Course: High Dimensional Probability: applications to statistical learning and data science.

Text(s): "High-Dimensional Probability" Roman Vershynin
"High-Dimensional Statistics" Martin J. Wainwright

Geography: Let n be sample size, d is the dimension of data

- Classical Asymptotics
  $n \nearrow +\infty$, d fixed
  Ex: Laws of large numbers, Central limit theorem
- High dimensional Asymptotics
  $(n,d) \nearrow +\infty$ while $\Psi(n,d) \to \alpha < \infty$, where Ψ is a "scaling function"
  E.g.: $\Psi(n,d) = \frac{d}{n}$
- Non-asymptotic Bounds ⟸ This class
  (n,d) is fixed. Probability of successful estimation/inference depends on (n,d)

See (W:1) for examples of the three viewpoints.

Lecture 1: Review (V:1)

Let X be a random variable on Probability space

Expectation and Variance

$$\mathbb{E}X, \quad \mathrm{Var}(X) = \mathbb{E}(X-\mathbb{E}X)^2 = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

Moment Generating Function:

$$M_X(t) = \mathbb{E}e^{tX}, \quad t \in \mathbb{R}$$

L^p-norm $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$, $p \in (0,\infty)$

Banach space:

$$L^p = \{X : \|X\|_p < \infty\}$$

Remark: L^2 is a Hilbert space

$$\langle X, Y \rangle_2 = \mathbb{E}[XY], \quad \|X\|_2 = \sqrt{\langle X, X\rangle} = \sqrt{\mathbb{E}[X^2]}$$

Then

$$\|X - \mathbb{E}X\|_2 = \sqrt{\mathrm{Var}(X)} \quad \text{and}$$

$$\mathrm{cov}(X,Y) := \mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y)$$
$$= \langle X - \mathbb{E}X, \; Y - \mathbb{E}Y \rangle_2$$

Limit Theorems

Thm (Strong Law of Large Numbers)
Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}X_i = \mu$. Then $S_n = \sum_{i=1}^{n} X_i$ satisfies

$$\frac{1}{n} \cdot S_n \to \mu \quad \text{almost surely}$$

$$\left[\lim_{n\to\infty} \frac{S_n}{n} = \mu \text{ w.p. } 1\right]$$

Intuition: $\mathrm{Var}(S_n) = \dfrac{\mathrm{Var}(X_i)}{n}$

Thm (Central Limit Theorem)
Let $X_1, X_2, \ldots$ be i.i.d. with

$$\mathbb{E}X_i = \mu, \quad \mathrm{Var}\, X_i = \sigma^2$$

Define $$Z_n = \frac{S_n - \mathbb{E}S_n}{\sqrt{\mathrm{Var}(S_n)}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

Then $$Z_n \to N(0,1) \quad \text{in distribution}$$

$$\left[\mathbb{P}[Z_n \geq t] \to \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx\right]$$

Chapter 1

Concentration Inequalities

- Chernoff Bound
- Sub-Gaussian RV (Hoeffding)
- Sub-Exponential RV (Bernstein)
- Johnson-Lindenstrauss
- McDiarmid inequality
- Robust Mean Estimation

Concentration Inequalities

Goal: Bounds of the form

$$\mathbb{P}\left[\frac{S_n}{n} \geq \mu + t\right] \leq \text{small}(n, t)$$

Prop: Let $Z \sim N(\mu, \sigma)$. Then

$$\mathbb{P}[Z \geq \mu + t] \leq e^{-t^2/2\sigma^2} \quad \forall t > 0$$

So would hope $\mathbb{P}\left[\frac{S_n}{n} \geq \mu + t\right] \leq C\, e^{-\frac{nt^2}{2\sigma^2}}$

Bad news from CLT:

$$\sup_t \left\{ \left| \mathbb{P}[Z_n > t] - \mathbb{P}[Z > t] \right| \right\}$$ can be $\Omega\left(\frac{1}{\sqrt{n}}\right)$ even for Bernoulli

where $Z \sim N(0,1)$

[See (V: 2.1) for example]

Chernoff Method: {Sub-Gaussian, Sub-Exponential} RV's

Lemma (Markov) For any non-negative X and $t>0$, we have

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}X}{t}$$

pf: $\mathbb{E}X = \mathbb{E}\, X \mathbb{1}_{\{X \geq t\}} + \mathbb{E}\, X \mathbb{1}_{\{X < t\}}$

$$\geq t\, \mathbb{E}\mathbb{1}_{\{X \geq t\}} + 0 = t\, \mathbb{P}[X \geq t]$$

Cor: (Chebycher) For any random variable X with $\mathrm{var}(X) \leq \sigma^2 < \infty$, we have

$$\mathbb{P}[\,|X - \mathbb{E}[X]| \geq t\,] \leq \frac{\sigma^2}{t^2} \quad \forall t > 0$$

More generally, suppose $\mu = \mathbb{E}X < \infty$. Then for all $\lambda \geq 0$, we have

$$\mathbb{P}[X - \mu \geq t] = \mathbb{P}[e^{\lambda(X-\mu)} \geq e^{\lambda t}] \overset{\text{Markov}}{\leq} \frac{\mathbb{E}\, e^{\lambda(X-\mu)}}{e^{\lambda t}}$$

$$\Rightarrow \log \mathbb{P}[X-\mu \geq t] \leq \inf_{\lambda \geq 0} \{ \log \mathbb{E} e^{\lambda(X-\mu)} - \lambda t \}$$

$$= -\sup_{\lambda \geq 0} \{ \lambda t - \log \mathbb{E} e^{\lambda(X-\mu)} \}$$

Define for any function $\psi: \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, the Fenchel conjugate

$$\psi^*(t) = \sup_{\lambda} \{ t\lambda - \psi(\lambda) \}$$

Let's look at the main example

$$\psi_X(\lambda) = \log \mathbb{E} e^{\lambda(X-\mu)}$$

For all $\lambda \in \mathbb{R}$, observe

$$\psi_X(\lambda) = \log \mathbb{E} e^{\lambda(X-\mu)} \geq \mathbb{E} \log e^{\lambda(X-\mu)} = 0$$

So when $\lambda < 0$ and $t \geq 0$, we have

$$\lambda t - \psi(\lambda) \leq 0 = 0 \cdot t - \psi(0).$$

Therefore for $t \geq 0$, equality holds:

$$\psi_X^*(t) = \sup_{\lambda \geq 0} \{ t\lambda - \psi(\lambda) \}$$

We therefore arrive at the bound:

Chernoff Bound:

$$\mathbb{P}[X-\mu \geq t] \leq \exp\left(-\psi_X^*(t)\right)$$

where $\psi_X(\lambda) := \log\left(\mathbb{E}\, e^{\lambda(X-\mu)}\right)$.

Ex: Let $X \sim N(\mu, \sigma^2)$. Then

$$\mathbb{E}\, e^{\lambda(X-\mu)} = e^{\frac{\sigma^2\lambda^2}{2}} \quad \forall \lambda \in \mathbb{R}.$$

So $\mathbb{P}[X \geq \mu + t] \leq e^{-\frac{t^2}{2\sigma^2}} \quad \forall t > 0$

Defn: X with mean $\mu = \mathbb{E}X$ is sub-Gaussian with parameter $\sigma > 0$ if

$$\mathbb{E}\, e^{\lambda(X-\mu)} \leq e^{\frac{\sigma^2\lambda^2}{2}} \quad \forall \lambda \in \mathbb{R}$$

If X is sub-Gaussian, so is $-X$.

$$\Rightarrow \mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

Ex: (Rademacher)

Let ε be a rademacher RV:

$$\mathbb{P}[\varepsilon = 1] = \mathbb{P}[\varepsilon = -1] = \frac{1}{2}.$$

$$\mathbb{E}[e^{\lambda\varepsilon}] = \frac{1}{2}\left(e^{-\lambda} + e^{\lambda}\right) = \cosh(\lambda)$$

Exercise: $\cosh(x) \le \exp\left(\frac{x^2}{2}\right) \quad \forall x \in \mathbb{R}$

So ε is sub-Gaussian with $\sigma = 1$.

Ex: (Bounded RV)

Suppose X is supported on $[a,b]$.

Then Jensen's inequality implies

$$\mathbb{E}\, e^{\lambda(X - \mathbb{E}_{X'}X')} \le \mathbb{E}_{X,X'}\, e^{\lambda(X-X')}$$

where X' is independent copy of X.

Observe $X - X' \sim \varepsilon(X - X')$

So $\mathbb{E}_{X,X'}\, e^{\lambda(X-X')} = \mathbb{E}_{X,X'}\mathbb{E}_{\varepsilon}\, e^{\varepsilon\cdot\lambda(X-X')}$

$$\le \mathbb{E}_{X,X'} \exp(\lambda^2(X-X')^2/2)$$

$$\le \exp(\lambda^2(b-a)^2/2)$$

A slightly more involved argument shows that X is sub-Gaussian with parameter $\frac{b-a}{2}$.

$$\Rightarrow \mathbb{P}[|X-\mu| \geq t] \leq 2\exp\left(-\frac{2t^2}{(b-a)^2}\right)$$

Lemma: [Sum Rule]

X_i are independent, σ_i-sub-Gaussian $\forall i=1,\dots,n$ $\Rightarrow \sum_{i=1}^{n} X_i$ is $\|\sigma\|_2$-sub-Gaussian

---

Cor: (Hoeffding)

Suppose $X_1, \dots, X_n$ are independent with $\mathbb{E}X_i = \mu_i$ and are σ_i-sub-Gaussian. Then

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i-\mu_i) \geq t\right] \leq \exp\left\{-\frac{t^2}{2\|\sigma\|_2^2}\right\}$$

$\Rightarrow$ If $\mu_i = \mu$, $\sigma_i = \sigma$ then

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + t\right] \leq \exp\left\{-\frac{nt^2}{2\sigma^2}\right\}$$

Subexponential RV:

Ex: Let $Z \sim N(0,1)$. Let's compute

$$\mathbb{E}\left[e^{\lambda(Z^2-1)}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(x^2-1)} e^{-x^2/2} dx$$

$$= \frac{e^{-\lambda}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(1-2\lambda)x^2/2} dx$$

$$= \begin{cases} \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} & \text{if } \lambda < \frac{1}{2} \\ +\infty, & \text{if } \lambda > \frac{1}{2} \end{cases}$$

Defn: X with mean $\mu = \mathbb{E}X$ is **subexponential with parameters (σ, α)** if

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}} \quad \forall |\lambda| \leq \frac{1}{\alpha}$$

Back to example $Z \sim N(0,1)$

$$\mathbb{E}\left[e^{\lambda(Z^2-1)}\right] \leq \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{\frac{4\lambda^2}{2}} \quad \forall |\lambda| < \frac{1}{4}$$

So Z^2 is $(2,4)$-subexponential.

Thm (Subexponential tail bound)
Let X be subexponential with (ν, α).
Then

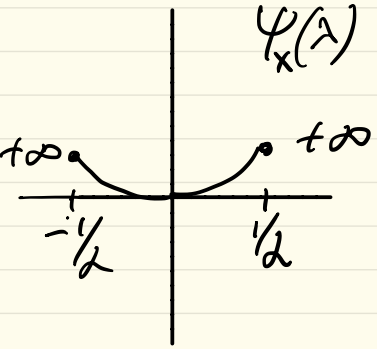
$$\mathbb{P}[X - \mu \geq t] \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}}, & \text{if } |t| \leq \frac{\nu^2}{\alpha} \\ e^{-\frac{t}{2\alpha}}, & \text{o.w.} \end{cases}$$

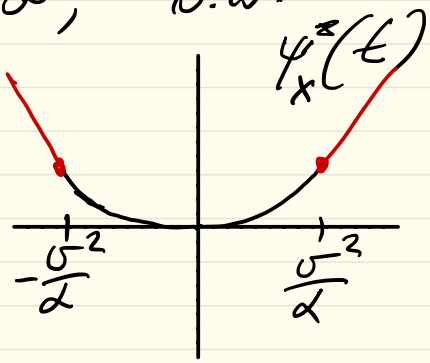
pf: Back to Chernoff

$$\log \mathbb{P}[X - \mu \geq t] \leq -\psi_X^*(t)$$

where $\psi_X(\lambda) = \log \mathbb{E} e^{\lambda(X-\mu)}$

$$= \begin{cases} \frac{\nu^2 \lambda^2}{2}, & \text{if } |\lambda| \leq \frac{1}{\alpha} \\ +\infty, & \text{o.w.} \end{cases}$$





Thm: (Bernstein) Let X be subexponential with parameter (σ, α) and mean $\mu = \mathbb{E}X$. Then

$$\mathbb{P}[|X-\mu| \geq t] \leq 2\exp\left[-\left(\frac{t^2}{\sigma^2} \wedge \frac{t}{\alpha}\right)/2\right]$$

minimum

$$t \mapsto \left(\frac{t^2}{\sigma^2} \wedge \frac{t}{\alpha}\right)/2$$

Lemma: [Sum Rule]

X_i are (σ_i, α_i)-subexp $\forall i = 1, \dots, n$ $\Rightarrow \sum_{i=1}^{n} X_i$ is $(\|\sigma\|_2, \|\alpha\|_\infty)$-subexp

Thm (Bernstein for sums) Let $X_1, \dots, X_n$ be independent subexponential with parameters (σ_i, α_i), and with mean $\mu_i = \mathbb{E}X_i$. Then

$$\mathbb{P}\left[\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| > t\right] \leq 2\exp\left[-\frac{1}{2}\left(\frac{t^2}{\|\sigma\|_2^2} \wedge \frac{t}{\|\alpha\|_\infty}\right)\right]$$

Thm: (Improved Bernstein for bounded RV)
Suppose $|X-\mu| \le b$, $\mathbb{E}(X-\mu)^2 = \sigma^2$. Then

$$\mathbb{E}\, e^{\lambda(X-\mu)} \le \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right) \quad \forall\, |\lambda| < \frac{1}{b}$$

Therefore

$$\text{(A)} \quad \mathbb{P}[|X-\mu| \ge t] \le 2e^{-\frac{t^2}{2(\sigma^2+bt)}}$$

pf: Taylor expansion:

$$\mathbb{E}\, e^{\lambda(X-\mu)} = \sum_{k=0}^{\infty} \lambda^k \frac{\mathbb{E}(X-\mu)^k}{k!} = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}(X-\mu)^k}{k!}$$

$$\le 1 + \sum_{k=2}^{\infty} \frac{\lambda^2\sigma^2 \cdot b^{k-2}\lambda^{k-2}}{2\cdot 3\cdot 4\cdots k}$$

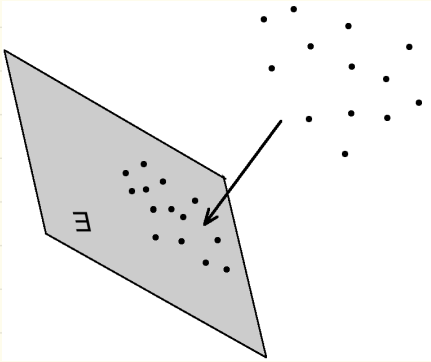
$$\le 1 + \frac{\lambda^2\sigma^2}{2}\cdot\frac{1}{1-b|\lambda|} \le \exp\left(\frac{\lambda^2\sigma^2}{2}\cdot\frac{1}{1-b|\lambda|}\right)$$

(A) Follows from Chernoff by setting $\lambda = \frac{t}{bt+\sigma^2} \in [0, \frac{1}{b})$ ☒

This is superior to Hoeffding when $\sigma < b$.

Application: Dimensionality Reduction

Given $u_1, \ldots, u_m \in \mathbb{R}^d$ with $m \ll d$, can one map $u_1, \ldots, u_m$ into a lower dimensional space with low distortion?



Thm: (Johnson-Lindenstrauss)

Fix $\varepsilon, \delta \in (0,1)$, a set $U \subseteq \mathbb{R}^d$ of m points and a number $n > \frac{16 \ln(m^2/\delta)}{\varepsilon^2}$.

Let $X \in \mathbb{R}^{n \times d}$ consist of i.i.d $N(0,1)$ entries.

Then with probability $1-\delta$, the map $f(u) = \frac{1}{\sqrt{n}} Xu$ satisfies

$$1-\varepsilon \leq \frac{\|f(u)-f(v)\|_2^2}{\|u-v\|_2^2} \leq 1+\varepsilon \quad \forall u,v \in U.$$

pf: Observe

$$\frac{\|Xu\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{n} \underbrace{\left\langle X_i, \frac{u}{\|u\|} \right\rangle^2}_{\text{i.i.d } N(0,1)}$$

(X_i: ith row of X)

$$\Rightarrow \frac{\|Xu\|_2^2}{\|u\|_2^2} \text{ is } (2\sqrt{n}, 4)\text{-subexponential}$$

$\Rightarrow$ Bernstein:

$$P\left[\left|\frac{\|Xu\|_2^2}{n\|u\|_2^2} - 1\right| > \varepsilon\right] \le 2\exp\left[-\left(\frac{n\varepsilon^2}{8} \wedge \frac{n\varepsilon}{8}\right)\right]$$

$$= 2\exp\left(-\frac{n\varepsilon^2}{8}\right) \quad \forall\, 0 \le \varepsilon \le 1$$

So for any i, j get

$$P\left[\frac{\|f(u_i - u_j)\|^2}{\|u_i - u_j\|_2^2} \notin [1-\varepsilon, 1+\varepsilon]\right] \le 2e^{-n\varepsilon^2/8}$$

Take union bound over $\binom{m}{2}$ pairs of points

$$2\binom{m}{2} e^{-n\varepsilon^2/8} \le m^2 e^{-n\varepsilon^2/8} = \delta$$ ∎

Question: What if $m = \infty$ but U has few "degrees of freedom"?

So far, we have focused on concetration of the average $\frac{1}{n}\sum_{i=1}^{n} X_i$.

Often one is interested in bounds

$$\mathbb{P}\left[\left|f(X_1,\dots,X_n) - \mathbb{E}f(X_1,\dots,X_n)\right| > t\right] \leq \text{small}(n,t)$$

where X_i are independent and $f: \mathbb{R}^n \to \mathbb{R}$ is some function.

Useful insight:

As long as $f(X_1,\dots,X_n)$ depends weakly on individual X_i, concentration holds!

Thm: (McDiarmid) Suppose that $f: \mathcal{X}^n \to \mathbb{R}$ has the bounded difference property: $\exists L_1, \dots, L_n$ such that

$$|f(x_1, \dots, x_k, \dots, x_n) - f(x_1, \dots, x_k', \dots, x_n)| \le L_k \quad \forall x, x' \in \mathcal{X}^n$$

Then for independent R's $X = (X_1, \dots, X_n)$ we have

$$\mathbb{P}[|f(X) - \mathbb{E}f(X)| > t] \le 2e^{-\frac{2t^2}{\|L\|_2^2}}$$

pf: We will use the Martingale method.

Define

$$Y_0 = \mathbb{E}f(X) \quad \text{and} \quad Y_i = \mathbb{E}[f(X) | X_1, \dots, X_i] \quad \forall i$$

Observe

$$Y_i = Y_0 + \sum_{j=0}^{i-1} \underbrace{(Y_{j+1} - Y_j)}_{=: D_{j+1}} = Y_0 + \sum_{j=1}^{i} D_j$$

and

$$\mathbb{E}[Y_i | X_1, \dots, X_{i-1}] = \mathbb{E}[f(X) | X_1, \dots X_{i-1}] = Y_{i-1}$$

$$\Rightarrow \mathbb{E}[D_{j+1} | X_1, \dots, X_i] = 0$$

$$\Rightarrow \mathbb{E}[e^{\lambda(f(X) - \mathbb{E}f(X))}] = \mathbb{E}[e^{\lambda(Y_n - Y_0)}] =$$

$$\mathbb{E}\left[e^{\lambda(f(X)-\mathbb{E}f(X))}\right] = \mathbb{E}\left[e^{\lambda(y_n - y_0)}\right]$$

$$= \mathbb{E}\left[e^{\lambda \sum_{j=1}^{n} D_j}\right]$$

$$= \mathbb{E}\left[e^{\lambda(y_{n-1}-y_0)} e^{\lambda D_n}\right]$$

$$= \mathbb{E}\left[e^{\lambda(y_{n-1}-y_0)} \mathbb{E}\left[e^{\lambda D_n} \mid X_1, \dots, X_{n-1}\right]\right]$$

Let X' differ from X in x_i. Then

$$\mathbb{E}\left[e^{\lambda D_i} \mid X_1, \dots, X_{i-1}\right] = \mathbb{E}\left[e^{\lambda(y_i - y_{i-1})} \mid X_1, \dots, X_{i-1}\right]$$

$$= \mathbb{E}\left[e^{\lambda \mathbb{E}[f(X) - f(X') \mid X_1, \dots, X_i]} \,\middle|\, X_1, \dots, X_{i-1}\right]$$

Jensen $\leq$

$$\mathbb{E}\left[e^{\lambda (f(X) - f(X'))} \,\middle|\, X_1, \dots, X_{i-1}\right]$$

$(f(X) - f(X'))$ bounded by L_i

$$\Rightarrow \mathbb{E}\left[e^{\lambda(f(X) - f(X')) \mid X_1, \dots, X_{i-1}}\right] \leq e^{\frac{\lambda^2 L_i^2}{8}}$$

So $\mathbb{E}\left[e^{\lambda(f(X) - \mathbb{E}f(X))}\right] \leq e^{\frac{\lambda^2}{8}\|L\|^2}$. Apply Chernoff.

