An introduction to chaining, and applications to sublinear algorithms

Jelani Nelson
Harvard

August 28, 2015

What's this talk about?

What's this talk about?

Given a collection of random variables $X_1, X_2, \ldots,$, we would like to say that $\max_i X_i$ is small with high probability. (Happens all over computer science, e.g. "Chernion" (Chernoff+Union) bound)

What's this talk about?

Given a collection of random variables $X_1, X_2, \ldots,$, we would like to say that $\max_i X_i$ is small with high probability. (Happens all over computer science, e.g. "Chernion" (Chernoff+Union) bound)

Today's topic: Beating the Union Bound

What's this talk about?

Given a collection of random variables $X_1, X_2, \ldots,$, we would like to say that $\max_i X_i$ is small with high probability. (Happens all over computer science, e.g. "Chernion" (Chernoff+Union) bound)

Today's topic: Beating the Union Bound

Disclaimer: This is an educational talk, about ideas which aren't mine.

A first example

- $T \subset B_{\ell_2^n}$

A first example

- $T \subset B_{\ell_2^n}$
- Random variables $(Z_x)_{x \in T}$

  $Z_x = \langle g, x \rangle$ for a vector g with i.i.d. $\mathcal{N}(0, 1)$ entries

A first example

- $T \subset B_{\ell_2^n}$
- Random variables $(Z_x)_{x\in T}$

  $Z_x = \langle g, x\rangle$ for a vector g with i.i.d. $\mathcal{N}(0,1)$ entries
- Define *gaussian mean width* $\mathfrak{g}(T) = \mathbb{E}_g \sup_{x\in T} Z_x$

A first example

- $T \subset B_{\ell_2^n}$
- Random variables $(Z_x)_{x \in T}$

  $Z_x = \langle g, x \rangle$ for a vector g with i.i.d. $\mathcal{N}(0,1)$ entries
- Define *gaussian mean width* $\mathfrak{g}(T) = \mathbb{E}_g \sup_{x \in T} Z_x$
- How can we bound $\mathfrak{g}(T)$?

A first example

- $T \subset B_{\ell_2^n}$
- Random variables $(Z_x)_{x \in T}$

  $Z_x = \langle g, x \rangle$ for a vector g with i.i.d. $\mathcal{N}(0, 1)$ entries
- Define *gaussian mean width* $\mathfrak{g}(T) = \mathbb{E}_g \sup_{x \in T} Z_x$
- How can we bound $\mathfrak{g}(T)$?
- **This talk:** four progressively tighter ways to bound $\mathfrak{g}(T)$, then applications of techniques to some TCS problems

Gaussian mean width bound 1: union bound

- $\mathfrak{g}(T) = \mathbb{E} \sup_{x \in T} Z_x = \mathbb{E} \sup_{x \in T} \langle g, x \rangle$

Gaussian mean width bound 1: union bound

- $\mathfrak{g}(T) = \mathbb{E} \sup_{x \in T} Z_x = \mathbb{E} \sup_{x \in T} \langle g, x \rangle$
- Z_x is a gaussian with variance one

Gaussian mean width bound 1: union bound

- $\mathfrak{g}(T) = \mathbb{E}\sup_{x\in T} Z_x = \mathbb{E}\sup_{x\in T}\langle g, x\rangle$
- Z_x is a gaussian with variance one

$$\mathbb{E}\sup_{x\in T} Z_x = \int_0^\infty \mathbb{P}(\sup_{x\in T} Z_x > u)du$$

Gaussian mean width bound 1: union bound

- $\mathfrak{g}(T) = \mathbb{E}\sup_{x\in T} Z_x = \mathbb{E}\sup_{x\in T}\langle g, x\rangle$
- Z_x is a gaussian with variance one

$$\begin{aligned}\mathbb{E}\sup_{x\in T} Z_x &= \int_0^\infty \mathbb{P}(\sup_{x\in T} Z_x > u)du \\ &= \int_0^{u_*} \underbrace{\mathbb{P}(\sup_{x\in T} Z_x > u)}_{\leq 1} du + \int_{u_*}^\infty \underbrace{\mathbb{P}(\sup_{x\in T} Z_x > u)}_{\leq |T|\cdot e^{-u^2/2}\text{ (union bound)}} du\end{aligned}$$

Gaussian mean width bound 1: union bound

- $\mathfrak{g}(T) = \mathbb{E} \sup_{x \in T} Z_x = \mathbb{E} \sup_{x \in T} \langle g, x \rangle$
- Z_x is a gaussian with variance one

$$
\begin{aligned}
\mathbb{E} \sup_{x \in T} Z_x &= \int_0^\infty \mathbb{P}(\sup_{x \in T} Z_x > u) du \\
&= \int_0^{u_*} \underbrace{\mathbb{P}(\sup_{x \in T} Z_x > u)}_{\leq 1} du + \int_{u_*}^\infty \underbrace{\mathbb{P}(\sup_{x \in T} Z_x > u)}_{\leq |T| \cdot e^{-u^2/2} \text{ (union bound)}} du \\
&\leq u_* + |T| \cdot e^{-u_*^2/2} \\
&\lesssim \sqrt{\log |T|} \text{ (set } u_* = \sqrt{2 \log |T|})
\end{aligned}
$$

Gaussian mean width bound 2: ε-net

- $\mathfrak{g}(T) = \mathbb{E}\sup_{x\in T}\langle g, x\rangle$
- Let S_ε be ε-net of (T, ℓ_2)

