S&DS 6020: High Dimensional Probability

Professor Zhou Fan

Scribe: Linghai Liu

Contents

1 Chernoff Bound, Sub-gaussian Variables, Martingale Method

Readings: §2.1-2.6 in [Ver18], §3.1-3.2 in [vH14].

1.1 Chernoff Bound

Example 1.1. $X_1, \dots, X_n$ are iid Bernoulli(p). Fix $t > 0$. By CLT, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq t\right) \approx 1 - \Phi\left(t/\sqrt{p(1-p)/n}\right) \approx \exp\left(-\frac{nt^2}{2p(1-p)}\right).$$

By Cramér's Theorem:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq t\right) \approx e^{-nD((p+t)\|p)},$$

where

$$D(q\|p) = (1-q)\log\frac{1-q}{1-p} + q\log\frac{q}{p}$$

for $t \asymp 1$ and large n. When $q \approx p$, the Taylor expansion is approximately $\frac{(q-p)^2}{2p(1-p)}$.

We also care about the regime in which t is chaning.

Can we not consider specific regime for t and consider all $t \in (0, 1-p)$ to obtain an asymptotic bound instead?

1. Chebyshev / Markov inequality

   $$\mathbb{P}(\bar{X} - p \geq t) \leq \mathbb{P}((\bar{X}-p)^2 \geq t^2) \leq \mathbb{E}[(\bar{X}-p)^2]/t^2 = \frac{p(1-p)}{nt^2}.$$

   This result is poor because we could have achieved exponential decay.

2. k-th moment.

   $$\mathbb{P}(\bar{X} - p \geq t) \leq \mathbb{P}(|\bar{X}-p| \geq t^k) \leq \frac{1}{t^k}\mathbb{E}|\bar{X}-p|^k \asymp \left(\frac{1}{t}\sqrt{\frac{p(1-p)}{n}}\right).$$

   It is hard to compute the k-th moment. Moreover this is still not exponential tail.

3. Moment / Cumulant generating function

   For any $\lambda \in \mathbb{R}$, we have

   $$\mathbb{E}\left[e^{\lambda(X_i-p)}\right] := e^{\psi(\lambda)} \iff \psi(\lambda) = \log\mathbb{E}\left[e^{\lambda(X_i-p)}\right].$$

   When $\lambda \geq 0$,

   $$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq t\right) &= \mathbb{P}\left(e^{\lambda(\sum_{i=1}^n X_i - p)} \geq e^{\lambda nt}\right) \\ &\leq e^{-\lambda nt}\mathbb{E}\left[e^{\lambda\sum_{i=1}^n (X_i-p)}\right] \\ &= \exp\left(-\lambda nt + n\psi(\lambda)\right).\end{aligned}$$

   Here, the cumulant generating function $\psi(\lambda)$ is

   $$\psi(\lambda) = \log\mathbb{E}\left[e^{\lambda(X_i-p)}\right] = -\lambda p + \log\mathbb{E}\left[e^{\lambda X_i}\right] = -\lambda p + \log(1 - p + pe^{\lambda}).$$

Optimize it over λ, we have:

$$\sup_{\lambda \geq 0} \{\lambda(t) - \psi(\lambda)\} = D(p+t\|p)) \Longrightarrow \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq t\right) \leq \exp(-nD(p+t\|p)).$$

We can abstract this argument to general random variables in the following:

Theorem 1.2 (Chernoff Bound)**.** Let $X_1, \ldots, X_n$ be iid random variables with $\mathbb{E}[X_1] = 0$, and let $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X_1}]$. Define the convex conjugate $\psi^*(t) = \sup_{\lambda \geq 0}\{\lambda t - \psi(\lambda)\}$. Then,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \leq \exp(-n\psi^*(t)), \quad \forall t \geq 0.$$

Example 1.3.

- $X_i \sim \text{Bernoulli}(p) - p$.

$$\psi(\lambda) = \log(1 - p + pe^{\lambda}) - \lambda p, \quad \psi^*(t) = D(p+t\|p).$$

- $X_i \sim \text{Poisson}(\theta) - \theta$.

$$\psi(\lambda) = \theta(e^{\lambda} - 1 - \lambda), \quad \psi^*(t) = (t+\theta)\log\frac{\theta + t}{\theta} - t.$$

- $X_i \sim \mathcal{N}(0, \sigma^2)$.

$$\psi(\lambda) = \frac{1}{2}\lambda^2\sigma^2, \quad \psi^*(t) = \sup_{\lambda \geq 0}\left\{\lambda t - \frac{1}{2}\lambda^2\sigma^2\right\} = \frac{t^2}{2\sigma^2}, \quad \lambda^* = \frac{t}{\sigma^2}.$$

In fact, the main idea here is to control the growth of the cumulant generating function. The notion of sub-Gaussian random variables formalizes exactly this kind of control, allowing us to obtain bounds even in the absence of explicit formulas for ψ.

