Dimensionality reduction via sparse matrices

Jelani Nelson
Harvard

June 17, 2014

based on joint works with Jean Bourgain (IAS), Daniel Kane (Stanford), Huy Nguyễn (Princeton)

Metric Johnson-Lindenstrauss lemma

Metric JL (MJL) Lemma, 1984

Every set of N points in Euclidean space can be embedded into $O(\varepsilon^{-2} \log N)$-dimensional Euclidean space so that all pairwise distances are preserved up to a $1 \pm \varepsilon$ factor.

Metric Johnson-Lindenstrauss lemma

Metric JL (MJL) Lemma, 1984

Every set of N points in Euclidean space can be embedded into $O(\varepsilon^{-2} \log N)$-dimensional Euclidean space so that all pairwise distances are preserved up to a $1 \pm \varepsilon$ factor.

Uses:

- Speed up geometric algorithms by first reducing dimension of input [Indyk, Motwani '98], [Indyk '01]
- Faster/streaming numerical linear algebra algorithms [Sarlós '06], [LWMRT '07], [Clarkson, Woodruff '09]
- Essentially equivalent to RIP matrices from compressed sensing [Baraniuk et al. '08], [Krahmer, Ward '11] (used for recovery of sparse signals)
- Volume-preserving embeddings (applications to projective clustering) [Magen '02]
- ...

How to prove the JL lemma

Distributional JL (DJL) lemma

Lemma

For any $0 < \varepsilon, \delta < 1/2$ *there exists a distribution* $\mathcal{D}_{\varepsilon,\delta}$ *on* $\mathbb{R}^{m \times n}$ *for* $m = O(\varepsilon^{-2} \log(1/\delta))$ *so that for any* u *of unit* ℓ_2 *norm*

$$\mathop{\mathbb{P}}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}} \left(\left| \|\Pi u\|_2^2 - 1 \right| > \varepsilon \right) < \delta.$$

How to prove the JL lemma

Distributional JL (DJL) lemma

Lemma

For any $0 < \varepsilon, \delta < 1/2$ *there exists a distribution* $\mathcal{D}_{\varepsilon,\delta}$ *on* $\mathbb{R}^{m \times n}$ *for* $m = O(\varepsilon^{-2} \log(1/\delta))$ *so that for any* u *of unit* ℓ_2 *norm*

$$\mathop{\mathbb{P}}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}} \left(\left| \|\Pi u\|_2^2 - 1 \right| > \varepsilon \right) < \delta.$$

Proof of MJL: Set $\delta = 1/N^2$ in DJL and u as the difference vector of some pair of points. Union bound over the $\binom{N}{2}$ pairs.

How to prove the JL lemma

Distributional JL (DJL) lemma

Lemma

For any $0 < \varepsilon, \delta < 1/2$ *there exists a distribution* $\mathcal{D}_{\varepsilon,\delta}$ *on* $\mathbb{R}^{m \times n}$ *for* $m = O(\varepsilon^{-2} \log(1/\delta))$ *so that for any* u *of unit* ℓ_2 *norm*

$$\mathop{\mathbb{P}}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}} \left(\left| \|\Pi u\|_2^2 - 1 \right| > \varepsilon \right) < \delta.$$

Proof of MJL: Set $\delta = 1/N^2$ in DJL and u as the difference vector of some pair of points. Union bound over the $\binom{N}{2}$ pairs.

Theorem (Alon, 2003)

For MJL, $m = \Omega((\varepsilon^{-2}/\log(1/\varepsilon)) \log N)$ *is required.*

Theorem (Jayram-Woodruff, 2011; Kane-Meka-N., 2011)

For DJL, $m = \Theta(\varepsilon^{-2} \log(1/\delta))$ *is optimal.*

Proving the distributional JL lemma

Older proofs

- [Johnson-Lindenstrauss, 1984], [Frankl-Maehara, 1988], [Dasgupta-Gupta, 2003]: Random rotation, then projection onto first m coordinates.
- [Indyk-Motwani, 1998]:
  Random matrix with independent Gaussian entries.
- [Achlioptas, 2001]: Independent ± 1 entries.
- [Clarkson-Woodruff, 2009]:
  $O(\log(1/\delta))$-wise independent ± 1 entries.
- [Arriaga-Vempala, 1999], [Matousek, 2008]:
  Independent entries having mean 0, variance $1/m$, and subGaussian tails

Proving the distributional JL lemma

Older proofs

- [Johnson-Lindenstrauss, 1984], [Frankl-Maehara, 1988], [Dasgupta-Gupta, 2003]: Random rotation, then projection onto first m coordinates.
- [Indyk-Motwani, 1998]:
  Random matrix with independent Gaussian entries.
- [Achlioptas, 2001]: Independent ± 1 entries.
- [Clarkson-Woodruff, 2009]:
  $O(\log(1/\delta))$-wise independent ± 1 entries.
- [Arriaga-Vempala, 1999], [Matousek, 2008]:
  Independent entries having mean 0, variance $1/m$, and subGaussian tails

Downside: Performing embedding is dense matrix-vector multiplication, $O(m \cdot \|u\|_0)$ time

Fast JL Transforms

- [Ailon-Chazelle, 2006]: $x \mapsto PHDx$, $O(n \log n + m^3)$ time

  P random+sparse, H Fourier, D has random ± 1 on diagonal
- Also follow-up works based on similar approach which improve the time while, for some, slightly increasing target dimension [Ailon, Liberty '08], [Ailon, Liberty '11], [Krahmer, Ward '11], [N., Price, Wootters '14], ...

Fast JL Transforms

- [Ailon-Chazelle, 2006]: $x \mapsto PHDx$, $O(n \log n + m^3)$ time

  P random+sparse, H Fourier, D has random ± 1 on diagonal
- Also follow-up works based on similar approach which improve the time while, for some, slightly increasing target dimension [Ailon, Liberty '08], [Ailon, Liberty '11], [Krahmer, Ward '11], [N., Price, Wootters '14], ...

Downside: Slow to embed sparse vectors: running time is $\Omega(\min\{m \cdot \|u\|_0, n \log n\})$.

Where Do Sparse Vectors Show Up?

- **Document as bag of words:** u_i = number of occurrences of word i. Compare documents using cosine similarity.

  n = lexicon size; most documents aren't dictionaries
- **Network traffic:** $u_{i,j}$ = #bytes sent from i to j

  $n = 2^{64}$ (2^{256} in IPv6); most servers don't talk to each other
- **User ratings:** $u_{i,j}$ is user i's score for movie j on Netflix

  $n = \#\text{movies}$; most people haven't rated all movies
- **Streaming:** u receives a stream of updates of the form: "add v to u_i". Maintaining Πu requires calculating $v \cdot \Pi e_i$.
- ...

Sparse JL transforms

One way to embed sparse vectors faster: use sparse matrices.

Sparse JL transforms

One way to embed sparse vectors faster: use sparse matrices.

$s = \#$non-zero entries per column in Π
(so embedding time is $s \cdot \|u\|_0$)

reference	value of s	type
[JL84], [FM88], [IM98], ...	$m \approx 4\varepsilon^{-2} \ln(1/\delta)$	dense
[Achlioptas01]	$m/3$	sparse Bernoulli
[WDALS09]	no proof	hashing
[DKS10]	$\tilde{O}(\varepsilon^{-1} \log^3(1/\delta))$	hashing
[KN10a], [BOR10]	$\tilde{O}(\varepsilon^{-1} \log^2(1/\delta))$	"
[KN12]	$O(\varepsilon^{-1} \log(1/\delta))$	modified hashing

Sparse JL transforms

One way to embed sparse vectors faster: use sparse matrices.

$s = \#$non-zero entries per column in Π
(so embedding time is $s \cdot \|u\|_0$)

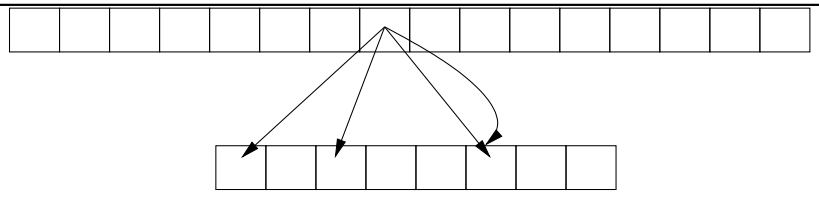
reference	value of s	type
[JL84], [FM88], [IM98], ...	$m \approx 4\varepsilon^{-2} \ln(1/\delta)$	dense
[Achlioptas01]	$m/3$	sparse Bernoulli
[WDALS09]	no proof	hashing
[DKS10]	$\tilde{O}(\varepsilon^{-1} \log^3(1/\delta))$	hashing
[KN10a], [BOR10]	$\tilde{O}(\varepsilon^{-1} \log^2(1/\delta))$	"
[KN12]	$O(\varepsilon^{-1} \log(1/\delta))$	modified hashing

[N., Nguyễn '13]: for any $m \leq \mathrm{poly}(1/\varepsilon) \cdot \log N$, $s = \Omega(\varepsilon^{-1} \log N / \log(1/\varepsilon))$ is required, even for metric JL, so [KN12] is optimal up to $O(\log(1/\varepsilon))$.

*[Thorup, Zhang '04] gives $m = O(\varepsilon^{-2}\delta^{-1}), s = 1$.

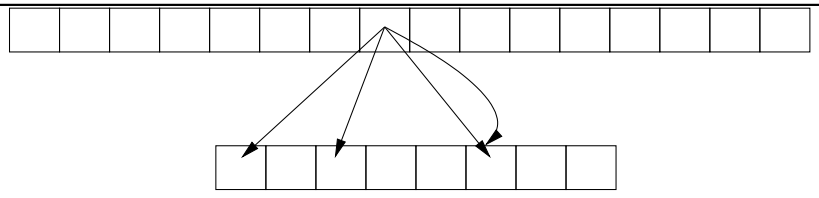
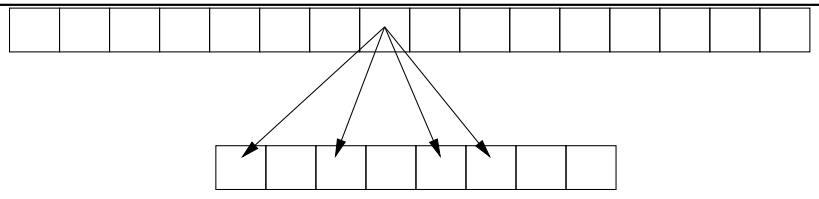
Sparse JL Constructions

[DKS, 2010]

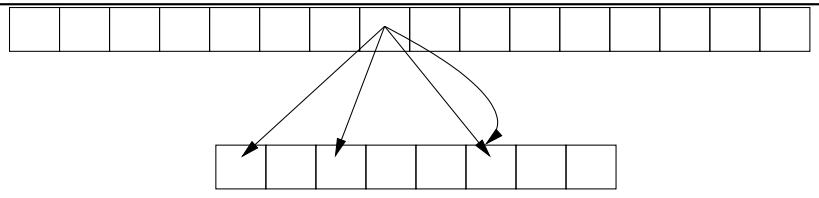
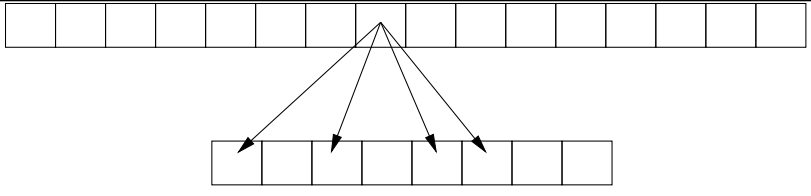
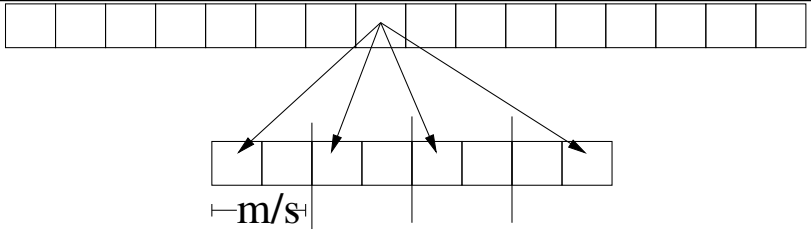


