Toward a unified theory of sparse dimensionality reduction in Euclidean space

Jean Bourgain (IAS Princeton)
Sjoerd Dirksen (RWTH Aachen)
Jelani Nelson (Harvard)

SPARS 2015, Cambridge, UK
July 9, 2015

Menu

Dimensionality reduction with Gaussian matrices

Sparse analog of Gordon's theorem

Applications

Dimensionality reduction with Gaussian matrices

Dimensionality reduction

Goal: map data $X \subset \mathbb{R}^n$ into $\mathbb{R}^m$ with $m \ll n$ using $f : X \to \mathbb{R}^m$.

Dimensionality reduction

Goal: map data $X \subset \mathbb{R}^n$ into $\mathbb{R}^m$ with $m \ll n$ using $f : X \to \mathbb{R}^m$.
Benefits:

- Smaller storage consumption;
- Speedup during data analysis;
- Cheaper transmission of data across computing clusters.

Dimensionality reduction

Goal: map data $X \subset \mathbb{R}^n$ into $\mathbb{R}^m$ with $m \ll n$ using $f : X \to \mathbb{R}^m$.
Benefits:

- Smaller storage consumption;
- Speedup during data analysis;
- Cheaper transmission of data across computing clusters.

Depending on the application we may want that

- f is linear, $f(x) = \Phi x$;
- f is oblivious to the data X;
- f preserves relevant structural information.

Dimensionality reduction

Goal: map data $X \subset \mathbb{R}^n$ into $\mathbb{R}^m$ with $m \ll n$ using $f : X \to \mathbb{R}^m$.
Benefits:

- Smaller storage consumption;
- Speedup during data analysis;
- Cheaper transmission of data across computing clusters.

Depending on the application we may want that

- f is linear, $f(x) = \Phi x$;
- f is oblivious to the data X;
- f preserves relevant structural information.

This talk: want ε-*isometry* $\Phi \in \mathbb{R}^{m\times n}$

$$(1-\varepsilon)\|x-y\|_2^2 \leq \|\Phi(x-y)\|_2^2 \leq (1+\varepsilon)\|x-y\|_2^2 \qquad \forall x, y \in X \quad (1)$$

Dimensionality reduction

Goal: map data $X \subset \mathbb{R}^n$ into $\mathbb{R}^m$ with $m \ll n$ using $f : X \to \mathbb{R}^m$.
Benefits:

- Smaller storage consumption;
- Speedup during data analysis;
- Cheaper transmission of data across computing clusters.

Depending on the application we may want that

- f is linear, $f(x) = \Phi x$;
- f is oblivious to the data X;
- f preserves relevant structural information.

This talk: want ε-*isometry* $\Phi \in \mathbb{R}^{m \times n}$

$$(1-\varepsilon)\|x-y\|_2^2 \leq \|\Phi(x-y)\|_2^2 \leq (1+\varepsilon)\|x-y\|_2^2 \qquad \forall x, y \in X \quad (1)$$

If $T = \{(x-y)/\|x-y\|_2 : x, y \in X\} \subset S^{n-1}$, want

$$\varepsilon_T := \sup_{x \in T} \left| \, \|\Phi x\|_2^2 - 1 \right| < \varepsilon. \qquad (2)$$

Gordon's theorem

Let $g(T) = \mathbb{E} \sup_{x \in T} \langle x, g \rangle$ be the *Gaussian width* of T.

Gordon's theorem

Let $g(T) = \mathbb{E}\sup_{x\in T}\langle x, g\rangle$ be the *Gaussian width* of T.

Theorem (Gordon '88)

Let $T \subset S^{n-1}$. If Φ is a $m \times n$ standard Gaussian matrix,

$$m \gtrsim \varepsilon^{-2}\ (g(T)^2 + \log(\eta^{-1})),$$

then $\frac{1}{\sqrt{m}}\Phi$ is an ε-isometry on T with probability $\geq 1-\eta$.

Gordon's theorem

Let $g(T) = \mathbb{E}\sup_{x\in T}\langle x, g\rangle$ be the *Gaussian width* of T.

Theorem (Gordon '88)

Let $T \subset S^{n-1}$. If Φ is a $m \times n$ standard Gaussian matrix,

$$m \gtrsim \varepsilon^{-2}\,(g(T)^2 + \log(\eta^{-1})),$$

then $\frac{1}{\sqrt{m}}\Phi$ is an ε-isometry on T with probability $\geq 1-\eta$.

- This is an 'instance-optimal' version of J-L lemma: $g(T)^2 \lesssim \log|T|$, but can be much smaller.

Gordon's theorem

Let $g(T) = \mathbb{E} \sup_{x \in T} \langle x, g \rangle$ be the *Gaussian width* of T.

Theorem (Gordon '88)

Let $T \subset S^{n-1}$. If Φ is a $m \times n$ standard Gaussian matrix,

$$m \gtrsim \varepsilon^{-2} \left(g(T)^2 + \log(\eta^{-1}) \right),$$

then $\frac{1}{\sqrt{m}} \Phi$ is an ε-isometry on T with probability $\geq 1 - \eta$.

- This is an 'instance-optimal' version of J-L lemma: $g(T)^2 \lesssim \log |T|$, but can be much smaller.
- If $T = k$-sparse vectors in S^{n-1}, then $g(T)^2 \lesssim k \log(n/k)$.

Gordon's theorem

Let $g(T) = \mathbb{E} \sup_{x \in T} \langle x, g \rangle$ be the *Gaussian width* of T.

Theorem (Gordon '88)

Let $T \subset S^{n-1}$. If Φ is a $m \times n$ standard Gaussian matrix,

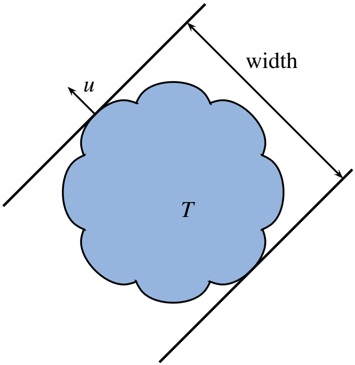
$$m \gtrsim \varepsilon^{-2} \left(g(T)^2 + \log(\eta^{-1}) \right),$$

then $\frac{1}{\sqrt{m}} \Phi$ is an ε-isometry on T with probability $\geq 1 - \eta$.

- This is an 'instance-optimal' version of J-L lemma: $g(T)^2 \lesssim \log |T|$, but can be much smaller.
- If $T = k$-sparse vectors in S^{n-1}, then $g(T)^2 \lesssim k \log(n/k)$.
- Theorem extends to random sign matrices: Klartag-Mendelson '05, Mendelson-Pajor-Tomczak-Jaegermann '07, see also D. '14.

Geometric interpretation

$$g(T-T) = \mathbb{E} \sup_{x \in T-T} \langle x, g \rangle = c_n \, \mathbb{E} \sup_{x \in T-T} \left\langle x, \frac{g}{\|g\|_2} \right\rangle, \quad c_n \sim \sqrt{n}.$$



$$\sup_{x \in T-T} \left\langle x, \frac{g}{\|g\|_2} \right\rangle = \text{width of } T \text{ in direction } u = g/\|g\|_2$$

Sketching constrained least squares

$A \in \mathbb{R}^{n\times d}$, $\mathcal{C} \subset \mathbb{R}^d$ closed convex set.