□

Recall if $X_1, \dots, X_n$ are independent σ-sub Gaussian with $\mathbb{E}X_i = \mu$, the Hoeffding implies that $\hat{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ satisfies

$$\mathbb{P}\left[|\hat{x} - \mu| \le t \right] \ge 1 - 2\exp\left(\frac{-nt^2}{2\sigma^2}\right)$$

or equivalently

$$\mathbb{P}\left[|\hat{x} - \mu| \le \sqrt{\frac{2\sigma^2 \ln\left(\frac{2}{\rho}\right)}{n}} \right] \ge 1 - \rho.$$

Can one achieve a similar guarantee without sub Gaussian assumption with a different estimator $\hat{x}$?

Answer: yes, almost!

Thm: (Median of means)

Consider $X \in \mathbb{R}$ with $\mathbb{E}X = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Let $x_1, \dots, x_n$ be i.i.d realizations of X. Subdivide into $k = 18 \log\left(\frac{1}{p}\right)$ bins and form the empirical means $\hat{x}_j$ for $j = 1, \dots, k$. Then $\hat{x} = \mathrm{median}(\hat{x}_1, \dots, \hat{x}_k)$ satisfies

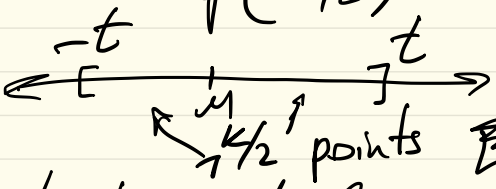
$$\mathbb{P}\left[|\hat{x} - \mu| \le \sqrt{\frac{27\sigma^2 \log\left(\frac{1}{p}\right)}{n}} \right] \ge 1 - p$$

pf: By Chebychev

$$\mathbb{P}\left[|\hat{X}_i - \mu| \ge \sqrt{\frac{36\sigma^2 k}{2n}} \right] \le \frac{\sigma^2}{\frac{36\sigma^2 k}{2n}} = \frac{2}{3} \quad \forall i.$$

Let $\mathbb{1}_i$ be indicator of this event. Then by Hoeffding,

$$\mathbb{P}\left[\frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_i > \frac{1}{2} \right] \ge 1 - \exp\left(-\frac{k}{18}\right)$$



in this event,

$$\Rightarrow |\hat{x} - \mu| \le \sqrt{\frac{36\sigma^2 k}{2n}}$$

Notice that in contrast to sub-Gaussian case, $\hat{x}$ depends on confidence level p.

Chapter 2

Random Vectors in High Dimensions
- Concentration of the norm
- Isotropy
- Similarity of Normal and Spherical
- Sub-Gaussian and Subexponential random vectors.

Two main results we'll prove:
1) Sub-Gaussian vectors concentrate around a sphere.
2) Two independent isotropic subGaussian random vectors are nearly orthogonal in high dimensions.

We next investigate the behavior of random vectors in high dimensions

Concentration of the ℓ_2 norm

Let $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ have independent σ-sub Gaussian coordinates with

$$\mathbb{E} X_i = 0 \quad \text{and} \quad \mathbb{E} X_i^2 = 1$$

What can we expect from

$$\|X\|_2^2 \quad \text{and} \quad \|X\|_2$$

Lemma: Suppose y is σ-sub Gaussian. Then y^2 is $(\sigma^2, 4\sigma^2)$ subexponential

pf sketch:

Step 1: Estimate $\mathbb{E}[|y|^r] \le r\, 2^{r/2} \sigma^r \Gamma(\frac{r}{2})$

using $\mathbb{E}[|y|^r] = \int_0^\infty \mathbb{P}[|y| > t^{1/r}]\, dt$

Step 2: Use Taylor Expansion

$$\mathbb{E}[e^{\lambda(y^2 - \mathbb{E}y^2)}] \le 1 + \sum_{r=2}^{\infty} \lambda^r 2^{r+1} \sigma^{2r} \le 1 + \frac{8\lambda^2\sigma^4}{1 - 2\lambda\sigma^2}$$

$$\le \exp(\dots) \quad \square$$

Cor: Let $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ have independent σ-sub Gaussian coordinates with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$

Then $\mathbb{P}\left[\left|\|X\|_2^2 - d\right| \geq td\right] \leq 2\exp\left(-\frac{d}{4\sigma^2}(t \wedge t^2)\right)$

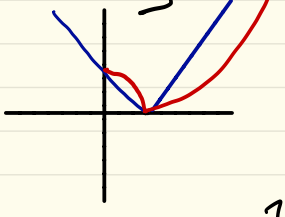
$$\mathbb{P}\left[\left|\|X\|_2 - \sqrt{d}\right| \geq t\sqrt{d}\right] \leq 2\exp\left(\frac{-dt^2}{4\sigma^2}\right)$$

pf: $\|X\|_2^2 = \sum_{i=1}^{d} X_i^2 \Rightarrow \mathbb{E}\|X\|_2^2 = d$

$\|X\|_2^2$ is $(\sigma\sqrt{d}, 4\sigma^2)$ subexponential.

Bernstein $\Rightarrow$

$$\mathbb{P}\left[\left|\frac{1}{d}\|X\|_2^2 - 1\right| \geq t\right] \leq 2\exp\left[-\frac{d}{4\sigma^2}(t \wedge t^2)\right]$$



Observation: for any $z \geq 0$, have

$|z-1| \geq t \Rightarrow |z^2 - 1| \geq t \vee t^2$.

So $\mathbb{P}\left[\left|\frac{1}{\sqrt{d}}\|X\|_2 - 1\right| \geq t\right] \leq \mathbb{P}\left[\left|\frac{1}{d}\|X\|_2^2 - 1\right| \geq t \vee t^2\right]$

$$\leq 2\exp\left(\frac{-dt^2}{4\sigma^2}\right) \quad \square$$

Isotropic Vectors

Recall for $X \in \mathbb{R}^d$, covariance

$$\mathrm{cov}(X) = \mathbb{E}((X-\mu)(X-\mu)^T)$$

where $\mu = \mathbb{E}X$.

Defn: A random vector $X \in \mathbb{R}^d$ with $\mathbb{E}X = 0$ is isotropic if

$$\Sigma(X) := \mathbb{E}XX^T = I_d$$

Remark: If $\Sigma = \Sigma(X)$ is invertible, then

$$Z := \Sigma^{-1/2}(X-\mu) \text{ is isotropic.}$$

Lemma: X is isotropic iff

$$\mathbb{E}\langle X, y\rangle^2 = \|y\|_2^2 \quad \forall y \in \mathbb{R}^d$$

pf: X is isotropic iff $\mathbb{E}XX^T = I$

iff $y^T \mathbb{E}XX^T y = y^T y$

iff $\mathbb{E}y^T XX^T y = \|y\|_2^2$

iff $\mathbb{E}\langle X, y\rangle^2 = \|y\|_2^2$ ☐

Thus if $\mathbb{E}X=0$, then X isotropic iff. marginal $\langle X, \frac{y}{\|y\|}\rangle$ has unit variance $\forall\, y \in \mathbb{R}^d$

Lemma: Let $X \in \mathbb{R}^d$ be isotropic. Then

$$\mathbb{E}\|X\|_2^2 = d.$$

Moreover, if X and Y are two independent isotropic vectors, then

$$\mathbb{E}\langle X, Y\rangle^2 = d$$

pf: First

$$\|X\|_2^2 = X^T X = \text{trace}(XX^T)$$

$$\Rightarrow \mathbb{E}\|X\|_2^2 = \text{trace}(I_d) = d.$$

Next $\mathbb{E}\langle X, Y\rangle^2 = \mathbb{E}_Y\left[\mathbb{E}_X \langle X, Y\rangle^2 \mid Y\right]$

$$= \mathbb{E}_Y \|Y\|_2^2 = d \qquad \square$$

Let X and Y be independent and isotropic

Then $\|X\| \sim \sqrt{d}$

$\|Y\| \sim \sqrt{d}$

$\therefore$ Almost orthogonal.

and $\left\langle \frac{X}{\|X\|}, \frac{Y}{\|Y\|} \right\rangle \sim \frac{1}{\sqrt{d}}$.

Can be made rigorous by assuming light tails.

Examples of isotropic RV:

1) Spherical $X \sim \mathrm{Unif}(\sqrt{d}\, S^{d-1})$ HW

2) Symmetric Bernoulli: $X \sim \mathrm{Unif}(\{-1,1\}^d)$

3) Any vector $X = (X_1, \dots, X_d)$, where X_i are independent, zero mean, unit variance.

4) Coordinate $\mathrm{Unif}(\{\sqrt{d}\, e_i\}_{i=1}^d)$

5) Gaussian $g = (g_1, \dots, g_d) \sim N(0, I_d)$

Recall this means g_i are i.i.d. $N(0,1)$

$\Rightarrow$ Density

$$p(x) = \prod_{i=1}^d p_i(x) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|x\|_2^2/2}.$$

$\Rightarrow N(0, I_d)$ is rotation invariant.

Exercise:

Let $g \sim N(0, I_d)$. Then

$$r := \|g\|_2 \quad \text{and} \quad \theta = \frac{g}{\|g\|_2}$$

are independent random variables and

$$\theta \sim \mathrm{Unif}(S^{d-1})$$

Defn: X in $\mathbb{R}^d$ is σ-subGaussian if $\langle X, u \rangle$ is σ-subGaussian $\forall u \in S^{d-1}$

Ex: Let $X = (X_1, \ldots, X_d)$ be RV with independent σ-subGaussian X_i. Then X is σ-subGaussian.

Ex: 1) $N(0, I_d)$ is 1-subGaussian.
2) $\mathrm{Unif}(\{-1, 1\}^d)$ is 1-subGaussian
3) $\mathrm{Unif}(\{\sqrt{d}\, e_i\}_{i=1}^d)$ is σ-subGaussian with $\sigma \asymp \sqrt{\frac{d}{\log(d)}}$

Way too big to be useful

4) $\mathrm{Unif}(\sqrt{d}\, S^{d-1})$ is c-subGaussian for a constant c.

Q: How to get high probability bound on $\|X - \mathbb{E}X\|$?

Idea: $\|\cdot\| = \sup_{y \in B^*} \langle x - \mathbb{E}X, y \rangle$

Chapter 3

Uniform Laws:

- Glivenko-Cantelli
- sample complexity of learning
- uniform law w\ Rademacher Complexity
- Upper Bounds on Rademacher Complexity:
  - Massart Lemma
  - Finite sample Glivenko-Cantelli
  - VC dimension
  - Dimension Independent generalization bounds
    - linear models
    - convex losses

Central Set-up:

Let $\mathcal{F}$ be a family of integrable functions on some probability space $(X, \Sigma, \mathbb{P})$.

Goal: Estimate

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathop{\mathbb{E}}_{x \sim \mathbb{P}} f(x) \right|$$

where $x_1, \ldots, x_n \overset{iid}{\sim} \mathbb{P}$.

Ex: (Emperical CDF)

Suppose we want to estimate the CDF

$$g(t) := \mathbb{P}[X \leq t].$$

We can use the empirical CDF

$$g_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, t]}(x_i)$$

Want to control

$$\|g_n - g\|_\infty := \sup_t |g_n(t) - g(t)|$$

This is exactly

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right|$$

where $\mathcal{F} = \{ \mathbb{1}_{(-\infty, t]}(\cdot) : t \in \mathbb{R} \}$.

Thm: (Glivenko-Cantelli) For any distribution

$$\| g_n - g \|_\infty \xrightarrow{a.s.} 0$$

[We'll prove a finite sample version soon]

Cor (Plug in Estimator) [HW]
Let $\gamma(\cdot)$ be a $\| \cdot \|_\infty$-continuous functional of CDF's. Then

$$\gamma(g_n) \to \gamma(g) \text{ in probability}$$

Ex: γ quantiles, median, mean
uniform distance to hypothesis' CDF.

Ex: (Statistical Learning)

Consider a family of probability measures

$$\{P_\theta : \theta \in \Omega\}$$

where Ω is some set.

Suppose we are given

$$X_1, \dots, X_n \overset{iid}{\sim} P_{\theta^*}$$

for some θ^* [may not be in Ω]

Fix a "goodness of fit" function

$$\theta \mapsto L_\theta(X)$$

Goal is to minimize <u>population risk</u>

$$\min_{\theta \in \Omega} R(\theta) := \mathbb{E}_{X \sim P_{\theta^*}}[L_\theta(X)]$$

One approach is to instead minimize the <u>Emperical Risk</u>:

$$\min_{\theta \in \Omega} R_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} L_\theta(X_i) \quad (\star)$$

Goal: Bound the excess risk

$$E(\hat{\theta}_n) := R(\hat{\theta}_n) - \inf_{\theta \in \Omega} R(\theta)$$

where $\hat{\theta}_n$ is the minimizer of ⊗

Ex: (Maximum Likelyhood)

Let p_θ be the density of P_θ

Define

$$L_\theta(x) = \log\left(\frac{p_{\theta^*}(x)}{p_\theta(x)}\right)$$

Then

$$\hat{\theta}_n \in \operatorname{argmax}_{\theta \in \Omega} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log(p_\theta(X_i)) \right\}$$

Ex: (Binary Classification)

Data $(X_i, y_i) \in \mathbb{R}^d \times \{-1, +1\}$ according to P.

We want to estimate $f: \mathbb{R}^d \to \{-1, +1\}$ solving

$$\min_{f \in T} P_{(X,y)}[f(X) \neq y] = \min_{f \in T} \underbrace{E_{(X,y) \sim P} \mathbb{1}\{f(X) \neq y\}}_{R(f)}$$

T: some function class

Try instead to solve

$$\min_{f \in T} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[f(x_i) \neq y_i]}_{R_n(f)}$$

Decomposition of Access Risk

$$E(\hat{\theta}_n, \theta^*) = R(\hat{\theta}_n) - \inf_{\theta \in \Omega} R(\theta)$$

$$= R(\hat{\theta}_n) - R_n(\hat{\theta}_n)$$

$$+ R_n(\hat{\theta}_n) - R_n(\theta_0)$$

$$+ R_n(\theta_0) - R(\theta_0)$$

where $\theta_0 = \operatorname{argmin}_{\theta \in \Omega} R(\theta)$

is approximation error. How to control?

is optimization error. If $\hat{\theta}_n$ is true minimizer of $R_n(\theta)$, then ≤ 0

easy to control approximation error:

$$\mathbb{E}_{\theta^*} R_n(\theta_0) = R(\theta_0)$$

$= R(\hat{\theta}_n) - R_n(\hat{\theta}_n)$

Main Difficulty: $\hat{\theta}_n$ is not fixed!

But we can estimate

$$\leq \sup_{\theta \in \Omega} |R_n(\theta) - R(\theta)|$$

$$= \sup_{\theta \in \Omega} \left| \frac{1}{n} \sum_{i=1}^{n} L_\theta(X_i) - \mathbb{E}_{X \sim P_\theta} L_\theta(X) \right|$$

So

$$\mathbb{E}(\hat{\theta}_n, \theta^*) \leq 2 \| P_n - P \|_{\mathcal{F}}$$

where $\mathcal{F} := \{ x \mapsto L_\theta(x) : \theta \in \Omega \}$

Remark:
- Function class $\mathcal{F}$ s.t. $\| P_n - P \|_{\mathcal{F}} \to 0$ in probability is called Glivenko-Cantelli
- If $\mathcal{F}$ is too big it may not be Glivenko-Cantelli (see Ex 4.7 in W)

Our first approach will be based on Rademacher Complexity.

Defn:

The Rademacher Complexity of a set $A \subseteq \mathbb{R}^n$, is the quantity

$$\mathcal{R}(A) := \mathbb{E}_\varepsilon \sup_{a \in A} \langle a, \varepsilon \rangle$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)$ are i.i.d. Rademacher random variables

Recall that support function of a set A is

$$\sigma_A(v) := \sup_{a \in A} \langle v, a \rangle$$

So

$$\mathcal{R}(A) = \mathbb{E}_\varepsilon \sigma_A(\varepsilon)$$

Defn: Consider a sequence of random independent $(X_1, \dots, X_n) \subseteq \mathcal{X}^n$ and a class of functions $\mathcal{F}$ on $\mathcal{X}$. The **Rademacher complexity** of $\mathcal{F}$ is

$$R_n(\mathcal{F}) := \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right|$$

$$= \mathbb{E}_X R(\mathcal{F}^I(X)/n)$$

$$\leq 2 \mathbb{E}_X R(\mathcal{F}(X)/n)$$

where $\mathcal{F}(X) = \{(f(X_1), \dots, f(X_n)) : f \in \mathcal{F}\}$

Thm: $\mathbb{E} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2 R_n(\mathcal{F})$

pf: Let $Y_1, \dots, Y_n \overset{iid}{\sim} \mathbb{P}$ and independent of $X_1, \dots, X_n$. Then

$$\mathbb{E} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)] \right| \right]$$

$$= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}_Y \left[\frac{1}{n} \sum (f(X_i) - f(Y_i)) \right] \right| \right]$$

$$= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}_Y \left[\frac{1}{n} \sum (f(x_i) - f(y_i)) \right] \right| \right]$$

$$\leq \mathbb{E}_X \mathbb{E}_Y \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum (f(x_i) - f(y_i)) \right|$$

$$= \mathbb{E}_{X,Y,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f(x_i) - f(y_i)) \right|$$

$$\leq 2 \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right| = 2 \mathcal{R}_n(\mathcal{F})$$ ∎

Thm: Suppose $\mathcal{F}$ is b-uniformly bounded, meaning

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)| \leq b \quad \forall f \in \mathcal{F}.$$

Then for any $n \in \mathbb{N}$, $t \geq 0$, it holds

$$\Pr\left(\|P_n - P\|_{\mathcal{F}} \leq 2 \mathcal{R}_n(\mathcal{F}) + t \right) \geq 1 - e^{-\frac{nt^2}{2b^2}}$$

pf: All we have to do is show the bounded difference property for the function $\|P_n - P\|_{\mathcal{F}}$ with $L_i = \frac{2b}{n}$.

Define $\bar{f}(x) = f(x) - Ef(x)$.

Then $\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \bar{f}(X_i) \right|$

As before, let x' differ from x only in i'th entry. Then

$$\left| \frac{1}{n} \sum_{i=1}^{n} \bar{f}(x_i) \right| - \sup_{g \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \bar{g}(x_i') \right|$$

$$\leq \left| \frac{1}{n} \sum_{i=1}^{n} \bar{f}(x_i) \right| - \left| \frac{1}{n} \sum_{i=1}^{n} \bar{f}(x_i') \right| \leq \frac{2b}{n}$$

Take $\sup_{f \in \mathcal{F}}$ and exchange x, x' ☑

Rademacher complexity characterizes the assymptotics of $\|P_n - P\|_{\mathcal{F}}$.

Prop $\frac{1}{2}\mathcal{R}_n(\bar{\mathcal{F}}) \le \mathbb{E}_X\left[\|P_n - P\|_{\mathcal{F}}\right] \le 2\mathcal{R}_n(\mathcal{F})$

where $\bar{\mathcal{F}} = \{f - \mathbb{E}f : f \in \mathcal{F}\}$.

[See Prop 4.1.1 in Wainwright]

Prop (HW) Suppose $\mathcal{F}$ is b-bounded. Then $\forall n \in \mathbb{N}, t > 0$, it holds

$$\Pr\left(\|P_n - P\|_{\mathcal{F}} \ge \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}} |\mathbb{E}f(x)|}{2\sqrt{n}} - t\right)$$

$$\ge 1 - e^{-\frac{nt^2}{2b^2}}$$

Goal: Bound the Rademacher Complexity of interesting sets.