$s = \tilde{\Theta}(\varepsilon^{-1} \log^2(1/\delta))$

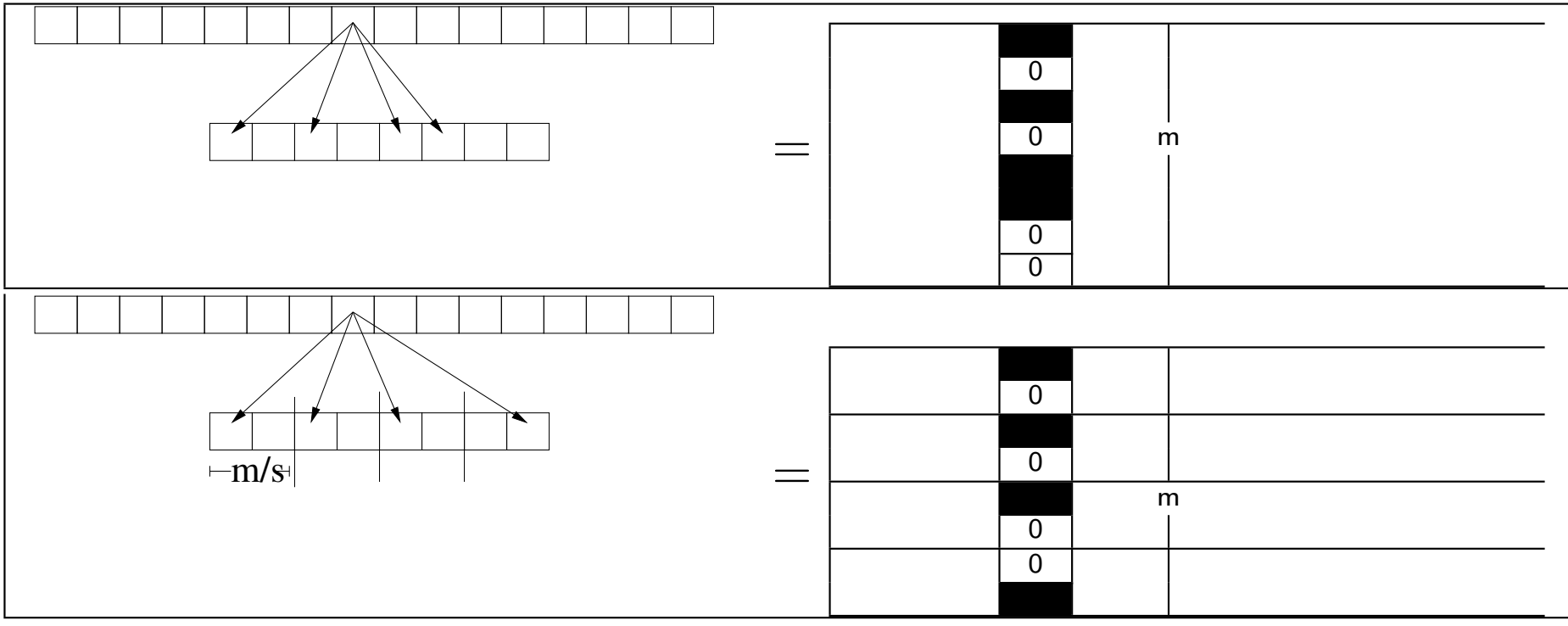
Sparse JL Constructions

[DKS, 2010] 	$s = \tilde{\Theta}(\varepsilon^{-1} \log^2(1/\delta))$
[KN12] 	$s = \Theta(\varepsilon^{-1} \log(1/\delta))$

Sparse JL Constructions

[DKS, 2010] 	$s = \tilde{\Theta}(\varepsilon^{-1} \log^2(1/\delta))$
[KN12] 	$s = \Theta(\varepsilon^{-1} \log(1/\delta))$
[KN12] 	$s = \Theta(\varepsilon^{-1} \log(1/\delta))$

Sparse JL Constructions (in matrix form)



Each black cell is $\pm 1/\sqrt{s}$ at random

One open problem ...

- [Achlioptas '01]:

$$\Pi_{i,j} = \begin{cases} +1/\sqrt{m/3}, & \text{w.p. } 1/6 \\ -1/\sqrt{m/3}, & \text{w.p. } 1/6 \\ 0, & \text{w.p. } 2/3 \end{cases}$$

- There is an expected $s = m/3$ non-zeroes per column.

One open problem ...

- [Achlioptas '01]:

$$\Pi_{i,j} = \begin{cases} +1/\sqrt{m/3}, & \text{w.p. } 1/6 \\ -1/\sqrt{m/3}, & \text{w.p. } 1/6 \\ 0, & \text{w.p. } 2/3 \end{cases}$$

- There is an expected $s = m/3$ non-zeroes per column.
- [Kane-N. '12]: Eliminate the variance in the number of non-zeroes per column. Force it to be *exactly* s.

  Surprisingly, **this helps**!

  (makes asymptotically smaller $s = O(\varepsilon m)$ possible)

One open problem ...

- [Achlioptas '01]:

$$\Pi_{i,j} = \begin{cases} +1/\sqrt{m/3}, & \text{w.p. } 1/6 \\ -1/\sqrt{m/3}, & \text{w.p. } 1/6 \\ 0, & \text{w.p. } 2/3 \end{cases}$$

- There is an expected $s = m/3$ non-zeroes per column.
- [Kane-N. '12]: Eliminate the variance in the number of non-zeroes per column. Force it to be *exactly* s.

  Surprisingly, **this helps**!

  (makes asymptotically smaller $s = O(\varepsilon m)$ possible)
- **Conjecture:** Error with exactly s non-zeroes per column is stochastically dominated by error with expected s per column.

Analysis

- In both constructions, can write $\Pi_{i,j} = \delta_{i,j}\sigma_{i,j}/\sqrt{s}$

$$\|\Pi u\|_2^2 - 1 = \frac{1}{s}\sum_{r=1}^{m}\sum_{i\neq j}\delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}u_i u_j = \sigma^T B\sigma$$

Analysis

- In both constructions, can write $\Pi_{i,j} = \delta_{i,j}\sigma_{i,j}/\sqrt{s}$

$$\|\Pi u\|_2^2 - 1 = \frac{1}{s}\sum_{r=1}^{m}\sum_{i\neq j}\delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}u_i u_j = \sigma^T B\sigma$$

$$B = \frac{1}{s}\cdot \begin{array}{|c|c|c|c|} \hline B_1 & 0 & \dots & 0 \\ \hline 0 & B_2 & \dots & 0 \\ \hline & & \ddots & \\ 0 & 0 & & 0 \\ \hline 0 & \dots & 0 & B_m \\ \hline \end{array}$$

- $(B_r)_{i,j} = \delta_{r,i}\delta_{r,j}u_i u_j$

Analysis

- In both constructions, can write $\Pi_{i,j} = \delta_{i,j}\sigma_{i,j}/\sqrt{s}$

$$\|\Pi u\|_2^2 - 1 = \frac{1}{s}\sum_{r=1}^{m}\sum_{i\neq j}\delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}u_i u_j = \sigma^T B\sigma$$

$$B = \frac{1}{s}\cdot\begin{array}{|c|c|c|c|}\hline B_1 & 0 & \dots & 0 \\ \hline 0 & B_2 & \dots & 0 \\ \hline & & \ddots & \\ 0 & 0 & & 0 \\ \hline 0 & \dots & 0 & B_m \\ \hline\end{array}$$

- $(B_r)_{i,j} = \delta_{r,i}\delta_{r,j}u_i u_j$
- $\mathbb{P}(|\|\Pi u\|^2 - 1| > \varepsilon) = \mathbb{P}(|\sigma^T B\sigma| > \varepsilon) < \varepsilon^{-\ell}\cdot\mathbb{E}\,|\sigma^T B\sigma|^\ell$.

  Use moment bound for quadratic forms, which depends on $\|B\|, \|B\|_F$ (Hanson-Wright inequality).

What next?

Natural "matrix extension" of sparse JL

[Kane, N. '12]

Theorem

Let $u \in \mathbb{R}^n$ be arbitrary, unit ℓ_2 norm, Π sparse sign matrix. Then

$$\mathbb{P}_{\Pi}\left(\left|\|\Pi u\|^2 - 1\right| > \varepsilon\right) < \delta$$

as long as

$$m \gtrsim \frac{\log(1/\delta)}{\varepsilon^2}, s \gtrsim \frac{\log(1/\delta)}{\varepsilon}, \ell = \log(1/\delta)$$

or

$$m \gtrsim \frac{1}{\varepsilon^2 \delta}, s = 1, \ell = 2 \ ([\text{Thorup, Zhang}'04])$$

Natural "matrix extension" of sparse JL

[Kane, N. '12]

Theorem

Let $u \in \mathbb{R}^{n\times 1}$ be arbitrary, o.n. cols, Π sparse sign matrix. Then

$$\mathbb{P}_{\Pi}(\|(\Pi u)^T(\Pi u) - I_1\| > \varepsilon) < \delta$$

as long as

$$m \gtrsim \frac{1+\log(1/\delta)}{\varepsilon^2}, s \gtrsim \frac{\log(1/\delta)}{\varepsilon}, \ell = \log(1/\delta)$$

or

$$m \gtrsim \frac{1^2}{\varepsilon^2\delta}, s = 1, \ell = 2$$

Natural "matrix extension" of sparse JL

[Kane, N. '12]

Theorem

Let $U \in \mathbb{R}^{n\times 1}$ be arbitrary, o.n. cols, Π sparse sign matrix. Then

$$\mathop{\mathbb{P}}_{\Pi}(\|(\Pi U)^T(\Pi U) - I_1\| > \varepsilon) < \delta$$

as long as

$$m \gtrsim \frac{1+\log(1/\delta)}{\varepsilon^2}, s \gtrsim \frac{\log(1/\delta)}{\varepsilon}, \ell = \log(1/\delta)$$

or

$$m \gtrsim \frac{1^2}{\varepsilon^2\delta}, s = 1, \ell = 2$$

Natural "matrix extension" of sparse JL

Conjecture

Theorem

Let $U \in \mathbb{R}^{n \times d}$ be arbitrary, o.n. cols, Π sparse sign matrix. Then

$$\mathop{\mathbb{P}}_{\Pi}(\|(\Pi U)^T(\Pi U) - I_d\| > \varepsilon) < \delta$$

as long as

$$m \gtrsim \frac{d + \log(1/\delta)}{\varepsilon^2}, s \gtrsim \frac{\log(d/\delta)}{\varepsilon}, \ell = \log(d/\delta)$$

or

$$m \gtrsim \frac{d^2}{\varepsilon^2 \delta}, s = 1, \ell = 2$$

Natural "matrix extension" of sparse JL

What we prove [N., Nguyễn '13]

Theorem

Let $U \in \mathbb{R}^{n \times d}$ be arbitrary, o.n. cols, Π sparse sign matrix. Then

$$\mathop{\mathbb{P}}_{\Pi}(\|(\Pi U)^T(\Pi U) - I_d\| > \varepsilon) < \delta$$

as long as

$$m \gtrsim \frac{d \cdot \log^c(d/\delta)}{\varepsilon^2}, s \gtrsim \frac{\log^c(d/\delta)}{\varepsilon} \text{ or } m \gtrsim \frac{d^{1.01}}{\varepsilon^2}, s \gtrsim \frac{1}{\varepsilon}$$

or

$$m \gtrsim \frac{d^2}{\varepsilon^2 \delta}, s = 1$$

Remarks

- [Clarkson, Woodruff '13] was first to show $m = d^2 \cdot \mathrm{polylog}(d/\varepsilon)/\varepsilon^2, s = 1$ bound via other methods
- $m = O(d^2/\varepsilon^2), s = 1$ also obtained by [Mahoney, Meng '13].
- $m = O(d^2/\varepsilon^2), s = 1$ also follows from [Thorup, Zhang '04] + [Kane, N. '12] (observed by Nguyễn)

Remarks

- [Clarkson, Woodruff '13] was first to show $m = d^2 \cdot \mathrm{polylog}(d/\varepsilon)/\varepsilon^2, s = 1$ bound via other methods
- $m = O(d^2/\varepsilon^2), s = 1$ also obtained by [Mahoney, Meng '13].
- $m = O(d^2/\varepsilon^2), s = 1$ also follows from [Thorup, Zhang '04] + [Kane, N. '12] (observed by Nguyễn)
- What does the "moment method" mean for matrices?

$$\mathbb{P}_{\Pi}(\|(\Pi U)^T(\Pi U) - I_d\| > \varepsilon) < \varepsilon^{-\ell} \cdot \mathbb{E}\,\|(\Pi U)^T(\Pi U) - I_d\|^{\ell}$$

$$\leq \varepsilon^{-\ell} \cdot \mathbb{E}\,\mathrm{tr}(((\Pi U)^T(\Pi U) - I_d)^{\ell})$$

- Classical "moment method" in random matrix theory; e.g. [Wigner, 1955], [Füredi, Komlós, 1981], [Bai, Yin, 1993]

Who cares about this matrix extension?