1.2 Subgaussian Variables

Definition 1.4 (subgaussian)**.** A mean-zero random variable X is σ^2-subgaussian if

$$\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2\sigma^2/2} \Longleftrightarrow \psi(\lambda) \leq \frac{1}{2}\lambda^2\sigma^2, \quad \forall \lambda \in \mathbb{R}.$$

Proposition 1.5. If X is mean-zero, σ^2-subgaussian, then

$$\mathbb{P}(|X| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}, \quad t \geq 0.$$

Proof. Let $\lambda \geq 0$. Note that

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda X}] \leq \exp(-\lambda t + \frac{\lambda^2\sigma^2}{2}).$$

Optimize over all $\lambda \geq 0$, we get $\mathbb{P}(X \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$ with $\lambda^* = \frac{t}{\sigma^2}$. For $\lambda \leq 0$,

$$\mathbb{P}(X \leq -t) = \mathbb{P}(e^{\lambda X} \geq e^{-\lambda t}) \leq e^{\lambda t}\mathbb{E}[e^{\lambda X}] \leq \exp(\lambda t + \frac{\lambda^2\sigma^2}{2}).$$

Optimize over all $\lambda \leq 0$, we get $\mathbb{P}(X \leq -t) \leq e^{-\frac{t^2}{2\sigma^2}}$ with $\lambda_* = -\frac{t}{\sigma^2}$. □

Remark 1.6. X is σ^2-subgaussian in the upper tail if $\mathbb{E}[e^{\lambda X}] \le e^{\lambda^2\sigma^2/2}$, $\forall \lambda \ge 0$. In this case,

$$\mathbb{P}(X \ge t) \le e^{-\frac{t^2}{2\sigma^2}}.$$

Theorem 1.7 (Hoeffding's Inequality)**.** If $X_1, \dots, X_n$ are independent, $\mathbb{E}X_i = 0$, and X_i is σ_i^2-subgaussian for each $1 \le i \le n$, then $\sum_{i=1}^n X_i$ is $\sum_{i=1}^n \sigma_i^2$-subgaussian, i.e.,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \ge t\right) \le 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Proof. We show that $\sum_{i=1}^n X_i$ by definition 1.4. $\forall \lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^n X_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda X_i}\right] \le \prod_{i=1}^n e^{\frac{1}{2}\lambda^2\sigma_i^2} = \exp\left(\frac{1}{2}\lambda^2 \sum_{i=1}^n \sigma_i^2\right).$$

□

Lemma 1.8 (Hoeffding)**.** If $X \in [a, b]$ with probability 1, then $X - \mathbb{E}[X]$ is $\frac{(b-a)^2}{4}$-subgaussian.

Proof. Without loss of generality, we let $\mathbb{E}X = 0$. For $\lambda \in \mathbb{R}$,

$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}], \quad \psi(0) = \log \mathbb{E}[e^0] = 0.$$
$$\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}, \quad \psi'(0) = \mathbb{E}X = 0.$$
$$\psi''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \frac{\mathbb{E}[Xe^{\lambda X}]^2}{\mathbb{E}[e^{\lambda X}]^2} = \mathrm{Var}(\tilde{X}) \le \frac{(b-a)^2}{4},$$

where for each measurable set A, $\mathbb{P}(\tilde{X} \in A) = \frac{\mathbb{E}[\mathbb{1}_A e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}$. Hence,

$$\psi(\lambda) = \int_0^\lambda \psi'(t)\mathrm{d}t = \int_0^\lambda \int_0^t \psi''(s)\mathrm{d}s\mathrm{d}t \le \int_0^\lambda \int_0^t \frac{(b-a)^2}{4}\mathrm{d}s\mathrm{d}t = \frac{\lambda^2(b-a)^2}{8}.$$

□

Corollary 1.9. Here we combine Hoeffding's Lemma 1.8 with Proposition 1.5. If independent random variables $X_i \in [a_i, b_i]$ with probability 1, then

$$\mathbb{P}\left(\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Example 1.10. If $X_1, \dots, X_n$ are i.i.d. Bernoulli(p), then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - p\right| \ge t\right) \le 2e^{-2nt^2}.$$

This is cruder than the exact Chernoff bound, especially for small p.