Sketching constrained least squares

$A \in \mathbb{R}^{n \times d}$, $\mathcal{C} \subset \mathbb{R}^d$ closed convex set. x_{opt} a minimizer of

$$\min \|Ax - b\|_2^2 \qquad \text{subject to} \qquad x \in \mathcal{C}. \tag{3}$$

Sketching constrained least squares

$A \in \mathbb{R}^{n\times d}$, $\mathcal{C} \subset \mathbb{R}^d$ closed convex set. x_{opt} a minimizer of

$$\min \|Ax - b\|_2^2 \qquad \text{subject to} \qquad x \in \mathcal{C}. \tag{3}$$

$\mathcal{C} = \mathbb{R}^d$ (unconstrained), $\mathcal{C} = B_{\ell_1}$ (LASSO$\to$sparse solution),
$\mathcal{C} = B_{\ell_1(\ell_2)}$ (group LASSO$\to$block sparse solution).

Sketching constrained least squares

$A \in \mathbb{R}^{n\times d}$, $\mathcal{C} \subset \mathbb{R}^d$ closed convex set. x_{opt} a minimizer of

$$\min \|Ax - b\|_2^2 \qquad \text{subject to} \qquad x \in \mathcal{C}. \tag{3}$$

$\mathcal{C} = \mathbb{R}^d$ (unconstrained), $\mathcal{C} = B_{\ell_1}$ (LASSO$\to$sparse solution),
$\mathcal{C} = B_{\ell_1(\ell_2)}$ (group LASSO$\to$block sparse solution).

Let $\Phi \in \mathbb{R}^{m\times n}$ be a sketching matrix. x_{S} a minimizer of the *sketched program*

$$\min \|\Phi Ax - \Phi b\|_2^2 \qquad \text{subject to} \qquad x \in \mathcal{C}.$$

Is x_S also a 'good' solution for the original?

Sketching constrained least squares

$A \in \mathbb{R}^{n \times d}$, $\mathcal{C} \subset \mathbb{R}^d$ closed convex set. x_{opt} a minimizer of

$$\min \|Ax - b\|_2^2 \qquad \text{subject to} \qquad x \in \mathcal{C}. \tag{3}$$

$\mathcal{C} = \mathbb{R}^d$ (unconstrained), $\mathcal{C} = B_{\ell_1}$ (LASSO$\to$sparse solution), $\mathcal{C} = B_{\ell_1(\ell_2)}$ (group LASSO$\to$block sparse solution).

Let $\Phi \in \mathbb{R}^{m \times n}$ be a sketching matrix. x_S a minimizer of the *sketched program*

$$\min \|\Phi Ax - \Phi b\|_2^2 \qquad \text{subject to} \qquad x \in \mathcal{C}.$$

Is x_S also a 'good' solution for the original?

First analysis by Sarlós '06 for $\mathcal{C} = \mathbb{R}^d$. Note: if Φ,A dense, then embedding time$>$time to solve (3).

Sparse analog of Gordon's theorem

Sparse Johnson-Lindenstrauss Transform (SJLT)

Dasgupta-Kumar-Sarlòs '10, Braverman-Ostrovsky-Rabani '10, Kane-Nelson '10,'14.

Sparse Johnson-Lindenstrauss Transform (SJLT)

Dasgupta-Kumar-Sarlòs '10, Braverman-Ostrovsky-Rabani '10, Kane-Nelson '10,'14.

- Start with $\Sigma \in \mathbb{R}^{m \times n}$, $\Sigma = (\sigma_{ij})$, σ_{ij} i.i.d. random signs.

Sparse Johnson-Lindenstrauss Transform (SJLT)

Dasgupta-Kumar-Sarlòs '10, Braverman-Ostrovsky-Rabani '10, Kane-Nelson '10,'14.

- Start with $\Sigma \in \mathbb{R}^{m \times n}$, $\Sigma = (\sigma_{ij})$, σ_{ij} i.i.d. random signs.
- For each column *independently*, select *exactly* s entries uniformly at random without replacement, put rest to 0 $\rightarrow$ random selectors $\delta_{ij} \in \{0, 1\}$.

Sparse Johnson-Lindenstrauss Transform (SJLT)

Dasgupta-Kumar-Sarlòs '10, Braverman-Ostrovsky-Rabani '10, Kane-Nelson '10,'14.

- Start with $\Sigma \in \mathbb{R}^{m \times n}$, $\Sigma = (\sigma_{ij})$, σ_{ij} i.i.d. random signs.
- For each column *independently*, select *exactly* s entries uniformly at random without replacement, put rest to $0 \rightarrow$ random selectors $\delta_{ij} \in \{0, 1\}$.
- SJLT$=\Phi$,

$$\Phi_{ij} = \frac{1}{\sqrt{s}} \sigma_{ij} \delta_{ij}.$$

Sparse Johnson-Lindenstrauss Transform (SJLT)

Dasgupta-Kumar-Sarlòs '10, Braverman-Ostrovsky-Rabani '10, Kane-Nelson '10,'14.

- Start with $\Sigma \in \mathbb{R}^{m \times n}$, $\Sigma = (\sigma_{ij})$, σ_{ij} i.i.d. random signs.
- For each column *independently*, select *exactly s* entries uniformly at random without replacement, put rest to $0 \to$ random selectors $\delta_{ij} \in \{0, 1\}$.
- SJLT$=\Phi$,

$$\Phi_{ij} = \frac{1}{\sqrt{s}} \sigma_{ij} \delta_{ij}.$$

Note: $\mathbb{E}\|\Phi x\|_2^2 = \|x\|_2^2$. Φx can be computed in time $O(s\|x\|_0)$.

Central question

How large do m and s need to be to get

$$\mathbb{E}\varepsilon_T = \mathbb{E}\sup_{x \in T} \left| \|\Phi x\|_2^2 - 1 \right| < \varepsilon?$$

Answer depends on appropriate 'complexity' parameter of T.

Central question

How large do m and s need to be to get

$$\mathbb{E}\varepsilon_T = \mathbb{E}\sup_{x\in T} \left| \|\Phi x\|_2^2 - 1 \right| < \varepsilon?$$

Answer depends on appropriate 'complexity' parameter of T.

Previously known:

- finite set: $m \gtrsim \varepsilon^{-2}\log|T|$, $s \gtrsim \varepsilon^{-1}\log|T| = \varepsilon m$.
  (Kane-Nelson '14)

Central question

How large do m and s need to be to get

$$\mathbb{E}\varepsilon_T = \mathbb{E}\sup_{x\in T} \left|\|\Phi x\|_2^2 - 1\right| < \varepsilon?$$

Answer depends on appropriate 'complexity' parameter of T.

Previously known:

- finite set: $m \gtrsim \varepsilon^{-2}\log|T|$, $s \gtrsim \varepsilon^{-1}\log|T| = \varepsilon m$. (Kane-Nelson '14)
- $T = E \cap S^{n-1}$, d-dim subspace E: $m \gtrsim \varepsilon^{-2} d \operatorname{polylog}(d)$, $s \gtrsim \varepsilon^{-1}\operatorname{polylog}(d)$. (Nelson-Nguyen '13)

Central question

How large do m and s need to be to get

$$\mathbb{E}\varepsilon_T = \mathbb{E}\sup_{x\in T} |\|\Phi x\|_2^2 - 1| < \varepsilon?$$

Answer depends on appropriate 'complexity' parameter of T.