Motivation for matrix extension of sparse JL

- $\|(\Pi U)^T(\Pi U) - I\| \leq \varepsilon$ equivalent to $\|\Pi x\| = (1 \pm \varepsilon)\|x\|$ for all $x \in V$, where V is the subspace spanned by the columns of U (up to changing ε by a factor of 2). "subspace embedding".

Motivation for matrix extension of sparse JL

- $\|(\Pi U)^T(\Pi U) - I\| \leq \varepsilon$ equivalent to $\|\Pi x\| = (1 \pm \varepsilon)\|x\|$ for all $x \in V$, where V is the subspace spanned by the columns of U (up to changing ε by a factor of 2). "subspace embedding".
- Subspace embeddings can be used to speed up algorithms for many numerical linear algebra problems on big matrices [Sarlós, 2006], [Dasgupta, Drineas, Harb, Kumar, Mahoney, 2008], [Clarkson, Woodruff, 2009], [Drineas, Magdon-Ismail, Mahoney, Woodruff, 2012], [Clarkson, Woodruff, 2013], [Clarkson, Drineas, Magdon-Ismail, Mahoney, Meng, Woodruff, 2013], [Woodruff, Zhang, 2013], ...

Motivation for matrix extension of sparse JL

- $\|(\Pi U)^T(\Pi U) - I\| \leq \varepsilon$ equivalent to $\|\Pi x\| = (1 \pm \varepsilon)\|x\|$ for all $x \in V$, where V is the subspace spanned by the columns of U (up to changing ε by a factor of 2). "subspace embedding".
- Subspace embeddings can be used to speed up algorithms for many numerical linear algebra problems on big matrices [Sarlós, 2006], [Dasgupta, Drineas, Harb, Kumar, Mahoney, 2008], [Clarkson, Woodruff, 2009], [Drineas, Magdon-Ismail, Mahoney, Woodruff, 2012], [Clarkson, Woodruff, 2013], [Clarkson, Drineas, Magdon-Ismail, Mahoney, Meng, Woodruff, 2013], [Woodruff, Zhang, 2013], ...
- Sparse Π: can multiply ΠA in $s \cdot \text{nnz}(A)$ time for big matrix A.

Numerical linear algebra

- $A \in \mathbb{R}^{n \times d}$, $n \gg d$, $rank(A) = r$

Numerical linear algebra

- $A \in \mathbb{R}^{n \times d}$, $n \gg d$, $rank(A) = r$

Classical numerical linear algebra problems

- Compute the **leverage scores** of A, i.e. the ℓ_2 norms of the n standard basis vectors when projected onto the subspace spanned by the columns of A.

Numerical linear algebra

- $A \in \mathbb{R}^{n \times d}$, $n \gg d$, $rank(A) = r$

Classical numerical linear algebra problems

- Compute the **leverage scores** of A, i.e. the ℓ_2 norms of the n standard basis vectors when projected onto the subspace spanned by the columns of A.
- **Least squares regression**: Given also $b \in \mathbb{R}^n$.

$$\text{Compute } x^* = \text{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_2$$

Numerical linear algebra

- $A \in \mathbb{R}^{n \times d}$, $n \gg d$, $rank(A) = r$

Classical numerical linear algebra problems

- Compute the **leverage scores** of A, i.e. the ℓ_2 norms of the n standard basis vectors when projected onto the subspace spanned by the columns of A.
- **Least squares regression**: Given also $b \in \mathbb{R}^n$.

$$\text{Compute } x^* = \text{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_2$$

- ℓ_p **regression ($p \in [1, \infty)$)**:

$$\text{Compute } x^* = \text{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_p$$

Numerical linear algebra

- $A \in \mathbb{R}^{n \times d}$, $n \gg d$, $rank(A) = r$

Classical numerical linear algebra problems

- Compute the **leverage scores** of A, i.e. the ℓ_2 norms of the n standard basis vectors when projected onto the subspace spanned by the columns of A.
- **Least squares regression**: Given also $b \in \mathbb{R}^n$.

$$\text{Compute } x^* = \text{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_2$$

- ℓ_p **regression ($p \in [1, \infty)$)**:

$$\text{Compute } x^* = \text{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_p$$

- **Low-rank approximation**: Given also an integer $1 \leq k \leq d$.

$$\text{Compute } A_k = \text{argmin}_{\text{rank}(B) \leq k} \|A - B\|_F$$

Numerical linear algebra

- $A \in \mathbb{R}^{n \times d}$, $n \gg d$, $rank(A) = r$

Classical numerical linear algebra problems

- Compute the **leverage scores** of A, i.e. the ℓ_2 norms of the n standard basis vectors when projected onto the subspace spanned by the columns of A.
- **Least squares regression**: Given also $b \in \mathbb{R}^n$.

$$\text{Compute } x^* = \text{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_2$$

- ℓ_p **regression ($p \in [1, \infty)$)**:

$$\text{Compute } x^* = \text{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_p$$

- **Low-rank approximation**: Given also an integer $1 \leq k \leq d$.

$$\text{Compute } A_k = \text{argmin}_{\text{rank}(B) \leq k} \|A - B\|_F$$

- **Preconditioning**: Compute $R \in \mathbb{R}^{d \times d}$ (for $d = r$) so that

$$\forall x \ \|ARx\|_2 \approx \|x\|_2$$

Computationally efficient solutions

Singular Value Decomposition

Theorem

Every matrix $A \in \mathbb{R}^{n \times d}$ of rank r can be written as

$$A = \underbrace{U}_{\substack{\text{orthonorm} \\ \text{columns} \\ n \times r}} \quad \underbrace{\Sigma}_{\substack{\text{diagonal} \\ \text{positive definite} \\ r \times r}} \quad \underbrace{V^T}_{\substack{\text{orthonorm} \\ \text{columns} \\ d \times r}}$$

Can compute SVD in $\tilde{O}(nd^{\omega-1})$ [Demmel, Dumitriu, Holtz, 2007].
$\omega < 2.373\ldots$ is the exponent of square matrix multiplication [Coppersmith, Winograd, 1987], [Stothers, 2010], [Vassilevska-Williams, 2012]

Computationally efficient solutions

$$A = \underbrace{U}_{\substack{\text{orthonorm} \\ \text{columns} \\ n \times r}} \quad \underbrace{\Sigma}_{\substack{\text{diagonal} \\ \text{positive definite} \\ r \times r}} \quad \underbrace{V^T}_{\substack{\text{orthonorm} \\ \text{columns} \\ d \times r}}$$

- **Leverage scores**: Output row norms of U.
- **Least squares regression**: Output $V\Sigma^{-1}U^T b$.
- **Low-rank approximation**: Output $U\Sigma_k V^T$.
- **Preconditioning**: Output $R = V\Sigma^{-1}$.

Computationally efficient solutions

$$A = \underbrace{U}_{\substack{\text{orthonorm} \\ \text{columns} \\ n \times r}} \quad \underbrace{\Sigma}_{\substack{\text{diagonal} \\ \text{positive definite} \\ r \times r}} \quad \underbrace{V^T}_{\substack{\text{orthonorm} \\ \text{columns} \\ d \times r}}$$

- **Leverage scores**: Output row norms of U.
- **Least squares regression**: Output $V\Sigma^{-1}U^T b$.
- **Low-rank approximation**: Output $U\Sigma_k V^T$.
- **Preconditioning**: Output $R = V\Sigma^{-1}$.

Conclusion: In time $\tilde{O}(nd^{\omega-1})$ we can compute the SVD then solve all the previously stated problems. Is there a faster way?

How to use subspace embeddings

Least squares regression: Let Π be a subspace embedding for the subspace spanned by b and the columns of A. Let $x^* = \operatorname{argmin} \|Ax - b\|$ and $\tilde{x} = \operatorname{argmin} \|\Pi Ax - \Pi b\|$. Then

How to use subspace embeddings

Least squares regression: Let Π be a subspace embedding for the subspace spanned by b and the columns of A. Let $x^* = \operatorname{argmin} \|Ax - b\|$ and $\tilde{x} = \operatorname{argmin} \|\Pi Ax - \Pi b\|$. Then

$$\|\Pi A\tilde{x} - \Pi b\| \leq \|\Pi Ax^* - \Pi b\|$$

How to use subspace embeddings

Least squares regression: Let Π be a subspace embedding for the subspace spanned by b and the columns of A. Let $x^* = \operatorname{argmin} \|Ax - b\|$ and $\tilde{x} = \operatorname{argmin} \|\Pi Ax - \Pi b\|$. Then

$$(1-\varepsilon)\|A\tilde{x}-b\| \leq \underbrace{\|\Pi A\tilde{x} - \Pi b\|}_{\|\Pi(A\tilde{x}-b)\|} \leq \|\Pi Ax^* - \Pi b\|$$

How to use subspace embeddings

Least squares regression: Let Π be a subspace embedding for the subspace spanned by b and the columns of A. Let $x^* = \operatorname{argmin} \|Ax - b\|$ and $\tilde{x} = \operatorname{argmin} \|\Pi Ax - \Pi b\|$. Then

$$(1-\varepsilon)\|A\tilde{x}-b\| \leq \|\Pi A\tilde{x}-\Pi b\| \leq \|\Pi Ax^*-\Pi b\| \leq (1+\varepsilon)\|Ax^*-b\|$$

$$\Rightarrow \|A\tilde{x} - b\| \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right) \cdot \|Ax^* - b\|$$

How to use subspace embeddings

Least squares regression: Let Π be a subspace embedding for the subspace spanned by b and the columns of A. Let $x^* = \operatorname{argmin} \|Ax - b\|$ and $\tilde{x} = \operatorname{argmin} \|\Pi Ax - \Pi b\|$. Then

$$(1-\varepsilon)\|A\tilde{x}-b\| \leq \|\Pi A\tilde{x}-\Pi b\| \leq \|\Pi Ax^*-\Pi b\| \leq (1+\varepsilon)\|Ax^*-b\|$$

$$\Rightarrow \|A\tilde{x} - b\| \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right) \cdot \|Ax^* - b\|$$

Computing SVD of ΠA takes time $\tilde{O}(md^{\omega-1})$, which is much faster than $\tilde{O}(nd^{\omega-1})$ since $m \ll n$.

How to use subspace embeddings

Least squares regression: Let Π be a subspace embedding for the subspace spanned by b and the columns of A. Let $x^* = \operatorname{argmin} \|Ax - b\|$ and $\tilde{x} = \operatorname{argmin} \|\Pi Ax - \Pi b\|$. Then

$$(1-\varepsilon)\|A\tilde{x}-b\| \leq \|\Pi A\tilde{x}-\Pi b\| \leq \|\Pi Ax^*-\Pi b\| \leq (1+\varepsilon)\|Ax^*-b\|$$

$$\Rightarrow \|A\tilde{x} - b\| \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right) \cdot \|Ax^* - b\|$$

Computing SVD of ΠA takes time $\tilde{O}(md^{\omega-1})$, which is much faster than $\tilde{O}(nd^{\omega-1})$ since $m \ll n$.

Note: [Sarlós '06] actually describes a more efficient reduction, where one only needs that (1) Π is a $(1 - 1/\sqrt{2})$-subspace embedding, and (2) $\|(\Pi U)^T(\Pi(Ax^* - b))\|_2 < \sqrt{\varepsilon} \cdot \|Ax^* - b\|$. Item (2) only requires $m \gtrsim d/\varepsilon, s \geq 1$ [Thorup, Zhang '04] + [Kane, N. '12].

