

Chaining methods in random dimensionality reduction

Sjoerd Dirksen (RWTH Aachen)

based on joint work with

Jean Bourgain (IAS Princeton)

Jelani Nelson (Harvard)

Workshop

'Chaining Methods and their Applications to Computer Science'

Harvard, June 22, 2016

Euclidean dimensionality reduction

Dimensionality reduction

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

Benefits:

- ▶ Smaller storage consumption;
- ▶ Speedup during data analysis;

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 $\Phi \in \mathbb{R}^{m \times n}$

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

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

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

Example applications: app. nearest neighbours, k -means clustering, constrained least squares, manifold learning.

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 an $m \times n$ standard Gaussian matrix,

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

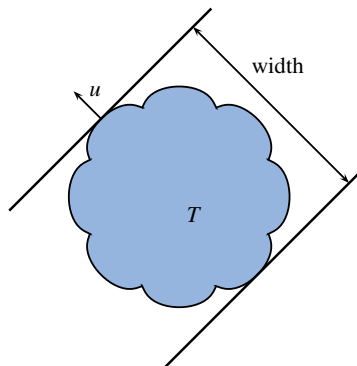
then $\frac{1}{\sqrt{m}}\Phi$ is an ε -isometry on T with probability $\geq 1 - \eta$.

- ▶ This is an 'instance-optimal' version of J-L lemma: $g(T)^2 \lesssim \log |T|$, but can be much smaller.
- ▶ 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, D. '15, Bednorz '14+, Mendelson '15+, Liaw-Mehrabian-Plan-Vershynin '16+.

All proofs use chaining based on Bernstein's inequality.

Geometric interpretation

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



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

A proof sketch

Theorem (Schechtman '06, Liaw et al. '16+)

If Φ is an isotropic, 1-subgaussian $m \times n$ matrix, then the process $Z_x = \frac{1}{\sqrt{m}} \|\Phi x\|_2 - \|x\|_2$ has subgaussian increments, i.e., for all $x, y \in \mathbb{R}^n$,

$$\mathbb{P}\left(|Z_x - Z_y| \geq \frac{u}{\sqrt{m}} \|x - y\|_2\right) \leq Ce^{-cu^2} \quad u \geq 0.$$

Theorem (Talagrand '01, D. '15)

$(X_t)_{t \in T}$ a stochastic process, where (T, d) is a metric space,

$$\mathbb{P}(|X_t - X_s| \geq ud(t, s)) \leq 2e^{-u^2} \quad \text{for all } s, t \in T, u \geq 0.$$

Let $p \geq 1$ and $t_0 \in T$. Then

$$\left(\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}|^p\right)^{1/p} \lesssim \gamma_2(T, d) + \text{Diam}_d(T) \sqrt{p}.$$

Tail bound via standard chaining

For $p \geq 1$ and $t_0 \in T$

$$\left(\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}|^p \right)^{1/p} \lesssim \gamma_2(T, d) + \text{Diam}_d(T) \sqrt{p}.$$

By Markov:

$$\mathbb{P} \left(\sup_{t \in T} (X_t - X_{t_0}) \gtrsim \gamma_2(T, d) + u \text{Diam}_d(T) \right) \leq e^{-u^2/2} \quad (u \geq 1).$$

To compare, consider $X_t = \langle t, g \rangle$, $t \in T$, $0 \in T$. Subgaussian with respect to $d(s, t) = \|s - t\|_2$. Know:

$$\mathbb{E} \sup_{t \in T} X_t \simeq \gamma_2(T, \|\cdot\|_2)$$

and by Gaussian concentration of Lipschitz functions,

$$\mathbb{P} \left(\left| \sup_{t \in T} X_t - \mathbb{E} \sup_{t \in T} X_t \right| \geq u \text{Diam}_{\ell_2}(T) \right) \leq 2e^{-u^2/2}.$$

L^p -bound via standard chaining

Idea of proof: $(T_n)_{n \geq 0}$ admissible, $\ell = \log_2(p)$. Split:

$$\begin{aligned} & (\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}|^p)^{1/p} \\ & \leq (\mathbb{E} \sup_{t \in T} |X_t - X_{\pi_\ell t}|^p)^{1/p} + (\mathbb{E} \sup_{t \in T} |X_{\pi_\ell t} - X_{t_0}|^p)^{1/p}. \end{aligned}$$