Previously known:

- finite set: $m \gtrsim \varepsilon^{-2}\log|T|$, $s \gtrsim \varepsilon^{-1}\log|T| = \varepsilon m$. (Kane-Nelson '14)
- $T = E \cap S^{n-1}$, d-dim subspace E: $m \gtrsim \varepsilon^{-2} d\,\mathrm{polylog}(d)$, $s \gtrsim \varepsilon^{-1}\mathrm{polylog}(d)$. (Nelson-Nguyen '13)
- $T = k$-sparse vectors in S^{n-1}, $\varepsilon = C$: if $m \gtrsim k\log(n/k)$ then must have $s \gtrsim m$. (Nelson-Nguyen '13)

New complexity parameter

Let $T \subset S^{n-1}$. Define the complexity parameter

$$\kappa(T) := \max_{q \leq \frac{m}{s} \log s} \Big\{ \frac{1}{\sqrt{qs}} \Big(\mathbb{E}_\eta \Big(\mathbb{E}_g \sup_{x \in T} \big| \sum_{j=1}^{n} \eta_j g_j x_j \big| \Big)^q \Big)^{1/q} \Big\},$$

where

- (g_j) are i.i.d. standard Gaussian;
- (η_j) i.i.d. Bernoulli with mean $qs/(m \log s)$.

New complexity parameter

Let $T \subset S^{n-1}$. Define the complexity parameter

$$\kappa(T) := \max_{q \leq \frac{m}{s} \log s} \Big\{ \frac{1}{\sqrt{qs}} \Big(\mathbb{E}_\eta \Big(\mathbb{E}_g \sup_{x \in T} \big| \sum_{j=1}^{n} \eta_j g_j x_j \big| \Big)^q \Big)^{1/q} \Big\},$$

where

- (g_j) are i.i.d. standard Gaussian;
- (η_j) i.i.d. Bernoulli with mean $qs/(m \log s)$.

Taking $q = \frac{m}{s} \log s$ shows

$$\frac{g(T)}{\sqrt{m \log s}} \leq \kappa(T).$$

New complexity parameter

Let $T \subset S^{n-1}$. Define the complexity parameter

$$\kappa(T) := \max_{q \leq \frac{m}{s}\log s} \Big\{ \frac{1}{\sqrt{qs}} \Big(\mathbb{E}_\eta \Big(\mathbb{E}_g \sup_{x \in T} \Big| \sum_{j=1}^n \eta_j g_j x_j \Big| \Big)^q \Big)^{1/q} \Big\},$$

where

- (g_j) are i.i.d. standard Gaussian;
- (η_j) i.i.d. Bernoulli with mean $qs/(m \log s)$.

Taking $q = \frac{m}{s}\log s$ shows

$$\frac{g(T)}{\sqrt{m \log s}} \leq \kappa(T).$$

Kahane's contraction principle shows

$$\kappa(T) \leq \frac{1}{\sqrt{s}} g(T).$$

Sparse analog of Gordon's theorem

Theorem (Bourgain-D.-Nelson '15)

Suppose that

$$m \gtrsim (\log m)^3 (\log n)^5 \cdot \frac{(g^2(T)+1)}{\varepsilon^2} \tag{4}$$

$$s \gtrsim (\log m)^6 (\log n)^4 \cdot \frac{1}{\varepsilon^2}. \tag{5}$$

Sparse analog of Gordon's theorem

Theorem (Bourgain-D.-Nelson '15)

Suppose that

$$m \gtrsim (\log m)^3 (\log n)^5 \cdot \frac{(g^2(T)+1)}{\varepsilon^2} \tag{4}$$

$$s \gtrsim (\log m)^6 (\log n)^4 \cdot \frac{1}{\varepsilon^2}. \tag{5}$$

Then $\mathbb{E}\sup_{x\in T} \left| \|\Phi x\|_2^2 - 1 \right| < \varepsilon$ *as long as* s, m *furthermore satisfy*

$$\kappa(T) < \frac{\varepsilon}{(\log m)^2 (\log n)^{5/2}}.$$

Sparse analog of Gordon's theorem

Theorem (Bourgain-D.-Nelson '15)

Suppose that

$$m \gtrsim (\log m)^3 (\log n)^5 \cdot \frac{(g^2(T)+1)}{\varepsilon^2} \tag{4}$$

$$s \gtrsim (\log m)^6 (\log n)^4 \cdot \frac{1}{\varepsilon^2}. \tag{5}$$

Then $\mathbb{E}\sup_{x\in T} |\|\Phi x\|_2^2 - 1| < \varepsilon$ *as long as* s, m *furthermore satisfy*

$$\kappa(T) < \frac{\varepsilon}{(\log m)^2 (\log n)^{5/2}}.$$

- (4) is redundant up to log-factors: set $q = (m \log s)/s$.

Sparse analog of Gordon's theorem

Theorem (Bourgain-D.-Nelson '15)

Suppose that

$$m \gtrsim (\log m)^3 (\log n)^5 \cdot \frac{(g^2(T)+1)}{\varepsilon^2} \tag{4}$$

$$s \gtrsim (\log m)^6 (\log n)^4 \cdot \frac{1}{\varepsilon^2}. \tag{5}$$

Then $\mathbb{E}\sup_{x \in T} |\|\Phi x\|_2^2 - 1| < \varepsilon$ *as long as* s, m *furthermore satisfy*

$$\kappa(T) < \frac{\varepsilon}{(\log m)^2 (\log n)^{5/2}}.$$

- (4) is redundant up to log-factors: set $q = (m \log s)/s$.
- Several log-factors are believed to be redundant. Not easy to remove!

Sparse analog of Gordon's theorem

Theorem (Bourgain-D.-Nelson '15)

Suppose that

$$m \gtrsim (\log m)^3 (\log n)^5 \cdot \frac{(g^2(T)+1)}{\varepsilon^2} \tag{4}$$

$$s \gtrsim (\log m)^6 (\log n)^4 \cdot \frac{1}{\varepsilon^2}. \tag{5}$$

Then $\mathbb{E}\sup_{x\in T} |\|\Phi x\|_2^2 - 1| < \varepsilon$ *as long as* s, m *furthermore satisfy*

$$\kappa(T) < \frac{\varepsilon}{(\log m)^2 (\log n)^{5/2}}.$$

- (4) is redundant up to log-factors: set $q = (m \log s)/s$.
- Several log-factors are believed to be redundant. Not easy to remove!

- Need to estimate $\kappa(T)$ to get explicit conditions on s, m. Use 'standard' tools from Banach space theory.

- Need to estimate $\kappa(T)$ to get explicit conditions on s, m. Use 'standard' tools from Banach space theory.
- (Qualitatively) unifies known results for specific sets.

- Need to estimate $\kappa(T)$ to get explicit conditions on s, m. Use 'standard' tools from Banach space theory.
- (Qualitatively) unifies known results for specific sets.
- New results for (infinite) unions of subspaces and manifolds.

- Need to estimate $\kappa(T)$ to get explicit conditions on s, m. Use 'standard' tools from Banach space theory.
- (Qualitatively) unifies known results for specific sets.
- New results for (infinite) unions of subspaces and manifolds.
- Proof ingredients: generic chaining for supremum of 2nd order chaos process (Krahmer-Mendelson-Rauhut '14), dual Sudakov inequality, entropy duality (Bourgain et al., '89), Maurey's lemma.