Gaussian mean width bound 2: ε-net

- $\mathfrak{g}(T) = \mathbb{E} \sup_{x \in T} \langle g, x \rangle$
- Let S_ε be ε-net of (T, ℓ_2)
- $\langle g, x \rangle = \langle g, x' \rangle + \langle g, x - x' \rangle$ $(x' = \mathrm{argmin}_{y \in T} \|x - y\|_2)$
  $\mathfrak{g}(T) \le \mathfrak{g}(S_\varepsilon) + \mathbb{E}_g \sup_{x \in T} \underbrace{\langle g, x - x' \rangle}_{\le \varepsilon \cdot \|g\|_2}$

Gaussian mean width bound 2: ε-net

- $\mathfrak{g}(T) = \mathbb{E} \sup_{x \in T} \langle g, x \rangle$
- Let S_ε be ε-net of (T, ℓ_2)
- $\langle g, x \rangle = \langle g, x' \rangle + \langle g, x - x' \rangle$ $(x' = \mathrm{argmin}_{y \in T} \|x - y\|_2)$
  $$\mathfrak{g}(T) \le \mathfrak{g}(S_\varepsilon) + \mathbb{E}_g \sup_{x \in T} \underbrace{\langle g, x - x' \rangle}_{\le \varepsilon \cdot \|g\|_2}$$
- $\lesssim \sqrt{\log |S_\varepsilon|} + \varepsilon (\mathbb{E}_g \|g\|_2^2)^{1/2}$
- $\lesssim \log^{1/2} \underbrace{\mathcal{N}(T, \ell_2, \varepsilon)}_{\text{smallest } \varepsilon-\text{net size}} + \varepsilon \sqrt{n}$

Gaussian mean width bound 2: ε-net

- $\mathfrak{g}(T) = \mathbb{E} \sup_{x \in T} \langle g, x \rangle$
- Let S_ε be ε-net of (T, ℓ_2)
- $\langle g, x \rangle = \langle g, x' \rangle + \langle g, x - x' \rangle$ $(x' = \mathrm{argmin}_{y \in T} \|x - y\|_2)$
  $\mathfrak{g}(T) \le \mathfrak{g}(S_\varepsilon) + \mathbb{E}_g \sup_{x \in T} \underbrace{\langle g, x - x' \rangle}_{\le \varepsilon \cdot \|g\|_2}$
- $\lesssim \sqrt{\log |S_\varepsilon|} + \varepsilon (\mathbb{E}_g \|g\|_2^2)^{1/2}$
- $\lesssim \log^{1/2} \underbrace{\mathcal{N}(T, \ell_2, \varepsilon)}_{\text{smallest } \varepsilon-\text{net size}} + \varepsilon \sqrt{n}$
- Choose ε to optimize bound; can never be worse than last slide (which amounts to choosing $\varepsilon = 0$)

Gaussian mean width bound 3: ε-net sequence

- S_k is a $(1/2^k)$-net of T, $k \geq 0$

  $\pi_k x$ is closest point in S_k to $x \in T$, $\Delta_k x = \pi_k x - \pi_{k-1} x$

Gaussian mean width bound 3: ε-net sequence

- S_k is a $(1/2^k)$-net of T, $k \geq 0$
  $\pi_k x$ is closest point in S_k to $x \in T$, $\Delta_k x = \pi_k x - \pi_{k-1} x$
- wlog $|T| < \infty$ (else apply this slide to ε-net of T for ε small)
- $\langle g, x \rangle = \langle g, \pi_0 x \rangle + \sum_{k=1}^{\infty} \langle g, \Delta_k x \rangle$

Gaussian mean width bound 3: ε-net sequence

- S_k is a $(1/2^k)$-net of T, $k \geq 0$

  $\pi_k x$ is closest point in S_k to $x \in T$, $\Delta_k x = \pi_k x - \pi_{k-1} x$
- wlog $|T| < \infty$ (else apply this slide to ε-net of T for ε small)
- $\langle g, x \rangle = \langle g, \pi_0 x \rangle + \sum_{k=1}^{\infty} \langle g, \Delta_k x \rangle$
- $\mathfrak{g}(T) \leq \underbrace{\mathbb{E}_g \sup_{x \in T} \langle g, \pi_0 x \rangle}_{0} + \sum_{k=1}^{\infty} \mathbb{E}_g \sup_{x \in T} \langle g, \Delta_k x \rangle$

Gaussian mean width bound 3: ε-net sequence

- S_k is a $(1/2^k)$-net of T, $k \geq 0$
  $\pi_k x$ is closest point in S_k to $x \in T$, $\Delta_k x = \pi_k x - \pi_{k-1} x$
- wlog $|T| < \infty$ (else apply this slide to ε-net of T for ε small)
- $\langle g, x \rangle = \langle g, \pi_0 x \rangle + \sum_{k=1}^{\infty} \langle g, \Delta_k x \rangle$
- $\mathfrak{g}(T) \leq \underbrace{\mathop{\mathbb{E}}_{g} \sup_{x \in T} \langle g, \pi_0 x \rangle}_{0} + \sum_{k=1}^{\infty} \mathbb{E}_g \sup_{x \in T} \langle g, \Delta_k x \rangle$
- $$\begin{aligned} |\{\Delta_k x : x \in T\}| &\leq \mathcal{N}(T, \ell_2, 1/2^k) \cdot \mathcal{N}(T, \ell_2, 1/2^{k-1}) \\ &\leq (\mathcal{N}(T, \ell_2, 1/2^k))^2 \end{aligned}$$