Lemma (Basic Calculus)
For any $A, B \subseteq \mathbb{R}^n$, $c \in \mathbb{R}$, $a_0 \in \mathbb{R}^n$ it holds

- $\mathcal{R}(cA) = |c| \mathcal{R}(A)$
- $\mathcal{R}(A+B) \leq \mathcal{R}(A) + \mathcal{R}(B)$
  with equality if A and B are convex
- $\mathcal{R}(A+a_0) = \mathcal{R}(A)$
- $\mathcal{R}(A) = \mathcal{R}(\text{conv}(A))$

Lemma: (Massart) Let $A = \{a_1, \dots, a_N\} \subseteq \mathbb{R}^n$.
Define $\bar{a} = \frac{1}{N} \sum_{i=1}^{N} a_i$. Then

$$\mathcal{R}(A) \leq \max_{a \in A} \|a - \bar{a}\| \cdot \sqrt{2 \log(N)}$$

pf: WLOG assume $\bar{a}=0$. Let $\lambda>0$ and $A'=\{\lambda a_1,\dots,\lambda a_N\}$. Then

$$R(A') = \mathbb{E}_\varepsilon \max_{a\in A'} \langle \varepsilon, a\rangle$$

$$= \mathbb{E}_\varepsilon \log\left(\max_{a\in A'} e^{\langle \varepsilon, a\rangle}\right)$$

$$\le \mathbb{E}_\varepsilon \log\left(\sum_{a\in A'} e^{\langle \varepsilon, a\rangle}\right)$$

Jensen

$$\le \log\left(\sum_{a\in A'} \mathbb{E}_\varepsilon e^{\langle \varepsilon, a\rangle}\right)$$

Observe

$$\mathbb{E}_\varepsilon e^{\langle \varepsilon, a\rangle} = \prod_{i=1}^{d} \mathbb{E} e^{\varepsilon_i a_i} \le \prod_{i=1} e^{\frac{a_i^2}{2}} = e^{\|a\|_2^2/2}$$

$$\le \log\left(\sum_{a\in A'} e^{\|a\|_2^2/2}\right)$$

$$\le \log\left(|A'| \max_{a\in A'} \left(e^{\|a\|_2^2/2}\right)\right)$$

$$= \log(|A'|) + \max_{a\in A'} \frac{\|a\|_2^2}{2}$$

Thus

$$\mathcal{R}(A) = \frac{1}{\lambda}\mathcal{R}(A') \le \frac{\log(|A|) + \max_{a\in A} \frac{\lambda^2 \|a\|^2}{2}}{\lambda}$$

Optimize over $\lambda > 0$. ∎

Defn: $\mathcal{F}$ has <u>polynomial discrimination of order $v \ge 1$</u> if $\forall n \in \mathbb{N}$ and $x_1, \dots, x_n \in \mathcal{X}$, we have

$$\text{card}\left(\mathcal{F}(x_1, \dots, x_n)\right) \le (n+1)^v.$$

Cor: Suppose $\mathcal{F}$ has polynomial discrimination of order v. Then

$$\mathcal{R}_n(\mathcal{F}) \le 4\, \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}\right] \cdot \sqrt{\frac{v \log(n+1)}{n}}$$

pf: Fix $x = (x_1, \dots, x_n)$. Then Massart implies

$$\mathcal{R}(\mathcal{F}(x)/n) \le 4 \cdot \frac{\max_{f\in\mathcal{F}} \sqrt{\sum_{i=1}^n f_i^2(x)}}{n} \cdot \sqrt{v \log(n+1)}$$

∎

In particular, if $\mathcal{F}$ is b-bounded, then

$$\boxed{\mathcal{R}_n(\mathcal{F}) \le 4b\sqrt{\frac{v\log(n+1)}{n}}}$$

Cor: (Glivenko-Cantelli)

Let $g(t) = \mathbb{P}[X \le t]$ be the CDF of X, and let $g_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \le t]$ where $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathbb{P}$. Then

$$\mathbb{P}\left[\|g_n - g\|_\infty \ge 8\sqrt{\frac{\log(n+1)}{n}} + \delta\right] \le \exp\left(-\frac{n\delta^2}{2}\right)$$

pf: Let $\mathcal{F} = \{\mathbb{1}_{(-\infty, t]} : t \in \mathbb{R}\}$.

Then

$$\|g_n - g\|_\infty = \sup_{t\in\mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \le t] - \mathbb{P}[X \le t]\right|$$

$$= \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)\right|$$

Since $\mathcal{F}$ is 1-bounded

$$\mathbb{P}\left[\|g_n - g\|_\infty \le 2\mathcal{R}_n(\mathcal{F}) + \delta\right] \ge 1 - \exp\left(-\frac{n\delta^2}{2}\right)$$

Observe $\mathcal{F}$ has poly discrimination with $v = 1$ ∎

Vapnik-Chervonenkis (VC) Theory

Method for bounding polynomial discrimination of $\{0,1\}$-valued $\mathcal{F}$.

Consider a class $\mathcal{F}$ of binary valued functions on $\mathcal{X}$.

Def: We say that $x=(x_1,\dots,x_n)$ is **shattered** by $\mathcal{F}$ if $\text{card}(\mathcal{F}(x))=2^n$.

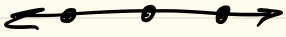
The **VC dimension** is

$$VC(\mathcal{F})=\sup\{n\in\mathbb{N}: \exists x\in\mathcal{X}^n \text{ shattered by } \mathcal{F}\}$$

Notation: If $\mathcal{F}=\{\mathbb{1}_S : S \text{ some sets}\}$ set $S(x):=\mathcal{F}(x)$ and $VC(S):=VC(\mathcal{F})$

Ex: $S_{left}=\{(-\infty,a]: a\in\mathbb{R}\} \Rightarrow VC(S_{left})=1$

$S_{two}=\{(a,b]: a,b\in\mathbb{R}, a<b\} \Rightarrow VC(S_{two})=2$



Thm (Sauer and Shelah)

For any $x = (x_1, \ldots, x_n)$ with $n \geq VC(S)$,

we have

$$\text{card}(S(x)) \leq \sum_{i=0}^{VC(S)} \binom{n}{i} \leq (n+1)^{VC(S)}$$

Therefore $\mathcal{F}$ has polynomial discrimination of order $VC(S)$ and

$$\mathcal{R}_n(\mathcal{F}) \leq 2\sqrt{\frac{VC(S) \cdot \log(n+1)}{n}}$$

$\log(n+1)$ can be removed.

[See Prop 4.18 in Wainwright for a proof.]

Examples: Let G be a class of functions.

For any $g: \mathcal{X} \to \mathbb{R}$ define

$$S_g = \{x \in \mathcal{X} \;\; g(x) \leq 0\}$$

$$S(G) = \{S_g : g \in G\}$$

Prop: Let G be a vector space of functions $g: \mathbb{R}^d \to \mathbb{R}$ with $\dim(G) < \infty$. Then

$$VC(S(G)) \leq \dim(G)$$

pf: Set $n = \dim(G) + 1$ and fix $x = \{x_1, \ldots, x_n\}$ with $x_i \in \mathcal{X}$. Define $L: G \to \mathbb{R}^n$ by

$$L(g) = (g(x_1), \ldots, g(x_n))$$

Since $n > \dim(G)$, there exists $0 \neq \gamma \in \mathbb{R}^n$ s.t. $\langle \gamma, L(g) \rangle = 0 \quad \forall g \in G$.

$$\Rightarrow \sum_{\{i : \gamma_i \leq 0\}} (-\gamma_i) g(x_i) = \sum_{\{i : \gamma_i > 0\}} \gamma_i g(x_i) \quad \forall g \in G$$

WLOG suppose $\gamma_i > 0$ for some i.

Suppose there were $g \in G$ such that S_g includes only $\{x_i : \gamma_i \leq 0\}$. Then

$$0 \geq LHS = RHS > 0$$

Contradiction ∎

Ex: (Halfspaces)

Define $S_{a,b} = \{x \in \mathbb{R}^d : \langle a, x \rangle + b \leq 0\}$

$S = \{S_{a,b} : a, b \in \mathbb{R}\}$

$G = \{x \mapsto \langle a, x \rangle + b : a, b \in \mathbb{R}\}$

Then

$VC(S) \leq \dim(G) = d+1.$

actually equality

Ex: (Balls)

Define $S_{a,b} = \{x \in \mathbb{R}^d : \|x-a\|_2 \leq b\}$

$S = \{S_{a,b} : a \in \mathbb{R}^d, b \geq 0\}$

Define $g_{a,b}(x) = \|x\|_2^2 - 2\langle a, x \rangle + \|a\|^2 - b^2$

$= \|x-a\|^2 - b^2$

Trick: Define $\Phi : \mathbb{R}^d \to \mathbb{R}^{d+2}$ by

$\Phi(x) = (1, x_1, \ldots, x_d, \|x\|_2^2)$

$g_c(x) = \langle c, \Phi(x) \rangle$ where $c \in \mathbb{R}^{d+2}$

Then

$$\left\{ g_{a,b} : \begin{matrix} a \in \mathbb{R} \\ b \geq 0 \end{matrix} \right\} \subseteq \left\{ g_c : c \in \mathbb{R}^d \right\}$$

vector space of dimension $d+2$.

$\Rightarrow VC(S) \leq d+2$

[Exact bound is $d+1$: harder to prove]

Rademacher Complexity and VC-dim often scale with the dimension of the ambient space $x \in \mathcal{X}$.

Thm: Consider $\min_{w \in W \subseteq \mathbb{R}^d} f(w) = \mathbb{E}_z f(w,z)$ where $\max_{w \in W} \|w\| \leq B$ and $f(\cdot, z)$ is L-Lipschitz

Then $\mathbb{E}\left[\sup_{w \in W} \left| \frac{1}{n}\sum_{i=1}^n f(w,z_i) - \underbrace{\mathbb{E} f(w,z)}_{\leq f(w)} \right|\right] \leq O\left(\sqrt{\frac{L^2 B^2 d \log(n)}{n}}\right)$ ← tight

[We'll prove this later!]

Dimension Independent Bound for Generalization

- linear models
- convexity

Linear Models:

Consider the problem

$$\min_{w \in W} \mathbb{E}_{(a,b)\sim P} \ell(\langle w, a \rangle, b)$$

To get generalization bounds we need to compute $\mathcal{R}_n(F)$ where

$$F = \{(a,b) \mapsto \ell(\langle w, a \rangle, b) : w \in W\}$$

Lemma: (Contraction)

Consider a set $A \subseteq \mathbb{R}^n$ and let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function.

Let $A' = \{(\phi_1(a_1), \dots, \phi_n(a_n)) : a \in A\}$

Then $$\mathcal{R}(A') \leq \rho \mathcal{R}(A)$$

pf: WLOG, assume $\rho = 1$. It suffices to assume $A' = \{(\phi_1(a_1), a_2, \dots, a_n) : a \in A\}$

Then

$$\mathcal{R}(A') = \mathbb{E}_\varepsilon\left[\sup_{a \in A'} \sum_{i=1}^n \varepsilon_i a_i\right]$$

$$= \frac{1}{2}\mathbb{E}_{\varepsilon_2,\dots,\varepsilon_n}\left[\sup_{a \in A}\left\{\phi_1(a_1) + \sum_{i=2}^n \varepsilon_i a_i\right\} + \sup_{\hat{a} \in A}\left\{-\phi_1(\hat{a}_1) + \sum_{i=2}^n \varepsilon_i \hat{a}_i\right\}\right]$$

$$= \frac{1}{2}\mathbb{E}_{\varepsilon_2,\dots,\varepsilon_n}\left[\sup_{a, \hat{a} \in A} \phi_1(a_1) - \phi_1(\hat{a}_1) + \sum_{i=2}^n \varepsilon_i (a_i + \hat{a}_i)\right]$$

$$\leq \frac{1}{2}\mathbb{E}_{\varepsilon_2,\dots,\varepsilon_n}\left[\sup_{a, \hat{a} \in A} a_1 - \hat{a}_1 + \sum_{i=2}^n \varepsilon_i (a_i + \hat{a}_i)\right]$$

$$= \mathcal{R}(A)$$

∎

Rademacher Complexity of linear classes

Lemma: Consider a set of vectors $x_1, \dots, x_n \in \mathbb{R}^d$ and define

$$A = \left\{ (\langle w, x_1 \rangle, \dots, \langle w, x_n \rangle) : \|w\|_2 \le 1 \right\}$$

Then $$\mathcal{R}(A) \le \sqrt{\sum_{i=1}^{n} \|x_i\|_2^2}$$

and therefore

$$\mathcal{R}_n\left(\left\{ x \mapsto \langle w, x \rangle : \|w\|_2 \le 1 \right\} \right) \le 2 \frac{\mathbb{E} \max_{i=1,\dots,n} \|x_i\|_2}{\sqrt{n}}$$

Pf:

$$\begin{aligned}
\mathcal{R}(A) &= \mathbb{E}_\varepsilon \sup_{a \in A} \sum_{i=1}^{m} \varepsilon_i a_i \\
&= \mathbb{E}_\varepsilon \sup_{\|w\|_2 \le 1} \sum_{i=1}^{m} \varepsilon_i \langle w, x_i \rangle \\
&= \mathbb{E}_\varepsilon \sup_{\|w\|_2 \le 1} \left\langle w, \sum_{i=1}^{m} \varepsilon_i x_i \right\rangle \\
&= \mathbb{E}_\varepsilon \left\| \sum_{i=1}^{m} \varepsilon_i x_i \right\|_2 \\
&\le \sqrt{\mathbb{E}_\varepsilon \left\| \sum_{i=1}^{m} \varepsilon_i x_i \right\|_2^2} \le \sqrt{\sum_{i=1}^{m} \|x_i\|_2^2}
\end{aligned}$$

□

Lemma: Consider a set of vectors $x_1, \dots, x_m \in \mathbb{R}^d$ and define

$$A = \{(\langle w, x_1 \rangle, \dots, \langle w, x_m \rangle) : \|w\|_1 \le 1\}$$

Then

$$R(A) \le \sqrt{2n \log(2d)} \cdot \max_{i=1,\dots,n} \|x_i\|_\infty$$

and therefore

$$R_n(\{x \mapsto \langle w, x \rangle : \|w\|_1 \le 1\}) \le \sqrt{8 \ln(d)} \cdot \frac{\mathbb{E} \max_{i=1,\dots,n} \|x_i\|_\infty}{\sqrt{n}}$$

Pf:

$$\begin{aligned} R(A) &= \mathbb{E}_\varepsilon \sup_{a \in A} \sum_{i=1}^m \varepsilon_i a_i \\ &= \mathbb{E}_\varepsilon \sup_{\|w\|_1 \le 1} \sum_{i=1}^m \varepsilon_i \langle w, x_i \rangle \\ &= \mathbb{E}_\varepsilon \sup_{\|w\|_1 \le 1} \left\langle w, \sum_{i=1}^m \varepsilon_i x_i \right\rangle \\ &= \mathbb{E}_\varepsilon \left\| \sum_{i=1}^m \varepsilon_i x_i \right\|_\infty \\ &= \mathbb{E}_\varepsilon \sup_{v \in \{\pm e_j\}_{j=1}^d} \sum_{i=1}^m \varepsilon_i \langle x_i, v \rangle \end{aligned}$$

$$\mathbb{E}_\varepsilon \sup_{v \in \{\pm e_j\}_{j=1}^d} \sum_{i=1}^n \varepsilon_i \langle x_i, v \rangle$$

$$= \mathcal{R}\left((\langle x_1, v\rangle, \dots, \langle x_n, v\rangle) : v \in \{\pm e_j\} \right)$$

$$\le \sqrt{2 \log(2d)} \cdot \max_{v \in \{\pm e_j\}} \| (\langle x_1, v\rangle, \dots, \langle x_n, v\rangle) \|_2$$

$$\le \sqrt{2n \log(2d)} \cdot \max_{i=1,\dots,n} \|x_i\|_\infty \quad \square$$

Back to $\min_{w \in W} \mathbb{E}_{(a,b)\sim P} \ell(\langle w, a\rangle, b)$ where $W = B_2$ or B_1

So if $\ell(\cdot, b)$ is ρ-Lipschitz $\forall b$, then the Rademacher bounds are

ℓ_2-case: $\rho \cdot \dfrac{\mathbb{E} \max_{i=1,\dots,n} \|a\|_2}{\sqrt{n}}$

ℓ_1-case: $\rho \cdot \sqrt{\log(d)} \cdot \dfrac{\mathbb{E} \max_{i=1,\dots,n} \|a_i\|_\infty}{\sqrt{n}}$

Convexity: Regret without uniform laws

Suppose we want to solve

$$\text{(*)} \quad \min_{w \in W} f(w) = \mathbb{E}_{x \sim P} f(w, x)$$

Let $S = \{x_1, \ldots, x_n\}$ be i.i.d from P, and let $A(S)$ be an output of an algorithm which aims to approximately solve (*). Let $S^i = (x_1, \ldots, x_{i-1}, x', x_{i+1}, \ldots, x_n)$ where $x' \sim P$ independent of $x_1, \ldots, x_n$

Which $A(S)$ generalizes?

Intuition: $f(A(S^i), x_i) - f(A(S), x_i)$ should not be big, otherwise overfitting

Thm:

$$\mathbb{E}_S \left[f(A(S)) - \frac{1}{n} \sum_{i=1}^{n} f(A(S), x_i) \right]$$

$$= \mathop{\mathbb{E}}_{\substack{(S, x') \sim P^{n+1} \\ i \sim U(m)}} \left[f(A(S^i), x_i) - f(A(S), x_i) \right]$$

Pf: For every i,

$$\mathbb{E}_S f(A(S)) = \mathbb{E}_{S,z'} f(A(S), x') = \mathbb{E}_{S,z'} f(A(S^i), x_i)$$

Observe

$$\mathbb{E}_S \frac{1}{n}\sum_{i=1}^n f(A(S), x_i) = \mathbb{E}_{S,i}\left[f(A(S), x_i)\right]$$

Defn: $A(\cdot)$ is *leave-one-out stable* with rate $\varepsilon(n)$ if

$$\mathbb{E}_{(S,x')\sim P^{n+1}, i\sim U(n)}\left[f(A(S^i), x_i) - f(A(S), x_i)\right] \le \varepsilon(n).$$

Henceforth, fix $\lambda > 0$ and we'll analyze

$$A(S) := \arg\min_{w\in W} \frac{1}{n}\sum_{i=1}^n f(w, x_i) + \frac{\lambda}{2}\|w\|^2.$$

Also suppose W and $f(\cdot, x_i)$ are convex.

Defn: A function $g: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is α-strongly convex if $g - \frac{\alpha}{2}\|\cdot\|^2$ is convex.

Lemma: If g is α-strongly convex, then it has a unique minimizer $\bar{w}$ and

$$g(w) - g(\bar{w}) \geq \frac{\alpha}{2}\|w - \bar{w}\|^2 \quad \forall w$$

Thm: Suppose W is convex, and $f(\cdot; x)$ is convex and ρ-Lipschitz. Then the rule

$$A(S) = \operatorname*{argmin}_{w \in W} \frac{1}{n}\sum_{i=1}^{n} f(w, x_i) + \frac{\lambda}{2}\|w\|^2$$

is leave-one-out-stable with rate $\frac{2\rho^2}{\lambda n}$

Therefore

$$\mathbb{E}_S\left[f(A(S)) - \frac{1}{n}\sum_{i=1}^{n} f(A(S), x_i)\right] \leq \frac{2\rho^2}{\lambda n}$$

pf: Define $f_S(w) := \frac{1}{n}\sum_{i=1}^{n} f(w, x_i) + \frac{\lambda}{2}\|w\|^2$.

$\Rightarrow f_S(w) - f_S(A(S)) \geq \frac{\lambda}{2}\|w - A(S)\|^2 \quad \forall w$

For all $w, v \in \mathcal{W}$ it holds

$$f_S(w) - f_S(v) = \frac{1}{n}\sum_{x \in S} f(w,x) + \frac{\lambda}{2}\|w\|^2 - \frac{1}{n}\sum_{x \in S} f(v,x) - \frac{\lambda}{2}\|v\|^2$$

$$= \frac{1}{n}\sum_{x \in S^i} f(w,x) + \frac{\lambda}{2}\|w\|^2 - \frac{1}{n}\sum_{x \in S^i} f(v,x) - \frac{\lambda}{2}\|v\|^2 + \frac{f(w,x_i) - f(v,x_i)}{n} + \frac{f(v,x') - f(w,x')}{n}$$

Setting $w = A(S^i)$, $v = A(S)$ get

$$f_S(A(S^i)) - f_S(A(S)) \leq f_{S^i}(A(S^i)) - f_{S^i}(A(S)) + \frac{2\rho\|A(S^i) - A(S)\|}{n}$$

$$\leq -\frac{\lambda}{2}\|A(S^i) - A(S)\|^2$$

$$\Rightarrow \|A(S^i) - A(S)\| \leq \frac{2\rho}{\lambda n} \quad \square$$

Cor: $\mathbb{E}_S f(A(S)) \le \min_W f + \frac{\lambda}{2}\|\bar{w}\|^2 + \frac{2\rho^2}{\lambda n}$

where $\bar{w}$ is any minimizer of f on W.