Proposition 1.11. The following are equivalent:

(a) $\exists K_1 > 0$ such that $\forall t \ge 0$, $\mathbb{P}(|X| \ge t) \le 2\exp(-t^2/K_1^2)$.

(b) $\exists K_2 > 0$ such that $\forall p \ge 1$, $||X||_{L^p} = (\mathbb{E}[|X|^p])^{\frac{1}{p}} \le K_2\sqrt{p}$.

(c) $\exists K_3 > 0$ such that $\mathbb{E}[e^{X^2/K_3^2}] \le 2$.

(d) If, in addition, $\mathbb{E}[X] = 0$, then also equivalent to

$$\exists K_4 > 0 \quad \text{s.t.} \quad \psi(\lambda) \le K_4\lambda^2, \quad \forall \lambda \in \mathbb{R}.$$

For the proof, we will use the integral identity:

Lemma 1.12 (Integral Identity)**.** If $Z \geq 0$, then $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \geq t)\mathrm{d}t$.

Proof. Notice $Z = \int_0^Z \mathrm{d}t = \int_0^\infty \mathbb{1}_{t \leq Z}\mathrm{d}t$ and take expectation on both sides. □

Proof of Proposition 1.11. We will show $(a) \Longrightarrow (b) \Longrightarrow (c) \Longrightarrow (a)$.

- $(a) \Longrightarrow (b)$. By Lemma 1.12, we have

$$\begin{aligned}
\mathbb{E}[|X|^p] = \int_0^\infty \mathbb{P}(|X|^p \geq t)\mathrm{d}t &= \int_0^\infty \mathbb{P}(|X|^p \geq u^p) p u^{p-1}\mathrm{d}u \\
&\leq \int_0^\infty 2e^{-u^2/K_1^2} p u^{p-1}\mathrm{d}u \\
&= \int_0^\infty 2e^{-v^2} p (K_1 v)^{p-1}\mathrm{d}v \\
&= K_1^{p-1} p \Gamma(p/2) \\
&\leq 3K_1^{p-1} p (p/2)^{p/2}.
\end{aligned}$$

- $(b) \Longrightarrow (c)$. For $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda^2 X^2}\right] &= \sum_{k=0}^\infty \frac{1}{k!}\lambda^{2k}\mathbb{E}[X^{2k}] \quad \text{Taylor expansion} \\
&\leq \sum_{k=0}^\infty \lambda^{2k}\left(\frac{e}{k}\right)^k (K_2\sqrt{2k})^{2k} \quad \text{by (b) and } k! \geq \left(\frac{k}{e}\right)^k \\
&= \sum_{k=0}^\infty (2eK_2^2\lambda^2)^k \leq 2 \quad \text{for } |\lambda| \leq \frac{1}{2K_2\sqrt{e}}.
\end{aligned}$$

- $(c) \Longrightarrow (a)$. For $t \geq 0$, by Markov inequality and part (c),

$$\mathbb{P}(|X| \geq t) = \mathbb{P}\left(e^{X^2/K_3^2} \geq e^{t^2/K_3^2}\right) \leq e^{-t^2/K_3^2}\mathbb{E}[e^{X^2/K_3^2}] \leq 2e^{-t^2/K_3^2}.$$

- Now we have $\mathbb{E}[X] = 0$. We are to show $(d) \Longrightarrow (a)$ and $(b) \Longrightarrow (d)$.
- $(d) \Longrightarrow (a)$. This is shown by Proposition 1.5.
- $(b) \Longrightarrow (d)$. Since $\mathbb{E}X = 0$, by Taylor expansion, we have

$$\mathbb{E}[e^{\lambda X}] = 1 + \sum_{k=2}^\infty \frac{\lambda^k}{k!}\mathbb{E}[X^k].$$

  For $k \geq 1$, $\mathbb{E}|\lambda X|^{2k+1} \leq \frac{1}{2}\left(\mathbb{E}|\lambda X|^{2k} + \mathbb{E}|\lambda X|^{2k+2}\right)$. This implies that

$$\begin{aligned}
e^{\psi(\lambda)} = \mathbb{E}[e^{\lambda X}] &\leq 1 + 2\sum_{k=1}^\infty \frac{\lambda^{2k}\mathbb{E}[X^{2k}]}{(2k)!} \leq \sum_{k=0}^\infty \frac{2^k\lambda^{2k}(K_2\sqrt{2k})^{2k}}{(2k)!} \quad \text{by (b)} \\
&\leq \sum_{k=0}^\infty \frac{(2K_2\lambda)^{2k}}{k!} \quad \text{by } (2k)! \geq (k!)^2, k! \geq (k/e)^k. \\
&= \exp(4K_2^2\lambda^2).
\end{aligned}$$

□

Definition 1.13 (Subgaussian Norm)**.** The subgaussian norm of a subgaussian random variable X is

$$||X||_{\psi_2} := \inf\{K > 0 \mid \mathbb{E}\exp(X^2/K^2) \leq 2\}.$$

In [Ver18], the author takes $||X||_{\psi_2} < \infty$ to be the definition of subgaussian if $\mathbb{E}X \neq 0$.