Gaussian mean width bound 3: ε-net sequence

- S_k is a $(1/2^k)$-net of T, $k \geq 0$
  $\pi_k x$ is closest point in S_k to $x \in T$, $\Delta_k x = \pi_k x - \pi_{k-1} x$
- wlog $|T| < \infty$ (else apply this slide to ε-net of T for ε small)
- $\langle g, x \rangle = \langle g, \pi_0 x \rangle + \sum_{k=1}^{\infty} \langle g, \Delta_k x \rangle$
- $\mathfrak{g}(T) \leq \underbrace{\mathop{\mathbb{E}}_{g} \sup_{x \in T} \langle g, \pi_0 x \rangle}_{0} + \sum_{k=1}^{\infty} \mathbb{E}_g \sup_{x \in T} \langle g, \Delta_k x \rangle$
- $|\{\Delta_k x : x \in T\}| \leq \mathcal{N}(T, \ell_2, 1/2^k) \cdot \mathcal{N}(T, \ell_2, 1/2^{k-1})$
  $\leq (\mathcal{N}(T, \ell_2, 1/2^k))^2$
- $\mathfrak{g}(T) \lesssim \sum_{k=1}^{\infty} (1/2^k) \cdot \log^{1/2} \mathcal{N}(T, \ell_2, 1/2^k)$
  $\lesssim \int_0^{\infty} \log^{1/2} \mathcal{N}(T, \ell_2, u) du$ (Dudley's theorem)

Gaussian mean width bound 4: generic chaining

- Again, wlog $|T| < \infty$. Define $T_0 \subseteq T_1 \subseteq \cdots \subseteq T_{k_*} = T$ $|T_0| = 1, |T_k| \leq 2^{2^k}$ (call such a sequence "admissible")

Gaussian mean width bound 4: generic chaining

- Again, wlog $|T| < \infty$. Define $T_0 \subseteq T_1 \subseteq \cdots \subseteq T_{k_*} = T$
  $|T_0| = 1, |T_k| \le 2^{2^k}$ (call such a sequence "admissible")
- Exercise: show Dudley's theorem is equivalent to
  $\mathfrak{g}(T) \lesssim \inf_{\{T_k\} \text{ admissible}} \sum_{k=1}^{\infty} 2^{k/2} \cdot \sup_{x \in T} d_{\ell_2}(x, T_k)$
  (should pick T_k to be the best $\varepsilon = \varepsilon(k)$ net of size 2^{2^k})

Gaussian mean width bound 4: generic chaining

- Again, wlog $|T| < \infty$. Define $T_0 \subseteq T_1 \subseteq \cdots \subseteq T_{k_*} = T$
  $|T_0| = 1, |T_k| \le 2^{2^k}$ (call such a sequence "admissible")
- Exercise: show Dudley's theorem is equivalent to
  $\mathfrak{g}(T) \lesssim \inf_{\{T_k\} \text{ admissible}} \sum_{k=1}^{\infty} 2^{k/2} \cdot \sup_{x \in T} d_{\ell_2}(x, T_k)$
  (should pick T_k to be the best $\varepsilon = \varepsilon(k)$ net of size 2^{2^k})
- Fernique'76*: can pull the $\sup_x$ *outside* the sum
- $\mathfrak{g}(T) \lesssim \inf_{\{T_k\}} \sup_{x \in T} \sum_{k=1}^{\infty} 2^{k/2} \cdot d_{\ell_2}(x, T_k) \stackrel{\text{def}}{=} \gamma_2(T, \ell_2)$

Gaussian mean width bound 4: generic chaining

- Again, wlog $|T| < \infty$. Define $T_0 \subseteq T_1 \subseteq \cdots \subseteq T_{k_*} = T$
  $|T_0| = 1, |T_k| \le 2^{2^k}$ (call such a sequence "admissible")
- Exercise: show Dudley's theorem is equivalent to
  $\mathfrak{g}(T) \lesssim \inf_{\{T_k\} \text{ admissible}} \sum_{k=1}^{\infty} 2^{k/2} \cdot \sup_{x \in T} d_{\ell_2}(x, T_k)$
  (should pick T_k to be the best $\varepsilon = \varepsilon(k)$ net of size 2^{2^k})
- Fernique'76*: can pull the $\sup_x$ *outside* the sum
- $\mathfrak{g}(T) \lesssim \inf_{\{T_k\}} \sup_{x \in T} \sum_{k=1}^{\infty} 2^{k/2} \cdot d_{\ell_2}(x, T_k) \stackrel{\text{def}}{=} \gamma_2(T, \ell_2)$
  $*$ equivalent upper bound proven by Fernique (who minimized some integral over all measures over T), but reformulated in terms of admissible sequences by Talgarand

Gaussian mean width bound 4: generic chaining

Proof of Fernique's bound

$$\mathfrak{g}(T) \le \underbrace{\mathop{\mathbb{E}}_{g} \sup_{x \in T} \langle g, \pi_0 x \rangle}_{0} + \mathop{\mathbb{E}}_{g} \sup_{x \in T} \sum_{k=1}^{\infty} \underbrace{\langle g, \Delta_k x \rangle}_{Y_k} \text{ (from before)}$$

- $\forall t$, $\mathbb{P}(Y_k > t 2^{k/2} \|\Delta_k x\|_2) \le e^{t^2 2^k / 2}$ (gaussian decay)

Gaussian mean width bound 4: generic chaining

Proof of Fernique's bound

$$\mathfrak{g}(T) \le \underbrace{\mathop{\mathbb{E}}_{g} \sup_{x \in T} \langle g, \pi_0 x \rangle}_{0} + \mathop{\mathbb{E}}_{g} \sup_{x \in T} \sum_{k=1}^{\infty} \underbrace{\langle g, \Delta_k x \rangle}_{Y_k} \text{ (from before)}$$