Back to the analysis

$$\mathbb{P}_{\Pi}\left(\left\|(\Pi U)^T(\Pi U) - I_d\right\| > \varepsilon\right) < \varepsilon^{-\ell} \cdot \mathbb{E}\,\mathrm{tr}(((\Pi U)^T(\Pi U) - I_d)^{\ell})$$

Analysis ($\ell = 2$)

$s = 1$, $m = O(d^2/\varepsilon^2)$

Want to understand $S - I$, $S = (\Pi U)^T(\Pi U)$

Analysis ($\ell = 2$)

$s = 1$, $m = O(d^2/\varepsilon^2)$

Want to understand $S - I$, $S = (\Pi U)^T(\Pi U)$

Let the columns of U be $u^1, \ldots, u^d$

Recall $\Pi_{i,j} = \delta_{i,j}\sigma_{i,j}/\sqrt{s}$

Analysis ($\ell = 2$)
$s = 1$, $m = O(d^2/\varepsilon^2)$

Want to understand $S - I$, $S = (\Pi U)^T(\Pi U)$

Let the columns of U be $u^1, \ldots, u^d$
Recall $\Pi_{i,j} = \delta_{i,j}\sigma_{i,j}/\sqrt{s}$

Some computations yield

$$(S - I)_{k,k'} = \frac{1}{s}\sum_{r=1}^{m}\sum_{i\neq j}\delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}u_i^k u_j^{k'}$$

Analysis ($\ell = 2$)
$s = 1$, $m = O(d^2/\varepsilon^2)$

Want to understand $S - I$, $S = (\Pi U)^T(\Pi U)$

Let the columns of U be $u^1, \ldots, u^d$
Recall $\Pi_{i,j} = \delta_{i,j}\sigma_{i,j}/\sqrt{s}$

Some computations yield

$$(S - I)_{k,k'} = \frac{1}{s}\sum_{r=1}^{m}\sum_{i\neq j}\delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}u_i^k u_j^{k'}$$

Computing $\mathbb{E}\,\mathrm{tr}((S - I)^2) = \mathbb{E}\,\|S - I\|_F^2$ is straightforward, and can show $\mathbb{E}\,\|S - I\|_F^2 \leq (d^2 + d)/m$

$$\mathbb{P}(\|S - I\| > \varepsilon) < \frac{1}{\varepsilon^2}\frac{d^2 + d}{m}$$

Analysis ($\ell = 2$)
$s = 1$, $m = O(d^2/\varepsilon^2)$

Want to understand $S - I$, $S = (\Pi U)^T(\Pi U)$

Let the columns of U be $u^1, \ldots, u^d$
Recall $\Pi_{i,j} = \delta_{i,j}\sigma_{i,j}/\sqrt{s}$

Some computations yield

$$(S - I)_{k,k'} = \frac{1}{s}\sum_{r=1}^{m}\sum_{i\neq j}\delta_{r,i}\delta_{r,j}\sigma_{r,i}\sigma_{r,j}u_i^k u_j^{k'}$$

Computing $\mathbb{E}\,\mathrm{tr}((S - I)^2) = \mathbb{E}\,\|S - I\|_F^2$ is straightforward, and can show $\mathbb{E}\,\|S - I\|_F^2 \leq (d^2 + d)/m$

$$\mathbb{P}(\|S - I\| > \varepsilon) < \frac{1}{\varepsilon^2}\frac{d^2 + d}{m}$$

Set $m \geq \delta^{-1}(d^2 + d)/\varepsilon^2$ for success probability $1 - \delta$

Analysis (large ℓ)

$s = O_\gamma(1/\varepsilon)$, $m = O(d^{1+\gamma}/\varepsilon^2)$

$$(S - I)_{k,k'} = \frac{1}{s} \sum_{r=1}^{m} \sum_{i \neq j} \delta_{r,i} \delta_{r,j} \sigma_{r,i} \sigma_{r,j} u_i^k u_j^{k'}$$

Analysis (large ℓ)
$s = O_\gamma(1/\varepsilon)$, $m = O(d^{1+\gamma}/\varepsilon^2)$

$$(S - I)_{k,k'} = \frac{1}{s} \sum_{r=1}^{m} \sum_{i \neq j} \delta_{r,i} \delta_{r,j} \sigma_{r,i} \sigma_{r,j} u_i^k u_j^{k'}$$

By induction, for any square matrix B and integer $\ell \geq 1$,

$$(B^\ell)_{i,j} = \sum_{\substack{i_1, \ldots, i_{\ell+1} \\ i_1 = i, i_{\ell+1} = j}} \prod_{t=1}^{\ell} B_{i_t, i_{t+1}}$$

Analysis (large ℓ)
$s = O_\gamma(1/\varepsilon)$, $m = O(d^{1+\gamma}/\varepsilon^2)$

$$(S - I)_{k,k'} = \frac{1}{s} \sum_{r=1}^{m} \sum_{i \neq j} \delta_{r,i} \delta_{r,j} \sigma_{r,i} \sigma_{r,j} u_i^k u_j^{k'}$$

By induction, for any square matrix B and integer $\ell \geq 1$,

$$(B^\ell)_{i,j} = \sum_{\substack{i_1,\ldots,i_{\ell+1} \\ i_1 = i, i_{\ell+1} = j}} \prod_{t=1}^{\ell} B_{i_t, i_{t+1}}$$

$$\Rightarrow \operatorname{tr}(B^\ell) = \sum_{\substack{i_1,\ldots,i_{\ell+1} \\ i_1 = i_{\ell+1}}} \prod_{t=1}^{\ell} B_{i_t, i_{t+1}}$$

Analysis (large ℓ)
$s = O_\gamma(1/\varepsilon)$, $m = O(d^{1+\gamma}/\varepsilon^2)$

$$\mathbb{E}\,\mathrm{tr}((S-I)^\ell) = \sum_{\substack{i_1 \neq j_1, \ldots, i_\ell \neq j_\ell \\ r_1, \ldots, r_\ell \\ k_1, \ldots, k_{\ell+1} \\ k_1 = k_{\ell+1}}} \left(\mathbb{E} \prod_{t=1}^{\ell} \delta_{r_t, i_t} \delta_{r_t, j_t} \right) \left(\mathbb{E} \prod_{t=1}^{\ell} \sigma_{r_t, i_t} \sigma_{r_t, j_t} \right) \prod_{t=1}^{\ell} u_{i_t}^{k_t} u_{j_t}^{k_{t+1}}$$

Analysis (large ℓ)
$s = O_\gamma(1/\varepsilon)$, $m = O(d^{1+\gamma}/\varepsilon^2)$

$$\mathbb{E}\,\mathrm{tr}((S-I)^\ell) = \sum_{\substack{i_1 \neq j_1, \ldots, i_\ell \neq j_\ell \\ r_1, \ldots, r_\ell \\ k_1, \ldots, k_{\ell+1} \\ k_1 = k_{\ell+1}}} \left(\mathbb{E} \prod_{t=1}^{\ell} \delta_{r_t, i_t} \delta_{r_t, j_t} \right) \left(\mathbb{E} \prod_{t=1}^{\ell} \sigma_{r_t, i_t} \sigma_{r_t, j_t} \right) \prod_{t=1}^{\ell} u_{i_t}^{k_t} u_{j_t}^{k_{t+1}}$$

The strategy: Associate each monomial in summation above with a graph, group monomials that have the same graph, and estimate the contribution of each graph then do some combinatorics

(a common strategy; see [Wigner, 1955], [Füredi, Komlós, 1981], [Bai, Yin, 1993])

Finite pointsets, then subspaces, now what?

(Linear) dimensionality reduction

- Given $T \subset \mathbb{R}^n$, want a $\Pi \in \mathbb{R}^{m \times n}$ such that

$$\forall u \in T,\ (1-\varepsilon)\|u\|_2 \leq \|\Pi u\|_2 \leq (1+\varepsilon)\|u\|_2$$

  with m as small as possible.

(Linear) dimensionality reduction

- Given $T \subset \mathbb{R}^n$, want a $\Pi \in \mathbb{R}^{m \times n}$ such that

$$\forall u \in T,\ (1-\varepsilon)\|u\|_2 \leq \|\Pi u\|_2 \leq (1+\varepsilon)\|u\|_2$$

  with m as small as possible.
- Above is equivalent to, assuming $T \subset S^{n-1}$,

$$\sup_{u \in T} \left| \|\Pi u\|_2^2 - 1 \right| < \varepsilon$$

(Linear) dimensionality reduction

- Given $T \subset \mathbb{R}^n$, want a $\Pi \in \mathbb{R}^{m \times n}$ such that

$$\forall u \in T,\ (1-\varepsilon)\|u\|_2 \leq \|\Pi u\|_2 \leq (1+\varepsilon)\|u\|_2$$

  with m as small as possible.
- Above is equivalent to, assuming $T \subset S^{n-1}$,

$$\sup_{u \in T} |\|\Pi u\|_2^2 - 1| < \varepsilon$$

- We consider Π random:

$$\mathop{\mathbb{E}}_{\Pi} \sup_{u \in T} |\|\Pi u\|_2^2 - 1| < \varepsilon$$

Applications

- **Finite T:** *Johnson Lindenstrauss (JL) lemma* has $T = \{(u_i - u_j)/\|u_i - u_j\|_2\}_{1 \le i < j \le N}$ [JL84].

  Applications to clustering, nearest neighbor search, ...

Applications

- **Finite T:** *Johnson Lindenstrauss (JL) lemma* has $T = \{(u_i - u_j)/\|u_i - u_j\|_2\}_{1 \le i < j \le N}$ [JL84].

  Applications to clustering, nearest neighbor search, ...
- **Linear subspace:** $T = E \cap S^{n-1}$, $E \subset \mathbb{R}^n$, $\dim(E) = d$.

  Subspace embeddings pioneered by [Sarlós '06]; faster least squares regression and low-rank approximation.

Applications

- **Finite T:** *Johnson Lindenstrauss (JL) lemma* has $T = \{(u_i - u_j)/\|u_i - u_j\|_2\}_{1 \le i < j \le N}$ [JL84].

  Applications to clustering, nearest neighbor search, ...
- **Linear subspace:** $T = E \cap S^{n-1}$, $E \subset \mathbb{R}^n$, $\dim(E) = d$.

  Subspace embeddings pioneered by [Sarlós '06]; faster least squares regression and low-rank approximation.

  Also applications to approximating leverage scores [DMMW'12], k-means clustering [BZMD'11], canonical correlation analysis [ABTZ'13], support vector machines [PBMD'13], ℓ_p regression [CDM+13,WZ13], ridge regression [LDFU'13].

Applications

- **Union of subspaces:** streaming approx of eigenvals [Andoni, Nguyễn'13], compressed sensing [Donoho'04, Candès-Tao'06]

Applications

- **Union of subspaces:** streaming approx of eigenvals [Andoni, Nguyễn'13], compressed sensing [Donoho'04, Candès-Tao'06]

  e.g. compressed sensing: $T = \{u \in \mathbb{R}^n : \|u\|_2 = 1, \|u\|_0 \leq k\}$, $\|u\|_0$ is support size. Can also be sparse over some other basis.

Applications

- **Union of subspaces:** streaming approx of eigenvals [Andoni, Nguyễn'13], compressed sensing [Donoho'04, Candès-Tao'06]

  e.g. compressed sensing: $T = \{u \in \mathbb{R}^n : \|u\|_2 = 1, \|u\|_0 \leq k\}$, $\|u\|_0$ is support size. Can also be sparse over some other basis.
- This is a union of $\binom{n}{k}$ subspaces

Applications

- **Union of subspaces:** streaming approx of eigenvals [Andoni, Nguyễn'13], compressed sensing [Donoho'04, Candès-Tao'06]

  e.g. compressed sensing: $T = \{u \in \mathbb{R}^n : \|u\|_2 = 1, \|u\|_0 \leq k\}$, $\|u\|_0$ is support size. Can also be sparse over some other basis.
- This is a union of $\binom{n}{k}$ subspaces
- Π for this T known as having *restricted isometry property (RIP)*. Given Πu for (near) sparse u, can (approximately) recover u in polynomial time [Candès, Tao'05]

Applications

- **Union of subspaces:** streaming approx of eigenvals [Andoni, Nguyễn'13], compressed sensing [Donoho'04, Candès-Tao'06]

  e.g. compressed sensing: $T = \{u \in \mathbb{R}^n : \|u\|_2 = 1, \|u\|_0 \leq k\}$, $\|u\|_0$ is support size. Can also be sparse over some other basis.
- This is a union of $\binom{n}{k}$ subspaces
- Π for this T known as having *restricted isometry property (RIP)*. Given Πu for (near) sparse u, can (approximately) recover u in polynomial time [Candès, Tao'05]

  Supporting $\ll \binom{n}{k}$ sparsity patterns also of interest ("model-based compressed sensing" [BCDH'10])

Applications

- **Smooth manifolds:** $\mathcal{M} = F(B_{\ell_2^d})$, for smooth $F : B_{\ell_2^d} \to \mathbb{R}^n$

Applications

- **Smooth manifolds:** $\mathcal{M} = F(B_{\ell_2^d})$, for smooth $F : B_{\ell_2^d} \to \mathbb{R}^n$
  Manifold learning: classifying images of handwritten digits [Hinton-Dayan-Revow'97], or human faces [Broomhead-Kirby'01]

Applications

- **Smooth manifolds:** $\mathcal{M} = F(B_{\ell_2^d})$, for smooth $F : B_{\ell_2^d} \to \mathbb{R}^n$
  Manifold learning: classifying images of handwritten digits [Hinton-Dayan-Revow'97], or human faces [Broomhead-Kirby'01]
- Idea: All valid drawings of the digit "2" lie on a low-dimensional manifold in $\mathbb{R}^{\sqrt{n}\times\sqrt{n}}$ (pixelated image). From samples, learn the manifold.
- [Baraniuk-Wakin'09] suggest reducing dimensionality using Π and doing learning in lower-dimensional space

Applications

- **Smooth manifolds:** $\mathcal{M} = F(B_{\ell_2^d})$, for smooth $F : B_{\ell_2^d} \to \mathbb{R}^n$
  Manifold learning: classifying images of handwritten digits [Hinton-Dayan-Revow'97], or human faces [Broomhead-Kirby'01]
- Idea: All valid drawings of the digit "2" lie on a low-dimensional manifold in $\mathbb{R}^{\sqrt{n}\times\sqrt{n}}$ (pixelated image). From samples, learn the manifold.
- [Baraniuk-Wakin'09] suggest reducing dimensionality using Π and doing learning in lower-dimensional space
- to preserve curve lengths on manifold, can set T to be tangent space of $\mathcal{M}$ (infinite union $T = (\bigcup_{u\in\mathcal{M}} E_u) \cap S^{n-1}$)

Euclidean dimensionality reduction: Optimality?