- Need to estimate $\kappa(T)$ to get explicit conditions on s, m. Use 'standard' tools from Banach space theory.
- (Qualitatively) unifies known results for specific sets.
- New results for (infinite) unions of subspaces and manifolds.
- Proof ingredients: generic chaining for supremum of 2nd order chaos process (Krahmer-Mendelson-Rauhut '14), dual Sudakov inequality, entropy duality (Bourgain et al., '89), Maurey's lemma.
- Simpler proof for subspace case. For general T: reduce (in many steps) to subspace case.

- Need to estimate $\kappa(T)$ to get explicit conditions on s, m. Use 'standard' tools from Banach space theory.
- (Qualitatively) unifies known results for specific sets.
- New results for (infinite) unions of subspaces and manifolds.
- Proof ingredients: generic chaining for supremum of 2nd order chaos process (Krahmer-Mendelson-Rauhut '14), dual Sudakov inequality, entropy duality (Bourgain et al., '89), Maurey's lemma.
- Simpler proof for subspace case. For general T: reduce (in many steps) to subspace case.
- $A \underset{*}{<} B$ means $A \leq \text{polylog}(n, m) B$.

Applications

Sparse subspace embedding

$E \subset \mathbb{R}^n$ is a d-dimensional *subspace*, $T = \{x \in E \ : \ \|x\|_2 = 1\}$,

Sparse subspace embedding

$E \subset \mathbb{R}^n$ is a d-dimensional *subspace*, $T = \{x \in E \ : \ \|x\|_2 = 1\}$,

$$\mu(E) = \max_{1 \le j \le n} \|P_E e_j\|_2 = \text{incoherence}\,.$$

Note: $\sqrt{d/n} \le \mu(E) \le 1$.

Sparse subspace embedding

$E \subset \mathbb{R}^n$ is a d-dimensional *subspace*, $T = \{x \in E \ : \ \|x\|_2 = 1\}$,

$$\mu(E) = \max_{1 \le j \le n} \|P_E e_j\|_2 = \text{incoherence}\,.$$

Note: $\sqrt{d/n} \le \mu(E) \le 1$. To get

$$\mathbb{E} \sup_{x \in T} \big|\ \|\Phi x\|_2^2 - 1\big| < \varepsilon$$

it is suffices if

$$m \gtrsim \varepsilon^{-2} d (\log m)^3, \qquad s \gtrsim \varepsilon^{-2} \mu(E)^2 (\log m)^3.$$

Sparse subspace embedding

$E \subset \mathbb{R}^n$ is a d-dimensional *subspace*, $T = \{x \in E \ : \ \|x\|_2 = 1\}$,

$$\mu(E) = \max_{1 \le j \le n} \|P_E e_j\|_2 = \text{incoherence}\,.$$

Note: $\sqrt{d/n} \le \mu(E) \le 1$. To get

$$\mathbb{E} \sup_{x \in T} \big|\ \|\Phi x\|_2^2 - 1\big| < \varepsilon$$

it is suffices if

$$m \gtrsim \varepsilon^{-2} d (\log m)^3, \qquad s \gtrsim \varepsilon^{-2} \mu(E)^2 (\log m)^3.$$

If $\mu(E) \lesssim \varepsilon (\log m)^{-3/2}$, then $s = 1$ suffices.

Sparse subspace embedding

$E \subset \mathbb{R}^n$ is a d-dimensional *subspace*, $T = \{x \in E \ : \ \|x\|_2 = 1\}$,

$$\mu(E) = \max_{1 \le j \le n} \|P_E e_j\|_2 = \text{incoherence}\,.$$

Note: $\sqrt{d/n} \le \mu(E) \le 1$. To get

$$\mathbb{E} \sup_{x \in T} \big| \, \|\Phi x\|_2^2 - 1 \big| < \varepsilon$$

it is suffices if

$$m \gtrsim \varepsilon^{-2} d (\log m)^3, \qquad s \gtrsim \varepsilon^{-2} \mu(E)^2 (\log m)^3.$$

If $\mu(E) \lesssim \varepsilon (\log m)^{-3/2}$, then $s = 1$ suffices.

A *random* subspace E has $\mu(E) \simeq \sqrt{d/n}$ w.h.p. for $d \gtrsim \log n$ (by JL lemma).

Sparse sketch for LASSO

$$\min \|Ax - b\|_2^2 \qquad \text{subject to} \qquad \|x\|_1 \le R,$$

x_{opt} a minimizer.

Sparse sketch for LASSO

$$\min \|Ax - b\|_2^2 \qquad \text{subject to} \qquad \|x\|_1 \leq R,$$

x_{opt} a minimizer.

$$\sigma_{\min,k} = \inf_{\|y\|_2=1,\ \|y\|_1 \leq 2\sqrt{k}} \|Ay\|_2.$$

Sparse sketch for LASSO

$$\min \|Ax - b\|_2^2 \qquad \text{subject to} \qquad \|x\|_1 \le R,$$

x_{opt} a minimizer.

$$\sigma_{\min,k} = \inf_{\|y\|_2=1,\ \|y\|_1 \le 2\sqrt{k}} \|Ay\|_2.$$

If x_{opt} is k-sparse and $\|x_{\text{opt}}\|_1 = R$,

Sparse sketch for LASSO

$$\min \|Ax - b\|_2^2 \qquad \text{subject to} \qquad \|x\|_1 \le R,$$

x_{opt} a minimizer.

$$\sigma_{\min,k} = \inf_{\|y\|_2=1,\ \|y\|_1 \le 2\sqrt{k}} \|Ay\|_2.$$

If x_{opt} is k-sparse and $\|x_{\text{opt}}\|_1 = R$,

$$m \gtrsim_* \varepsilon^{-2} k \sigma_{\min,k}^{-2} \max_j \|A_j\|_2^2, \tag{6}$$

$$s \gtrsim_* \varepsilon^{-2} k \sigma_{\min,k}^{-2} \max_{i,j} |A_{ij}|^2,$$

then with high probability

$$\|Ax_S - b\|_2^2 \le \frac{1}{(1-\varepsilon)^2} \|Ax_{\text{opt}} - b\|_2^2.$$

Pilanci-Wainwright '14 showed for *Gaussian* Φ:

$$m \gtrsim \varepsilon^{-2} k \sigma_{\min,k}^{-2} \max_j \|A_j\|_2^2.$$

(6) also suffices for 'Fast' JLT.

Union of subspaces, model-based CS

Θ collection of N d-dimensional subspaces $E \subset \mathbb{R}^n$.

$$T = \bigcup_{E \in \Theta} \{x \in E : \|x\|_2 = 1\}.$$

Union of subspaces, model-based CS

Θ collection of N d-dimensional subspaces $E \subset \mathbb{R}^n$.

$$T = \bigcup_{E \in \Theta} \{x \in E : \|x\|_2 = 1\}.$$

$$\alpha = \sup_{E \in \Theta} \max_j \|P_E e_j\|_2 = \text{largest incoherence in } \Theta.$$

Note: $\sqrt{d/n} \leq \alpha \leq 1$.