- $\forall t$, $\mathbb{P}(Y_k > t 2^{k/2} \|\Delta_k x\|_2) \le e^{t^2 2^k / 2}$ (gaussian decay)
- $\mathbb{P}(\exists x, k \;\; Y_k > t 2^{k/2} \|\Delta_k x\|_2) \le \sum_k (2^{2^k})^2 e^{-t^2 2^k / 2}$

Gaussian mean width bound 4: generic chaining

Proof of Fernique's bound

$$\mathfrak{g}(T) \le \underbrace{\mathop{\mathbb{E}}_{g} \sup_{x \in T} \langle g, \pi_0 x \rangle}_{0} + \mathop{\mathbb{E}}_{g} \sup_{x \in T} \sum_{k=1}^{\infty} \underbrace{\langle g, \Delta_k x \rangle}_{Y_k} \text{ (from before)}$$

- $\forall t$, $\mathbb{P}(Y_k > t 2^{k/2} \|\Delta_k x\|_2) \le e^{t^2 2^k / 2}$ (gaussian decay)
- $\mathbb{P}(\exists x, k \;\; Y_k > t 2^{k/2} \|\Delta_k x\|_2) \le \sum_k (2^{2^k})^2 e^{-t^2 2^k / 2}$

$$\mathop{\mathbb{E}}_{g} \sup_{x \in T} \sum_k Y_k = \int_0^\infty \mathbb{P}(\sup_{x \in T} \sum_k Y_k > u) du$$

Gaussian mean width bound 4: generic chaining

$$\mathop{\mathbb{E}}_{g} \sup_{x \in T} \sum_k Y_k = \int_0^\infty \mathbb{P}(\sup_{x \in T} \sum_k Y_k > u) du$$

Gaussian mean width bound 4: generic chaining

$$\mathop{\mathbb{E}}_{g} \sup_{x \in T} \sum_k Y_k = \int_0^\infty \mathbb{P}(\sup_{x \in T} \sum_k Y_k > u) du$$

$$= \gamma_2(T, \ell_2) \cdot \int_0^\infty \mathbb{P}(\sup_{x \in T} \sum_k Y_k > t \sup_{x \in T} \sum_k 2^{k/2} \|\Delta_k x\|_2) dt$$

(change of variables: $u = t \sup_{x \in T} \sum_k 2^{k/2} \|\Delta_k x\|_2 \simeq t\gamma_2(T, \ell_2)$)

Gaussian mean width bound 4: generic chaining

$$\mathop{\mathbb{E}}_{g} \sup_{x \in T} \sum_k Y_k = \int_0^\infty \mathbb{P}(\sup_{x \in T} \sum_k Y_k > u) du$$

$$= \gamma_2(T, \ell_2) \cdot \int_0^\infty \mathbb{P}(\sup_{x \in T} \sum_k Y_k > t \sup_{x \in T} \sum_k 2^{k/2} \|\Delta_k x\|_2) dt$$

(change of variables: $u = t \sup_{x \in T} \sum_k 2^{k/2} \|\Delta_k x\|_2 \simeq t\gamma_2(T, \ell_2)$)

$$\leq \gamma_2(T, \ell_2) \cdot [2 + \int_2^\infty \left(\sum_{k=1}^\infty (2^{2^k})^2 e^{-t^2 2^k / 2} \right) dt]$$

$$\simeq \gamma_2(T, \ell_2)$$

Gaussian mean width bound 4: generic chaining

$$\mathop{\mathbb{E}}_{g} \sup_{x\in T} \sum_k Y_k = \int_0^\infty \mathbb{P}(\sup_{x\in T} \sum_k Y_k > u) du$$

$$= \gamma_2(T, \ell_2) \cdot \int_0^\infty \mathbb{P}(\sup_{x\in T} \sum_k Y_k > t \sup_{x\in T} \sum_k 2^{k/2} \|\Delta_k x\|_2) dt$$

(change of variables: $u = t \sup_{x\in T} \sum_k 2^{k/2} \|\Delta_k x\|_2 \simeq t\gamma_2(T, \ell_2)$)

$$\leq \gamma_2(T, \ell_2) \cdot [2 + \int_2^\infty \left(\sum_{k=1}^\infty (2^{2^k})^2 e^{-t^2 2^k/2} \right) dt]$$

$$\simeq \gamma_2(T, \ell_2)$$

- Conclusion: $\mathfrak{g}(T) \lesssim \gamma_2(T, \ell_2)$

Gaussian mean width bound 4: generic chaining

$$\mathop{\mathbb{E}}_{g} \sup_{x\in T} \sum_k Y_k = \int_0^\infty \mathbb{P}(\sup_{x\in T} \sum_k Y_k > u) du$$

$$= \gamma_2(T, \ell_2) \cdot \int_0^\infty \mathbb{P}(\sup_{x\in T} \sum_k Y_k > t \sup_{x\in T} \sum_k 2^{k/2} \|\Delta_k x\|_2) dt$$

$$(\text{change of variables: } u = t \sup_{x\in T} \sum_k 2^{k/2} \|\Delta_k x\|_2 \simeq t\gamma_2(T, \ell_2))$$

$$\leq \gamma_2(T, \ell_2) \cdot [2 + \int_2^\infty \left(\sum_{k=1}^\infty (2^{2^k})^2 e^{-t^2 2^k/2} \right) dt]$$

$$\simeq \gamma_2(T, \ell_2)$$

- Conclusion: $\mathfrak{g}(T) \lesssim \gamma_2(T, \ell_2)$
- Talagrand: $\mathfrak{g}(T) \simeq \gamma_2(T, \ell_2)$ (won't show today) ("Majorizing measures theorem")

Are these bounds really different?