Can we do better than the JL lemma?

Euclidean dimensionality reduction: Optimality?

Can we do better than the JL lemma?

- Unfortunately, not by much if any.
  [Alon '03]: Let $T = \{e_1, \ldots, e_n\} \cup \{(e_i - e_j)/\sqrt{2}\}_{i \neq j}$

Euclidean dimensionality reduction: Optimality?

Can we do better than the JL lemma?

- Unfortunately, not by much if any.

  [Alon '03]: Let $T = \{e_1, \ldots, e_n\} \cup \{(e_i - e_j)/\sqrt{2}\}_{i \neq j}$
- Any embedding of T with distortion $1 + \varepsilon$ requires $m \gtrsim \frac{\log n}{\varepsilon^2 \log(1/\varepsilon)}$ (off from JL lemma by just $\log(1/\varepsilon)$)

Euclidean dimensionality reduction: "Beyond Worst-Case Analysis"

- Suppose $\Pi_{i,j} \sim \mathcal{N}(0, 1/m)$ (i.i.d. gaussian entries)

Euclidean dimensionality reduction: "Beyond Worst-Case Analysis"

- Suppose $\Pi_{i,j} \sim \mathcal{N}(0, 1/m)$ (i.i.d. gaussian entries)
- **Gordon's theorem (1988):** suffices to set $m \sim (g^2(T) + 1)/\varepsilon^2$

  where $g(T) = \mathbb{E}_g \sup_{u \in T} \langle g, u \rangle$ (g a gaussian vector)

Euclidean dimensionality reduction: "Beyond Worst-Case Analysis"

- Suppose $\Pi_{i,j} \sim \mathcal{N}(0, 1/m)$ (i.i.d. gaussian entries)
- **Gordon's theorem (1988):** suffices to set $m \sim (g^2(T) + 1)/\varepsilon^2$

  where $g(T) = \mathbb{E}_g \sup_{u \in T} \langle g, u \rangle$ (g a gaussian vector)
- $g(T) \leq \sqrt{\log |T|}$ always (union bound), but can be much smaller if the vectors in T are well-clusterable

Euclidean dimensionality reduction: "Beyond Worst-Case Analysis"

- Suppose $\Pi_{i,j} \sim \mathcal{N}(0, 1/m)$ (i.i.d. gaussian entries)
- **Gordon's theorem (1988):** suffices to set $m \sim (g^2(T) + 1)/\varepsilon^2$

  where $g(T) = \mathbb{E}_g \sup_{u \in T} \langle g, u \rangle$ (g a gaussian vector)
- $g(T) \leq \sqrt{\log |T|}$ always (union bound), but can be much smaller if the vectors in T are well-clusterable
- **[Klartag, Mendelson '05]:** Same result if $\Pi_{i,j} \sim \pm 1/\sqrt{m}$

Euclidean dimensionality reduction: "Beyond Worst-Case Analysis"

- Suppose $\Pi_{i,j} \sim \mathcal{N}(0, 1/m)$ (i.i.d. gaussian entries)
- **Gordon's theorem (1988):** suffices to set $m \sim (g^2(T) + 1)/\varepsilon^2$

  where $g(T) = \mathbb{E}_g \sup_{u \in T} \langle g, u \rangle$ (g a gaussian vector)
- $g(T) \leq \sqrt{\log |T|}$ always (union bound), but can be much smaller if the vectors in T are well-clusterable
- **[Klartag, Mendelson '05]:** Same result if $\Pi_{i,j} \sim \pm 1/\sqrt{m}$
- Same problem as before: Π is dense.

Sparse JL: Optimality

- **JL:** $m \lesssim \varepsilon^{-2} \log |T|, s \lesssim \varepsilon^{-1} \log |T|$ [Kane, N. '12]

Sparse JL: Optimality

- **JL:** $m \lesssim \varepsilon^{-2} \log |T|, s \lesssim \varepsilon^{-1} \log |T|$ [Kane, N. '12]
- **Near-optimal:**
  $m \lesssim \varepsilon^{-100} \log |T|$ requires $s \gtrsim \varepsilon^{-1} \frac{\log |T|}{\log(1/\varepsilon)}$ [N., Nguyễn '13]

Sparse JL:
"Beyond Worst-Case Analysis"

Question: Can we obtain a Gordon-type theorem to understand how to set m, s as a function of T?

What we show [Bourgain-N. '14]

set T	our m	our s	previous m	previous s	ref
$\lvert T\rvert < \infty$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	[JL'84]
$\lvert T\rvert < \infty,\ \lVert u\rVert_\infty \le \alpha$	$\log \lvert T\rvert$	$\lceil \alpha \log \lvert T\rvert \rceil^2$	$\log \lvert T\rvert$	$\lceil \alpha \log \lvert T\rvert \rceil^2$	[M'08]
$E, \dim(E) \le d$	d	1	d	1	[NN'13]
$S_{n,k}$	$k \log(n/k)$	$k \log(n/k)$	$k \log(n/k)$	$k \log(n/k)$	[CT'05]
$HS_{n,k}$	$k \log(n/k)$	1	$k \log(n/k)$	1	[RV'08]*
$\lvert\Theta\rvert < \infty$ $\forall E \in \Theta, \dim(E) \le d$	$d + \log \lvert\Theta\rvert$	$\log \lvert\Theta\rvert$	$d + (\log \lvert\Theta\rvert)^6$	$(\log \lvert\Theta\rvert)^3$	[NN'13]
$\lvert\Theta\rvert < \infty$ $\forall E \in \Theta, \dim(E) \le d$ $\forall j,\ E\ \lVert P_E e_j\rVert_2 \le \alpha$	$d + \log \lvert\Theta\rvert$	$\lceil \alpha \log \lvert\Theta\rvert \rceil^2$	—	—	—
$\lvert\Theta\rvert$ infinite $\forall E \in \Theta, \dim(E) \le d$	see paper	see paper (non-trivial)	d	m	[D'14]
$\mathcal{M}$ a smooth manifold $\forall E \in TM, j\ \lVert P_E e_j\rVert_2 \le \alpha$	d	$\lceil \alpha d \rceil^2$	d	d	[D'14]

What we show [Bourgain-N. '14]

set T	our m	our s	previous m	previous s	ref
$\lvert T\rvert < \infty$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	[JL'84]
$\lvert T\rvert < \infty,\ \lVert u\rVert_\infty \le \alpha$	$\log \lvert T\rvert$	$\lceil \alpha \log \lvert T\rvert \rceil^2$	$\log \lvert T\rvert$	$\lceil \alpha \log \lvert T\rvert \rceil^2$	[M'08]
$E, \dim(E) \le d$	d	1	d	1	[NN'13]
$S_{n,k}$	$k \log(n/k)$	$k \log(n/k)$	$k \log(n/k)$	$k \log(n/k)$	[CT'05]
$HS_{n,k}$	$k \log(n/k)$	1	$k \log(n/k)$	1	[RV'08]*
$\lvert\Theta\rvert < \infty$ $\forall E \in \Theta, \dim(E) \le d$	$d + \log \lvert\Theta\rvert$	$\log \lvert\Theta\rvert$	$d + (\log \lvert\Theta\rvert)^6$	$(\log \lvert\Theta\rvert)^3$	[NN'13]
$\lvert\Theta\rvert < \infty$ $\forall E \in \Theta, \dim(E) \le d$ $\forall j, E\ \lVert P_E e_j\rVert_2 \le \alpha$	$d + \log \lvert\Theta\rvert$	$\lceil \alpha \log \lvert\Theta\rvert \rceil^2$	—	—	—
$\lvert\Theta\rvert$ infinite $\forall E \in \Theta, \dim(E) \le d$	see paper	see paper (non-trivial)	d	m	[D'14]
$\mathcal{M}$ a smooth manifold $\forall E \in TM, j\ \lVert P_E e_j\rVert_2 \le \alpha$	d	$\lceil \alpha d \rceil^2$	d	d	[D'14]

All bounds shown hide $\mathrm{poly}(\varepsilon^{-1} \log n)$ **factors**. One row is blank in previous work due to no non-trivial results being previously known.

What we show [Bourgain-N. '14]

set T	our m	our s	previous m	previous s	ref
$\lvert T\rvert < \infty$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	$\log \lvert T\rvert$	[JL'84]
$\lvert T\rvert < \infty,\ \lVert u\rVert_\infty \le \alpha$	$\log \lvert T\rvert$	$\lceil \alpha \log \lvert T\rvert \rceil^2$	$\log \lvert T\rvert$	$\lceil \alpha \log \lvert T\rvert \rceil^2$	[M'08]
$E, \dim(E) \le d$	d	1	d	1	[NN'13]
$S_{n,k}$	$k \log(n/k)$	$k \log(n/k)$	$k \log(n/k)$	$k \log(n/k)$	[CT'05]
$HS_{n,k}$	$k \log(n/k)$	1	$k \log(n/k)$	1	[RV'08]*
$\lvert\Theta\rvert < \infty$ $\forall E \in \Theta, \dim(E) \le d$	$d + \log \lvert\Theta\rvert$	$\log \lvert\Theta\rvert$	$d + (\log \lvert\Theta\rvert)^6$	$(\log \lvert\Theta\rvert)^3$	[NN'13]
$\lvert\Theta\rvert < \infty$ $\forall E \in \Theta, \dim(E) \le d$ $\forall j, E\ \lVert P_E e_j\rVert_2 \le \alpha$	$d + \log \lvert\Theta\rvert$	$\lceil \alpha \log \lvert\Theta\rvert \rceil^2$	—	—	—
$\lvert\Theta\rvert$ infinite $\forall E \in \Theta, \dim(E) \le d$	see paper	see paper (non-trivial)	d	m	[D'14]
$\mathcal{M}$ a smooth manifold $\forall E \in TM, j\ \lVert P_E e_j\rVert_2 \le \alpha$	d	$\lceil \alpha d \rceil^2$	d	d	[D'14]

All bounds shown hide $\mathrm{poly}(\varepsilon^{-1} \log n)$ **factors**. One row is blank in previous work due to no non-trivial results being previously known.

Question we studied: What's going on here?