Union of subspaces, model-based CS

Θ collection of N d-dimensional subspaces $E \subset \mathbb{R}^n$.

$$T = \bigcup_{E \in \Theta} \{x \in E : \|x\|_2 = 1\}.$$

$$\alpha = \sup_{E \in \Theta} \max_j \|P_E e_j\|_2 = \text{largest incoherence in } \Theta.$$

Note: $\sqrt{d/n} \le \alpha \le 1$. Then

$$\mathbb{E} \sup_{x \in T} |\ \|\Phi x\|_2^2 - 1| < \varepsilon$$

if (incoherent case)

$$m \underset{*}{>} \varepsilon^{-2}(d + \log N), \qquad s \underset{*}{>} \varepsilon^{-2}(1 + (\alpha \log N)^2)$$

Union of subspaces, model-based CS

Θ collection of N d-dimensional subspaces $E \subset \mathbb{R}^n$.

$$T = \bigcup_{E \in \Theta} \{x \in E : \|x\|_2 = 1\}.$$

$$\alpha = \sup_{E \in \Theta} \max_j \|P_E e_j\|_2 = \text{largest incoherence in } \Theta.$$

Note: $\sqrt{d/n} \leq \alpha \leq 1$. Then

$$\mathbb{E} \sup_{x \in T} | \ \|\Phi x\|_2^2 - 1| < \varepsilon$$

if (incoherent case)

$$m \gtrsim_* \varepsilon^{-2}(d + \log N), \qquad s \gtrsim_* \varepsilon^{-2}(1 + (\alpha \log N)^2)$$

or (coherent case)

$$m \gtrsim_* \varepsilon^{-2}(d + \log N), \qquad s \gtrsim_* \varepsilon^{-2}(1 + \log N)$$

Union of subspaces, model-based CS

Θ collection of N d-dimensional subspaces $E \subset \mathbb{R}^n$.

$$T = \bigcup_{E \in \Theta} \{x \in E : \|x\|_2 = 1\}.$$

$$\alpha = \sup_{E \in \Theta} \max_j \|P_E e_j\|_2 = \text{largest incoherence in } \Theta.$$

Note: $\sqrt{d/n} \leq \alpha \leq 1$. Then

$$\mathbb{E} \sup_{x \in T} | \, \|\Phi x\|_2^2 - 1| < \varepsilon$$

if (incoherent case)

$$m \gtrsim_* \varepsilon^{-2}(d + \log N), \qquad s \gtrsim_* \varepsilon^{-2}(1 + (\alpha \log N)^2)$$

or (coherent case)

$$m \gtrsim_* \varepsilon^{-2}(d + \log N), \qquad s \gtrsim_* \varepsilon^{-2}(1 + \log N)$$

Previous best: $m \gtrsim_* \varepsilon^{-2}(d + \log N)$, $s = m$ (Blumensath-Davies '09), $m \gtrsim_* \varepsilon^{-2} d \cdot (\log N)^6$, $s \gtrsim_* \varepsilon^{-1} (\log N)^3$ (Nelson-Nguyen '13).

Manifolds

$\mathcal{M} \subset \mathbb{R}^n$ d-dimensional manifold of the form $\mathcal{M} = F(B_{\ell_2^d})$,

Manifolds

$\mathcal{M} \subset \mathbb{R}^n$ d-dimensional manifold of the form $\mathcal{M} = F(B_{\ell_2^d})$,

- F is bi-Lipschitz: $\|F(x) - F(y)\|_2 \simeq \|x - y\|_2, \ \forall x, y$;
- $\|DF(x) - DF(y)\|_{2\to 2} \lesssim \|x - y\|_2$ for $x, y \in B_{\ell_2^d}$.

Manifolds

$\mathcal{M} \subset \mathbb{R}^n$ d-dimensional manifold of the form $\mathcal{M} = F(B_{\ell_2^d})$,

- F is bi-Lipschitz: $\|F(x) - F(y)\|_2 \simeq \|x - y\|_2, \ \forall x, y$;
- $\|DF(x) - DF(y)\|_{2\to 2} \lesssim \|x - y\|_2$ for $x, y \in B_{\ell_2^d}$.

For $x \in \mathcal{M}$, E_x = tangent plane at x. $T\mathcal{M} = \cup_{x\in\mathcal{M}} E_x$ =tangent bundle.

Manifolds

$\mathcal{M} \subset \mathbb{R}^n$ d-dimensional manifold of the form $\mathcal{M} = F(B_{\ell_2^d})$,

- F is bi-Lipschitz: $\|F(x) - F(y)\|_2 \simeq \|x - y\|_2, \ \forall x, y$;
- $\|DF(x) - DF(y)\|_{2\to 2} \lesssim \|x - y\|_2$ for $x, y \in B_{\ell_2^d}$.

For $x \in \mathcal{M}$, E_x = tangent plane at x. $T\mathcal{M} = \cup_{x\in\mathcal{M}} E_x$ =tangent bundle.

$$\alpha = \max_{x\in\mathcal{M}} \max_{1\le j\le n} \|P_{E_x} e_j\|_2 = \text{largest incoherence}\,.$$

Manifolds

$\mathcal{M} \subset \mathbb{R}^n$ d-dimensional manifold of the form $\mathcal{M} = F(B_{\ell_2^d})$,

- F is bi-Lipschitz: $\|F(x) - F(y)\|_2 \simeq \|x - y\|_2, \ \forall x, y$;
- $\|DF(x) - DF(y)\|_{2\to 2} \lesssim \|x - y\|_2$ for $x, y \in B_{\ell_2^d}$.

For $x \in \mathcal{M}$, E_x = tangent plane at x. $T\mathcal{M} = \cup_{x\in\mathcal{M}} E_x$ =tangent bundle.

$$\alpha = \max_{x\in\mathcal{M}} \max_{1\le j\le n} \|P_{E_x} e_j\|_2 = \text{largest incoherence}\,.$$

Suppose

$$m \underset{*}{>} \varepsilon^{-2} d, \qquad s \underset{*}{>} \varepsilon^{-2}(1 + (\alpha d)^2).$$

Manifolds

$\mathcal{M} \subset \mathbb{R}^n$ d-dimensional manifold of the form $\mathcal{M} = F(B_{\ell_2^d})$,

- F is bi-Lipschitz: $\|F(x) - F(y)\|_2 \simeq \|x - y\|_2, \ \forall x, y$;
- $\|DF(x) - DF(y)\|_{2\to 2} \lesssim \|x - y\|_2$ for $x, y \in B_{\ell_2^d}$.

For $x \in \mathcal{M}$, E_x = tangent plane at x. $T\mathcal{M} = \cup_{x\in\mathcal{M}} E_x$ =tangent bundle.

$$\alpha = \max_{x\in\mathcal{M}} \max_{1\leq j\leq n} \|P_{E_x} e_j\|_2 = \text{largest incoherence}\,.$$

Suppose

$$m \underset{*}{>} \varepsilon^{-2} d, \qquad s \underset{*}{>} \varepsilon^{-2}(1 + (\alpha d)^2).$$

Then, with large probability, for all $x, y \in \mathcal{M}$

$$(1-\varepsilon)\rho_{\mathcal{M}}(x,y) \leq \rho_{\Phi(\mathcal{M})}(\Phi x, \Phi y) \leq (1+\varepsilon)\rho_{\mathcal{M}}(x,y).$$

Manifolds

$\mathcal{M} \subset \mathbb{R}^n$ d-dimensional manifold of the form $\mathcal{M} = F(B_{\ell_2^d})$,

- F is bi-Lipschitz: $\|F(x) - F(y)\|_2 \simeq \|x - y\|_2$, $\forall x, y$;
- $\|DF(x) - DF(y)\|_{2\to 2} \lesssim \|x - y\|_2$ for $x, y \in B_{\ell_2^d}$.