Therefore under optimal choice $\lambda = \sqrt{\frac{4\rho^2}{n\|\bar{w}\|^2}}$

get

$$\mathbb{E}_S(f(A(S)) \le \min f + 2\sqrt{\frac{\rho^2\|\bar{w}\|^2}{n}}$$

pf: $\mathbb{E}_S f(A(S)) = \mathbb{E}_S\left[\frac{1}{n}\sum_{x\in S} f(A(S), x)\right]$

$$+ \mathbb{E}_S\left[f(A(S)) - \frac{1}{n}\sum_{x\in S} f(A(S), x)\right]$$

$$\le \mathbb{E}_S\left[\frac{1}{n}\sum_{x\in S} f(A(S), x)\right] + \frac{2\rho^2}{\lambda n}$$

For any w, we have

$$\mathbb{E}_S\Big[\underbrace{\frac{1}{n}\sum_{x\in S} f(A(S), x) + \frac{\lambda}{2}\|A(S)\|^2}_{f_S(A(S))}\Big] \le \mathbb{E}_S f_S(w)$$

$$\le f(w) + \frac{\lambda}{2}\|w\|^2$$

$$\Rightarrow \mathbb{E}_S\left[\frac{1}{n}\sum_{z\in S} f(A(S), z)\right] \le \min f + \frac{\lambda}{2}\|\bar{w}\|^2$$

□

Chapter 4

Metric Entropy and Matrix Concentration

- Nets, covering numbers, and metric entropy
- Eigenvalues and Singular values of random matrices
- Matrix Concentration
- Operator norm vs. row/column norms
- Applications:
  - community detection,
  - (sparse) covariance matrix estimation
  - matrix completion

Let (T, ρ) be a metric space

Ex: • $(\mathbb{R}^d, \|\cdot\|_2)$ with $\rho(x,y) = \|x-y\|_2$

• $(\{0,1\}^d, \rho_H)$ with Hamming metric

$$\rho_H(\theta, \tilde{\theta}) = \frac{1}{d} \sum_{j=1}^{d} \mathbb{1}[\theta_j \neq \tilde{\theta}_j]$$

• $L^2(\mu, [0,1])$ with

$$\|f-g\|_2 = \sqrt{\int_0^1 (f(x)-g(x))^2 \, d\mu(x)}$$

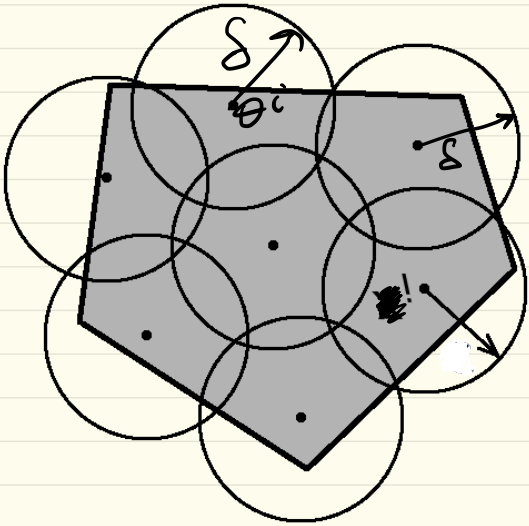
• $C[0,1]$ with

$$\|f-g\|_\infty = \sup_{x \in [0,1]} |f(x) - g(x)|$$

Defn: A δ-cover of a set T with respect to a metric ρ is a set $\{\theta^1, \ldots, \theta^N\} \subset T$ such that

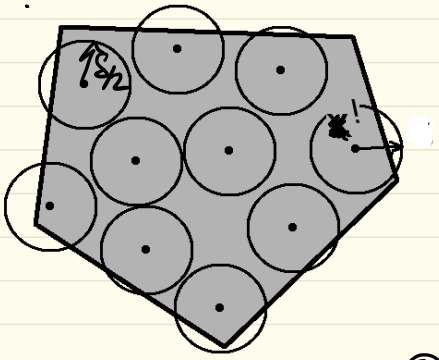
$\forall \theta \in T \; \exists i \in \{1, \ldots, N\}$ s.t. $\rho(\theta, \theta^i) \leq \delta$

The δ-covering number $N(\delta; T, \rho)$ is the cardinality of the smallest δ-cover. The quantity $\log(N(\delta; T, \rho))$ is called the metric entropy.



Defn: A δ-packing of T with respect to ρ is a set $\{\theta^1, \ldots, \theta^m\} \subseteq T$ such that $\rho(\theta^i, \theta^j) > \delta \quad \forall i,j \in \{1,2,\ldots,m\}$

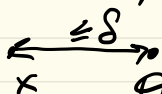
The δ-packing number $M(\delta; T, \rho)$ is the cardinality of the largest δ-packing.



Lemma:

$$M(2\delta; T, \rho) \overset{@}{\leq} N(\delta; T, \rho) \overset{ⓑ}{\leq} M(\delta; T, \rho)$$

Pf: ⓑ $M(\delta, T, \rho)$



$= M(\delta, T, \rho) \geq N(\delta, T, \rho)$

@ $M(2\delta, T, \rho)$

No δ Ball can cover θ^i, θ^j simultaneously. ☆

Lemma: Let $\|\cdot\|$, $\|\cdot\|'$ be norms on $\mathbb{R}^d$ and set

$$B = \{x : \|x\| \le 1\}$$

$$B' = \{x : \|x\|' \le 1\}$$

Then

$$\left(\frac{1}{\delta}\right)^d \frac{\mathrm{vol}(B)}{\mathrm{vol}(B')} \overset{(a)}{\le} N(\delta; B, \|\cdot\|') \overset{(b)}{\le} \frac{\mathrm{vol}\left(\frac{2}{\delta}B + B'\right)}{\mathrm{vol}(B')}$$

Remark: If $B' \subseteq B$, then simplifies

$$N(\delta; B, \|\cdot\|') \le \left(1 + \frac{2}{\delta}\right)^d \frac{\mathrm{vol}(B)}{\mathrm{vol}(B')}$$

pf of lemma: If $\{\theta^1, \dots, \theta^n\}$ is a δ-covering of B, then $B \subseteq \bigcup_{i=1}^n \{\theta^i + \delta B'\}$

$\Rightarrow \mathrm{vol}(B) \le N\delta^d \mathrm{vol}(B') \Rightarrow$ (a).

Let $\{\theta^1, \dots, \theta^m\}$ be a maximal $\delta/2$-packing of B. $\Rightarrow$ must also be δ-covering.

Taking into account:

$\{x + \frac{\delta}{2}B'\} \subseteq B + \frac{\delta}{2}B'$

are disjoint

get:

$$M \operatorname{vol}\left(\tfrac{\delta}{2}B'\right) \le \operatorname{vol}\left(B + \tfrac{\delta}{2}B'\right)$$

$$\left(\tfrac{\delta}{2}\right)^d M \operatorname{vol}(B') \quad \left(\tfrac{\delta}{2}\right)^d \operatorname{vol}\left(\tfrac{2}{\delta}B + B'\right)$$

∎

Cor: If $\|\cdot\|' = \|\cdot\|$, then

$$d \log\left(\tfrac{1}{\delta}\right) \le \log N(\delta; B, \|\cdot\|) \le d \log\left(1 + \tfrac{2}{\delta}\right)$$

Metric entropy can be very large...

Ex: Define

$$\mathcal{F}_L([0,1]^d) = \{f : [0,1]^d \to \mathbb{R} : f(0) = 0,\ |f(x) - f(x')| \le L\|x - x'\|_\infty \ \forall x, x' \in [0,1]^d\}$$

Then $\log N(\delta; \mathcal{F}_L([0,1]^d), \|\cdot\|_\infty) \asymp \left(\frac{L}{\delta}\right)^d$

See ex 5.10 in Wainwright.

One of the main uses of covering numbers is for random matrix theory.

Review: Any $A \in \mathbb{R}^{m \times n}$ can be written as

$$A = \sum_{i=1}^{r} s_i u_i v_i^T = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_r \\ | & & | \end{bmatrix} \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_r \end{bmatrix} \begin{bmatrix} - & v_1 & - \\ & \vdots & \\ - & v_r & - \end{bmatrix}$$

where $r = \text{rank}(A)$, $s_i \geq 0$, $\|u_i\|_2 = \|v_i\| = 1$

- Decomposition is called a singular value decomposition.
- s_i are the singular values. We define $s_{r+1}, \ldots, s_n = 0$. and arrange

$$s_1 \geq s_2 \geq s_3 \geq \cdots \geq s_n \geq 0$$

- $s_i(A) = \sqrt{\lambda_i(AA^T)} = \sqrt{\lambda_i(A^TA)}$ $\forall i = 1, \ldots, r$.

The set $\mathbb{R}^{m \times n}$ admits the inner-product

$$\langle A, B \rangle := \sum_{i,j} A_{ij} B_{ij} = \text{tr}(A^T B)$$

and the norms

$$\|A\|_F := \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\langle A, A \rangle} = \|s(A)\|_2$$

$$\|A\|_2 = \sup_{\|x\| \leq 1} \|Ax\|_2 = s_1(A) = \|s(A)\|_\infty$$

Remark: $s_1(A)$ and $s_n(A)$ are respectively the smallest $\bar{m}$ and largest $\underline{m}$ with

$$\underline{m}\|x\|_2 \le \|Ax\|_2 \le \bar{m}\|x\|_2 \quad \forall x \in \mathbb{R}^n$$

Thm (Eckart-Young-Mirsky)
Let $A = \sum_{i=1}^{n} s_i u_i v_i^T$ be an SVD of A
Then the matrix $A = \sum_{i=1}^{r} s_i u_i v_i^T$
solves the problem

$$\min_{A' : \operatorname{rank}(A') \le r} \|A - A'\|$$

where $\|\cdot\|$ is either $\|\cdot\|_F$ or $\|\cdot\|_2$.

We next aim to show that for $A \in \mathbb{R}^{m \times n}$ with independent subGaussian entries, it holds:

$$\|A\|_2 \lesssim \sqrt{m} + \sqrt{n}$$

with high probability.

Lemma: Let $A \in \mathbb{R}^{m \times n}$ and $\varepsilon \in [0, \sqrt{2}-1)$. Then for any ε-nets $\mathcal{N}$ of S^{n-1} and $\mathcal{M}$ of S^{m-1}, it holds:

$$\sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \overset{@}{\leq} \|A\|_2 \overset{(b)}{\leq} \frac{1}{1-2\varepsilon-\varepsilon^2} \sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle$$

pf: @ is clear. To see (b), suppose $\|A\|_2 = \langle Ax, y \rangle$ for some $x \in S^{n-1}$, $y \in S^{m-1}$.

$\exists \hat{x} \in \mathcal{N}, \hat{y} \in \mathcal{M}$ with $\|x - \hat{x}\| \leq \varepsilon$, $\|y - \hat{y}\| \leq \varepsilon$

$$\Rightarrow \langle Ax, y \rangle = \langle A\hat{x} - A(\hat{x}-x), \hat{y} - (\hat{y}-y) \rangle$$

$$= \langle A\hat{x}, \hat{y} \rangle - \langle A\hat{x}, \hat{y}-y \rangle - \langle A(\hat{x}-x), \hat{y} \rangle + \langle A(\hat{x}-x), \hat{y}-y \rangle$$

$$\leq \langle A\hat{x}, \hat{y} \rangle + 2\varepsilon \|A\|_2 + \|A\|_2 \varepsilon^2$$

$$\Rightarrow \|A\|_2 \leq \frac{1}{1-2\varepsilon-\varepsilon^2} \langle A\hat{x}, \hat{y} \rangle \quad \blacksquare$$

Thm: Let $A \in \mathbb{R}^{m\times n}$, where A_{ij} are independent mean-zero σ_{ij}-subGaussian. Then $\exists$ constant $C>0$ s.t.

$$\mathbb{P}\left[\|A\|_2 \le C\cdot\left(\max_{i,j}\sigma_{ij}\right)\cdot\left(\sqrt{m}+\sqrt{n}+t\right)\right] \ge 1-2e^{-t^2}$$

[Remark: when $A_{ij}\sim N(0,1)$, optimal constant is $C=1$]

pf: Step 1: Approximation

We know $N\left(S^{n-1}, \varepsilon, \|\cdot\|_2\right) \le \left(1+\frac{2}{\varepsilon}\right)^n$

$N\left(S^{m-1}, \varepsilon, \|\cdot\|_2\right) \le \left(1+\frac{2}{\varepsilon}\right)^m$

$\Rightarrow$ Set $\varepsilon=\frac{1}{4}$. There are nets

$$\text{card}(\mathcal{N}) \le 9^n, \quad \text{card}(\mathcal{M}) \le 9^m$$

$\overset{\text{Lemma}}{\Rightarrow}$ $\|A\|_2 \le 3\cdot \max_{x\in\mathcal{N}, y\in\mathcal{M}} \langle Ax, y\rangle$

Step 2: Concentration

The RV $\langle Ax, y\rangle = \sum_{i,j} A_{ij} x_i y_j$ satisfies

$$e^{\lambda \langle Ax,y\rangle} = \prod_{i,j} e^{(\lambda x_i y_j) A_{ij}} \le \prod_{i,j} e^{\lambda^2 \sigma_{ij}^2 x_i^2 y_j^2}$$

$$= e^{\lambda^2 \left(\max_{ij} \sigma_{ij}\right)^2 \sum_{i,j} x_i^2 y_j^2} = e^{\lambda^2 \max_{i,j} \sigma_{ij}^2}$$

So Hoeffding $\Rightarrow$

$$P\left[\langle Ax,y\rangle \ge u\right] \le e^{\frac{-u^2}{2\max \sigma_{ij}^2}} \quad \forall u > 0$$

$$\Rightarrow P\left[\max_{\substack{x\in\mathcal{N} \\ y\in\mathcal{M}}} \langle Ax,y\rangle \ge u\right] \le 9^{n+m} e^{\frac{-u^2}{2\max_{ij} \sigma_{ij}^2}}$$

Choose $u = C\left(\max_{ij} \sigma_{ij}\right)\left(\sqrt{n} + \sqrt{m} + t\right)$

Then

$$RHS \le 9^{n+m} \cdot e^{-C(n+m+t^2)}$$

$$= e^{\ln(9)(n+m) - C(n+m+t^2)} \le e^{-t^2}$$

by choosing C large ∎

Cor: (HW) $\mathbb{E}\|A\|_2 \le C(\max_{ij} \sigma_{ij})(\sqrt{m}+\sqrt{n})$

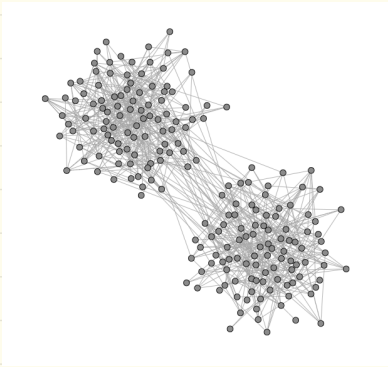
Obvious analogues hold for symmetric matrices.

Application: Community Detection

Defn: (Stochastic Block Model)

Divide n vertices into two sets Q_1, Q_2 of size $n/2$ each. Build the graph

$$\mathbb{P}[ij \text{ is an edge}] = \begin{cases} p & \text{if } i,j \in Q_1 \text{ or } i,j \in Q_2 \\ q & \text{o.w} \end{cases}$$



The goal is to find the communities from the realization of the graph.

Let A be the adjacency matrix. Write it as

$$A = D + R \quad \text{where}$$

$$D = \mathbb{E}A = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right]$$

Exercise: (HW)

Show that D has rank 2 and

$$\lambda_1 = \left(\frac{p+q}{2}\right)n, \quad u_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} / \sqrt{n}$$

$$\lambda_2 = \left(\frac{p-q}{2}\right)n, \quad u_2 = \begin{bmatrix} 1 \\ \vdots \\ -1 \end{bmatrix} / \sqrt{n}$$

If we knew u_2, we would be able to identify the communities.

Signal to noise ratio (SNR): $\frac{\|D\|_2}{\|R\|_2}$

We know $\|D\|_2 = \lambda_1 \sim n$

$$\mathbb{P}\left[\|R\| \le C\sqrt{n}\right] \ge 1 - 4e^{-n}$$

So $SNR \sim \sqrt{n}$

$\Rightarrow$ problem should be solvable.

A natural idea is to use spectral clustering.

Alg: (Spectral Clustering)

Input: graph G

1) Compute adjacency matrix A

2) Compute eigenvector $V_2(A)$ corresponding to second largest eigenvalue of A.

3) Set $Q_1 = \{i : [V_2(A)]_i > 0\}$

$Q_2 = \{i : [V_2(A)]_i \le 0\}$.

Perturbation Results:

Thm: (Weyl's inequality)
For any symmetric $X, Y \in \mathbb{R}^{n\times n}$, it holds

$$\max_{i=1,\dots,n} |\lambda_i(X) - \lambda_i(Y)| \leq \|X - Y\|_2$$

Thm (Davis-Kahan)
Let $X, Y \in \mathbb{R}^{n\times n}$ be symmetric. Fix an index i and assume

$$\min_{j : j \neq i} |\lambda_i(X) - \lambda_j(X)| = \delta > 0$$

Then

$$\sin \angle (v_i(X), v_i(Y)) \leq \frac{2\|X - Y\|_2}{\delta}$$

Remark: This implies

$$\min_{\theta \in \{\pm 1\}} \|v_i(X) - \theta v_i(\theta)\|_2 \leq \frac{2^{3/2}\|X - Y\|_2}{\delta}$$

Let's compute the eigengap

$$\delta = \min\{\lambda_2, \lambda_1 - \lambda_2\} = \min\left\{\frac{p-q}{2}, q\right\} \cdot n$$

$$\Downarrow$$

$$\mu \cdot n$$

$$\Rightarrow \min_{\theta \in \{\pm 1\}} \|V_2(D) - \theta V_2(A)\|_2 \leq \frac{\|D-A\|_2}{\mu n} \leq \frac{\|R\|_2}{\mu n}$$

$$\Rightarrow \mathbb{P}\left[\min_{\theta \in \{\pm 1\}} \|V_2(D) - \theta V_2(A)\|_2 \leq \frac{c}{\mu\sqrt{n}}\right] \geq 1 - 4e^{-n}.$$

Since all coordinates of $\theta V_2(D)$ are $\pm \frac{1}{\sqrt{n}}$, the signs can only disagree in at most $\frac{c^2}{\mu^2}$ coordinates.

Thm: Set $\mu := \min\{q, \frac{p-q}{2}\}$. Then with probability $1 - 4e^{-n}$, the spectral clustering algorithm identifies the communities correctly up to $\left(\frac{c}{\mu}\right)^2$ misclassified vertices.

We next obtain two-sided bounds on the full spectrum

$$\sqrt{m} - C\sqrt{n} \le s_i(A) \le \sqrt{m} + C\sqrt{n}$$

and we relax independence of entries to independence of the rows.

Thm: Let $A \in \mathbb{R}^{m \times n}$ be a matrix with rows A_i that are independent and σ_i-sub-Gaussian isotropic. Then

$$\sqrt{m} - CK^2(\sqrt{n}+t) \le s_n(A) \le s_1(A) \le \sqrt{m} + CK^2(\sqrt{n}+t)$$

w.p. $1 - 2\exp(-t^2)$, where $K = \max_i \sigma_i$

pf: We first show

$$\left\| \frac{1}{m} A^T A - I_n \right\|_2 \le K^2 \max\{\delta, \delta^2\} \text{ where}$$

$$\delta = C\left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\right)$$

Step 1: As before, there exist a $\frac{1}{4}$-cover $\mathcal{N}$ of S^{n-1} with $\text{card}(\mathcal{N}) \le 9^n$.

For $B = \frac{1}{m} A^T A - I$ have $\|B\|_2 = \sup_{v \in S^{n-1}} \langle Bv, v \rangle$

If v achieves sup. Then $\exists w$ with $\|w - v\| \le \frac{1}{4}$ and

$$\|B\|_2 = \langle Bw - B(w-v),\ w - (w-v) \rangle$$

$$= \langle Bw, w \rangle - 2\langle Bw, w-v \rangle + \langle B(w-v), w-v \rangle$$

$$\le \sup_{w \in \mathcal{N}} \langle Bw, w \rangle + \frac{1}{4}\|B\|_2 + \frac{1}{16}\|B\|$$

$$\Rightarrow \|B\|_2 \le 2 \sup_{w \in \mathcal{N}} \langle Bw, w \rangle$$

$$= 2\left(\sup_{w \in \mathcal{N}} \frac{1}{m} \|Aw\|_2^2 - 1 \right)$$

Step 2: Concentration

Fix $w \in \mathcal{N}$ and write

$$\|Aw\|_2^2 = \sum_{i=1}^{m} \langle A_{i,\cdot}, w \rangle^2 =: \sum_{i=1}^{m} X_i^2$$

$\Rightarrow X_i$ are independent, 6_i-subGaussian with $\mathbb{E} X_i^2 = 1$. Therefore, Bernstein

$$\mathbb{P}\left[\left| \frac{1}{m} \|Aw\|_2^2 - 1 \right| \ge t \right] \le 2 \exp\left[-\frac{1}{2} \left(\frac{mt^2}{K^4} \wedge \frac{mt}{4K^2} \right) \right]$$

$$2\exp\left[-\frac{1}{2}\left(\frac{mt^2}{K^4} \wedge \frac{mt}{4K^2}\right)\right] = 2\exp\left[-c_1 m\left[\frac{t^2}{K^4} \wedge \frac{t}{K^2}\right]\right]$$

Set $t = K^2 \max\{\delta, \delta^2\}$

$$= 2\exp\{-c_1 m \delta^2\}$$

$\leq$

Step 3: Union bound

$$\mathbb{P}\left[\max_{w \in N} \left|\frac{1}{m}\|Aw\|_2^2 - 1\right| \geq K^2 \max\{\delta, \delta^2\}\right]$$

$$\leq 9^n \cdot 2\exp\{-c_1 C^2 (n + t^2)\} \leq 2e^{-t^2}$$

if C is big enough.