- $\gamma_2(T, \ell_2)$: $\inf_{\{T_k\}} \sup_{x \in T} \sum_{k=1}^{\infty} 2^{k/2} \cdot d_{\ell_2}(x, T_k)$
- **Dudley:** $\inf_{\{T_k\}} \sum_{k=1}^{\infty} 2^{k/2} \cdot \sup_{x \in T} d_{\ell_2}(x, T_k)$
  $\simeq \int_0^{\infty} \log^{1/2} \mathcal{N}(T, \ell_2, u) du$

Are these bounds really different?

- $\gamma_2(T, \ell_2)$**:** $\inf_{\{T_k\}} \sup_{x \in T} \sum_{k=1}^{\infty} 2^{k/2} \cdot d_{\ell_2}(x, T_k)$
- **Dudley:** $\inf_{\{T_k\}} \sum_{k=1}^{\infty} 2^{k/2} \cdot \sup_{x \in T} d_{\ell_2}(x, T_k)$
  $\simeq \int_0^{\infty} \log^{1/2} \mathcal{N}(T, \ell_2, u) du$
- Dudley not optimal: $T = B_{\ell_1^n}$

Are these bounds really different?

- $\gamma_2(T, \ell_2)$**:** $\inf_{\{T_k\}} \sup_{x \in T} \sum_{k=1}^{\infty} 2^{k/2} \cdot d_{\ell_2}(x, T_k)$
- **Dudley:** $\inf_{\{T_k\}} \sum_{k=1}^{\infty} 2^{k/2} \cdot \sup_{x \in T} d_{\ell_2}(x, T_k)$
  $\simeq \int_0^{\infty} \log^{1/2} \mathcal{N}(T, \ell_2, u) du$
- Dudley not optimal: $T = B_{\ell_1^n}$
- $\sup_{x \in B_{\ell_1^n}} \langle g, x \rangle = \|g\|_\infty$, so $\mathfrak{g}(T) \simeq \sqrt{\log n}$
- **Exercise:** Come up with admissible $\{T_k\}$ yielding $\gamma_2 \lesssim \sqrt{\log n}$ (must exist by majorizing measures)

Are these bounds really different?

- $\gamma_2(T, \ell_2)$: $\inf_{\{T_k\}} \sup_{x \in T} \sum_{k=1}^{\infty} 2^{k/2} \cdot d_{\ell_2}(x, T_k)$
- **Dudley:** $\inf_{\{T_k\}} \sum_{k=1}^{\infty} 2^{k/2} \cdot \sup_{x \in T} d_{\ell_2}(x, T_k)$
  $\simeq \int_0^{\infty} \log^{1/2} \mathcal{N}(T, \ell_2, u) du$
- Dudley not optimal: $T = B_{\ell_1^n}$
- $\sup_{x \in B_{\ell_1^n}} \langle g, x \rangle = \|g\|_{\infty}$, so $\mathfrak{g}(T) \simeq \sqrt{\log n}$
- **Exercise:** Come up with admissible $\{T_k\}$ yielding $\gamma_2 \lesssim \sqrt{\log n}$ (must exist by majorizing measures)
- Dudley: $\log \mathcal{N}(B_{\ell_1^n}, \ell_2, u) \simeq (1/u^2) \log n$ for u not too small (consider just covering $(1/u^2)$-sparse vectors with u^2 in each coordinate). Dudley can only give $\mathfrak{g}(B_{\ell_1^n}) \lesssim \log^{3/2} n$.

Are these bounds really different?

- $\gamma_2(T,\ell_2)$: $\inf_{\{T_k\}} \sup_{x\in T} \sum_{k=1}^{\infty} 2^{k/2} \cdot d_{\ell_2}(x, T_k)$
- **Dudley:** $\inf_{\{T_k\}} \sum_{k=1}^{\infty} 2^{k/2} \cdot \sup_{x\in T} d_{\ell_2}(x, T_k)$
  $\simeq \int_0^\infty \log^{1/2} \mathcal{N}(T,\ell_2,u)du$
- Dudley not optimal: $T = B_{\ell_1^n}$
- $\sup_{x\in B_{\ell_1^n}} \langle g, x\rangle = \|g\|_\infty$, so $\mathfrak{g}(T) \simeq \sqrt{\log n}$
- **Exercise:** Come up with admissible $\{T_k\}$ yielding $\gamma_2 \lesssim \sqrt{\log n}$ (must exist by majorizing measures)
- Dudley: $\log \mathcal{N}(B_{\ell_1^n}, \ell_2, u) \simeq (1/u^2) \log n$ for u not too small (consider just covering $(1/u^2)$-sparse vectors with u^2 in each coordinate). Dudley can only give $\mathfrak{g}(B_{\ell_1^n}) \lesssim \log^{3/2} n$.
- Simple vanilla ε-net argument gives $\mathfrak{g}(B_{\ell_1^n}) \lesssim \mathrm{poly}(n)$.

High probability

- So far just talked about $\mathfrak{g}(T) = \mathbb{E}_g \sup_{x \in T} Z_x$

  But what if we want to know $\sup_{x \in T} Z_x$ is small whp, not just in expectation?

High probability

- So far just talked about $\mathfrak{g}(T) = \mathbb{E}_g \sup_{x \in T} Z_x$

  But what if we want to know $\sup_{x \in T} Z_x$ is small whp, not just in expectation?