For $x \in \mathcal{M}$, E_x = tangent plane at x. $T\mathcal{M} = \cup_{x\in\mathcal{M}} E_x$ =tangent bundle.

$$\alpha = \max_{x\in\mathcal{M}} \max_{1\leq j\leq n} \|P_{E_x} e_j\|_2 = \text{largest incoherence}\,.$$

Suppose

$$m \gtrsim_* \varepsilon^{-2} d, \qquad s \gtrsim_* \varepsilon^{-2}(1 + (\alpha d)^2).$$

Then, with large probability, for all $x, y \in \mathcal{M}$

$$(1-\varepsilon)\rho_{\mathcal{M}}(x,y) \leq \rho_{\Phi(\mathcal{M})}(\Phi x, \Phi y) \leq (1+\varepsilon)\rho_{\mathcal{M}}(x,y).$$

If $\alpha \geq 1/\sqrt{d}$, then necessarily $s \geq d$ (existence of 'bad' manifold).

Manifolds

$\mathcal{M} \subset \mathbb{R}^n$ d-dimensional manifold of the form $\mathcal{M} = F(B_{\ell_2^d})$,

- F is bi-Lipschitz: $\|F(x) - F(y)\|_2 \simeq \|x - y\|_2, \ \forall x, y$;
- $\|DF(x) - DF(y)\|_{2\to 2} \lesssim \|x - y\|_2$ for $x, y \in B_{\ell_2^d}$.

For $x \in \mathcal{M}$, $E_x =$ tangent plane at x. $T\mathcal{M} = \cup_{x\in\mathcal{M}} E_x =$tangent bundle.

$$\alpha = \max_{x\in\mathcal{M}} \max_{1\leq j\leq n} \|P_{E_x} e_j\|_2 = \text{largest incoherence}\,.$$

Suppose

$$m \underset{*}{>} \varepsilon^{-2} d, \qquad s \underset{*}{>} \varepsilon^{-2}(1 + (\alpha d)^2).$$

Then, with large probability, for all $x, y \in \mathcal{M}$

$$(1 - \varepsilon)\rho_{\mathcal{M}}(x, y) \leq \rho_{\Phi(\mathcal{M})}(\Phi x, \Phi y) \leq (1 + \varepsilon)\rho_{\mathcal{M}}(x, y).$$

If $\alpha \geq 1/\sqrt{d}$, then necessarily $s \geq d$ (existence of 'bad' manifold).
Previous best $m \gtrsim \varepsilon^{-2} d$, $s = m$ (D. '14)

Proof overview of Main Theorem

Step 1: rewrite as 2nd order chaos process

Recall $\Phi = (\Phi_{ij})$, $\Phi_{ij} = \frac{1}{\sqrt{s}}\delta_{ij}\sigma_{ij}$.

Step 1: rewrite as 2nd order chaos process

Recall $\Phi = (\Phi_{ij})$, $\Phi_{ij} = \frac{1}{\sqrt{s}}\delta_{ij}\sigma_{ij}$.Want

$$\mathbb{E}_{\delta,\sigma} \sup_{x \in T} \left| \, \|\Phi x\|_2^2 - 1 \right| = \mathbb{E}_{\delta,\sigma} \sup_{x \in T} \left| \, \|\Phi x\|_2^2 - \mathbb{E}\|\Phi x\|_2^2 \right| < \varepsilon.$$

Step 1: rewrite as 2nd order chaos process

Recall $\Phi = (\Phi_{ij})$, $\Phi_{ij} = \frac{1}{\sqrt{s}}\delta_{ij}\sigma_{ij}$.Want

$$\mathbb{E}_{\delta,\sigma} \sup_{x\in T} \left| \, \|\Phi x\|_2^2 - 1 \right| = \mathbb{E}_{\delta,\sigma} \sup_{x\in T} \left| \, \|\Phi x\|_2^2 - \mathbb{E}\|\Phi x\|_2^2 \right| < \varepsilon.$$

For $x \in \mathbb{R}^n$ can write $\Phi x = A_{\delta,x}\sigma$, where

$$A_{\delta,x} := \frac{1}{\sqrt{s}} \begin{bmatrix} -x^{(\delta_{1,\cdot})}- & 0 & \cdots & 0 \\ 0 & -x^{(\delta_{2,\cdot})}- & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -x^{(\delta_{m,\cdot})}- \end{bmatrix}. \qquad (7)$$

for $x_j^{(\delta_{i,\cdot})} = \delta_{ij} x_j$.

Step 1: rewrite as 2nd order chaos process

Recall $\Phi = (\Phi_{ij})$, $\Phi_{ij} = \frac{1}{\sqrt{s}}\delta_{ij}\sigma_{ij}$.Want

$$\mathbb{E}_{\delta,\sigma} \sup_{x\in T} \big|\, \|\Phi x\|_2^2 - 1\big| = \mathbb{E}_{\delta,\sigma} \sup_{x\in T} \big|\, \|\Phi x\|_2^2 - \mathbb{E}\|\Phi x\|_2^2\big| < \varepsilon.$$

For $x \in \mathbb{R}^n$ can write $\Phi x = A_{\delta,x}\sigma$, where

$$A_{\delta,x} := \frac{1}{\sqrt{s}} \begin{bmatrix} -x^{(\delta_{1,\cdot})}- & 0 & \cdots & 0 \\ 0 & -x^{(\delta_{2,\cdot})}- & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -x^{(\delta_{m,\cdot})}- \end{bmatrix}. \qquad (7)$$

for $x_j^{(\delta_{i,\cdot})} = \delta_{ij}x_j$. Then

$$\sup_{x\in T} \big|\|\Phi x\|_2^2 - \mathbb{E}\|\Phi x\|_2^2\big| = \sup_{x\in T} \big|\|A_{\delta,x}\sigma\|_2^2 - \mathbb{E}\|A_{\delta,x}\sigma\|_2^2\big|.$$

Step 2: estimate chaos process

Theorem (Krahmer-Mendelson-Rauhut '14)

Let $\mathcal{A} \subset \mathbb{R}^{m \times n}$, $\sigma_1, \ldots, \sigma_n$ *independent random signs.*
$\|A\|$ =*operator norm,* $\|A\|_F$ =*Frobenius norm,*
$d_F(\mathcal{A}) = \sup_{A \in \mathcal{A}} \|A\|_F$.