Finally result follows from

Lemma: Suppose $A \in \mathbb{R}^{m \times n}$ and $\delta > 0$ satisfy

$$\|A^T A - I\|_2 \leq \max\{\delta, \delta^2\}$$

Then $1 - \delta \leq s_n(A) \leq s_1(A) \leq 1 + \delta$ ☐

Cor: (HW)
$$\mathbb{E}\left\|\frac{1}{m}A^TA - I\right\|_2 \le CK^2\left(\sqrt{\frac{n}{m}} + \frac{n}{m}\right)$$

Covariance Estimation

Assume $X \in \mathbb{R}^n$ and $\mathbb{E}X = 0$

Define $\Sigma = \mathbb{E}XX^T$. Let's estimate Σ by the sample covariance

$$\Sigma_m := \frac{1}{m}\sum_{i=1}^{m} X_i X_i^T$$

Clearly for $\Sigma - \Sigma_m$ to be small we need $m > n$. We'll see $m \sim n$ suffice

Thm: Suppose $X \in \mathbb{R}^n$ satisfies

$$\mathbb{E}_X e^{\lambda\langle X, w\rangle} \le e^{K^2\lambda^2\langle \Sigma w, w\rangle} \quad \forall w \in \mathbb{R}^n$$

for some $K > 0$.

Then

$$\mathbb{E}\frac{\|\Sigma_m - \Sigma\|_2}{\|\Sigma\|_2} \le CK^2\left(\sqrt{\frac{n}{m}} + \frac{n}{m}\right)$$

pf. Fix random $X, X_1, \dots, X_m$. Define

$$Z = \Sigma^{-1/2} X \qquad Z_i := \Sigma^{-1/2} X_i$$

Recall Z, Z_i are independent and isotropic. Moreover for any $w \in S^{n-1}$, have

$$\mathbb{E} e^{\lambda \langle Z, w \rangle} = \mathbb{E} e^{\lambda \langle \Sigma^{-1/2} X, w \rangle}$$
$$= \mathbb{E} e^{\lambda \langle X, \Sigma^{-1/2} w \rangle} \le e^{\frac{1}{2}\lambda^2 K^2 \langle \Sigma (\Sigma^{-1/2} w), \Sigma^{-1/2} w \rangle}$$
$$= e^{\frac{1}{2}\lambda^2 K^2}$$

So Z and Z_i are K-sub-Gaussian

$$\Rightarrow \|\Sigma_m - \Sigma\|_2 = \left\| \Sigma^{1/2} \left(\frac{1}{m} \sum_{i=1}^m Z_i Z_i^T - I_n \right) \Sigma^{1/2} \right\|_2$$
$$\le \|\Sigma\|_2 \cdot \left\| \frac{1}{m} \sum_{i=1}^n Z_i Z_i^T - I \right\|_2$$

Define A whose rows are Z_i. Then

$$\frac{1}{m} A^T A - I = \frac{1}{m} \sum_{i=1}^m Z_i Z_i^T - I. \qquad \blacksquare$$

So can get

$$\mathbb{E}\,\|\Sigma_m - \Sigma\|_2 \le \varepsilon\,\|\Sigma\|_2$$

as long as $m \sim \frac{n}{\varepsilon^2}$

Exc: (HW) Actually get the high probability estimate

$$\|\Sigma_m - \Sigma\|_2 \le CK^2\left(\sqrt{\frac{n+t}{m}} + \frac{n+t}{m}\right)\|\Sigma\|_2$$

with probability $1 - 2e^{-t}$

Remark: In particular, setting $t = \delta m$ get

$$\frac{\|\Sigma_m - \Sigma\|_2}{\|\Sigma\|_2} \le CK^2\left(\sqrt{\frac{n}{m}} + \frac{n}{m} + \delta\right)$$

with probability $1 - 2e^{-\delta m}$

Clustering (Gaussian Mixture Model)

Generate m random points in $\mathbb{R}^n$ as follows.

Flip a fair coin

- if heads choose $x_i \sim N(\mu, I)$
- if tails choose $x_i \sim N(-\mu, I)$

Suppose we are given m points drawn from the Gaussian Mixture Model.

Goal: Identify which points belong to which cluster.

Spectral Clustering:

Input: $x_1, \ldots, x_m \in \mathbb{R}^n$

Output: a partition of points into 2 clusters.

1) Compute $\Sigma_m = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T$
2) Compute eigenvector of Σ_m corresponding to $\lambda_1(\Sigma)$.
3) If $\langle v, x_i \rangle > 0$, put x_i in first community otherwise, put x_i in second.

Thm: Let $\delta > 0$ be s.t. $\|\mu\|_2 \geq C\sqrt{\log\left(\frac{1}{\delta}\right)}$.
Suppose $m \geq \left(\frac{n}{\|\mu\|_2}\right)^c$ where c is a constant. Then with probability $1 - 4e^{-n}$, the algorithm identifies communities correctly up to δm misclassified vertices.

pf sketch: Note $X = \varepsilon\mu + g$ where ε is a Rademacher RV and $g \sim N(0, I)$

$\Rightarrow \mathbb{E}\, XX^T = \mathbb{E}(\mu\mu^T + gg^T) = I + \mu\mu^T$

$\Rightarrow \mathbb{E}\, XX^T$ has eigenvalues $1 + \|\mu\|^2, 1, \dots, 1$

Eigenvector corresponding to $1 + \|\mu\|^2$ is μ

So from the e-vector corresponding to the largest eigenvalue, you can "guess" at the communities by checking the sign of $\langle x_i, \mu \rangle$.

You do the rest in your HW ☺

Our next goal is to investigate tails of matrices generalizing bounds

$$\left\| \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T - \Sigma \right\|_2$$ from before.

Notation:

- $\mathbb{S}^d$ are $d \times d$ symmetric matrices
- $\mathbb{S}_+^d = \{ Q \in \mathbb{S}^d : \langle Qx, x \rangle \geq 0 \ \forall x \in \mathbb{R}^d \}$
- $\gamma_1(X) \geq \gamma_2(X) \geq \dots \geq \gamma_d(X)$ are the eigenvalues of $X \in \mathbb{S}^d$
- $\mathbb{O}^d = \{ U \in \mathbb{R}^{d \times d} : U^T U = I \}$.

Any function $f : \mathbb{R} \to \mathbb{R}$ defines a function $f : \mathbb{S}^d \to \mathbb{S}^d$ by

$$f(Q) = U \operatorname{diag}(f(\gamma_1(Q)), \dots, f(\gamma_d(Q))) U^T$$

where $Q = U \operatorname{diag}(\gamma(Q)) U^T$ is any eigenvalue decomposition of Q.

[The choice of U does not matter (check!)]

Note

$$\gamma(f(Q)) = \{ f(\gamma_j(Q)) : j = 1, \ldots, d \}$$

Matrix exponential: e^Q

Matrix Logarithm: $\log Q$ defined on $\mathbb{S}^d_{++}$

The Löwner ordering on $\mathbb{S}^d$

$$X \succeq Y \iff Y - X \in \mathbb{S}^d_+$$

Lemma: (HW) $Q_2 \succeq Q_1 \succ 0 \Rightarrow \log(Q_2) \succeq \log(Q_1)$

[e^Q is not monotone]

Lemma: If $f: \mathbb{R} \to \mathbb{R}$ is continuous and non-decreasing, then

$$Q \succeq R \Rightarrow \operatorname{tr}(f(Q)) \geq \operatorname{tr}(f(R))$$

Lemma: $I + A \preceq e^A$

Thm (Lieb): For any $H \in \mathbb{S}^d$, the function $f: \mathbb{S}^d_{++} \to \mathbb{R}$ given by $f(A) = \operatorname{tr}(e^{H + \log(A)})$ is concave.

Defn: Moment generating function of a random matrix $Q \in \mathbb{S}^d$ is the function $\Psi_Q : \mathbb{R} \to \mathbb{S}^d$ given by

$$\Psi_Q(\lambda) = \mathbb{E} e^{\lambda Q} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[Q^k]$$

Defn: A zero-mean $Q \in \mathbb{S}^d$ is V-subGaussian (with $V \in \mathbb{S}^d_+$) if

$$\Psi_Q(\lambda) \preceq e^{\frac{\lambda^2 V}{2}} \quad \forall \lambda \in \mathbb{R}$$

Ex: Suppose $Q = \epsilon A$ where ϵ is a Rademacher RV and $A \in \mathbb{S}^d$ is fixed.

$$\mathbb{E} e^{\lambda Q} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}(\epsilon A)^k = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} A^{2k}$$

$$\preceq \sum_{k=1}^{\infty} \frac{1}{k!} \left(\frac{\lambda^2 A^2}{2} \right)^k$$

$$= e^{\frac{\lambda^2 A^2}{2}}$$

So Q is A^2-subGaussian.

Exercise (HW)

If $Q = gA$, where $g \in \mathbb{R}$ is a symmetric distribution 6-sub Gaussian variable, then Q is 6^2A^2 sub-Gaussian.

Ex: Suppose $Q = \varepsilon A$ where ε is Rademacher and A is random with $\|A\|_2 \le b$. Then

$$\psi_Q(\lambda) = \mathbb{E}e^{\lambda Q} = \mathbb{E}_A \mathbb{E}_\varepsilon e^{\varepsilon A} \overset{\text{(why?)}}{\le} \mathbb{E}_A e^{\frac{\lambda^2 A^2}{2}} \le e^{\frac{\lambda^2 b^2 I}{2}}$$

So Q is $b^2 I$-sub Gaussian.

Defn: A zero-mean random matrix Q is subexponential with parameters (V, α) if

$$\psi_Q(\lambda) \le e^{\frac{\lambda^2 V}{2}} \quad \forall |\lambda| \le \frac{1}{\alpha}.$$

Ex: $Q = \varepsilon g^2 A$ where ε is Rademacher and $g \sim N(0,1)$ independent.

[You'll compute V and α in HW]

Lemma: Suppose zero-mean $Q \in \mathcal{S}^d$ satisfies $\|Q\|_2 \le b$. Then we have

$$\psi_Q(\lambda) \lesssim \exp\left(\frac{\lambda^2 \mathrm{Var}(Q)}{2(1-b|\lambda|)}\right) \quad \forall |\lambda| \le \frac{1}{b}$$

pf: $\mathbb{E} e^{\lambda Q} = \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E} Q^k}{k!} =$

$$= I + \frac{\lambda^2 \mathrm{Var}(Q)}{2} + \sum_{j=3}^{\infty} \frac{\lambda^j \mathbb{E} Q^j}{j!}$$

$$\lesssim I + \frac{\lambda^2 \mathrm{Var}(Q)}{2} + \sum_{j=3}^{\infty} \frac{\lambda^2 \lambda^{j-2} \mathbb{E} Q^2 b^{j-2}}{j!}$$

$$\lesssim I + \frac{\lambda^2 \mathrm{Var}(Q)}{2} + \frac{\lambda^2 \mathrm{Var}(Q)}{2} \sum_{j=3}^{\infty} \lambda^{j-2} b^{j-2}$$

$$= I + \frac{\lambda^2 \mathrm{Var}(Q)}{2(1-b|\lambda|)} \lesssim \exp\left(\frac{\lambda^2 \mathrm{Var}(Q)}{2(1-b|\lambda|)}\right)$$ ∎

Matrix Chernoff:

Lemma: Let $Q \in \mathbb{S}^d$ be zero mean with ψ_Q existing on $(-a, a)$. Then for any $t > 0$, it holds:

$$\mathbb{P}[\gamma_1(Q) \geq t] \leq \operatorname{tr}(\psi_Q(\lambda)) e^{-\lambda t} \quad \forall_{\lambda \in [0,a)}$$

Consequently

$$\mathbb{P}[\|Q\|_2 \geq t] \leq 2 \operatorname{tr}(\psi_Q(\lambda)) e^{-\lambda t} \quad \forall_{\lambda \in [0,a)}$$

pf: We begin as in scalar case:

$$\begin{aligned}
\mathbb{P}[\gamma_1(Q) \geq t] &= \mathbb{P}[e^{\gamma_1(\lambda Q)} \geq e^{\lambda t}] \\
&= \mathbb{P}[\gamma_1(e^{\lambda Q}) \geq e^{\lambda t}] \\
&\leq \mathbb{E}[\gamma_1(e^{\lambda Q})] e^{-\lambda t} \\
&\leq \mathbb{E}[\operatorname{tr} e^{\lambda Q}] e^{-\lambda t} \\
&\leq \operatorname{tr}[\psi_Q(\lambda)] e^{-\lambda t} \quad \square
\end{aligned}$$

Lemma: Let $Q_1, \dots, Q_n \in \mathbb{S}^n$ be independent with $\Psi_{Q_i}(\lambda)$ existing for all λ in an interval J. Define $S_n = \sum_{i=1}^n Q_i$. Then

$$\operatorname{tr}\left(\Psi_{S_n}(\lambda)\right) \le \operatorname{tr}\left(e^{\sum_{i=1}^n \log \Psi_{Q_i}(\lambda)}\right) \quad \forall \lambda \in J$$

[**Note:** together with the previous lemma, we deduce

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_2 \ge t\right] \le 2\operatorname{tr}\left(e^{\sum_{i=1}^n \log \Psi_{Q_i}(\lambda)}\right) \cdot e^{-\lambda t n}$$

]

pf: Define $G(\lambda) := \operatorname{tr}(\Psi_{S_n}(\lambda))$. Then

$$G(\lambda) = \operatorname{tr}\left(\mathbb{E}\, e^{\lambda S_{n-1} + \log \exp(\lambda Q_n)}\right)$$

$$= \mathbb{E}_{S_{n-1}} \mathbb{E}_{Q_n} \operatorname{tr}\left(e^{\lambda S_{n-1} + \log \exp(\lambda Q_n)}\right)$$

Lieb

$$\le \mathbb{E}_{S_{n-1}} \operatorname{tr}\left(e^{\lambda S_{n-1} + \log \Psi_{Q_n}(\lambda)}\right)$$

$$\le \dots \le \operatorname{tr}\left(e^{\sum_{i=1}^n \log \Psi_{Q_i}(\lambda)}\right) \quad \square$$

Thm (Hoeffding)
Let $Q_1, \dots, Q_n \in \mathbb{S}^d$ be independent, zero-mean, and V_i-subGaussian. Then

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_2 \geq t\right] \leq 2\,\mathrm{rank}\left(\sum_{i=1}^n V_i\right) e^{\frac{-nt^2}{2\sigma^2}}$$

$$\leq 2d\, e^{\frac{-nt}{2\sigma^2}} \qquad \forall t > 0,$$

where $\sigma^2 = \left\|\frac{1}{n}\sum_{i=1}^n V_i\right\|_2$.

pf: Define $V = \sum_{i=1}^n V_i$. Suppose first $\mathrm{rank}\, V = d$. Then

$$\sum_{i=1}^n \log \Psi_{Q_i}(\lambda) \preceq \frac{\lambda^2}{2}\sum_{i=1}^n V_i$$

because $\log(\cdot)$ is matrix monotone. The function $t \mapsto e^t$ is increasing

$$\Rightarrow \mathrm{tr}\left(e^{\sum_{i=1}^n \log \Psi_{Q_i}(\lambda)}\right) \leq \mathrm{tr}\left(e^{\frac{\lambda^2}{2}\sum_{i=1}^n V_i}\right)$$

Chernoff

$$\Rightarrow \mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Q_i\right\|_2 \geq t\right] \leq 2\,\mathrm{tr}\left(e^{\frac{\lambda^2}{2}V}\right)e^{-\lambda n t}$$

Note $\mathrm{tr}\left(e^{A}\right) \leq d\,e^{\|A\|_2}$ for any $A \in \mathbb{R}^d$

$$\Rightarrow \leq 2d\,e^{\frac{\lambda^2}{2}n\sigma^2 - \lambda n t}$$

Optimize over $\lambda \Rightarrow \lambda = \frac{t}{\sigma^2}$. Done.

If $r := \mathrm{rank}(V) < d$, we can form an eigenvalue decomposition

$V = UDU^T$ where $U \in \mathbb{R}^{d \times r}$

Then replace Q by

$\hat{Q} = U^T Q U \in \mathbb{R}^{r \times n}$. ∎

Thm: Let $Q_1, \dots, Q_n \in S^d$ be zero-mean, independent, with $\|Q_i\|_2 \le b$. Then

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_2 \ge t\right] \le 2\,\mathrm{rank}(V)\exp\left(-\frac{nt^2}{2\sigma^2 + bt}\right)$$

$$\le 2\,\mathrm{rank}(V)\exp\left(-\frac{nt^2}{4\sigma^2} \wedge \frac{nt}{2b}\right)$$

where (HW) $V = \sum_{i=1}^n \mathbb{E}Q_i^2$ and $\sigma^2 = \frac{1}{n}\|V\|_2$.

Cor: Under same assumptions,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n Q_i\right\|_2 \le \sqrt{\frac{\sigma^2 \log(d)}{n}} + \frac{b\log(d)}{n}$$

The Bernstein bound can be extended to rectangular matrices.

Idea: Given a sequence $A_1, \dots, A_n \in \mathbb{R}^{d_1 \times d_2}$ form $Q_i := \begin{bmatrix} 0_{d_1 \times d_2} & A_i \\ A_i^T & 0_{d_1 \times d_2} \end{bmatrix}$

Lemma: (HW) $\|Q_i\|_2 = \|A_i\|$ and

$$\sigma^2 \le \max\left\{\left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[A_i A_i^T]\right\|_2, \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[A_i A_i^T]\right\|_2\right\}$$

and therefore

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n A_i\right\|_2 \ge t\right] \le 2(d_1+d_2)e^{-\frac{nt^2}{2(\sigma^2+bt)}}$$

Covariance Estimation for general distributions

Let $X \in \mathbb{R}^d$ be a random vector with $\mathbb{E}X = 0$. Define $\Sigma = \mathbb{E}XX^T$ and $\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ where X_i are i.i.d realizations of X.

[If X_i were Σ, K subGaussian, we proved $\frac{\mathbb{E}\|\Sigma_n - \Sigma\|_2}{\|\Sigma\|_2} \le K^2\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)$]

Thm: Assume for some $K>0$, have

$$\|X\|_2 \le K\sqrt{\operatorname{tr}\Sigma} \qquad \left[=\sqrt{\mathbb{E}\|X\|_2^2}\right]$$

Then

$$\frac{\mathbb{E}\|\Sigma_n-\Sigma\|_2}{\|\Sigma\|_2} \le C\left(\sqrt{\frac{K^2 r\log(d)}{n}}+\frac{K^2 r\log(d)}{n}\right)$$

where $r := \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|_2} \le d$

pf: Matrix Bernstein

$$\mathbb{E}\|\Sigma_n-\Sigma\|_2 = \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n (x_i x_i^T-\Sigma)\right\|_2$$

$$\le C\left(\sqrt{\sigma^2\log(d)/n}+M\log(d)/n\right)$$

where $\sigma^2 = \|\mathbb{E}(xx^T-\Sigma)^2\|_2$ and $M \ge \|xx^T-\Sigma\|_2$

Let's estimate σ^2 and M.

$$\mathbb{E}(xx^T-\Sigma)^2 = \mathbb{E}(xx^T)^2-\Sigma^2 \lesssim \mathbb{E}(xx^T)^2$$

and

$$(xx^T)^2 = \|x\|_2^2 xx^T \lesssim K^2(\operatorname{tr}\Sigma)\,xx^T$$

Take expectations $\Rightarrow \mathbb{E}(xx^T)^2 \lesssim K^2\operatorname{tr}(\Sigma)\Sigma$

$$\Rightarrow \sigma^2 \le K^2\operatorname{tr}(\Sigma)\|\Sigma\|_2$$

Finally

$$\|xx^T - \Sigma_1\|_2 \le \|x\|_2^2 + \|\Sigma_1\|_2$$
$$= K^2 \operatorname{tr}\Sigma_1 + \|\Sigma\|_2$$
$$\le 2K^2 \operatorname{tr}\Sigma_1 =: M$$

So

$$\mathbb{E}\|\Sigma_n - \Sigma\|_2 \le C\left(\sqrt{\frac{K^2 \operatorname{tr}(\Sigma_1)\|\Sigma\|_2 \log(d)}{n}} + \frac{2K^2 \operatorname{tr}(\Sigma_1)\log(d)}{n}\right)$$

Simplifying completes the proof. □

Sparse covariance estimation

If we know a priori that Σ_1 is sparse, we should be able to estimate it using fewer observations. Goal: replace $\|\Sigma_n - \Sigma_1\|_2 \le \sqrt{\frac{d}{n}}$ by $\|\hat{\Sigma}_n - \Sigma_1\|_2 \le \sqrt{\frac{\ell \log(d)}{n}}$ where "ℓ" is the sparsity level.