Our main contribution [Bourgain-N. '14]

$T \subseteq S^{n-1}$, $\Pi \in \mathbb{R}^{m \times n}$ an SJLT. Suffices to set

Our main contribution [Bourgain-N. '14]

$T \subseteq S^{n-1}$, $\Pi \in \mathbb{R}^{m \times n}$ an SJLT. Suffices to set

$$m \gg g^2(T) + 1, \qquad s \gg 1$$

such that

$$\max_{q \leq \frac{m}{s} \log s} \left\{ \frac{1}{\sqrt{qs}} \left(\mathop{\mathbb{E}}_{\eta} \left(\mathop{\mathbb{E}}_{g} \sup_{u \in T} \left| \sum_{j=1}^{n} \eta_j u_j g_j \right| \right)^q \right)^{1/q} \right\} \ll 1$$

Our main contribution [Bourgain-N. '14]

$T \subseteq S^{n-1}$, $\Pi \in \mathbb{R}^{m \times n}$ an SJLT. Suffices to set

$$m \gg g^2(T) + 1, \qquad s \gg 1$$

such that

$$\max_{q \leq \frac{m}{s} \log s} \left\{ \frac{1}{\sqrt{qs}} \left(\mathop{\mathbb{E}}_{\eta} \left(\mathop{\mathbb{E}}_{g} \sup_{u \in T} \left| \sum_{j=1}^{n} \eta_j u_j g_j \right| \right)^q \right)^{1/q} \right\} \ll 1$$

$\ll$ and $\gg$ hide $(\log n)/\varepsilon$ factors. (η_j) i.i.d. Bernoulli of expectation $qs/(m \log s)$. We abbreviate as

$$\max_{q \leq \frac{m}{s} \log s} \left\{ \frac{1}{\sqrt{qs}} \left\| \sup_{u \in T} \sum_{j=1}^{n} \eta_j u_j g_j \right\|_{L_\eta^q L_g^1} \right\} \ll 1$$

Fast JL note

Although we study sparse JL transforms, our results also apply to the "Fast JL transforms" of [Ailon, Chazelle '06] and follow-ups.

Fast JL note

Although we study sparse JL transforms, our results also apply to the "Fast JL transforms" of [Ailon, Chazelle '06] and follow-ups.

$$\Phi = \underbrace{\begin{pmatrix} 0 & 0 & \dots & 1 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \dots & 0 & 0 \end{pmatrix}}_{S} \times \underbrace{\begin{pmatrix} & & H & & \end{pmatrix}}_{\text{Fourier}} \times \underbrace{\begin{pmatrix} \pm 1 & 0 & \dots & 0 & 0 \\ 0 & \pm 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \pm 1 & 0 \\ 0 & 0 & \dots & 0 & \pm 1 \end{pmatrix}}_{D}$$

- S a sampling matrix (exactly one non-zero per row)
- D a diagonal matrix with signs on diagonal

Fast JL note

Although we study sparse JL transforms, our results also apply to the "Fast JL transforms" of [Ailon, Chazelle '06] and follow-ups.

$$\Phi = \underbrace{\begin{pmatrix} 0 & 0 & \dots & 1 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \dots & 0 & 0 \end{pmatrix}}_{S} \times \underbrace{\begin{pmatrix} & & H & & \end{pmatrix}}_{\text{Fourier}} \times \underbrace{\begin{pmatrix} \pm 1 & 0 & \dots & 0 & 0 \\ 0 & \pm 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \pm 1 & 0 \\ 0 & 0 & \dots & 0 & \pm 1 \end{pmatrix}}_{D}$$

- S a sampling matrix (exactly one non-zero per row)
- D a diagonal matrix with signs on diagonal
- Imagine replacing S with a SJLT Π, so $\Phi = \Pi HD$

Fast JL note

Although we study sparse JL transforms, our results also apply to the "Fast JL transforms" of [Ailon, Chazelle '06] and follow-ups.

$$\Phi = \underbrace{\begin{pmatrix} 0 & 0 & \dots & 1 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & \dots & 0 & 0 \end{pmatrix}}_{S} \times \underbrace{\begin{pmatrix} & & H & & \end{pmatrix}}_{\text{Fourier}} \times \underbrace{\begin{pmatrix} \pm 1 & 0 & \dots & 0 & 0 \\ 0 & \pm 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \pm 1 & 0 \\ 0 & 0 & \dots & 0 & \pm 1 \end{pmatrix}}_{D}$$

- S a sampling matrix (exactly one non-zero per row)
- D a diagonal matrix with signs on diagonal
- Imagine replacing S with a SJLT Π, so $\Phi = \Pi HD$
- Want Φ to preserve all of T; can view as Π preserving HDT

An example new application

Model-based compressed sensing

- Our result implies sparse Π for model-based RIP when # sparsity patterns is $2^{o(k)}$

An example new application

Model-based compressed sensing

- Our result implies sparse Π for model-based RIP when # sparsity patterns is $2^{o(k)}$
- Example: **block-sparsity**. u divided into n/b blocks of size b each. Each block either "on" or "off".

An example new application

Model-based compressed sensing

- Our result implies sparse Π for model-based RIP when # sparsity patterns is $2^{o(k)}$
- Example: **block-sparsity**. u divided into n/b blocks of size b each. Each block either "on" or "off".
- Sparsity k means at most k/b blocks are on.

An example new application

Model-based compressed sensing

- Our result implies sparse Π for model-based RIP when # sparsity patterns is $2^{o(k)}$
- Example: **block-sparsity**. u divided into n/b blocks of size b each. Each block either "on" or "off".
- Sparsity k means at most k/b blocks are on.
- $\log |\Theta| = \log \binom{n/b}{k/b} \sim (k/b) \log(n/k)$
- Thus $m \ll k + (k/b) \log(n/k), s \ll (k/b) \log(n/k)$
- Non-trivial sparsity s for $b \gg \mathrm{polylog}(n)$

Main theorem: proof outline

- Want to bound $\mathbb{E}_{\Pi} \sup_{u \in T} |\|\Pi u\|_2^2 - 1|$

Main theorem: proof outline

- Want to bound $\mathbb{E}_\Pi \sup_{u \in T} |\|\Pi u\|_2^2 - 1|$
- Can write $\Pi_{i,j} = \delta_{i,j}\sigma_{i,j}/\sqrt{s}$

  $\delta_{i,j} \in \{0,1\}$, $\sigma_{i,j} = \pm 1$

Main theorem: proof outline

- Want to bound $\mathbb{E}_\Pi \sup_{u \in T} |\|\Pi u\|_2^2 - 1|$
- Can write $\Pi_{i,j} = \delta_{i,j}\sigma_{i,j}/\sqrt{s}$

  $\delta_{i,j} \in \{0,1\}$, $\sigma_{i,j} = \pm 1$
- Can write $\|\Pi u\|_2^2 = \|A_{u,\delta}\sigma\|_2^2$

$$A_{u,\delta} = \frac{1}{\sqrt{s}} \cdot \begin{bmatrix} u_1^{(1)} & \cdots & u_n^{(1)} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & u_1^{(2)} & \cdots & u_n^{(2)} & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & u_1^{(m)} & \cdots & u_n^{(m)} \end{bmatrix}$$

where $u_j^{(i)} = \delta_{i,j} u_j$. Define $\mathcal{A}_\delta = \{A_{u,\delta} : u \in T\}$.

Main theorem: proof outline

- Want to bound $\mathbb{E}_\Pi \sup_{u \in T} |\|\Pi u\|_2^2 - 1|$
- Can write $\Pi_{i,j} = \delta_{i,j}\sigma_{i,j}/\sqrt{s}$

  $\delta_{i,j} \in \{0,1\}$, $\sigma_{i,j} = \pm 1$
- Can write $\|\Pi u\|_2^2 = \|A_{u,\delta}\sigma\|_2^2$

$$A_{u,\delta} = \frac{1}{\sqrt{s}} \cdot \begin{bmatrix} u_1^{(1)} & \cdots & u_n^{(1)} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & u_1^{(2)} & \cdots & u_n^{(2)} & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & u_1^{(m)} & \cdots & u_n^{(m)} \end{bmatrix}$$

  where $u_j^{(i)} = \delta_{i,j} u_j$. Define $\mathcal{A}_\delta = \{A_{u,\delta} : u \in T\}$.
- **Want:** $\mathbb{E}_\delta \mathbb{E}_\sigma \sup_{A \in \mathcal{A}_\delta} |\|A\sigma\|_2^2 - \mathbb{E}\|A\sigma\|_2^2| < \varepsilon$
- Can handle using chaining and tools from Banach space theory

Proof outline

$$\mathop{\mathbb{E}}_{\delta}\mathop{\mathbb{E}}_{\sigma}\sup_{A\in\mathcal{A}_\delta}\left|\|A\sigma\|_2^2-\mathbb{E}\,\|A\sigma\|_2^2\right|$$

Proof outline

$$\mathop{\mathbb{E}}_{\delta}\mathop{\mathbb{E}}_{\sigma}\sup_{A\in\mathcal{A}_\delta}\left|\|A\sigma\|_2^2-\mathbb{E}\,\|A\sigma\|_2^2\right|$$

We're in luck!

Proof outline

$$\mathop{\mathbb{E}}_{\delta}\mathop{\mathbb{E}}_{\sigma}\sup_{A\in\mathcal{A}_\delta}\left|\|A\sigma\|_2^2-\mathbb{E}\,\|A\sigma\|_2^2\right|$$

We're in luck!

Theorem (Krahmer, Mendelson, Rauhut '11)

For any collection $\mathcal{A}$ of matrices and σ a Rademacher vector

$$\begin{aligned}&\mathop{\mathbb{E}}_{\sigma}\sup_{A\in\mathcal{A}}\left|\|A\sigma\|_2^2-\mathbb{E}\,\|A\sigma\|_2^2\right|\\&\quad\lesssim\gamma_2^2(\mathcal{A},\|\cdot\|)+\gamma_2(\mathcal{A},\|\cdot\|)\cdot d_F(\mathcal{A})+d_F(\mathcal{A})\cdot d_{\ell_2\to\ell_2}(\mathcal{A})\end{aligned}$$

$d_X(Y)$ is the radius of Y under the norm $\|\cdot\|_X$

Proof outline

$$\mathop{\mathbb{E}}_{\delta}\mathop{\mathbb{E}}_{\sigma}\sup_{A\in\mathcal{A}_\delta}\left|\|A\sigma\|_2^2-\mathbb{E}\,\|A\sigma\|_2^2\right|$$

We're in luck!

Theorem (Krahmer, Mendelson, Rauhut '11)

For any collection $\mathcal{A}$ of matrices and σ a Rademacher vector

$$\begin{aligned}&\mathop{\mathbb{E}}_{\sigma}\sup_{A\in\mathcal{A}}\left|\|A\sigma\|_2^2-\mathbb{E}\,\|A\sigma\|_2^2\right|\\&\quad\lesssim \gamma_2^2(\mathcal{A},\|\cdot\|)+\gamma_2(\mathcal{A},\|\cdot\|)\cdot d_F(\mathcal{A})+d_F(\mathcal{A})\cdot d_{\ell_2\to\ell_2}(\mathcal{A})\end{aligned}$$

$d_X(Y)$ is the radius of Y under the norm $\|\cdot\|_X$

γ_2 and Talagrand's "generic chaining"

- What is γ_2? Doesn't matter for us except for three things:

γ_2 and Talagrand's "generic chaining"

- What is γ_2? Doesn't matter for us except for three things:

(1) $\gamma_2(Y, \|\cdot\|_2) \sim g(Y)$ for any Y
("majorizing measures") [Fernique '76], [Talagrand '01]

(2) $\gamma_2(Y, \|\cdot\|_X) \ll \sup_{\eta>0} \eta \cdot [\log \mathcal{N}(Y, \|\cdot\|_X, \eta)]^{1/2}$ for any norm

(3) $\gamma_2(Y, \|\cdot\|_X)$ is a function only of pairwise distances in Y under the $\|\cdot\|_X$ norm

γ_2 and Talagrand's "generic chaining"

- What is γ_2? Doesn't matter for us except for three things:

(1) $\gamma_2(Y, \|\cdot\|_2) \sim g(Y)$ for any Y
("majorizing measures") [Fernique '76], [Talagrand '01]

(2) $\gamma_2(Y, \|\cdot\|_X) \ll \sup_{\eta>0} \eta \cdot [\log \mathcal{N}(Y, \|\cdot\|_X, \eta)]^{1/2}$ for any norm

(3) $\gamma_2(Y, \|\cdot\|_X)$ is a function only of pairwise distances in Y under the $\|\cdot\|_X$ norm

- Here $\mathcal{N}(Y, \|\cdot\|_X, \eta)$ is the minimum number of $\|\cdot\|_X$-balls of radius η centered at points in Y required to cover Y
(usually called "covering" or "entropy" numbers)

Bounding the right hand side

Fix δ.