Since $|T_\ell| \leq 2^{2^\ell}$,

$$\begin{aligned} (\mathbb{E} \sup_{t \in T} |X_{\pi_\ell t} - X_{t_0}|^p)^{1/p} & \leq \left(\sum_{t \in T_\ell} \mathbb{E} |X_t - X_{t_0}|^p \right)^{1/p} \\ & \leq |T_\ell|^{1/p} \sup_{t \in T_\ell} (\mathbb{E} |X_t - X_{t_0}|^p)^{1/p} \\ & \leq 2\sqrt{p} \sup_{t \in T_\ell} d(t, t_0). \end{aligned}$$

Can show using generic chaining:

$$\left(\mathbb{E} \sup_{t \in T} |X_t - X_{\pi_\ell t}|^p \right)^{1/p} \lesssim \gamma_2(T, d).$$

Finalizing the proof sketch

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

Theorem

Let $T \subset S^{n-1}$. If Φ is an $m \times n$ standard Gaussian or random sign matrix and

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

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

Proof.

Combine the two theorems and use Markov: w.p. $\geq 1 - e^{-u^2}$

$$\begin{aligned} & \sup_{x \in T} \left| \frac{1}{\sqrt{m}} \|\Phi x\|_2 - \|x\|_2 \right| \\ &= \sup_{x \in T} |Z_x - Z_0| \\ &\lesssim \frac{\gamma_2(T \cup \{0\}, \|\cdot\|_2)}{\sqrt{m}} + \frac{u}{\sqrt{m}} \text{Diam}_{\ell_2}(T \cup \{0\}) \\ &\lesssim (g(T) + u)/\sqrt{m} \end{aligned}$$

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

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

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

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

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

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).

Fast and sparse Johnson-Lindenstrauss transforms

Fast Johnson-Lindenstrauss transforms (FJLT)

Ailon-Chazelle '06, Ailon-Liberty '09, '13, Vybíral '11,
Krahmer-Ward '11.

- ▶ F = discrete Fourier transform.
- ▶ $\sigma_1, \dots, \sigma_n$ independent random signs, $D_\sigma = \text{diag}((\sigma_i)_{i=1}^n)$.
- ▶ $\theta_1, \dots, \theta_n$ independent random selectors, $\mathbb{P}(\theta_i = 1) = m/n$.
- ▶ FJLT:

$$\Psi = \sqrt{\frac{n}{m}} D_\theta F D_\sigma.$$

Note: $\mathbb{E}\|\Psi x\|_2^2 = \|x\|_2^2$. Ψx can be computed in time $O(n \log n)$.

Oymak-Recht-Soltanolkotabi '15 proved version of Gordon's theorem for FJLT → *Mahdi's talk tomorrow*.

Sparse Johnson-Lindenstrauss Transform (SJLT)

Achlioptas '03, 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 = Φ ,

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

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

Sparse analog of Gordon's theorem

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 : $m \gtrsim \varepsilon^{-2} d \text{ polylog}(d)$,
 $s \gtrsim \varepsilon^{-1} \text{ polylog}(d)$. (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

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

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

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

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

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)

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

$$s >_* \frac{1}{\varepsilon^2}, \quad \kappa(T) \leq_* \varepsilon.$$

- ▶ 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:

- ▶ Simpler proof for subspace case. For general T : reduce (in many steps) to subspace case.

Proof ingredients:

- ▶ generic chaining for supremum of 2nd order chaos process (Krahmer-Mendelson-Rauhut '14)
- ▶ Tools to control covering numbers: Maurey's lemma, dual Sudakov inequality, entropy duality (Bourgain et al., '89).

Applications

Sparse subspace embedding

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

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

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

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

it suffices if

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

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

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

x_{opt} a minimizer.

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

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

$$\begin{aligned} m &> \varepsilon^{-2} k \sigma_{\min,k}^{-2} \max_j \|A_j\|_2^2, \\ s &> \varepsilon^{-2} k \sigma_{\min,k}^{-2} \max_{i,j} |A_{ij}|^2, \end{aligned} \tag{6}$$

then with high probability

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

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 $E \subset \mathbb{R}^n$.