For $\lambda > 0$, define the hard-thresholding operator

$$T_\lambda(u) := \begin{cases} u, & \text{if } |u| > \lambda \\ 0, & \text{o.w.} \end{cases}$$

As a surrogate for sparsity we will use $\|A\|_2$ where

$$A_{ij} \begin{cases} 1, & \text{if } i=j \text{ or } \Sigma_{ij} \neq 0 \\ 0, & \text{o.w.} \end{cases}$$

Lemma: If Σ has at most s nonzero entries per row, then $\|A\|_2 \leq s$.

pf: $A + (s-1)I$ is diagonally dominant and therefore positive semidefinite.

$\Rightarrow |\lambda_j(A)| \leq s-1$.

Let u be an eigenvector of A corresponding to $\gamma_1(A)$ and let j index its maximal entry. WLOG $u_j > 0$

Then $\gamma_1 u_j = (Au)_j = \langle A_j, u \rangle \leq s u_j$ ∎

Thm: Let $\{X_i\}_{i=1}^n$ be i.i.d zero mean with covariance Σ, and suppose X_{ij} is σ-subGaussian. Then as long as $n > \log(d)$ setting

$$\lambda_n = \sigma^2\left(8\sqrt{\frac{\log(d)}{n}} + \delta\right)$$

it holds

$$\mathbb{P}\left[\frac{\|T_{\lambda_n}(\Sigma_n) - \Sigma\|_2}{\sigma^2} \geq 2\|A\|_2\left(8\sqrt{\frac{\log(d)}{n}} + \delta\right)\right] \leq 8e^{-\frac{n}{16}\min\{\delta, \delta^2\}}$$

We proceed with the following lemmas.

Lemma: Suppose $\|\Sigma_n - \Sigma\|_{max} \leq \lambda_n$ (coordinate wise)
Then $|T_{\lambda_n}(\Sigma_n) - \Sigma| \leq 2\lambda_n A$ (coordinate wise)

pf: $A^{ij} = 0 \Rightarrow |\Sigma_n^{ij}| \leq \lambda_n \Rightarrow T_{\lambda_n}(\Sigma_n^{ij}) = 0$

$A^{ij} \neq 0 \Rightarrow |T_{\lambda_n}(\Sigma_n^{ij}) - \Sigma^{ij}| \leq |T_{\lambda_n}(\Sigma_n^{ij}) - \Sigma_n^{ij}| + |\Sigma_n^{ij} - \Sigma^{ij}|$

$\leq 2\lambda_n$ ∎

Lemma: For any symmetric matrices $0 \leq A \leq B$, it holds $\|A\|_2 \leq \|B\|_2$.

pf: Observe

$$\langle a_i, a_j \rangle \leq \langle b_i, b_j \rangle$$

$$\Rightarrow \quad A^2 \leq B^2$$

Since A^2 and B^2 are positive semidefinite, $\|A^2\|_2 = \gamma_1(A^2)$

$$\|B^2\|_2 = \gamma_1(B^2)$$

The Perron-Frobenius theorem guarantees that there exists an eigenvector $u \geq 0$ corresponding to $\gamma_1(A^2)$. In this symmetric setting, this follows directly from:

$$\gamma_1(A^2) = \sup_{\|u\| \leq 1} \sum (A^2)_{ij} u_i u_j \leq \sup_{\|u\| \leq 1} \sum (A^2)_{ij} |u_i| |u_j| \leq \gamma_1(A^2)$$

Therefore

$$\gamma_1(A) \leq \langle A^2 u, u \rangle \leq \sup_{\|v\|_2 \leq 1} \langle B^2 v, v \rangle = \|B^2\|_2$$

$$\Rightarrow \quad \|A\|_2^2 = \|A^2\|_2 \leq \|B^2\|_2 \leq \|B\|_2^2. \qquad \square$$

So $|\Sigma_n - \Sigma|_{max} \le \lambda_n$

$\Rightarrow \|T_{\lambda_n}(\Sigma_n) - \Sigma\|_2 \le 2\lambda_n \|A\|_2$

Lemma:

$$\mathbb{P}\left[\frac{|\Sigma_n - \Sigma|_{max}}{\sigma^2} \ge t\right] \le 8d^2 e^{-\frac{n}{16}\min\{t, t^2\}}$$

pf: Set $\Delta_n := \Sigma_n - \Sigma$. Observe

$\Delta_n^{ii} = \frac{1}{n}\sum_{k=1}^n (X_k)_i^2 - \mathbb{E}X_i^2$. Apply Bernstein

If $i \neq j$, then

$$2\Delta_n^{ij} = \frac{1}{n}\sum_{k=1}^n 2(X_k)_i (X_k)_j - 2\Sigma_{ij}$$

$$= \frac{1}{n}\sum_{k=1}^n \Big(\big[(X_k)_i + (X_k)_j\big]^2 - \big[\Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij}\big] - (X_k)_i^2 - (X_k)_j^2 - \Sigma_{ii} - \Sigma_{jj} \Big)$$

Apply Bernstein again + union bound. ∎

Notice the bound

$$\|T_{\chi_n}(\Sigma_n) - \Sigma\|_2 \leq C\|A\|_2\sqrt{\frac{\log(d)}{n}}$$

can be $\sim d\sqrt{\frac{\log d}{n}}$, if Σ is dense, with small entries. This is much worse than our bounds from before.

Operator vs row norm:

For any $A \in S^d$, it holds

$$\|A\|_2 \geq \max_i \|A_i\|_2$$

Thm: Let $A \in S^d$ whose entries on and above the diagonal are independent mean zero. Then

$$\mathbb{E}\|A\|_2 \leq C\sqrt{\log(d)} \cdot \mathbb{E}\max_i \|A_i\|_2$$

The proof is based on Bernstein.

pf: Write

$$A = \sum_{i \le j} Z_{ij} \quad \text{where}$$

$$Z_{ij} = \begin{cases} A_{ij}(e_i e_j^T + e_j e_i^T), & \text{if } i \neq j \\ A_{ii}, & \text{if } i = j \end{cases}$$

$$\mathbb{E}\|A\|_2 = \mathbb{E}_{\varepsilon,Z}\left\| \sum_{i \le j} Z_{ij} - \mathbb{E}_{i \le j} Z'_{ij} \right\|_2$$

$$\le \mathbb{E}_{\varepsilon,Z}\left\| \sum_{i \le j} \varepsilon_{ij}(Z_{ij} - Z'_{ij}) \right\|_2$$

$$\le 2\,\mathbb{E}_{\varepsilon,Z}\left\| \sum_{i \le j} \varepsilon_{ij} Z_{ij} \right\|_2$$

Recall $\varepsilon_{ij} Z_{ij}$ is Z_{ij}^2-subGaussian.

Therefore with $\sigma^2 = \left\| \sum_{ij} Z_{ij}^2 \right\|_2$ get

$$\mathbb{P}\left[\left\| \sum_{i \le j} \varepsilon_{ij} Z_{ij} \right\|_2 \ge t \right] \le 2d \cdot \exp\left(-\frac{t^2}{2\sigma^2} \right)$$

$$= \exp\left(\log(2d) - \frac{t^2}{2\sigma^2} \right)$$

set $t = \sqrt{2\sigma^2 \log(2d) + u^2}$

$$\Rightarrow \mathbb{P}\left[\left\| \sum_{i \le j} \varepsilon_{ij} Z_{ij} \right\| \ge \sqrt{2\sigma^2 \log(2d)} + u \right] \le e^{-\frac{u^2}{2\sigma^2}}$$

Integrating, get

$$\mathbb{E}_\varepsilon \left\| \sum_{i,j} \varepsilon_{ij} Z_{ij} \right\| \le C\sqrt{\log d} \cdot \underbrace{\sqrt{\left\| \sum_{i \le j} Z_{ij}^2 \right\|_2}}_{\sigma}$$

Next evaluate $\mathbb{E}_Z \sqrt{\left\| \sum_{ij} Z_{ij} \right\|_2}$

Quick computation shows

$$Z_{ij}^2 = \begin{cases} A_{ij}^2 (e_i e_i^T + e_j e_j^T), & i < j \\ A_{ii}^2 e_i e_i^T, & i = j. \end{cases}$$

↑ diagonal.

$$\Rightarrow \sum_{i \le j} Z_{ij}^2 = \begin{bmatrix} \|A_1\|_2^2 & & 0 \\ & \|A_2\|_2^2 & \\ & & \ddots \\ 0 & & \|A_d\|_2^2 \end{bmatrix}$$

$$\Rightarrow \sqrt{\left\| \sum_{i \le j} Z_{ij}^2 \right\|_2} = \max_{i=1,\dots,d} \|A_i\|_2$$

∎

Cor (HW) Let $A \in \mathbb{R}^{m \times n}$ whose entries are independent mean-zero. Then

$$\mathbb{E}\|A\|_2 \leq C\sqrt{\log(m+n)}\left(\mathbb{E}\max_i \|A_i\|_2 + \mathbb{E}\max_j \|A^j\|_2\right)$$

Application: Matrix Completion.

Consider $X \in \mathbb{R}^{d \times d}$ with $\text{rank}(X) = r << d$.

Suppose each entry X_{ij} is revealed independently with probability $p \in (0,1)$

So we get to see $Y \in \mathbb{R}^{d \times d}$ with

$$Y_{ij} = \delta_{ij} X_{ij} \quad \text{where } \delta_{ij} \sim \text{Ber}(p)$$

Thus we get to see

$$m := p \cdot d^2 \text{ entries on average}$$

Thm: Let $\hat{X}$ be a best rank-r approximation to $\rho^{-1}Y$. Then

$$\mathbb{E}\,\frac{1}{d}\|\hat{X}-X\|_F \le C\sqrt{\frac{rd\log(d)}{m}}\,\|X\|_{\max}$$

as long as $m \gtrsim \log(d)$.

So in order to ensure

$$\mathbb{E}\sqrt{\frac{1}{d^2}\sum_{i,j}(\hat{X}_{ij}-X_{ij})^2} \le \varepsilon\,\|X\|_{\max},$$

it suffices to see $m = \frac{1}{\varepsilon^2} rd\log(d)$ entries on average.

pf: First bound error is operator norm

$$\|\hat{X}-X\|_2 \le \|\hat{X}-\rho^{-1}Y\|_2 + \|\rho^{-1}Y - X\|_2$$

$$\le 2\|\rho^{-1}Y - X\|_2 = \frac{2}{\rho}\underbrace{\|Y-\rho X\|_2}_{\text{easy to understand}}$$

$$(Y - pX)_{ij} = (\delta_{ij} - p) X_{ij}$$

↑ independent mean-zero

$$\Rightarrow \mathbb{E}\|Y - pX\|_2 \le C\sqrt{\log(n)} \cdot \left(\mathbb{E}\max_i \|(Y-pX)_i\|_2 + \mathbb{E}\max_j \|(Y-pX)^j\|_2\right)$$

Observe

$$\|(Y-pX)_i\|_2^2 = \sum_j (Y-pX)_{ij}^2 = \sum_j (\delta_{ij} - p)^2 X_{ij}^2$$

$$\le \left(\sum_j (\delta_{ij} - p)^2\right) \|X\|_{max}^2$$

Can bound (HW) $\mathbb{E}\max_i \sum_j (\delta_{ij} - p)^2 \le Cpd$

using concentration and union bound.

$$\Rightarrow \mathbb{E}\|Y - pX\|_2 \le C\sqrt{pd\log(d)}\,\|X\|_{max}$$

$$\Rightarrow \mathbb{E}\|\hat{X} - X\|_2 \le C\sqrt{\frac{d\log(d)}{p}}\,\|X\|_{max}$$

Step 2: Pass to Frobenius.

$$\|\hat{X} - X\|_F \leq \sqrt{2r}\,\|\hat{X} - X\|_2$$

($\hat{X}$: rank $2r$)

$$\Rightarrow \mathbb{E}\,\frac{1}{d}\|\hat{X} - X\|_F \leq C\sqrt{\frac{rd\log(d)}{pd^2}}\,\|X\|_{\max}.$$

(p: m)

∎

Chapter 5

Quadratic forms and mean estimation

- mean estimation for $N(\mu, \Sigma)$
- Hanson-Wright inequality
- norm concentration w/out isotropy
- projections of random vectors.

As motivation, let's find the sample complexity of estimating the mean of a multivariate normal.

Thm: Suppose $x_i \overset{\text{i.i.d.}}{\sim} N(\mu, \Sigma)$. Then

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right\|_2 \le \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \sqrt{\frac{2\log\left(\frac{2}{\delta}\right)\|\Sigma\|_2}{n}}\right] \ge 1-\delta$$

Read Lugosi-Mendelson survey! $\forall \delta > 0$

pf: Fix a matrix $A \in \mathbb{S}^d$ and $z \sim N(0, I)$

Then $\|Az\|_2^2 = z^T A^T A z = \sum_{i,j} (A^TA)_{ij} z_i z_j$

Write the eigenvector-decomposition

$$A = \sum_{i=1}^{d} \gamma_i u_i u_i^T$$

Then

$$z^T A^T A z = z^T \left(\sum_{i=1}^{d} \gamma_i^2 u_i u_i^T\right) z = \sum_{i=1}^{d} \gamma_i^2 \langle u_i, z\rangle^2$$

$\langle u_i, z\rangle \sim N(0,1)$

Notice $\{\langle u_i, z\rangle\}_{i=1}^{m}$ are independent.

Bernstein $\Rightarrow$

$$\mathbb{P}\left[\left|\|Az\|_2^2 - \|A\|_F^2\right| > t\right]$$

$$\leq 2\exp\left[-\frac{1}{2}\left(\frac{t^2}{\sum_i \gamma_i^4} \wedge \frac{4t}{\|A\|_2^2}\right)\right]$$

$z \sim N(0, I)$

Finally set

$$A = \frac{1}{\sqrt{n}}\Sigma^{1/2}, \qquad z = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \Sigma^{-1/2}(X_i - \mu)$$

Then

$$\mathbb{P}\left[\left|\left\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right\|^2 - \frac{\operatorname{tr}(\Sigma)}{n}\right| > t\right]$$

$$\leq 2\exp\left[-\frac{1}{2}\left(\frac{t^2}{\|\Sigma\|_F^2/n} \wedge \frac{t}{\|\Sigma\|_2}\right)n\right]$$

∎

We will now generalize results of this type to sub-Gaussians. Namely we want to understand

$$\left| x^T A x - \mathbb{E} x^T A x \right|$$

where the coordinates of $x = (x_1, \dots, x_n)$ are independent and subGaussian.

Thm (Hanson-Wright)
Let $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ be a random vector with x_i independent, mean zero, and σ-subGaussian. Let $A \in \mathbb{R}^{n \times n}$ be arbitrary. Then

$$\mathbb{P}\left[\left| x^T A x - \mathbb{E} x^T A x \right| \geq t \right] \leq 2 \exp\left(-c \min\left(\frac{t^2}{\sigma^4 \|A\|_F^2}, \frac{t}{\sigma^2 \|A\|_2} \right) \right)$$

See Vershynin 6.1-6.2 for the proof

Cor: Let $B \in \mathbb{R}^{m \times n}$ be a fixed matrix. and let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ have independent σ-subGaussian coordinates with $\mathbb{E} X_i = 0$ and $\mathbb{E} X_i^2 = 1$

Then

$$\mathbb{P}\left[\left| \|BX\|_2 - \|B\|_F \right| > t \right] \leq 2 e^{-\frac{ct^2}{\sigma^4 \|B\|_2^2}}$$

pf: Apply Hanson-Wright with $A = B^T B$. Compute

- $X^T A X = \|BX\|_2^2$, $\mathbb{E} X^T A X = \mathbb{E}\, \mathrm{tr}(A X X^T) = \mathrm{tr}(B^T B) = \|B\|_F^2$
- $\|A\|_2 = \|B\|_2^2$, $\|A\|_F = \|B^T B\|_F \leq \|B\|_2 \cdot \|B\|_F$

$$\Rightarrow \mathbb{P}\left[\left| \|BX\|_2^2 - \|B\|_F^2 \right| \geq u \right] \leq 2 \exp\left(-\frac{c}{\sigma^4} \left(\frac{u^2}{\|B\|_2^2 \|B\|_F^2} \wedge \frac{u}{\|B\|_2^2} \right) \right)$$

$\Rightarrow$ continue like in the concentration of the norm $\|X\|$. ∎

Cor: Let $E \subset \mathbb{R}^n$ be a subspace with $\dim(E) = d$. Let $X = (X_1, \dots, X_n)$ have independent σ-subGaussian coordinates with $\mathbb{E}X_i = 0$, $\mathbb{E}X_i^2 = 1$.

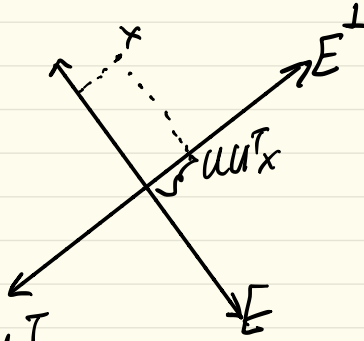
Then

$$\mathbb{P}\left[\left|d(X,E) - \sqrt{n-d}\right| > t\right] \le 2\exp\left(\frac{-ct^2}{\sigma^4}\right)$$

pf: Let $U \in \mathbb{R}^{n \times (n-d)}$ have as columns an orthonormal basis for $E^\perp$

Then $\text{Proj}_{E^\perp} = UU^T$ and

$$\text{dist}(X, E) = \|UU^T X\|_2$$



Then

$$\mathbb{E}\,\text{dist}^2(X, E^\perp) = \mathbb{E}\, X^T UU^T UU^T X$$
$$= \mathbb{E} X^T UU^T X = \text{tr}(UU^T) = n-d$$

$$\Rightarrow \mathbb{E}\,\text{dist}(X, E^\perp) \le \sqrt{n-d}.$$ Apply previous result $B = UU^T$ ∎

Chapter 6

Gaussian Processes

- Concentration of Lipschitz functions of the uniform spherical random vector
- Basics of Gaussian processes.
- Slepian's Inequality.
- Sudakov-Fernique Inequality

In this chapter, we explore a number of constructions built from gaussians. We begin with the following

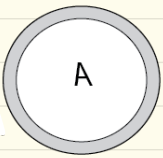
Thm (*): Let $f: \sqrt{n}S^{n-1} \to \mathbb{R}$ be Lipschitz with constant L_f and let $x \sim \text{Unif}(\sqrt{n}S^{n-1})$. Then

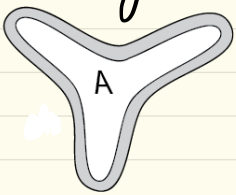
$$\mathbb{P}[|f(x) - \mathbb{E}f(x)| \geq t] \leq 2\exp\left(-\frac{ct^2}{L_f^2}\right)$$

The proof will be based on the isoparametric inequality

Thm: Let $\varepsilon > 0$. Then among all sets $A \in S^{n-1}$ with prescribed area $\mu^{n-1}(A)$, the spherical caps minimize the area of the blow-up

$$A_\varepsilon = \{x \in S^{n-1} : \exists y \in A \text{ s.t. } \|x-y\|_2 \leq \varepsilon\}$$





Recall that $\sqrt{n}\,S^{n-1}$ is a natural scaling of the sphere because $X \sim \mathrm{Unif}(\sqrt{n}\,S^{n-1})$ is isotropic and c-subGaussian for a constant c.

[Vershynin Thm 3.4.6: $X \sim \sqrt{n}\,\frac{g}{\|g\|_2}$ where $g \sim N(0, I)$ and $\|g\|_2 \sim \sqrt{n}$ w.h.p.]