$$\mathop{\mathbb{E}}_{\sigma} \sup_{A \in \mathcal{A}_\delta} \left| \|A\sigma\|_2^2 - \mathbb{E}\,\|A\sigma\|_2^2 \right|$$
$$\lesssim \gamma_2^2(\mathcal{A}_\delta, \|\cdot\|) + \gamma_2(\mathcal{A}_\delta, \|\cdot\|) \cdot d_F(\mathcal{A}_\delta) + d_F(\mathcal{A}_\delta) \cdot d_{\ell_2 \to \ell_2}(\mathcal{A}_\delta)$$

Bounding the right hand side

Fix δ.

$$\mathop{\mathbb{E}}_{\sigma} \sup_{A \in \mathcal{A}_\delta} \left| \|\|A\sigma\|_2^2 - \mathbb{E}\,\|A\sigma\|_2^2 \right|$$
$$\lesssim \gamma_2^2(\mathcal{A}_\delta, \|\cdot\|) + \gamma_2(\mathcal{A}_\delta, \|\cdot\|) \cdot d_F(\mathcal{A}_\delta) + d_F(\mathcal{A}_\delta) \cdot d_{\ell_2 \to \ell_2}(\mathcal{A}_\delta)$$

- $\gamma_2(\mathcal{A}_\delta, \|\cdot\|)$ depends only on $\|\cdot\|$-distances
- **Facts:** (1) $A_{u,\delta} - A_{w,\delta} = A_{u-w,\delta}$,
  (2) $\|A_{u,\delta}\| = \frac{1}{\sqrt{s}} \cdot \max_{1 \le i \le m} \|u^{(i)}\|_2 \stackrel{\text{def}}{=} \|\|u\|\|$

Bounding the right hand side

Fix δ.

$$\mathop{\mathbb{E}}_{\sigma} \sup_{A \in \mathcal{A}_\delta} \left| \|\|A\sigma\|_2^2 - \mathbb{E}\,\|A\sigma\|_2^2 \right|$$
$$\lesssim \gamma_2^2(\mathcal{A}_\delta, \|\cdot\|) + \gamma_2(\mathcal{A}_\delta, \|\cdot\|) \cdot d_F(\mathcal{A}_\delta) + d_F(\mathcal{A}_\delta) \cdot d_{\ell_2 \to \ell_2}(\mathcal{A}_\delta)$$

- $\gamma_2(\mathcal{A}_\delta, \|\cdot\|)$ depends only on $\|\cdot\|$-distances
- **Facts:** (1) $A_{u,\delta} - A_{w,\delta} = A_{u-w,\delta}$,
  (2) $\|A_{u,\delta}\| = \frac{1}{\sqrt{s}} \cdot \max_{1 \le i \le m} \|u^{(i)}\|_2 \stackrel{\text{def}}{=} |||u|||$
- Therefore $\|A_{u,\delta} - A_{w,\delta}\| = |||u - w|||$

Bounding the right hand side

Fix δ.

$$\mathop{\mathbb{E}}_{\sigma} \sup_{A \in \mathcal{A}_\delta} \left| \|\|A\sigma\|_2^2 - \mathbb{E}\,\|A\sigma\|_2^2 \right|$$
$$\lesssim \gamma_2^2(\mathcal{A}_\delta, \|\cdot\|) + \gamma_2(\mathcal{A}_\delta, \|\cdot\|) \cdot d_F(\mathcal{A}_\delta) + d_F(\mathcal{A}_\delta) \cdot d_{\ell_2 \to \ell_2}(\mathcal{A}_\delta)$$

- $\gamma_2(\mathcal{A}_\delta, \|\cdot\|)$ depends only on $\|\cdot\|$-distances
- **Facts:** (1) $A_{u,\delta} - A_{w,\delta} = A_{u-w,\delta}$,
  (2) $\|A_{u,\delta}\| = \frac{1}{\sqrt{s}} \cdot \max_{1 \le i \le m} \|u^{(i)}\|_2 \stackrel{\text{def}}{=} |||u|||$
- Therefore $\|A_{u,\delta} - A_{w,\delta}\| = |||u - w|||$
- Therefore $\gamma_2(\mathcal{A}_\delta, \|\cdot\|) = \gamma_2(T, |||\cdot|||)$

Bounding the right hand side

Fix δ.

$$\mathop{\mathbb{E}}_{\sigma} \sup_{A \in \mathcal{A}_\delta} \left| \|A\sigma\|_2^2 - \mathbb{E}\,\|A\sigma\|_2^2 \right|$$
$$\lesssim \gamma_2^2(T, \|\|\cdot\|\|) + \gamma_2(T, \|\|\cdot\|\|) \cdot d_F(\mathcal{A}_\delta) + d_F(\mathcal{A}_\delta) \cdot d_{\ell_2 \to \ell_2}(\mathcal{A}_\delta)$$

- **Facts:** (1) $A_{u,\delta} - A_{w,\delta} = A_{u-w,\delta}$,
  (2) $\|A_{u,\delta}\| = \frac{1}{\sqrt{s}} \cdot \max_{1 \le i \le m} \|u^{(i)}\|_2 = \|\|u\|\|$

Bounding the right hand side

Fix δ.

$$\mathop{\mathbb{E}}_{\sigma} \sup_{A \in \mathcal{A}_\delta} \left| \|A\sigma\|_2^2 - \mathbb{E}\,\|A\sigma\|_2^2 \right|$$
$$\lesssim \gamma_2^2(T, \|\|\cdot\|\|) + \gamma_2(T, \|\|\cdot\|\|) \cdot d_F(\mathcal{A}_\delta) + d_F(\mathcal{A}_\delta) \cdot d_{\ell_2 \to \ell_2}(\mathcal{A}_\delta)$$

- **Facts:** (1) $A_{u,\delta} - A_{w,\delta} = A_{u-w,\delta}$,
  (2) $\|A_{u,\delta}\| = \frac{1}{\sqrt{s}} \cdot \max_{1 \le i \le m} \|u^{(i)}\|_2 = \|\|u\|\|$
- Since $T \subset S^{n-1}$, $\|A_{u,\delta}\| \le \frac{1}{\sqrt{s}} \cdot \|u\|_2 = \frac{1}{\sqrt{s}}$
- Also $\|A_{u,\delta}\|_F^2 = \frac{1}{s} \sum_{i=1}^m \sum_{j=1}^n \delta_{i,j} u_j^2 = \|u\|_2^2 = 1$

Bounding the right hand side

Fix δ.

$$\mathop{\mathbb{E}}_{\sigma} \sup_{A \in \mathcal{A}_\delta} \left| \|A\sigma\|_2^2 - \mathbb{E}\,\|A\sigma\|_2^2 \right|$$
$$\lesssim \gamma_2^2(T, |||\cdot|||) + \gamma_2(T, |||\cdot|||) \cdot d_F(\mathcal{A}_\delta) + d_F(\mathcal{A}_\delta) \cdot d_{\ell_2 \to \ell_2}(\mathcal{A}_\delta)$$

- **Facts:** (1) $A_{u,\delta} - A_{w,\delta} = A_{u-w,\delta}$,
  (2) $\|A_{u,\delta}\| = \frac{1}{\sqrt{s}} \cdot \max_{1 \le i \le m} \|u^{(i)}\|_2 = |||u|||$
- Since $T \subset S^{n-1}$, $\|A_{u,\delta}\| \le \frac{1}{\sqrt{s}} \cdot \|u\|_2 = \frac{1}{\sqrt{s}}$
- Also $\|A_{u,\delta}\|_F^2 = \frac{1}{s} \sum_{i=1}^m \sum_{j=1}^n \delta_{i,j} u_j^2 = \|u\|_2^2 = 1$
- So RHS is at most $\gamma_2^2(T, |||\cdot|||) + \gamma_2(T, |||\cdot|||) + \frac{1}{\sqrt{s}}$

Warmup for main theorem: case of a linear subspace

- $T = E$, the intersection of a d-dim. linear subspace and S^{n-1}

Warmup for main theorem: case of a linear subspace

- $T = E$, the intersection of a d-dim. linear subspace and S^{n-1}
- Want to bound $\mathbb{E}_\delta \gamma_2(T, \|\|\cdot\|\|)$

Warmup for main theorem: case of a linear subspace

- $T = E$, the intersection of a d-dim. linear subspace and S^{n-1}
- Want to bound $\mathbb{E}_\delta \gamma_2(T, \|\|\cdot\|\|)$
- One of our facts:
  $\gamma_2(T, \|\|\cdot\|\|) \ll \sup_{\eta>0} \eta \cdot [\log \mathcal{N}(T, \|\|\cdot\|\|, \eta)]^{1/2}$

Warmup for main theorem: case of a linear subspace

- $T = E$, the intersection of a d-dim. linear subspace and S^{n-1}
- Want to bound $\mathbb{E}_\delta \gamma_2(T, \|\|\cdot\|\|)$
- One of our facts:
  $\gamma_2(T, \|\|\cdot\|\|) \ll \sup_{\eta>0} \eta \cdot [\log \mathcal{N}(T, \|\|\cdot\|\|, \eta)]^{1/2}$
- Another stroke of luck. **Dual Sudakov minoration** [Bourgain, Lindenstrauss, Milman '89], [Pajor, Tomczak-Jaegermann '86]

$$\sup_{\eta>0} \eta \cdot [\log \mathcal{N}(B_E, \|\|\cdot\|\|, \eta)]^{1/2} \lesssim \mathop{\mathbb{E}}_{g} \|\|Ug\|\|$$

  where g is a gaussian vector (very specific to linear subspaces)
- Here columns of $U \in \mathbb{R}^{n\times d}$ form orthonormal basis for E

Warmup for main theorem: case of a linear subspace

Let $U^{(i)}$ be U but where the jth row is multiplied by $\delta_{i,j}$

$$\mathop{\mathbb{E}}_{g} |||Ug||| \quad = \quad \frac{1}{\sqrt{s}} \mathop{\mathbb{E}}_{g} \max_{1 \leq i \leq m} \|U^{(i)} g\|_2$$

Warmup for main theorem: case of a linear subspace

Let $U^{(i)}$ be U but where the jth row is multiplied by $\delta_{i,j}$

$$\begin{aligned}
\mathop{\mathbb{E}}_{g} |||Ug||| &= \frac{1}{\sqrt{s}} \mathop{\mathbb{E}}_{g} \max_{1 \le i \le m} \|U^{(i)} g\|_2 \\
&\le \frac{1}{\sqrt{s}} \left[\max_{1 \le i \le m} \mathop{\mathbb{E}}_{g} \|U^{(i)} g\|_2 + \mathop{\mathbb{E}}_{g} \max_{1 \le i \le m} \left| \|U^{(i)} g\|_2 - \mathop{\mathbb{E}}_{g} \|U^{(i)} g\|_2 \right| \right]
\end{aligned}$$

Warmup for main theorem: case of a linear subspace

Let $U^{(i)}$ be U but where the jth row is multiplied by $\delta_{i,j}$

$$\begin{aligned}
\mathop{\mathbb{E}}_{g} |||Ug||| &= \frac{1}{\sqrt{s}} \mathop{\mathbb{E}}_{g} \max_{1\le i\le m} \|U^{(i)}g\|_2 \\
&\le \frac{1}{\sqrt{s}} \left[\max_{1\le i\le m} \mathop{\mathbb{E}}_{g} \|U^{(i)}g\|_2 + \mathop{\mathbb{E}}_{g} \max_{1\le i\le m} \left| \|U^{(i)}g\|_2 - \mathop{\mathbb{E}}_{g} \|U^{(i)}g\|_2 \right| \right] \\
&\le \frac{1}{\sqrt{s}} \left[\max_{1\le i\le m} \mathop{\mathbb{E}}_{g} \|U^{(i)}g\|_2 + \left(\sum_{i=1}^{m} \mathop{\mathbb{E}}_{g} \left| \|U^{(i)}g\|_2 - \mathop{\mathbb{E}}_{g} \|U^{(i)}g\|_2 \right|^p \right)^{1/p} \right]
\end{aligned}$$

Warmup for main theorem: case of a linear subspace

Let $U^{(i)}$ be U but where the jth row is multiplied by $\delta_{i,j}$

$$\begin{aligned}
\mathop{\mathbb{E}}_{g} |||Ug||| &= \frac{1}{\sqrt{s}} \mathop{\mathbb{E}}_{g} \max_{1\le i\le m} \|U^{(i)}g\|_2 \\
&\le \frac{1}{\sqrt{s}} \left[\max_{1\le i\le m} \mathop{\mathbb{E}}_{g} \|U^{(i)}g\|_2 + \mathop{\mathbb{E}}_{g} \max_{1\le i\le m} \left| \|U^{(i)}g\|_2 - \mathop{\mathbb{E}}_{g} \|U^{(i)}g\|_2 \right| \right] \\
&\le \frac{1}{\sqrt{s}} \left[\max_{1\le i\le m} \mathop{\mathbb{E}}_{g} \|U^{(i)}g\|_2 + \left(\sum_{i=1}^{m} \mathop{\mathbb{E}}_{g} \left| \|U^{(i)}g\|_2 - \mathop{\mathbb{E}}_{g} \|U^{(i)}g\|_2 \right|^p \right)^{1/p} \right] \\
&\le \frac{1}{\sqrt{s}} \left[\max_{1\le i\le m} \|U^{(i)}\|_F + \sqrt{\log m} \max_{1\le i\le m} \|U^{(i)}\| \right]
\end{aligned}$$

(last step: Chose $p = \log m$ and used gaussian concentration of Lipschitz functions [Pisier '86])

Warmup for main theorem: case of a linear subspace

Now must bound

$$\frac{1}{\sqrt{s}} \cdot \mathop{\mathbb{E}}_{\delta}\left[\max_{1\le i\le m}\|U^{(i)}\|_F + \sqrt{\log m}\max_{1\le i\le m}\|U^{(i)}\|\right] \tag{1}$$

$$\le \frac{1}{\sqrt{s}}\cdot\left[\left(\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\delta}\|U^{(i)}\|_F^p\right)^{1/p} + \sqrt{\log m}\left(\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\delta}\|U^{(i)}\|^p\right)^{1/p}\right] \tag{2}$$

choosing $p = \log m$ again as usual so that $\ell_p \sim \ell_\infty$.