Step 2: estimate chaos process

Theorem (Krahmer-Mendelson-Rauhut '14)

Let $\mathcal{A} \subset \mathbb{R}^{m\times n}$*,* $\sigma_1, \ldots, \sigma_n$ *independent random signs.*
$\|A\|$ *=operator norm,* $\|A\|_F$ *=Frobenius norm,*
$d_F(\mathcal{A}) = \sup_{A\in\mathcal{A}} \|A\|_F$*. Then,*

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

Step 2: estimate chaos process

Theorem (Krahmer-Mendelson-Rauhut '14)

Let $\mathcal{A} \subset \mathbb{R}^{m \times n}$, $\sigma_1, \ldots, \sigma_n$ independent random signs.
$\|A\|$ =operator norm, $\|A\|_F$ =Frobenius norm,
$d_F(\mathcal{A}) = \sup_{A \in \mathcal{A}} \|A\|_F$. Then,

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

- $\gamma_2(\mathcal{A}, \|\cdot\|)$=measure of metric complexity of $\mathcal{A}$.
- $\gamma_2(\mathcal{A}, \|\cdot\|) \lesssim I_2(\mathcal{A}, \|\cdot\|)$, where

$$I_2(\mathcal{A}, \|\cdot\|) = \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{A}, \|\cdot\|, t)}\, dt$$

Step 2: estimate chaos process

Theorem (Krahmer-Mendelson-Rauhut '14)

Let $\mathcal{A} \subset \mathbb{R}^{m\times n}$*,* $\sigma_1, \ldots, \sigma_n$ *independent random signs.*
$\|A\|$ *=operator norm,* $\|A\|_F$ *=Frobenius norm,*
$d_F(\mathcal{A}) = \sup_{A\in\mathcal{A}} \|A\|_F$*. Then,*

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

- $\gamma_2(\mathcal{A}, \|\cdot\|)$=measure of metric complexity of $\mathcal{A}$.
- $\gamma_2(\mathcal{A}, \|\cdot\|) \lesssim I_2(\mathcal{A}, \|\cdot\|)$, where

$$I_2(\mathcal{A}, \|\cdot\|) = \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{A}, \|\cdot\|, t)}\, dt$$

- $\mathcal{N}(\mathcal{A}, \|\cdot\|, t)$=minimal number of balls with radius t in $\|\cdot\|$ needed to cover $\mathcal{A}$.

Step 2: estimate chaos process

Theorem (Krahmer-Mendelson-Rauhut '14)

Let $\mathcal{A} \subset \mathbb{R}^{m\times n}$, $\sigma_1, \ldots, \sigma_n$ independent random signs.
$\|A\|$ =operator norm, $\|A\|_F$ =Frobenius norm,
$d_F(\mathcal{A}) = \sup_{A\in\mathcal{A}} \|A\|_F$. Then,

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

- $\gamma_2(\mathcal{A}, \|\cdot\|)$=measure of metric complexity of $\mathcal{A}$.
- $\gamma_2(\mathcal{A}, \|\cdot\|) \lesssim I_2(\mathcal{A}, \|\cdot\|)$, where

$$I_2(\mathcal{A}, \|\cdot\|) = \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{A}, \|\cdot\|, t)}\, dt$$

- $\mathcal{N}(\mathcal{A}, \|\cdot\|, t)$=minimal number of balls with radius t in $\|\cdot\|$ needed to cover $\mathcal{A}$.
- I_2 is called a (Dudley) entropy integral.

Step 2: estimate chaos process

$$\sup_{x\in T} |\|\Phi x\|_2^2 - 1| = \sup_{x\in T} |\|A_{\delta,x}\sigma\|_2^2 - \mathbb{E}\|A_{\delta,x}\sigma\|_2^2|.$$

Apply Theorem for $\mathcal{A} = \{A_{\delta,x} \ : \ x \in T\}$.

Step 2: estimate chaos process

$$\sup_{x\in T} |\|\Phi x\|_2^2 - 1| = \sup_{x\in T} |\|A_{\delta,x}\sigma\|_2^2 - \mathbb{E}\|A_{\delta,x}\sigma\|_2^2|.$$

Apply Theorem for $\mathcal{A} = \{A_{\delta,x} \ : \ x \in T\}$. Note:

$$\|A_{\delta,x} - A_{\delta,y}\| = \|x - y\|_\delta := \frac{1}{\sqrt{s}} \max_{1\le i\le m} \Big(\sum_{j=1}^{n} \delta_{ij}(x_j - y_j)^2\Big)^{1/2}$$

so

$$\mathcal{N}(\mathcal{A}, \|\cdot\|, t) = \mathcal{N}(T, \|\cdot\|_\delta, t).$$

Step 2: estimate chaos process

$$\sup_{x \in T} \left| \|\Phi x\|_2^2 - 1 \right| = \sup_{x \in T} \left| \|A_{\delta,x}\sigma\|_2^2 - \mathbb{E}\|A_{\delta,x}\sigma\|_2^2 \right|.$$

Apply Theorem for $\mathcal{A} = \{A_{\delta,x} \ : \ x \in T\}$. Note:

$$\|A_{\delta,x} - A_{\delta,y}\| = \|x - y\|_\delta := \frac{1}{\sqrt{s}} \max_{1 \le i \le m} \Big(\sum_{j=1}^{n} \delta_{ij}(x_j - y_j)^2 \Big)^{1/2}$$

so

$$\mathcal{N}(\mathcal{A}, \|\cdot\|, t) = \mathcal{N}(T, \|\cdot\|_\delta, t).$$

Also

$$\|A_{\delta,x}\|_F = \|x\|_2 \qquad \Rightarrow \qquad d_F(\mathcal{A}) = 1.$$

Therefore, Theorem implies that

$$\mathbb{E}_\sigma \sup_{x \in T} \Big| \|\Phi x\|_2^2 - 1 \Big| \lesssim I_2^2(T, \|\cdot\|_\delta) + I_2(T, \|\cdot\|_\delta).$$

Step 2: estimate chaos process

$$\sup_{x\in T} \left|\|\Phi x\|_2^2 - 1\right| = \sup_{x\in T} \left|\|A_{\delta,x}\sigma\|_2^2 - \mathbb{E}\|A_{\delta,x}\sigma\|_2^2\right|.$$

Apply Theorem for $\mathcal{A} = \{A_{\delta,x} \ : \ x \in T\}$. Note:

$$\|A_{\delta,x} - A_{\delta,y}\| = \|x-y\|_\delta := \frac{1}{\sqrt{s}} \max_{1\le i\le m} \Big(\sum_{j=1}^n \delta_{ij}(x_j - y_j)^2\Big)^{1/2}$$

so

$$\mathcal{N}(\mathcal{A}, \|\cdot\|, t) = \mathcal{N}(T, \|\cdot\|_\delta, t).$$

Also

$$\|A_{\delta,x}\|_F = \|x\|_2 \qquad \Rightarrow \qquad d_F(\mathcal{A}) = 1.$$

Therefore, Theorem implies that

$$\mathbb{E}_\sigma \sup_{x\in T} \Big|\|\Phi x\|_2^2 - 1\Big| \lesssim I_2^2(T, \|\cdot\|_\delta) + I_2(T, \|\cdot\|_\delta).$$

$$\mathbb{E}_\delta\mathbb{E}_\sigma \sup_{x\in T} \Big|\|\Phi x\|_2^2 - 1\Big| \lesssim \mathbb{E}_\delta I_2^2(T, \|\cdot\|_\delta) + \mathbb{E}_\delta I_2(T, \|\cdot\|_\delta).$$

Covering number estimates, subspace case

Step 3: estimate covering numbers, subspace case

Finally, need to bound:

$$\mathbb{E}_\delta I_2^2(T, \|\cdot\|_\delta) = \mathbb{E}_\delta\Big(\int_0^\infty \sqrt{\mathcal{N}(T, \|\cdot\|_\delta, t)}\, dt\Big)^2.$$

Step 3: estimate covering numbers, subspace case

Finally, need to bound:

$$\mathbb{E}_\delta I_2^2(T, \|\cdot\|_\delta) = \mathbb{E}_\delta\Big(\int_0^\infty \sqrt{\mathcal{N}(T, \|\cdot\|_\delta, t)}\, dt\Big)^2.$$