- Usual approach: bound $\mathbb{E}_g \sup_{x \in T} Z_x^p$ for large p and do Markov ("moment method")

  Can bound moments using chaining too; see (Dirksen'13)

Applications in computer science

- Fast RIP matrices (Candès, Tao'06), (Rudelson, Vershynin'06), (Cheragchi, Guruswami, Velingker'13), (N., Price, Wootters'14), (Bourgain'14), (Haviv, Regev'15)
- Fast JL (Ailon, Liberty'11), (Krahmer, Ward'11), (Bourgain, Dirksen, N.'15), (Oymak, Recht, Soltanolkotabi'15)
- Instance-wise JL bounds (Gordon'88), (Klartag, Mendelson'05), (Mendelson, Pajor, Tomczak-Jaegermann'07), (Dirksen'14)
- Approximate nearest neighbor (Indyk, Naor'07)
- Deterministic algorithm to estimate graph cover time (Ding, Lee, Peres'11)
- List-decodability of random codes (Wootters'13), (Rudra, Wootters'14)
- ...

A chaining result for quadratic forms

Theorem

[Krahmer, Mendelson, Rauhut'14] Let $\mathcal{A} \subset \mathbb{R}^{n \times n}$ be a family of matrices, and let $\sigma_1, \ldots, \sigma_n$ be independent subgaussians. Then

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

(Δ_X is diameter under X-norm)

A chaining result for quadratic forms

Theorem

[Krahmer, Mendelson, Rauhut'14] Let $\mathcal{A} \subset \mathbb{R}^{n\times n}$ be a family of matrices, and let $\sigma_1, \ldots, \sigma_n$ be independent subgaussians. Then

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

(Δ_X is diameter under X-norm)

Won't show proof today, but it is similar to bounding $\mathfrak{g}(T)$ (with some extra tricks). See `http://people.seas.harvard.edu/~minilek/madalgo2015/`, Lecture 3.

Instance-wise bounds for JL

Corollary (Gordon'88, Klartag-Mendelson'05, Mendelson, Pajor, Tomczak-Jaegermann'07, Dirksen'14)

For $T \subseteq S^{n-1}$ and $0 < \varepsilon < 1/2$, let $\Pi \in \mathbb{R}^{m \times n}$ have independent subgaussian independent entries with mean zero and variance $1/m$ for $m \gtrsim {(\mathfrak{g}^2(T)+1)}/{\varepsilon^2}$. Then

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

Proof of Gordon's theorem

- For $x \in T$ let A_x denote the $m \times mn$ matrix:

$$A_x = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & x_1 & \cdots & x_n \end{bmatrix}.$$

Instance-wise bounds for JL

Proof of Gordon's theorem

- For $x \in T$ let A_x denote the $m \times mn$ matrix:

$$A_x = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & x_1 & \cdots & x_n \end{bmatrix}.$$

- Then $\|\Pi x\|_2^2 = \|A_x \sigma\|_2^2$, where σ is formed by concatenating rows of Π (multiplied by $\sqrt{m}$).

Instance-wise bounds for JL

Proof of Gordon's theorem

- For $x \in T$ let A_x denote the $m \times mn$ matrix:

$$A_x = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & x_1 & \cdots & x_n \end{bmatrix}.$$

- Then $\|\Pi x\|_2^2 = \|A_x \sigma\|_2^2$, where σ is formed by concatenating rows of Π (multiplied by $\sqrt{m}$).
- $\|A_x - A_y\| = \|A_{x-y}\| = (1/\sqrt{m}) \cdot \|x - y\|_2$
  $\Rightarrow \gamma_2(\mathcal{A}_T, \|\cdot\|_{\ell_2 \to \ell_2}) = \gamma_2(T, \ell_2) \simeq \mathfrak{g}(T)$

Instance-wise bounds for JL

Proof of Gordon's theorem

- For $x \in T$ let A_x denote the $m \times mn$ matrix:

$$A_x = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & x_1 & \cdots & x_n \end{bmatrix}.$$

- Then $\|\Pi x\|_2^2 = \|A_x \sigma\|_2^2$, where σ is formed by concatenating rows of Π (multiplied by $\sqrt{m}$).
- $\|A_x - A_y\| = \|A_{x-y}\| = ({}^1/\sqrt{m}) \cdot \|x - y\|_2$
  $\Rightarrow \gamma_2(\mathcal{A}_T, \|\cdot\|_{\ell_2 \to \ell_2}) = \gamma_2(T, \ell_2) \simeq \mathfrak{g}(T)$
- $\Delta_F(\mathcal{A}_T) = 1$, $\Delta_{\ell_2 \to \ell_2}(\mathcal{A}_T) = 1/\sqrt{m}$

Instance-wise bounds for JL

Proof of Gordon's theorem

- For $x \in T$ let A_x denote the $m \times mn$ matrix:

$$A_x = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & x_1 & \cdots & x_n \end{bmatrix}.$$

- Then $\|\Pi x\|_2^2 = \|A_x \sigma\|_2^2$, where σ is formed by concatenating rows of Π (multiplied by $\sqrt{m}$).
- $\|A_x - A_y\| = \|A_{x-y}\| = (1/\sqrt{m}) \cdot \|x - y\|_2$
  $\Rightarrow \gamma_2(\mathcal{A}_T, \|\cdot\|_{\ell_2 \to \ell_2}) = \gamma_2(T, \ell_2) \simeq \mathfrak{g}(T)$
- $\Delta_F(\mathcal{A}_T) = 1$, $\Delta_{\ell_2 \to \ell_2}(\mathcal{A}_T) = 1/\sqrt{m}$
- Thus $\mathbb{E}_\Pi \sup_{x \in T} |\|\Pi x\|_2^2 - 1| \lesssim \mathfrak{g}^2(T)/m + \mathfrak{g}(T)/\sqrt{m} + 1/\sqrt{m}$