Lemma: Fix $A \subset \sqrt{n}\,S^{n-1}$ with $\mu^{n-1}(A) \geq \frac{1}{2}$ where μ^{n-1} denotes the normalized area on $\sqrt{n}\,S^{n-1}$. Then

$$\mu^{n-1}(A_t) \geq 1 - 2e^{-ct^2} \qquad \forall t \geq 0.$$

pf: Define

$$H := \{x \in \sqrt{n}\,S^{n-1} : x_1 \leq 0\}$$

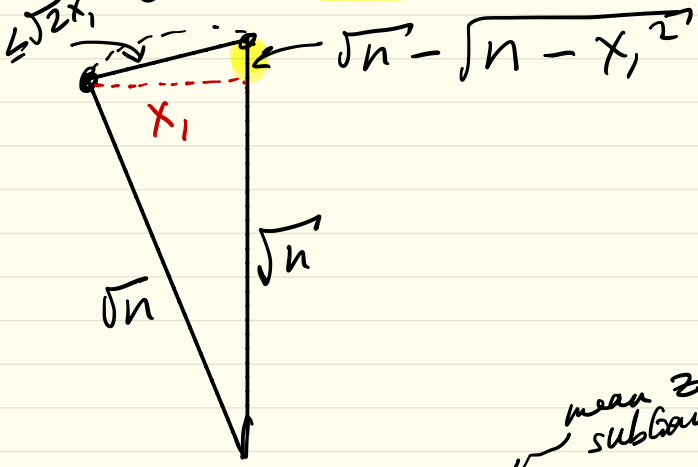
By assumption $\mu^{n-1}(A) \geq \frac{1}{2} = \mu^{n-1}(H)$

$\Rightarrow \mu^{n-1}(A_t) \geq \mu^{n-1}(H_t)$ ← let's compute

$$\mu^{n-1}(H_t) = \mathbb{P}[X \in H_t]$$

Notice

$$H_t \supseteq \left\{ x \in \sqrt{n}\, S^{n-1} : x_1 \le \frac{t}{\sqrt{2}} \right\}$$



$$\Rightarrow \mu^{n-1}(H_t) \ge \mathbb{P}\left[X_1 \le \frac{t}{\sqrt{2}}\right] \ge 1 - 2\exp(-ct^2)$$

(mean zero subGaussian)

pf of Thm ⊗:

Let M denote a median of $f(X)$, namely

$$\mathbb{P}[f(X) \le M] \ge \frac{1}{2} \quad \text{and} \quad \mathbb{P}[f(X) \ge M] \le \frac{1}{2}$$

Define

$$A := \{x \in \sqrt{n}S^{n-1} : f(x) \le M\}$$

Since $\mathbb{P}[X \in A] \ge \frac{1}{2}$, we just showed

$$\mathbb{P}[X \in A_t] \ge 1 - 2\exp(-ct^2)$$

Notice

$$\mathbb{P}[X \in A_t] \le \mathbb{P}[f(X) \le M + L_f t]$$

since for $\forall x \in A_t \; \exists y \in A$ s.t.

$$f(x) \le f(y) + L_f \|x - y\| \le M + L_f t$$

So $\mathbb{P}[f(X) \le M + L_f t] \le 1 - 2\exp(-ct^2)$

Replace f by $-f$ proves the lower deviation inequality. Finally, the centering inequality (Vershynin 2.6.8)

$$\Rightarrow \|f(X) - \mathbb{E}f(X)\|_{\psi_2} = \|(f(X) - M) - \mathbb{E}(f(X) - M)\|_{\psi_2}$$

$$\le C\|f(X) - M\|_{\psi_2} \le CL_f$$

∎

Cor: If $f: S^{n-1} \to \mathbb{R}$ is Lipschitz and $X \sim \text{Unif}(S^{n-1})$, then

$$\mathbb{P}[|f(X) - \mathbb{E}f(X)| \geq t] \leq 2\exp\left(-\frac{cnt^2}{L_f^2}\right)$$

This result extends to many other settings

Thm: Let $\varepsilon > 0$. Then among all sets $A \in \mathbb{R}^n$ with prescribed Gaussian measure $\gamma_n(A)$, the half-spaces minimize the Gaussian measure $\gamma_n(A_\varepsilon)$

Thm: (HW) Consider a $X \sim N(0, I)$ and a Lipschitz $f: \mathbb{R}^n \to \mathbb{R}$. Then

$$\mathbb{P}[|f(X) - \mathbb{E}f(X)| > t] \leq \exp\left(-\frac{ct^2}{L_f^2}\right)$$

See Vershynin 5.2 for other examples of probability spaces on which functions concentrate.

Typical results require $\underline{X}$ to be very symmetric. One important exception is the following.

Thm (Talagrand)
Consider a random $\underline{X} = (X_1, \dots, X_n) \in \mathbb{R}^n$ whose coordinates are independent and satisfy $|X_i| \leq 1$ almost surely. Then for any convex Lipschitz $f : [0,1]^n \to \mathbb{R}$ it holds

$$\mathbb{P}\left[|f(\underline{X}) - \mathbb{E} f(\underline{X})| \geq t \right] \leq 2 \exp\left(- \frac{c n t^2}{L_f^2} \right)$$

[See Boucheron, Lugosi, Massart for a proof]

Defn: A random process is a collection of random variables $\{X_t : t \in T\}$ indexed by some set $t \in T$.

Let $\{X_t\}_{t \in T}$ be a random process. Assume for simplicity $\mathbb{E} X_t = 0 \ \forall t \in T$. Define the covariance function

$$\Sigma(t,s) := \operatorname{cov}(X_t, X_s) = \mathbb{E} X_t X_s$$

and the increments

$$d(t,s) := \sqrt{\mathbb{E}(X_t - X_s)^2}$$

Rem: $d(\cdot,\cdot)$ is a pseudometric on T.

Rem: $d^2(t,s) = \Sigma(t,t) + \Sigma(s,s) - 2\Sigma(t,s)$

Defn: $\{X_t\}_{t\in T}$ is called a Gaussian process if for any finite $T_0 \subset T$, the random vector $(X_t)_{t\in T_0}$ has a Gaussian distribution.

Main example is the canonical Gaussian Process indexed by $T \subseteq \mathbb{R}^n$:

$$X_t = \langle g, t\rangle \quad \forall t \in T$$

where $g \sim N(0, I)$.

Notice

$$d(t,s) = \sqrt{\mathbb{E}\langle g, t-s\rangle^2} = \sqrt{\mathbb{E}\,\mathrm{tr}\left((t-s)(t-s)^T gg^T\right)} = \|t-s\|_2$$

Our goal is to estimate

$$\mathbb{E}\sup_{t\in T} X_t$$

or more formally

$$\sup_{T_0 \subset T \text{ finite}} \mathbb{E}\sup_{t\in T_0} X_t.$$

Thm (Slepian's inequality)

Let $\{X_t\}_{t\in T}$, $\{Y_t\}_{t\in T}$ be mean-zero Gaussian processes. Assume $\forall t,s\in T$:

$$\mathbb{E}X_t^2 = \mathbb{E}Y_t^2 \quad \text{and} \quad \mathbb{E}(X_t - X_s)^2 \le \mathbb{E}(Y_t - Y_s)^2$$

Then

$$\mathbb{P}\left[\sup_{t\in T} X_t \ge \delta\right] \le \mathbb{P}\left[\sup_{t\in T} Y_t \ge \delta\right] \quad \forall \delta$$

Therefore

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t.$$

Proof Strategy:

Assume T is finite and set $X=(X_t)$, $Y=(Y_t)$,

$$Z(u) = \sqrt{u}\,X + \sqrt{1-u}\,Y \quad \text{for } u\in[0,1].$$

We will show that

$u \longmapsto \mathbb{P}\left[\max_i Z_i(u) \ge \delta\right]$ is nonincreasing

Lemma (HW)

Let $X \sim N(0, \Sigma)$. Then for any differentiable $f: \mathbb{R}^n \to \mathbb{R}$, if holds:

$$\mathbb{E}\, X f(X) = \Sigma \cdot \mathbb{E}\, \nabla f(X)$$

or equivalently

$$\mathbb{E}\, X_i f(X) = \sum_{j=1}^{n} \Sigma_{ij}\, \mathbb{E}\, \frac{\partial f}{\partial x_j}(X) \quad \forall i = 1, \dots, n$$

Lemma: (Interpolation) Consider independent $X \sim N(0, \Sigma^X)$, $Y \sim N(0, \Sigma^Y)$, and set

$$Z(u) = \sqrt{u}\, X + \sqrt{1-u}\, Y \quad \text{for } u \in [0,1].$$

Then for any twice-differentiable $f: \mathbb{R}^n \to \mathbb{R}$ if holds:

$$\frac{d}{du}\, \mathbb{E}\, f(Z(u)) = \frac{1}{2} \left\langle \Sigma^X - \Sigma^Y, \mathbb{E}\, \nabla^2 f(Z(u)) \right\rangle \quad \forall u \in (0,1)$$

pf: Chain rule:

$$\frac{d}{du}\, \mathbb{E}\, f(Z(u)) = \mathbb{E} \left\langle \frac{d}{du} Z(u), \nabla f(Z(u)) \right\rangle$$

$$= \frac{1}{2\sqrt{u}}\, \underbrace{\mathbb{E} \langle X, \nabla f(Z(u)) \rangle}_{(1)} - \frac{1}{2\sqrt{1-u}}\, \underbrace{\mathbb{E} \langle Y, \nabla f(Z(u)) \rangle}_{(2)}$$

$$① = \sum_{i=1}^{n} \mathbb{E} X_i \frac{\partial f}{\partial x_i}(\sqrt{u}X + \sqrt{1-u}\,y)$$

$$= \mathbb{E}_y \sum_{i=1}^{n} \mathbb{E}_X X_i \frac{\partial f}{\partial x_i}(\sqrt{u}X + \sqrt{1-u}\,y)$$

$$= \mathbb{E}_y \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \Sigma^X_{ij} \mathbb{E}_X \sqrt{u} \cdot \frac{\partial^2 f}{\partial x_i \partial x_j}(z(u)) \right)$$

$$= \sqrt{u}\, \mathbb{E} \langle \Sigma^X, \nabla^2 f(z(u)) \rangle$$

Similar computation gives

$$② = \sqrt{1-u}\, \mathbb{E} \langle \Sigma^Y, \nabla^2 f(z(u)) \rangle$$

Lemma: Let $X \sim N(0, \Sigma^X)$, $y \sim N(0, \Sigma^Y)$ be independent. Assume

$$\mathbb{E} X_i^2 = \mathbb{E} y_i^2, \quad \mathbb{E}(X_i - X_j)^2 \le \mathbb{E}(y_i - y_j)^2 \quad \forall i,j$$

Consider $f: \mathbb{R}^n \to \mathbb{R}$ with $[\nabla^2 f]_{ij} \ge 0 \ \forall i \ne j$

Then $\mathbb{E} f(X) \ge \mathbb{E} f(y)$

pf: Notice

$$\Sigma^X_{ii} = \Sigma^Y_{ii} \quad \text{and} \quad \Sigma^X_{ij} \geq \Sigma^Y_{ij}$$

We can assume X and Y are independent. Apply the previous lemma. ∎

Finally, we are ready to prove Slepian's inequality.

Thm (Slepian's inequality)

Let $\{X_t\}_{t\in T}$, $\{Y_t\}_{t\in T}$ be mean-zero Gaussian processes. Assume $\forall t,s \in T$:

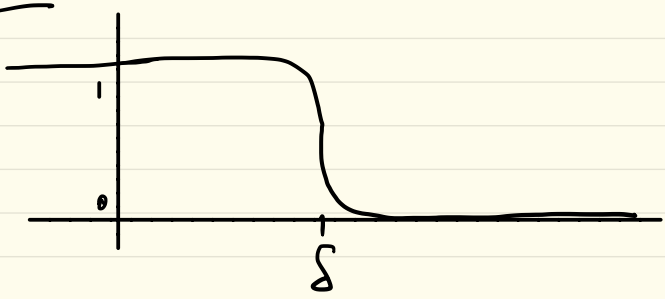
$$\mathbb{E}X_t^2 = \mathbb{E}Y_t^2 \quad \text{and} \quad \mathbb{E}(X_t - X_s)^2 \leq \mathbb{E}(Y_t - Y_s)^2$$

Then

$$\mathbb{P}\left[\sup_{t\in T} X_t \geq \delta\right] \leq \mathbb{P}\left[\sup_{t\in T} Y_t \geq \delta\right] \quad \forall \delta$$

Therefore $\mathbb{E}\sup_{t\in T} X_t \leq \mathbb{E}\sup_{t\in T} Y_t$.

pf: Let $h: \mathbb{R} \to [0,1]$ be the function



We can ensure h is a C^2-smooth approximation to $\mathbb{1}_{(-\infty, \delta)}$.

Define $f: \mathbb{R}^n \to \mathbb{R}$ by $f(x) = h(x_1) \cdots h(x_n)$

[Note f approximates $\mathbb{1}_{\{\max_i x_i \leq \delta\}}$]

Note $$\frac{\partial^2 f}{\partial x_i \partial x_j} = \underbrace{h'(x_i)}_{0} \cdot \underbrace{h'(x_j)}_{0} \prod_{k \neq i,j} h(x_k) \geq 0$$

So

$$\mathbb{E} f(X) \geq \mathbb{E} f(Y)$$

Pass to the limit [e.g. dominated convergence]

∎

Thm (Sudakov-Fernique)

Let $\{X_t\}_{t\in T}, \{Y_t\}_{t\in T}$ be two mean-zero Gaussian processes. Assume

$$\mathbb{E}(X_t - X_s)^2 \le \mathbb{E}(Y_t - Y_s)^2 \quad \forall s,t \in T.$$

Then

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t$$

pf: As before assume $X \sim N(0, \Sigma^X)$, $Y \sim N(0, \Sigma^Y)$

Fix $\beta > 0$ and define

$$f_\beta(x) = \frac{1}{\beta}\log\sum_{i=1}^n e^{\beta x_i}$$

You can check

$$\max_{i=1,\dots,n} x_i \le f_\beta(x) \le \max_{i=1,\dots,n} x_i + \frac{\log n}{\beta}$$

Apply Gaussian interpolation to deduce

$$\frac{d}{du}\mathbb{E} f_\beta(Z(u)) \le 0 \quad \text{(HW)}$$

□

Cor: (Gaussian Contraction Inequality)
Fix a set $T \subset \mathbb{R}^n$ and let $g_1, \dots, g_n \overset{iid}{\sim} N(0,1)$
Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz.
Then

$$\mathbb{E} \sup_{t \in T} \sum_{i=1}^n g_i \phi_i(t_i) \leq \mathbb{E} \sup_{t \in T} \sum_{i=1}^n g_i t_i$$

pf: Apply Sudakov-Fernique. ∎

We will now use Sudakov-Fernique to prove a useful lower bound on

$$\sup_{t \in T} X_t$$

Recall: $d(t,s) = \sqrt{\mathbb{E}(X_t - X_s)^2}$
defines a pseudo-metric on T.

Thm: (Sudakov's minoration inequality)
Let $\{X_t\}_{t\in T}$ be a mean-zero Gaussian process. Then

$$\mathbb{E}\sup_{t\in T} X_t \geq C\varepsilon\sqrt{\log_2 N(T,d,\varepsilon)} \quad \forall \varepsilon > 0,$$

where $N(T,d,\varepsilon)$ denotes the covering number of T is $d(\cdot,\cdot)$ metric.

pf: Suppose $N := N(T,d,\varepsilon)$ is finite. You'll consider the infinite case for HW.
Let M be a maximal ε-packing. Then we know $|M| \geq N$. It suffices to show

$$\mathbb{E}\sup_{t\in M} X_t \geq C\varepsilon\sqrt{\log N}$$

Define $\{Y_t\}_{t\in M}$ by $Y_t = \frac{\varepsilon}{\sqrt{2}} g_t$ where $g_t \sim N(0,1)$ independent.

Fix $t, s \in M$. Then

$$\mathbb{E}(X_t - X_s)^2 = d(t,s)^2 \geq \varepsilon^2$$

while

$$\mathbb{E}(Y_t - Y_s)^2 = \frac{\varepsilon^2}{2}\mathbb{E}(g_t - g_s)^2 = \varepsilon^2$$

Therefore Sudakov-Fernique

$$\Rightarrow \mathbb{E}\sup_{t\in M} X_t \geq \mathbb{E}\sup_{t\in M} Y_t \qquad \text{HW1}$$

$$= \frac{\varepsilon}{\sqrt{2}}\,\mathbb{E}\max_{t\in M} g_t \geq c\varepsilon\sqrt{\log M}$$

We next study a particular type of a Gaussian process, which is an analogue of the Rademacher complexity.

Defn: The Gaussian width of a set $T \subseteq \mathbb{R}^n$ is $w(T) := \mathbb{E}\sup_{x\in T}\langle g, x\rangle$ where $g \sim N(0, I)$.

Properties:

(1) $w(T) < \infty$ iff T is bounded.

(2) Affine Invariance:

$$w(UT+y) = w(T)$$

for any orthogonal $U \in \mathbb{R}^{n\times n}$ and $y \in \mathbb{R}^n$.

(3) $w(\text{conv}(T)) = w(T)$

(4) $w(T+S) = w(T) + w(S)$

$w(aT) = |a| w(T)$ $\forall a \in \mathbb{R}$.

(5) $w(T) = \frac{1}{2} w(T-T) = \frac{1}{2} \mathbb{E}_{x,y\in T} \langle g, x-y\rangle$

(6) $\frac{1}{\sqrt{2\pi}} \text{diam}(T) \le w(T) \le \frac{\sqrt{n}}{2} \text{diam}(T)$

(7) For any 1-Lipschitz $\varphi_i : \mathbb{R} \to \mathbb{R}$ it holds $w(\varphi \circ T) \le w(T)$

pf: (1)-(4) follow from defn of support functions. To see (5) observe

$$w(T) = \frac{1}{2}(w(T) + w(T)) = \frac{1}{2}(w(T) + w(-T)) \overset{(4)}{=} \frac{1}{2} w(T-T).$$

To see (6), fix $x, y \in T$. Then

$$\omega(T) \geq \frac{1}{2}\mathbb{E}\max(\langle x-y, g\rangle, \langle y-x, g\rangle)$$
$$= \frac{1}{2}\mathbb{E}|\langle x-y, g\rangle| = \frac{1}{2}\sqrt{\frac{2}{\pi}}\|x-y\|$$

Take $\sup_{x,y\in T}$.

Next

$$\omega(T) = \frac{1}{2}\mathbb{E}_g \sup_{x,y\in T}\langle g, x-y\rangle$$
$$\leq \frac{1}{2}\sup_{x,y\in T}\|x-y\|_2 \cdot \mathbb{E}_g\|g\| \leq \frac{\sqrt{n}}{2}\operatorname{diam}(T)$$

(7) is the contraction inequality we already proved.

Lemma: For any $A \in \mathbb{R}^{m\times n}$, it holds

$$\omega(AT) \leq \|A\|_2\,\omega(T)$$

[Hint: singular value decomposition + contraction]

Thm: (Gaussian vs Rademacher Complexity)
For any set $T \subseteq \mathbb{R}^d$, it holds:

$$\frac{\omega(T)}{2\sqrt{\log d}} \le \mathcal{R}(T) \le \sqrt{\frac{\pi}{2}}\,\omega(T)$$

pf:
$$\mathcal{R}(T) = \mathbb{E}_\varepsilon \sup_{x\in T} \sum_{i=1}^d \varepsilon_i x_i$$
$$= \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_\varepsilon \sup_{x\in T} \sum_{i=1}^d \varepsilon_i \cdot \mathbb{E}|g_i| x_i$$
$$\le \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{\varepsilon,g} \sup_{x\in T} \sum_{i=1}^d \varepsilon_i |g_i| x_i$$
$\varepsilon_i|g_i| \overset{\mathcal{L}}{=} g_i$
$$= \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_g \sup_{x\in T} \sum_{i=1}^d g_i x_i$$
$$= \sqrt{\frac{\pi}{2}}\,\omega(T)$$

Conversely
$$\omega(T) = \mathbb{E}_g \sup_{x\in T} \sum_{i=1}^d g_i x_i$$
$$= \mathbb{E}_{g,\varepsilon} \sup_{x\in T} \sum_{i=1}^d \varepsilon_i g_i x_i$$

$$= \mathbb{E}_g \mathcal{R}(g \circ T) \overset{\text{Contraction}}{\leq} \mathbb{E}_g \max_{i=1\ldots d} |g_i| \mathcal{R}(T)$$

$$\leq 2\sqrt{\log d}\, \mathcal{R}(T) \quad \square$$

There is a close cousin of Gaussian width called spherical width.

Defn: The <u>spherical width</u> of $T \subseteq \mathbb{R}^d$ is $s(T) = \mathbb{E}_\theta \sup_{x \in T} \langle \theta, x \rangle$

where $\theta \sim \text{Unif}(S^{n-1})$

Lemma: $(\sqrt{n} - C)\, s(T) \leq w(T) \leq (\sqrt{n} + C)\, s(T)$

pf: $w(T) = \mathbb{E}_g \sup_{x \in T} \langle g, x \rangle = \mathbb{E}_g \sup_{x \in T} \langle \|g\| \cdot \frac{g}{\|g\|}, x \rangle$

$$= \mathbb{E}_g \|g\|_2 \cdot s(T)$$

Let's compare the sizes of B_2^d, B_1^d, B_∞^d.