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

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

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

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

if (incoherent case)

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

or (coherent case)

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

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

Manifolds

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

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

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

$$T\mathcal{M} = \cup_{x \in \mathcal{M}} E_x = \text{tangent bundle}.$$

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

Suppose

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

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).

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

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

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

$$A_{\delta,x} := \frac{1}{\sqrt{s}} \begin{bmatrix} x \circ \delta_1 & 0 & \cdots & 0 \\ 0 & x \circ \delta_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & x \circ \delta_m \end{bmatrix}. \quad (7)$$

for $(x \circ \delta_i)_j = \delta_{ij}x_j$. Then

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

Step 2: estimate chaos process

Theorem (Krahmer-Mendelson-Rauhut '14, D. '15)

Let $\mathcal{A} \subset \mathbb{R}^{m \times n}$, $\sigma_1, \dots, \sigma_n$ independent random signs.

$\|A\|$ = operator norm, $\|A\|_F$ = Frobenius norm,

$d_F(\mathcal{A}) = \sup_{A \in \mathcal{A}} \|A\|_F$. Then,

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

For any $u \geq 1$,

$$\mathbb{P}_\sigma \left(\sup_{A \in \mathcal{A}} \left| \|A\sigma\|_2^2 - \mathbb{E}_\sigma \|A\sigma\|_2^2 \right| \gtrsim \gamma_2^2(\mathcal{A}, \|\cdot\|) + d_F(\mathcal{A})\gamma_2(\mathcal{A}, \|\cdot\|) + \sqrt{u}d_4^2(\mathcal{A}) + ud_{\ell_2 \rightarrow \ell_2}^2(\mathcal{A}) \right) \leq e^{-u}$$

where

$$d_4^2(\mathcal{A}) = \sup_{A \in \mathcal{A}} \|A^*A\|_F, \quad d_{\ell_2 \rightarrow \ell_2}^2(\mathcal{A}) = \sup_{A \in \mathcal{A}} \|A^*A\|.$$

Step 2: estimate chaos process

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

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

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

so

$$\gamma_2(\mathcal{A}, \|\cdot\|) = \gamma_2(T, \|\cdot\|_\delta).$$

Also

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

Therefore, Theorem implies that

$$\mathbb{E}_\sigma \sup_{x \in T} \left| \|\Phi x\|_2^2 - 1 \right| \lesssim \gamma_2^2(T, \|\cdot\|_\delta) + \gamma_2(T, \|\cdot\|_\delta).$$

$$\mathbb{E}_\delta \mathbb{E}_\sigma \sup_{x \in T} \left| \|\Phi x\|_2^2 - 1 \right| \lesssim \mathbb{E}_\delta \gamma_2^2(T, \|\cdot\|_\delta) + \mathbb{E}_\delta \gamma_2(T, \|\cdot\|_\delta).$$

Proof idea 1: decoupling

Theorem (Krahmer-Mendelson-Rauhut '14)

Let $\mathcal{A} \subset \mathbb{R}^{m \times n}$, $\sigma_1, \dots, \sigma_n$ independent random signs.

$\|A\|$ = operator norm, $\|A\|_F$ = Frobenius norm,

$d_F(\mathcal{A}) = \sup_{A \in \mathcal{A}} \|A\|_F$. Then,

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

Decouple: for σ' independent copy of σ

$$\begin{aligned} \mathbb{E}_\sigma \sup_{A \in \mathcal{A}} \left| \|A\sigma\|_2^2 - \mathbb{E} \|A\sigma\|_2^2 \right| &= \mathbb{E}_\sigma \sup_{A \in \mathcal{A}} \left| \sum_{i \neq j} \sigma_i \sigma_j (A^* A)_{ij} \right| \\ &\simeq \mathbb{E}_{\sigma, \sigma'} \sup_{A \in \mathcal{A}} \left| \sum_{i, j} \sigma_i \sigma'_j (A^* A)_{ij} \right| \\ &= \mathbb{E}_{\sigma, \sigma'} \sup_{A \in \mathcal{A}} |\sigma^* A^* A \sigma'| \end{aligned}$$