Instance-wise bounds for JL

Proof of Gordon's theorem

- For $x \in T$ let A_x denote the $m \times mn$ matrix:

$$A_x = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & x_1 & \cdots & x_n & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & x_1 & \cdots & x_n \end{bmatrix}.$$

- Then $\|\Pi x\|_2^2 = \|A_x \sigma\|_2^2$, where σ is formed by concatenating rows of Π (multiplied by $\sqrt{m}$).
- $\|A_x - A_y\| = \|A_{x-y}\| = (1/\sqrt{m}) \cdot \|x - y\|_2$
  $\Rightarrow \gamma_2(\mathcal{A}_T, \|\cdot\|_{\ell_2 \to \ell_2}) = \gamma_2(T, \ell_2) \simeq \mathfrak{g}(T)$
- $\Delta_F(\mathcal{A}_T) = 1$, $\Delta_{\ell_2 \to \ell_2}(\mathcal{A}_T) = 1/\sqrt{m}$
- Thus $\mathbb{E}_\Pi \sup_{x \in T} |\|\Pi x\|_2^2 - 1| \lesssim \mathfrak{g}^2(T)/m + \mathfrak{g}(T)/\sqrt{m} + 1/\sqrt{m}$
- Set $m \gtrsim (\mathfrak{g}^2(T)+1)/\varepsilon^2$

Consequences of Gordon's theorem

$m \gtrsim (\mathfrak{g}^2(T)+1)/\varepsilon^2$

- $|T| < \infty$: $\mathfrak{g}^2(T) \lesssim \log|T|$ (JL)
- T a d-dim subspace: $\mathfrak{g}^2(T) \simeq d$ (subspace embeddings)
- T all k-sparse vectors: $\mathfrak{g}^2(T) \simeq k\log(n/k)$ (RIP)

Consequences of Gordon's theorem

$m \gtrsim {}^{(\mathfrak{g}^2(T)+1)}/_{\varepsilon^2}$

- $|T| < \infty$: $\mathfrak{g}^2(T) \lesssim \log|T|$ (JL)
- T a d-dim subspace: $\mathfrak{g}^2(T) \simeq d$ (subspace embeddings)
- T all k-sparse vectors: $\mathfrak{g}^2(T) \simeq k\log(n/k)$ (RIP)
- more applications to constrained least squares, manifold learning, model-based compressed sensing, ...

  (see (Dirksen'14) and (Bourgain, Dirksen, N.'15))

Chaining isn't just for gaussians

Chaining without gaussians: RIP (Rudelson, Vershynin'06)

"Restricted isometry property" useful in compressed sensing.
$T = \{x : \|x\|_0 \leq k, \|x\|_2 = 1\}$.

Theorem (Candès-Tao'06, Donoho'06, Candés'08)

If Π *satisfies* (ε_*, k)*-RIP for* $\varepsilon_* < \sqrt{2} - 1$ *then there is a linear program which, given* Πx *and* Π *as input, recovers* $\tilde{x}$ *in polynomial time such that* $\|x - \tilde{x}\|_2 \leq O(1/\sqrt{k}) \cdot \min_{\|y\|_0 \leq k} \|x - y\|_1$.

Chaining without gaussians: RIP (Rudelson, Vershynin'06)

"Restricted isometry property" useful in compressed sensing.
$T = \{x : \|x\|_0 \leq k, \|x\|_2 = 1\}$.

Theorem (Candès-Tao'06, Donoho'06, Candés'08)

If Π satisfies (ε_*, k)-RIP for $\varepsilon_* < \sqrt{2} - 1$ then there is a linear program which, given Πx and Π as input, recovers $\tilde{x}$ in polynomial time such that $\|x - \tilde{x}\|_2 \leq O(1/\sqrt{k}) \cdot \min_{\|y\|_0 \leq k} \|x - y\|_1$.

Of interest to show sampling rows of discrete Fourier matrix is RIP

Chaining without gaussians: RIP (Rudelson, Vershynin'06)

- (Unnormalized) Fourier matrix F, rows: $z_1^*, \ldots, z_n^*$
- $\delta_1, \ldots, \delta_n$ independent Bernoulli with expectation m/n

Chaining without gaussians: RIP (Rudelson, Vershynin'06)

- (Unnormalized) Fourier matrix F, rows: $z_1^*, \ldots, z_n^*$
- $\delta_1, \ldots, \delta_n$ independent Bernoulli with expectation m/n
- Want

$$\mathop{\mathbb{E}}_{\delta} \sup_{\substack{T \subset [n] \\ |T| \leq k}} \| I_T - \frac{1}{m} \sum_{i=1}^{n} \delta_i z_i^{(T)} z_i^{(T)^*} \| < \varepsilon$$

Chaining without gaussians: RIP (Rudelson, Vershynin'06)

$$\text{LHS} = \mathop{\mathbb{E}}_{\delta} \sup_{\substack{T \subset [n] \\ |T| \le k}} \| \overbrace{\mathop{\mathbb{E}}_{\delta'} \frac{1}{m} \sum_{i=1}^{n} \delta'_i z_i^{(T)} z_i^{(T)^*}}^{I_T} - \frac{1}{m} \sum_{i=1}^{n} \delta_i z_i^{(T)} z_i^{(T)^*} \|$$