If $E \subset \mathbb{R}^n$ is a d-dimensional *subspace*, $T = B_E$, then up to $\log d$

$$I_2(B_E, \|\cdot\|_\delta) \ll \sup_{t>0} t\sqrt{\log \mathcal{N}(B_E, \|\cdot\|_\delta, t)}.$$

Step 3: estimate covering numbers, subspace case

Finally, need to bound:

$$\mathbb{E}_\delta I_2^2(T, \|\cdot\|_\delta) = \mathbb{E}_\delta\Big(\int_0^\infty \sqrt{\mathcal{N}(T, \|\cdot\|_\delta, t)}\, dt\Big)^2.$$

If $E \subset \mathbb{R}^n$ is a d-dimensional *subspace*, $T = B_E$, then up to $\log d$

$$I_2(B_E, \|\cdot\|_\delta) \ll \sup_{t>0} t\sqrt{\log \mathcal{N}(B_E, \|\cdot\|_\delta, t)}.$$

Let $U \in \mathbb{R}^{n\times d}$ have columns forming an orthonormal basis for E. Dual Sudakov inequality (Pajor-Tomczak-Jaegermann '86, Bourgain-Lindenstrauss-Milman '89) states

$$\sup_{t>0} t\sqrt{\log \mathcal{N}(B_E, \|\cdot\|_\delta, t)} \lesssim \mathbb{E}_g \|Ug\|_\delta$$

for g standard Gaussian.

Step 3: estimate covering numbers, subspace case

Finally, need to bound:

$$\mathbb{E}_\delta I_2^2(T, \|\cdot\|_\delta) = \mathbb{E}_\delta\Big(\int_0^\infty \sqrt{\mathcal{N}(T, \|\cdot\|_\delta, t)}\,dt\Big)^2.$$

If $E \subset \mathbb{R}^n$ is a d-dimensional *subspace*, $T = B_E$, then up to $\log d$

$$I_2(B_E, \|\cdot\|_\delta) \ll \sup_{t>0} t\sqrt{\log \mathcal{N}(B_E, \|\cdot\|_\delta, t)}.$$

Let $U \in \mathbb{R}^{n\times d}$ have columns forming an orthonormal basis for E. Dual Sudakov inequality (Pajor-Tomczak-Jaegermann '86, Bourgain-Lindenstrauss-Milman '89) states

$$\sup_{t>0} t\sqrt{\log \mathcal{N}(B_E, \|\cdot\|_\delta, t)} \lesssim \mathbb{E}_g\|Ug\|_\delta$$

for g standard Gaussian. Estimate $\mathbb{E}_\delta(\mathbb{E}_g\|Ug\|_\delta)^2$ with Gaussian concentration for Lipschitz functions + non-commutative Khintchine (Lust-Piquard, Pisier '91) to get

$$\mathbb{E}_\delta I_2^2(B_E, \|\cdot\|_\delta) \lesssim \frac{(d + \log m)\log^2 m}{m} + \frac{\mu(E)^2 log^3 m}{s}.$$

Covering number estimates, general case

Step 3: estimate covering numbers, general case

For general T cannot use dual Sudakov. Reduce (in many steps) to subspace case!

Step 3: estimate covering numbers, general case

For general T cannot use dual Sudakov. Reduce (in many steps) to subspace case!

- Use *entropy duality* (Bourgain et al. '89) to estimate $\log \mathcal{N}(T, \|\cdot\|_\delta, t)$ in terms of

Step 3: estimate covering numbers, general case

For general T cannot use dual Sudakov. Reduce (in many steps) to subspace case!

- Use *entropy duality* (Bourgain et al. '89) to estimate $\log \mathcal{N}(T, \|\cdot\|_\delta, t)$ in terms of

$$\log \mathcal{N}(\mathrm{conv}(\cup_{i=1}^m B_{J_i}), |||\cdot|||_T, \sqrt{s}t), \tag{8}$$

where B_{J_i}=unit ball in $\mathrm{span}\{e_j : \delta_{ij} = 1\}$, $|||x|||_T = \sup_{y\in T} |\langle x, y\rangle|$.

Step 3: estimate covering numbers, general case

For general T cannot use dual Sudakov. Reduce (in many steps) to subspace case!

- Use *entropy duality* (Bourgain et al. '89) to estimate $\log \mathcal{N}(T, \|\cdot\|_\delta, t)$ in terms of
$$\log \mathcal{N}(\mathrm{conv}(\cup_{i=1}^m B_{J_i}), |||\cdot|||_T, \sqrt{s}t), \tag{8}$$
where B_{J_i}=unit ball in $\mathrm{span}\{e_j : \delta_{ij} = 1\}$, $|||x|||_T = \sup_{y \in T} |\langle x, y\rangle|$.
- Use *Maurey's lemma* to get rid of convex hull: estimate (8) in terms of

Step 3: estimate covering numbers, general case

For general T cannot use dual Sudakov. Reduce (in many steps) to subspace case!

- Use *entropy duality* (Bourgain et al. '89) to estimate $\log \mathcal{N}(T, \|\cdot\|_\delta, t)$ in terms of

$$\log \mathcal{N}(\mathrm{conv}(\cup_{i=1}^m B_{J_i}), |||\cdot|||_T, \sqrt{s}t), \tag{8}$$

  where B_{J_i}=unit ball in $\mathrm{span}\{e_j : \delta_{ij} = 1\}$, $|||x|||_T = \sup_{y \in T} |\langle x, y\rangle|$.
- Use *Maurey's lemma* to get rid of convex hull: estimate (8) in terms of

$$\max_{k \lesssim 1/t^2} \max_{A \subset [m] \,:\, |A|=k} \log \mathcal{N}\Big(\frac{1}{k}\sum_{i \in A} B_{J_i}, |||\cdot|||_T, t\Big). \tag{9}$$

Step 3: estimate covering numbers, general case

For general T cannot use dual Sudakov. Reduce (in many steps) to subspace case!

- Use *entropy duality* (Bourgain et al. '89) to estimate $\log \mathcal{N}(T, \|\cdot\|_\delta, t)$ in terms of

$$\log \mathcal{N}(\mathrm{conv}(\cup_{i=1}^m B_{J_i}), \|\!|\cdot|\!\|_T, \sqrt{s}t), \tag{8}$$

  where B_{J_i}=unit ball in span$\{e_j : \delta_{ij} = 1\}$,
  $\|\!|x|\!\|_T = \sup_{y \in T} |\langle x, y\rangle|$.
- Use *Maurey's lemma* to get rid of convex hull: estimate (8) in terms of

$$\max_{k \lesssim 1/t^2} \max_{A \subset [m]\,:\,|A|=k} \log \mathcal{N}(\frac{1}{k}\sum_{i\in A} B_{J_i}, \|\!|\cdot|\!\|_T, t). \tag{9}$$

- For $A \subset [m]$ set $U_{\alpha,A} = \{j \in [n] : \sum_{i\in A}\delta_{ij} \simeq 2^\alpha\}$. Estimate (9) by

$$\max_{k \lesssim 1/t^2} \max_{A \subset [m]\,:\,|A|=k} \sum_\alpha \log \mathcal{N}\Big(B_{U_\alpha}, \|\!|\cdot|\!\|_T, \sqrt{\frac{k}{2^\alpha}}\frac{t}{\log m}\Big).$$