Warmup for main theorem: case of a linear subspace

Now must bound

$$\frac{1}{\sqrt{s}} \cdot \mathop{\mathbb{E}}_{\delta}\left[\max_{1\le i\le m}\|U^{(i)}\|_F + \sqrt{\log m}\max_{1\le i\le m}\|U^{(i)}\|\right] \tag{1}$$

$$\le \frac{1}{\sqrt{s}}\cdot\left[\left(\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\delta}\|U^{(i)}\|_F^p\right)^{1/p} + \sqrt{\log m}\left(\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\delta}\|U^{(i)}\|^p\right)^{1/p}\right] \tag{2}$$

choosing $p = \log m$ again as usual so that $\ell_p \sim \ell_\infty$.

- $\mathbb{E}_\delta \|U^{(i)}\|_F^p$ is basically the Chernoff bound
- $\mathbb{E}_\delta \|U^{(i)}\|^p$ can be handled by non-commutative Khintchine inequality [Lust-Piquard '86], [Lust-Piquard, Pisier '91]

Warmup for main theorem: case of a linear subspace

Now must bound

$$\frac{1}{\sqrt{s}} \cdot \mathop{\mathbb{E}}_{\delta}\left[\max_{1\le i\le m}\|U^{(i)}\|_F + \sqrt{\log m}\max_{1\le i\le m}\|U^{(i)}\|\right] \tag{1}$$

$$\le \frac{1}{\sqrt{s}}\cdot\left[\left(\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\delta}\|U^{(i)}\|_F^p\right)^{1/p} + \sqrt{\log m}\left(\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\delta}\|U^{(i)}\|^p\right)^{1/p}\right] \tag{2}$$

choosing $p = \log m$ again as usual so that $\ell_p \sim \ell_\infty$.

- $\mathbb{E}_\delta \|U^{(i)}\|_F^p$ is basically the Chernoff bound
- $\mathbb{E}_\delta \|U^{(i)}\|^p$ can be handled by non-commutative Khintchine inequality [Lust-Piquard '86], [Lust-Piquard, Pisier '91]

Going through calculations gives that we can set:

$$m \lesssim \frac{d(\log d)^2 + (\log m)(\log d)^2}{\varepsilon^2} \ll \frac{d}{\varepsilon^2}$$

$$s \lesssim 1 + (\log d)^2(\log m)^2 \cdot \frac{\max_{1\le j\le n}\|P_E e_j\|_2^2}{\varepsilon^2}$$

Can even set $s = 1$ if "leverage scores" small (true for random E)

Back to main theorem

General theorem

$T \subseteq S^{n-1}$ arbitrary. Suffices to set

$$m \gg g^2(T) + 1, \qquad s \gg 1$$

such that

$$\max_{q \leq \frac{m}{s} \log s} \left\{ \frac{1}{\sqrt{qs}} \| \sup_{u \in T} \sum_{j=1}^{n} \eta_j u_j g_j \|_{L_\eta^q L_g^1} \right\} \ll 1$$

General theorem

$T \subseteq S^{n-1}$ arbitrary. Suffices to set

$$m \gg g^2(T) + 1, \qquad s \gg 1$$

such that

$$\max_{q \leq \frac{m}{s} \log s} \left\{ \frac{1}{\sqrt{qs}} \| \sup_{u \in T} \sum_{j=1}^{n} \eta_j u_j g_j \|_{L_\eta^q L_g^1} \right\} \ll 1$$

Proof ingredients

- Dual Sudakov minoration (already seen, [BLM'89], [PTJ'86]): $\sup_{\eta>0} \eta \cdot [\log \mathcal{N}(E, \|\|\cdot\|\|, \eta)]^{1/2} \leq \mathbb{E}_g \, \|\|Ug\|\|$
- Maurey's empirical method [Pisier '81], [Carl '85]
- Duality of entropy numbers [Bourgain, Pajor, Szarek, Tomczak-Jaegermann '89]

General theorem: proof outline

- Recall: $\gamma_2(Y, \|\cdot\|_X) \ll \sup_{\eta>0} \eta \cdot [\log \mathcal{N}(Y, \|\cdot\|_X, \eta)]^{1/2}$
- So we have to bound $\log \mathcal{N}(T, |||\cdot|||, \eta)$

General theorem: proof outline

- Recall: $\gamma_2(Y, \|\cdot\|_X) \ll \sup_{\eta>0} \eta \cdot [\log \mathcal{N}(Y, \|\cdot\|_X, \eta)]^{1/2}$
- So we have to bound $\log \mathcal{N}(T, |||\cdot|||, \eta)$
- Using duality of entropy numbers [Bourgain, Pajor, Szarek, Tomczak-Jaegermann '89], can bound

$$\log \mathcal{N}(T, |||\cdot|||, \eta) \ll \log \mathcal{N}(\mathrm{conv}(B_{J_i}), \|\cdot\|_T, \frac{1}{8}\sqrt{s}\eta)$$

where $\|z\|_T = \sup_{u \in T} |\langle u, z \rangle|$.

General theorem: proof outline

- Recall: $\gamma_2(Y, \|\cdot\|_X) \ll \sup_{\eta>0} \eta \cdot [\log \mathcal{N}(Y, \|\cdot\|_X, \eta)]^{1/2}$
- So we have to bound $\log \mathcal{N}(T, |||\cdot|||, \eta)$
- Using duality of entropy numbers [Bourgain, Pajor, Szarek, Tomczak-Jaegermann '89], can bound

$$\log \mathcal{N}(T, |||\cdot|||, \eta) \ll \log \mathcal{N}(\mathrm{conv}(B_{J_i}), \|\cdot\|_T, \frac{1}{8}\sqrt{s}\eta)$$

  where $\|z\|_T = \sup_{u \in T} |\langle u, z \rangle|$.
- Ultimately want to use dual Sudakov; we're not there yet

General theorem: proof outline

- Recall: $\gamma_2(Y, \|\cdot\|_X) \ll \sup_{\eta>0} \eta \cdot [\log \mathcal{N}(Y, \|\cdot\|_X, \eta)]^{1/2}$
- So we have to bound $\log \mathcal{N}(T, |||\cdot|||, \eta)$
- Using duality of entropy numbers [Bourgain, Pajor, Szarek, Tomczak-Jaegermann '89], can bound

$$\log \mathcal{N}(T, |||\cdot|||, \eta) \ll \log \mathcal{N}(\mathrm{conv}(B_{J_i}), \|\cdot\|_T, \frac{1}{8}\sqrt{s}\eta)$$

  where $\|z\|_T = \sup_{u \in T} |\langle u, z \rangle|$.
- Ultimately want to use dual Sudakov; we're not there yet
- **Maurey's lemma:** A tool to bound covering numbers of convex combinations of spaces

$$\log \mathcal{N}(\mathrm{conv}(B_{J_i}), \|\cdot\|_T, \epsilon) \lesssim \frac{\log m}{\epsilon^2} + \log \frac{1}{\epsilon} \max_{k \le \frac{1}{\epsilon^2}} \max_{|A|=k} \log \mathcal{N}(\frac{1}{k} \sum_{i \in A} B_{J_i}, \|\cdot\|_T, \epsilon)$$

General theorem: proof outline

- Define $U_\alpha = U_\alpha(\delta) = \{j = 1, \ldots, n : \sum_{i \in A} \delta_{i,j} \sim 2^\alpha\}$

General theorem: proof outline

- Define $U_\alpha = U_\alpha(\delta) = \{j = 1, \ldots, n : \sum_{i \in A} \delta_{i,j} \sim 2^\alpha\}$
- Can show $\frac{1}{k} \sum_{i \in A} B_{J_i} \subset \sum_\alpha \frac{1}{\sqrt{k}} 2^{\alpha/2} E_\alpha$

  where E_α is the linear subspace of vectors supported on U_α

General theorem: proof outline

- Define $U_\alpha = U_\alpha(\delta) = \{j = 1, \ldots, n : \sum_{i \in A} \delta_{i,j} \sim 2^\alpha\}$
- Can show $\frac{1}{k} \sum_{i \in A} B_{J_i} \subset \sum_\alpha \frac{1}{\sqrt{k}} 2^{\alpha/2} E_\alpha$

  where E_α is the linear subspace of vectors supported on U_α
- Therefore

  $\log \mathcal{N}(\frac{1}{k} \sum_{i \in A} B_{J_i}, \|\cdot\|_T, \epsilon) \lesssim \sum_\alpha \log \mathcal{N}(\frac{1}{\sqrt{k}} 2^{\alpha/2} E_\alpha, \|\cdot\|_T, \frac{\epsilon}{\log m})$

General theorem: proof outline

- Define $U_\alpha = U_\alpha(\delta) = \{j = 1, \ldots, n : \sum_{i \in A} \delta_{i,j} \sim 2^\alpha\}$
- Can show $\frac{1}{k} \sum_{i \in A} B_{J_i} \subset \sum_\alpha \frac{1}{\sqrt{k}} 2^{\alpha/2} E_\alpha$

  where E_α is the linear subspace of vectors supported on U_α
- Therefore

  $\log \mathcal{N}(\frac{1}{k} \sum_{i \in A} B_{J_i}, \|\cdot\|_T, \epsilon) \lesssim \sum_\alpha \log \mathcal{N}(\frac{1}{\sqrt{k}} 2^{\alpha/2} E_\alpha, \|\cdot\|_T, \frac{\epsilon}{\log m})$
- Dual Sudakov minoration! (. . . plus some calculations)

Conclusion

Open Problems

- **OPEN:** Prove conjecture: to get subsp. embedding with prob. $1-\delta$, can set $m = O((d + \log(1/\delta))/\varepsilon^2), s = O(\log(d/\delta)/\varepsilon)$. Easier: obtain this m with $s = m$ **via moment method**. ([Cohen '14] has made progress via other methods)
- **OPEN:** Show that the tradeoff $m = O(d^{1+\gamma}/\varepsilon^2)$, $s = \mathrm{poly}(1/\gamma) \cdot 1/\varepsilon$ is optimal for any distribution over subspace embeddings (the poly is probably linear; some progress in [N., Nguyễn '14])
- **OPEN:** Show that $m = \Omega(d^2/\varepsilon^2)$ is optimal for $s = 1$ Partial progress: [N., Nguyễn '13] shows $m = \Omega(d^2)$
- **OPEN:** Improve parameters of [Bourgain, N. '14]: s should be $\gg 1/\varepsilon^2$, not $1/\varepsilon^2$, and remove log factors.
- **OPEN:** Give a full tradeoff curve for s, m for general theorem
- **OPEN:** Usual: m, s "big enough" $\Rightarrow$ T preserved $\Rightarrow$ application works Can we skip the middle step to get better bounds? e.g. for k-means clustering?