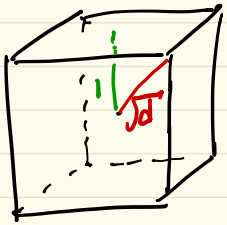
	$\log N(\cdot, \delta)$	$\omega(\cdot)$	R_I	R_c	# vertices
B_2^d	$d \log(\frac{1}{\delta})$	$\sqrt{d}$	—	—	—
B_1^d	$\lesssim \frac{\log d}{\delta^2}$ ②	$\sqrt{\log d}$ ①	$\frac{1}{\sqrt{d}}$	1	d
B_∞^d	$\lesssim \frac{d^2}{\delta^2}$ ④	d ③	1	$\sqrt{d}$	2^d

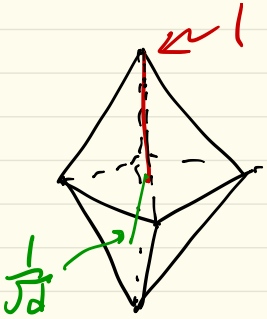
pf: ① $\mathbb{E}_g \sup_{x \in B_1} \langle g, x \rangle = \mathbb{E}_g \|g\|_\infty \simeq \sqrt{\log d}$

② Sudakov $\Rightarrow \omega(B_1^d) \geq \max_\delta c\delta \sqrt{\log N(B_1^d, \delta)}$

③ $\omega(B_\infty^d) = \mathbb{E}_g \|g\|_1 = \sqrt{\frac{2}{\pi}} d$

④ Sudakov





$\omega(B_2) \simeq \omega(\text{circumscribed ball})$ $\omega(B_1) \simeq \omega(\text{inscribed ball})$

Algebraic Dimension is highly unstable

Def'n: Define the **stable dimension** of T by

$$d(T) = \frac{h(T-T)^2}{\operatorname{diam}(T)^2} \sim \frac{w(T)^2}{\operatorname{diam}(T)^2}$$

where $h(S) = \mathbb{E}_g \sup_{x \in S} \langle g, x \rangle$

Lemma: (HW)
$$2w(T) \le h(T-T) \le 2C\, w(T)$$

Lemma: $d(T) \le \dim(\operatorname{span} T)$

pf: We can assume $T \subseteq \operatorname{Span}\{e_1, \dots, e_k\}$ where $k = \dim(T)$. Then

$$h(T-T)^2 = \mathbb{E}_g \sup_{x \in T-T} \langle g, x \rangle^2 \le \mathbb{E}_{g \sim N(0, I_n)} \|g\|^2 \cdot \operatorname{diam}(T)^2.$$

$$= k \operatorname{diam}(T)^2 \quad \square$$

Exercise (HW): Let $A \in \mathbb{R}^{m \times n}$. Then

$$d(AB_2^d) = \frac{\|A\|_F^2}{\|A\|_2^2}$$

- Stable rank of matrix
- changes gradually with perturbations

Random Projections of Sets

Thm: Let $T \subset \mathbb{R}^n$ and let P be a projection onto a uniformly random m-dimensional subspace. Then w.p. $1 - 2e^{-m}$, we have

$$\operatorname{diam}(PT) \leq C\left(s(T) + \sqrt{\frac{m}{n}} \operatorname{diam}(T)\right)$$

Lemma (HW): Let P be a projection in $\mathbb{R}^n$ onto a uniformly random m-dimensional subspace. Let Q be an $m \times n$ orthogonal matrix drawn uniformly. Then

(a) For any $x \in \mathbb{R}^n$, it holds: $\|Px\|_2$ and $\|Qx\|_2$ have the same distribution.

(b) For any z with $\|z\| = 1$, it holds $Q^T z \sim \mathrm{Unif}(S^{n-1})$

pf of thm: Covering argument. Can replace P by Q by Lemma. Can assume $\mathrm{Dom}(T) \subseteq S^{m-1}$

Approximation: Let $\mathcal{N}$ be $\frac{1}{2}$-net of S^{m-1}. Then $|\mathcal{N}| \leq 5^m$.

Then

$$\operatorname{diam}(QT) \le \sup_{x \in T-T} \|Qx\|_2$$

$$= \sup_{x \in T-T} \sup_{z \in S^{m-1}} \langle Qx, z \rangle$$

$$\le 2 \sup_{x \in T-T} \left[\sup_{z \in \mathcal{N}} \langle x, Q^T z \rangle \right]$$

$$\le 2 \sup_{z \in \mathcal{N}} \left[\sup_{x \in T-T} \langle x, Q^T z \rangle \right]$$

Fix $z \in \mathcal{N}$. Then $Q^T z \sim \text{Unif}(S^{n-1})$.

Note $\mathbb{E}_Q \sup_{x \in T-T} \langle x, Q^T z \rangle = s(T-T) = 2s(T)$

Concentration: The function $\theta \mapsto \sup_{x \in T-T} \langle \theta, x \rangle$ is 1-Lipschitz on S^{n-1} $\Rightarrow$ Concentration

$$\mathbb{P}\left[\sup_{x \in T-T} \langle x, Q^T z \rangle \ge 2s(T) + t \right] \le 2 \exp(-cnt^2)$$

Union bound:

$$\mathbb{P}\Big[\max_{z\in\mathcal{N}} \sup_{x\in T-T} \langle Q^T z, x\rangle \geq 2s(T) + t\Big]$$

$$\leq 5^m \cdot 2\exp(-cnt^2)$$

$$= 2\exp(m\log(5) - cnt^2)$$

set $t = c'\sqrt{\frac{m}{n}}$

$$\leq e^{-m}. \qquad \blacksquare$$

Exercise (HW) For all bounded sets:

$$\mathbb{E}\,\mathrm{diam}(PT) \geq C\left(s(T) + \sqrt{\frac{m}{n}}\,\mathrm{diam}(T)\right)$$

The phase transition:

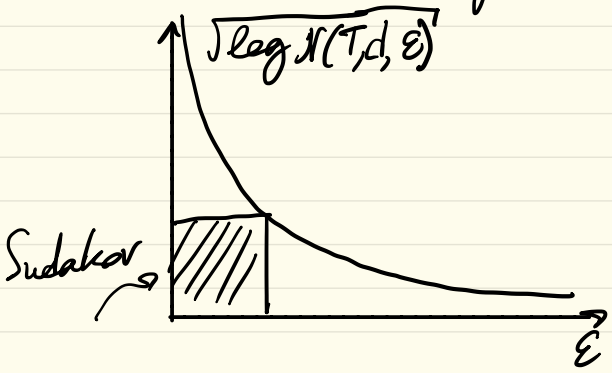
$$\mathrm{diam}(PT) \leq C \begin{cases} \sqrt{\frac{m}{n}}\,\mathrm{diam}(T), & \text{if } m \geq d(T) \\ s(T), & \text{if } m \leq d(T) \end{cases}$$

Chaining

We now prove the following upper bound

$$\mathbb{E} \sup_{t \in T} X_t \leq CK \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon$$

on sub Gaussian processes.



Defn: Consider a random process $\{X_t\}_{t \in T}$ on a metric space (T,d) if $\exists K \geq 0$ s.t.

$$\|X_t - X_s\|_{\psi_2} \leq K\, d(t,s) \quad \forall t,s$$

($\|\cdot\|_{\psi_2}$: sub Gaussian norm / canonical)

Ex: If X_t is a Gaussian process and $d(t,s) = \|t-s\|_2$, then $K=1$.

Thm: (Discrete version)
Let $\{X_t\}$ be a mean-zero subGaussian process on (T,d). Then

$$\mathbb{E}\sup_{t\in T} X_t \le CK\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T,d,2^{-k})}$$

[**Idea**: Let $\pi(\cdot)$ be project on ε-net repeat
$\Rightarrow \mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} X_{\pi(t)} + \mathbb{E}\sup_{t\in T}\left[X_t - X_{\pi_t}\right]$]

pf: WLOG assume $K=1$, T finite.
Set $\varepsilon_k = 2^{-k}$, $k\in\mathbb{Z}$
and let T_k be an ε_k-net of T.
Since T is finite $\exists k_0, k_\infty$ s.t.
$T_{k_0} = \{t_0\}$ for some $t_0\in T$
$T_{k_\infty} = T$
For $t\in T$, let $\pi_k(t)\in T_k$ satisfy
$d(t,\pi_k(t)) \le \varepsilon_k$.

Since $\mathbb{E} X_{t_0} = 0$, we can write

$$\mathbb{E} \sup_{t \in T} X_t = \mathbb{E} \sup_{t \in T} \left[X_t - X_{t_0} \right]$$

Write

$$X_t - X_{t_0} = \sum_{k=k_0+1}^{k_\infty} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \right)$$

So

$$\mathbb{E} \sup_{t \in T} (X_t - X_{t_0}) \leq \sum_{k=k_0+1}^{k_\infty} \mathbb{E} \sup_{t \in T} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \right)$$

supremum is over $|T_k| \cdot |T_{k-1}| \leq |T_k|^2$

For fixed t, have

$$\begin{aligned} \| X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \|_{\psi_2} &\leq d(\pi_k(t), \pi_{k-1}(t)) \\ &\leq d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \\ &\leq \varepsilon_k + \varepsilon_{k-1} \\ &\leq 2 \varepsilon_{k-1} \end{aligned}$$

So $\mathbb{E} \sup_{t \in T} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \right) = C \varepsilon_{k-1} \sqrt{\log |T_k|}$

Exercise 2.3.10. (HW)

So

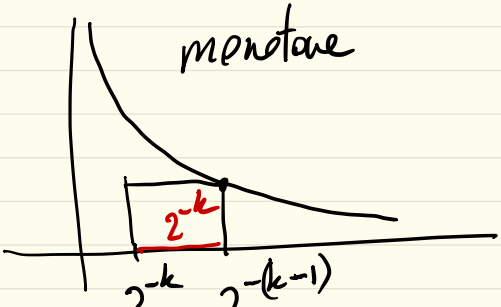
$$\mathbb{E} \sup_{t \in T} X_t - X_0 \leq C_1 \sum_{k=k_0}^{k_0 - 1} 2^{-k} \sqrt{\log N(T, d, 2^{-k})}$$

∎

Thm: (Dudley's Integral)
Let $\{X_t\}_{t \in T}$ be a mean zero random process on (T, d) with subGaussian increments. Then

$$\mathbb{E} \sup_{t \in T} X_t \leq CK \int_0^\infty \sqrt{\log N(T, d, \varepsilon)} \, d\varepsilon$$

pf:



∎

Thm (HW)
Let $\{X_t\}_{t \in T}$ be a subGaussian process on (T,d), possibly not centered. Then w.p. $1 - 2\exp(-\delta^2)$ have:

$$\sup_{t,s \in T} |X_t - X_s| \leq CK\left(\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon + \delta \cdot \mathrm{diam}(T) \right)$$

Thm: (Two-sided Sudakov)
Let $\{X_t\}_{t \geq 0}$ be a canonical Gaussian process on a set T. Define

$$\varphi(T) = \sup_{\varepsilon \geq 0} \varepsilon \sqrt{\log N(T,\varepsilon)}$$

Then

$$c_2\, \varphi(T) \leq w(T) \leq c_1 (\log n)\, \varphi(T)$$

[Read in Vershynin]

Improved Uniform Laws

Let $\mathcal{F}$ be a class of indicator functions with VC-dimension v.
We showed

$$\mathcal{R}_n(\mathcal{F}) \leq 2\sqrt{\frac{v\log(n+1)}{n}}$$

Let's get a better bound on

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_x \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)$$

Define $Z_f := \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(x_i)$ and consider the process $\{Z_f\}_{f\in\mathcal{F}}$

Let's compute

$$\|Z_f - Z_g\|_{\psi_2} = \frac{1}{\sqrt{n}}\left\|\sum_{i=1}^n \varepsilon_i (f(x_i) - g(x_i))\right\|_{\psi_2}$$

$$\leq \frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^n (f(x_i) - g(x_i))^2}\,\varepsilon \frac{1}{\sqrt{n}}\|f-g\|_2.$$

So Dudley

$$\Rightarrow \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_i \varepsilon_i f(x_i)\right| \le \frac{c}{\sqrt{n}} \int_0^2 \sqrt{\log N(\mathcal{F}, \delta, \|\cdot\|_2)}\, d\delta$$

Thm: $N(\mathcal{F}, \delta, \|\cdot\|_2) \le \left(\frac{2}{\delta}\right)^{cv}$

[Vershynin Thm 8.3.18]

So continuing

$$\le \frac{c}{\sqrt{n}} \int_0^2 \sqrt{v \ln\left(\frac{2}{\varepsilon}\right)}\, d\varepsilon \le c\sqrt{\frac{v}{n}}$$

So $$\boxed{\mathcal{R}_n(\mathcal{F}) \le c\sqrt{\frac{v}{n}}}$$

Thm: (Improved Glivenko-Cantelli)
Let $X_1, \dots, X_n$ be iid random variables with CDF F. Define $F_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \le t\}}$
Then $\mathbb{E}\|F_n - F\|_\infty \le \frac{c}{\sqrt{n}}$

pf: Let $\mathcal{F} = \{\mathbb{1}_{(-\infty, t]} : t \in \mathbb{R}\}$

$\Rightarrow vc(\mathcal{F}) = 1.$ ∎

Another Example:

Define

$$\mathcal{F} := \{ f : [0,1] \to [0,1] \mid \|f\|_{Lip} \le L \}$$

Thm: Let $X_1, \dots, X_n$ be iid random variables in $[0,1]$. Then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X) \right| \le \frac{CL}{\sqrt{n}}$$

pf: WLOG assume $L = 1$.

Again, let's upper bound

$$R_n(\mathcal{F}) = \mathbb{E}_X \underbrace{\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i)}$$

Define $Z_f = \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i)$

Then

$$\|Z_f - Z_g\|_{\psi_2} = \frac{1}{n} \left\| \sum_{i=1}^n \varepsilon_i (f(X_i) - g(X_i)) \right\|_{\psi}$$

$$\le \frac{1}{n} \sqrt{\sum_{i=1}^n (f(X_i) - g(X_i))^2} \le \frac{1}{\sqrt{n}} \|f - g\|_\infty$$

Dudley $\Rightarrow$ (diameter $= 1$)

$$\mathbb{E}\sup_{f\in\mathcal{F}} |Z_f| \le \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)}\, d\varepsilon.$$

Lemma (HW): $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le \left(\frac{2}{\varepsilon}\right)^{\frac{2}{\varepsilon}}$

$$\Rightarrow \mathbb{E}\sup_{f\in\mathcal{F}} |Z_f| \le \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\frac{1}{\varepsilon}\log\left(\frac{2}{\varepsilon}\right)}\, d\varepsilon$$

$$\le \frac{\hat{C}}{\sqrt{n}} \qquad \blacksquare$$

So far we know that

$$\sup_{\varepsilon>0} \varepsilon\sqrt{\log N(T,\varepsilon,\|\cdot\|_2)} \le \mathbb{E}\sup_{t\in T} X_t \le K\int_0^1 \sqrt{\log N(T,\varepsilon,d)}\, d\varepsilon$$

($\le$: Gaussian; $\le$: SubGaussian)

Question: Is there a tighter bound that separates ① the subGaussian parameter? ② size of T.

Yes!

Recall the key bound from chaining

$$\text{(*)} \quad \mathbb{E} \sup_{t \in T} X_t \le c \sum_{k=k_0}^{k_\infty} \varepsilon_{k-1} \sqrt{\log |T_k|}$$

where T_k are ε_k-nets of T.

$\Rightarrow$ We chose ε_k followed by T_k.

Let's switch the order.

Fix subset $T_1, T_2, \dots, T_k, \dots \subset T$ such that

$|T_0| = 1$ and $|T_k| \le 2^{2^k}$ for $k \ge 1$

The sequence $(T_k)_{k=0}^{\infty}$ is called <u>admissable</u>

Define $\varepsilon_k = \sup_{t \in T} d(t, T_k)$

Then T_k is an ε_k-net of T.

(*) becomes

$$\mathbb{E} \sup_{t \in T} X_t \le c \sum_{k=0}^{\infty} 2^{k/2} \sup_{t \in T} d(t, T_k)$$

Defn (Talagrand's γ_2 functional)
Let (T,d) be a metric space. Define

$$\gamma_2(T,d) = \inf_{(T_k)} \sup_{t\in T} \sum_{k=0}^{\infty} 2^{k/2} d(t,T_k)$$

where $\inf_{(T_k)}$ is taken over all admissable sequences.

Thm (Generic chaining bound)
Let $\{X_t\}_{t\in T}$ be a mean zero subGaussian process satisfying

$$\|X_t - X_s\|_{\psi_2} \le K d(t,s).$$

Then

$$\mathbb{E} \sup_{t\in T} X_t \le CK \gamma_2(T,d)$$

pf: WLOG assume $K=1$.

We do the chaining as before

$$t_0 \to \pi_0(t) \to \pi_1(t) \to \cdots \to \pi_N(t) = t$$

$$\Rightarrow X_t - X_0 = \sum_{k=1}^{N} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \right)$$

Fix k, t. We know

$$\| X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \|_{\psi_2} \le d(\pi_k(t), \pi_{k-1}(t))$$

$$\Rightarrow \mathbb{P}\left[|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| \le cu\, 2^{k/2}\, d(\pi_k(t), \pi_{k-1}(t)) \right]$$

$$\ge 1 - 2\exp(-8u^2 2^k)$$

Take union bound over

$$|T_k|\,|T_{k-1}| \le |T_k|^2 = 2^{2^{k+1}} \qquad k \in \mathbb{N}$$

So get uniform bound w.p.

$$1 - \sum_{k=1}^{\infty} 2^{2^{k+1}} \cdot 2\exp(-8u^2 2^k) \ge 1 - 2\exp(-cu^2)$$

(check)

for $u > \hat{c}$.

In this event, get

$$|X_t - X_0| \le C u \sum_{k=1}^{\infty} 2^{k/2} d(\pi_k(t), \pi_{k-1}(t))$$

$$\le C u \underbrace{\sum_{k=1}^{\infty} 2^{k/2} d(t, T_k)}_{\gamma_2(T,d)}$$

$$\Rightarrow \sup_{t \in T} |X_t - X_0| \le C u \gamma_2(T,d)$$

Now integrate the tail! ☑

Remarkably $\gamma_2(T,d)$ is also a lower bound

Thm (Talagrand) Let $\{X_t\}_{t \in T}$ be a Gaussian Process on T. Define $d(t,s) = \|X_t - X_s\|_2$. Then

$$c\, \gamma_2(T,d) \le \mathbb{E} \sup_{t \in T} X_t \le C \gamma_2(T,d)$$

For us the most important consequence is the following generalization of Sudakov-Fernique comparison inequality.

Thm: (Talagrand Comparison Inequality) Let $\{X_t\}_{t\in T}$ be a mean-zero random process on T and let $\{Y_t\}_{t\in T}$ be a Gaussian process. Assume

$$\|X_t - X_s\|_{\psi_2} \le K \|Y_t - Y_s\|_2$$

Then

$$\mathbb{E} \sup_{t\in T} X_t \le CK \, \mathbb{E} \sup_{t\in T} Y_t.$$

pf: Define $d(t,s) = \|Y_t - Y_s\|_2$. Then

$$\mathbb{E} \sup_{t\in T} X_t \le CK \gamma_2(T,d) \le \hat{C} K \, \mathbb{E} \sup_{t\in T} Y_t. \quad \square$$

(upper bound in Talagrand; lower bound in Talagrand)

Cor: Let $\{X_t\}_{t \in T}$ be a mean-zero random process on $T \subseteq \mathbb{R}^n$. Assume

$$\|X_t - X_s\|_{\psi_2} \leq K \|t-s\|_2$$

Then $$\mathbb{E} \sup_{t \in T} X_t \leq CK\, w(T)$$

pf: Define $Y_t = \langle g, t \rangle$ where $g \in N(0, I)$. Apply previous thm. ☐

Thm: (Sub Gaussian Chavet Inequality) Let $A \in \mathbb{R}^{m \times n}$ be a random matrix with independent mean zero K-sub Gaussian parameters. Let $T \subseteq \mathbb{R}^n$, $S \subset \mathbb{R}^m$ be arbitrary bounded sets. Then

$$\mathbb{E} \sup_{x \in T, y \in S} \langle Ax, y \rangle \leq CK\,(w(T)\,\mathrm{rad}(S) + \mathrm{rad}(T)\,w(S))$$

pf: WLOG assume $K=1$. Define

$X_{u,v} := \langle Au, v \rangle$ for $u \in T$, $v \in S$.

Compute

$$\|X_{uv} - X_{wz}\|_{\psi_2} = \Big\| \sum_{i,j} A_{ij} (u_i v_j - w_i z_j) \Big\|_{\psi_2}$$

$$\lesssim \Big(\sum_{i,j} |u_i v_j - w_i z_j|^2 \Big)^{1/2}$$

$$= \|uv^T - wz^T\|_F$$

$$= \|u(v-z)^T + (u-w)z^T\|_F$$

$$\le \|u(v-z)^T\|_F + \|(u-w)z^T\|_F$$

$$= \|v-z\|_2 \|u\|_2 + \|u-w\|_2 \|z\|_2$$

$$\le \|v-z\|_2 \cdot \mathrm{rad}(T) + \|u-w\|_2 \cdot \mathrm{rad}(S)$$

Define $Y_{uv} := \langle g, u \rangle \mathrm{rad}(T) + \langle h, v \rangle \mathrm{rad}\, S$

where $g \sim N(0, I)$, $h \sim N(0, I)$

Then $\|Y_{uv} - Y_{wz}\|^2 = \|u-w\|_2^2 \mathrm{rad}(T)^2 + \|v-z\|_2^2 \mathrm{rad}(S)^2$

So

$$\|X_{uv} - X_{wz}\|_{\psi_2} \le c \|y_{uv} - y_{wz}\|_2$$

Talagrand

$$\Rightarrow \mathbb{E} \sup_{u \in T, v \in S} X_{uv} \le c \mathbb{E} \sup_{u \in T, v \in S} y_{uv}$$

$$= \mathbb{E} \sup_{u \in T} \langle g, u \rangle \operatorname{rad}(S)$$

$$+ \mathbb{E} \sup_{v \in S} \langle h, v \rangle \operatorname{rad}(T)$$

$$= w(T) \operatorname{rad}(S) + w(S) \operatorname{rad}(T).$$

□