Chaining without gaussians: RIP (Rudelson, Vershynin'06)

$$\text{LHS} = \mathop{\mathbb{E}}_{\delta} \sup_{\substack{T\subset[n]\\ |T|\le k}} \| \overbrace{\mathop{\mathbb{E}}_{\delta'} \frac{1}{m}\sum_{i=1}^{n} \delta'_i z_i^{(T)} z_i^{(T)^*}}^{I_T} - \frac{1}{m}\sum_{i=1}^{n} \delta_i z_i^{(T)} z_i^{(T)^*} \|$$

$$\le \frac{1}{m} \mathop{\mathbb{E}}_{\delta,\delta'} \sup_{T} \| \sum_{i=1}^{n} (\delta'_i - \delta_i) z_i^{(T)} z_i^{(T)^*} \| \text{ (Jensen)}$$

Chaining without gaussians: RIP (Rudelson, Vershynin'06)

$$\text{LHS} = \mathop{\mathbb{E}}_{\delta} \sup_{\substack{T\subset[n]\\ |T|\le k}} \| \overbrace{\mathop{\mathbb{E}}_{\delta'} \frac{1}{m}\sum_{i=1}^{n} \delta_i' z_i^{(T)} z_i^{(T)^*}}^{I_T} - \frac{1}{m}\sum_{i=1}^{n} \delta_i z_i^{(T)} z_i^{(T)^*} \|$$

$$\le \frac{1}{m} \mathop{\mathbb{E}}_{\delta,\delta'} \sup_{T} \| \sum_{i=1}^{n} (\delta_i' - \delta_i) z_i^{(T)} z_i^{(T)^*} \| \text{ (Jensen)}$$

$$= \sqrt{\frac{\pi}{2}} \cdot \frac{1}{m} \mathop{\mathbb{E}}_{\delta,\delta',\sigma} \sup_{T} \| \mathop{\mathbb{E}}_{g} \sum_{i=1}^{n} |g_i| \sigma_i (\delta_i' - \delta_i) z_i^{(T)} z_i^{(T)^*} \|$$

Chaining without gaussians: RIP (Rudelson, Vershynin'06)

$$\mathrm{LHS} = \mathop{\mathbb{E}}_{\delta} \sup_{\substack{T\subset[n]\\|T|\le k}} \| \overbrace{\mathop{\mathbb{E}}_{\delta'} \frac{1}{m}\sum_{i=1}^{n} \delta'_i z_i^{(T)} z_i^{(T)^*}}^{I_T} - \frac{1}{m}\sum_{i=1}^{n} \delta_i z_i^{(T)} z_i^{(T)^*} \|$$

$$\le \frac{1}{m} \mathop{\mathbb{E}}_{\delta,\delta'} \sup_{T} \| \sum_{i=1}^{n} (\delta'_i - \delta_i) z_i^{(T)} z_i^{(T)^*} \| \text{ (Jensen)}$$

$$= \sqrt{\frac{\pi}{2}} \cdot \frac{1}{m} \mathop{\mathbb{E}}_{\delta,\delta',\sigma} \sup_{T} \| \mathop{\mathbb{E}}_{g} \sum_{i=1}^{n} |g_i| \sigma_i (\delta'_i - \delta_i) z_i^{(T)} z_i^{(T)^*} \|$$

$$\le \sqrt{2\pi} \cdot \frac{1}{m} \mathop{\mathbb{E}}_{\delta,g} \sup_{T} \| \sum_{i=1}^{n} g_i \delta_i z_i^{(T)} z_i^{(T)^*} \| \text{ (Jensen+triangle ineq)}$$

Chaining without gaussians: RIP (Rudelson, Vershynin'06)

$$\text{LHS} = \mathop{\mathbb{E}}_{\delta} \sup_{\substack{T\subset[n]\\ |T|\le k}} \| \overbrace{\mathop{\mathbb{E}}_{\delta'} \frac{1}{m}\sum_{i=1}^{n} \delta'_i z_i^{(T)} z_i^{(T)^*}}^{I_T} - \frac{1}{m}\sum_{i=1}^{n} \delta_i z_i^{(T)} z_i^{(T)^*} \|$$

$$\le \frac{1}{m} \mathop{\mathbb{E}}_{\delta,\delta'} \sup_{T} \| \sum_{i=1}^{n} (\delta'_i - \delta_i) z_i^{(T)} z_i^{(T)^*} \| \text{ (Jensen)}$$

$$= \sqrt{\frac{\pi}{2}} \cdot \frac{1}{m} \mathop{\mathbb{E}}_{\delta,\delta',\sigma} \sup_{T} \| \mathop{\mathbb{E}}_{g} \sum_{i=1}^{n} |g_i| \sigma_i (\delta'_i - \delta_i) z_i^{(T)} z_i^{(T)^*} \|$$

$$\le \sqrt{2\pi} \cdot \frac{1}{m} \mathop{\mathbb{E}}_{\delta,g} \sup_{T} \| \sum_{i=1}^{n} g_i \delta_i z_i^{(T)} z_i^{(T)^*} \| \text{ (Jensen+triangle ineq)}$$

$$\simeq \frac{1}{m} \mathop{\mathbb{E}}_{\delta} \mathop{\mathbb{E}}_{g} \sup_{x\in B_2^{n,k}} | \sum_{i=1}^{n} g_i \delta_i \langle z_i, x\rangle^2 | \text{ (gaussian mean width!)}$$

The End

June 22nd+23rd: workshop on concentration of measure / chaining at Harvard, after STOC'16. Details+website forthcoming.