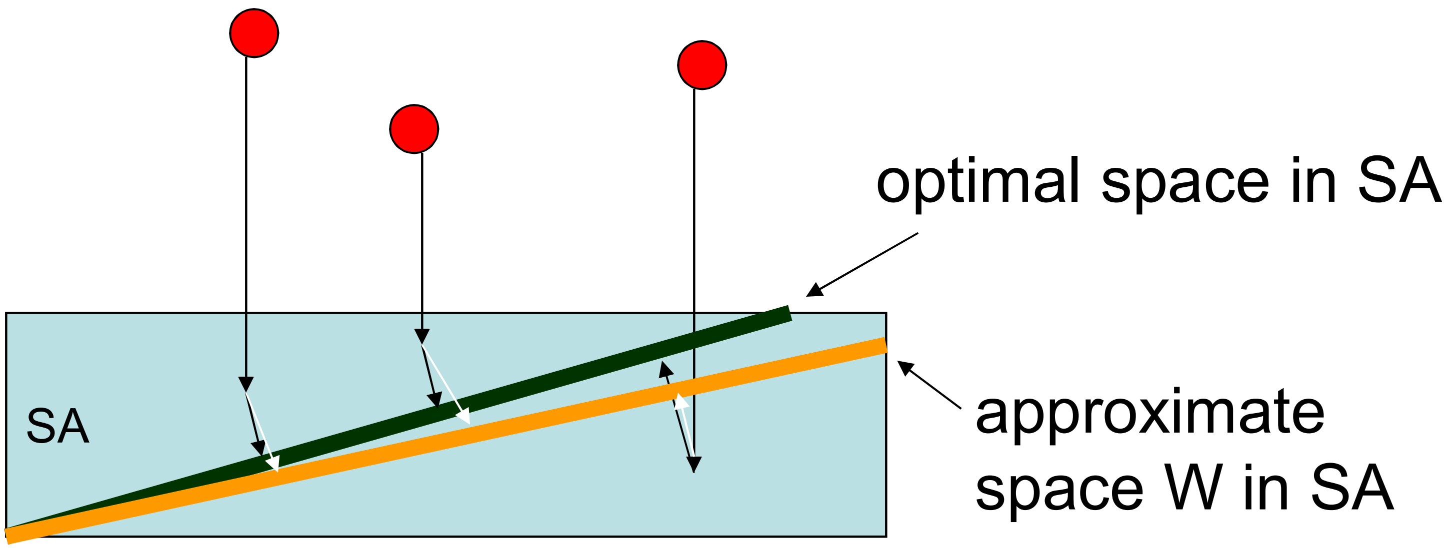


cost $\leq (1+\varepsilon)|A-A_k|_F$

[KVW] protocol

- Phase 2:

- Find an approximately optimal space W inside of SA



$$\text{cost} \leq (1+\varepsilon)^2 |A-A_k|_F$$

[BWZ] Protocol

- Main Problem: communication is O(skd/ε) + poly(sk/ε), but we want O(skd) + poly(sk/ε) communication!

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

- Let A be an n x d matrix

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

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

[BWZ] Protocol

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

- Let S be a k/ε^2 x n projection-cost preserving sketch
- Let T be a d x k/ε^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/ε^2x k matrix U of top k left singular vectors of SAT

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 $U^T SA$, 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,

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

- Since S is a projection-cost preserving sketch, there is a scalar c ≥ 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) + poly(sk/ε) 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 x 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 $|Ax-b|_1 = \Sigma |b_i - \langle A_{i*}, x \rangle|$
- 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 x d matrix A, can choose an n x 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 = $\{z \in R^d : |z|_{Q,1} \leq 1\}$ 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 = $\{z \in R^d : z^T F z \leq 1\}$
  - $\left(z^T F z\right)^{.5} \leq |z|_{Q,1} \leq \sqrt{d}\left(z^T F z\right)^{.5}$
  - $F = GG^T$ since F defines an ellipsoid
- Define $U = QG^{-1}$

Well-Conditioned Bases

- Recall $\mathrm{U} = \mathrm{QG}^{-1}$ where

  $(\mathrm{z}^{\mathrm{T}}\mathrm{Fz})^{.5} \leq |\mathrm{z}|_{\mathrm{Q},1} \leq \sqrt{\mathrm{d}}(\mathrm{z}^{\mathrm{T}}\mathrm{Fz})^{.5}$ and $\mathrm{F} = \mathrm{GG}^{\mathrm{T}}$

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

- $\mathrm{z}^{\mathrm{T}}\mathrm{Fz} \quad = \left(\mathrm{x}^{\mathrm{T}}(\mathrm{G}^{-1})^{\mathrm{T}}\mathrm{G}^{\mathrm{T}}\mathrm{G}\,(\mathrm{G}^{-1})\mathrm{x}\right) = \mathrm{x}^{\mathrm{T}}\mathrm{x} = |\mathrm{x}|_2^2$

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

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