Proof idea 2: bilinear chaining

Bilinear chaining:

$$\begin{aligned} & \sigma^*(A^*A - \pi_0(A)^*\pi_0(A))\sigma' \\ &= \sum_{n \geq 1} \sigma^* \pi_n(A)^* \pi_n(A) \sigma' - \sigma^* \pi_{n-1}(A)^* \pi_{n-1}(A) \sigma' \\ &= \sum_{n \geq 1} \sigma^* (\pi_n(A) - \pi_{n-1}(A))^* \pi_n(A) \sigma' \\ &+ \sum_{n \geq 1} \sigma^* \pi_{n-1}(A)^* (\pi_n(A) - \pi_{n-1}(A)) \sigma' =: S_1(A) + S_2(A). \end{aligned}$$

By Hoeffding's inequality, w.p. $\geq 1 - 2 \exp(-u^2)$,

$$\begin{aligned} & |\sigma^* (\pi_n(A) - \pi_{n-1}(A))^* \pi_n(A) \sigma'| \\ & \leq u \|(\pi_n(A) - \pi_{n-1}(A))^* \pi_n(A) \sigma'\|_2 \\ & \leq u \|\pi_n(A) - \pi_{n-1}(A)\| \|\pi_n(A) \sigma'\|_2 \\ & \leq u \|\pi_n(A) - \pi_{n-1}(A)\| \sup_{A \in \mathcal{A}} \|A \sigma'\|_2. \end{aligned}$$

Completing proof sketch

The generic chaining argument yields:

$$\mathbb{E}_\sigma \sup_{A \in \mathcal{A}} \sum_{n \geq 1} |\sigma^*(\pi_n(A) - \pi_{n-1}(A))^* \pi_n(A) \sigma'| \lesssim \gamma_2(\mathcal{A}, \|\cdot\|) \sup_{A \in \mathcal{A}} \|A\sigma'\|_2.$$

Hence

$$\begin{aligned} \mathbb{E}_{\sigma, \sigma'} \sup_{A \in \mathcal{A}} S_1(A) &\lesssim \gamma_2(\mathcal{A}, \|\cdot\|) \mathbb{E}_{\sigma'} \sup_{A \in \mathcal{A}} \|A\sigma'\|_2 \\ &\leq \gamma_2(\mathcal{A}, \|\cdot\|) (\mathbb{E}_\sigma \sup_{A \in \mathcal{A}} \|A\sigma\|_2^2)^{1/2}. \end{aligned}$$

Same inequality for S_2 . This leads to a quadratic inequality for

$$\left(\mathbb{E}_\sigma \sup_{A \in \mathcal{A}} \left| \|A\sigma\|_2^2 - \mathbb{E} \|A\sigma\|_2^2 \right| \right)^{1/2}.$$

Solving it yields the result.

Step 2: estimate chaos process

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

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

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

so

$$\gamma_2(\mathcal{A}, \|\cdot\|) = \gamma_2(T, \|\cdot\|_\delta).$$

Also

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

Therefore, Theorem implies that

$$\mathbb{E}_\sigma \sup_{x \in T} \left| \|\Phi x\|_2^2 - 1 \right| \lesssim \gamma_2^2(T, \|\cdot\|_\delta) + \gamma_2(T, \|\cdot\|_\delta).$$

$$\mathbb{E}_\delta \mathbb{E}_\sigma \sup_{x \in T} \left| \|\Phi x\|_2^2 - 1 \right| \lesssim \mathbb{E}_\delta \gamma_2^2(T, \|\cdot\|_\delta) + \mathbb{E}_\delta \gamma_2(T, \|\cdot\|_\delta).$$

Covering number estimates, subspace case

Step 3: estimate covering numbers, subspace case

Finally, need to bound:

$$\mathbb{E}_\delta \gamma_2^2(T, \|\cdot\|_\delta) \lesssim \mathbb{E}_\delta \left(\int_0^\infty \sqrt{\log \mathcal{N}(T, \|\cdot\|_\delta, t)} dt \right)^2.$$

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!

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

$$\log \mathcal{N}(\text{conv}(\cup_{i=1}^m B_{J_i}), \|\cdot\|_T, \sqrt{st}), \quad (8)$$

where B_{J_i} = unit ball in $\text{span}\{e_j : \delta_{ij} = 1\}$,

$$\|\|x\|\|_T = \sup_{y \in T} |\langle x, y \rangle|.$$

- ▶ Use *Maurey's lemma* to get rid of convex hull: estimate (8) in terms of

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

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

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