1.3 Martingale Method

How to show concentration of nonlinear functions $f(X_1, \ldots, X_n)$?

Idea: let $\mathbb{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_i] = M_i$ and

$$f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] = M_n - M_0 = \sum_{i=1}^{n}(M_i - M_{i-1}) := \sum_{i=1}^{n}\Delta_i,$$

where $\{M_i\}_{i=1}^n$ is a martingale, i.e., $\mathbb{E}[M_i \mid X_1, \ldots, X_{i-1}] = M_{i-1}$.

Theorem 1.14 (Azuma-Hoeffding)**.** Suppose $X_1, \ldots, X_n$ are independent, and $\forall\lambda \geq 0$,

$$\mathbb{E}[e^{\lambda\Delta_i} \mid X_1, \ldots, X_{i-1}] \leq e^{\lambda^2\sigma_i^2/2} \quad \text{a.s.},$$

then $\sum_{i=1}^n \Delta_i$ is $\sum_{i=1}^n \sigma_i^2$-subgaussian, so

$$\mathbb{P}\left[|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right), \quad \forall t \geq 0.$$

Proof. For any $\lambda \geq 0$ and $k \in \{1, \ldots, n\}$,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda\sum_{i=1}^k \Delta_i}\right] &= \mathbb{E}\left[\mathbb{E}\left[e^{\lambda\sum_{i=1}^k \Delta_i} \mid X_1, \ldots, X_{k-1}\right]\right] \\
&= \mathbb{E}\left[e^{\lambda\sum_{i=1}^{k-1} \Delta_i}\mathbb{E}\left[e^{\lambda\Delta_k} \mid X_1, \ldots, X_{k-1}\right]\right] \\
&\leq e^{\lambda^2\sigma_k^2}\mathbb{E}\left[e^{\lambda\sum_{i=1}^{k-1} \Delta_i}\right].
\end{aligned}$$

By induction, $\mathbb{E}\left[e^{\lambda\sum_{i=1}^n \Delta_i}\right] \leq \exp\left(\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2\right)$. □

Corollary 1.15 (Bounded Difference Inequality)**.** For each $i \in \{1, \ldots, n\}$, let

$$||D_i f||_\infty := \sup_{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n}\left[\sup_z f(X_1, \ldots, z, \ldots, X_n) - \inf_z f(X_1, \ldots, z, \ldots, X_n)\right].$$

If $X_1, \ldots, X_n$ are independent, then

$$\mathbb{P}\left[|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n ||D_i f||_\infty^2}\right), \quad \forall t \geq 0.$$

Proof. Use the same definition of Δ_i as in Theorem 1.14:

$$\Delta_i := \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_i] - \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_{i-1}].$$

$$\begin{aligned}
\implies \Delta_i &\geq \mathbb{E}\left[\inf_z f(X_1, \ldots, z, \ldots X_n) - f(X_1, \ldots, X_n) \mid X_1, \ldots, X_{i-1}\right] := A_i \\
\Delta_i &\leq \mathbb{E}\left[\sup_z f(X_1, \ldots, z, \ldots, X_n) - f(X_1, \ldots, X_n) \mid X_1, \ldots, X_{i-1}\right] := B_i.
\end{aligned}$$

So $\Delta_i \in [A_i, B_i]$ with probability 1, and $B_i - A_i \le ||D_i f||_\infty$ by definition. Using Hoeffdings Lemma 1.8,

$$\mathbb{E}\left[e^{\lambda \Delta_i} \mid X_1 \ldots, X_{i-1}\right] \le \frac{||D_i f||_\infty^2}{4}.$$

The corollary then follows from Azuma-Hoeffding Theorem 1.14. □

Example 1.16 (Rademacher Complexity)**.** Let $\varepsilon_1, \ldots, \varepsilon_n$ be iid Rademacher random variables, i.e., $\mathbb{P}(\varepsilon_i = \pm 1) = \frac{1}{2}$. Let $T \subseteq \mathbb{R}^n$, $f(\varepsilon_1, \ldots, \varepsilon_n) = \sup_{t \in T} \varepsilon^\top t$. Then $||D_i f||_\infty = 2 \sup_{t \in T} |t_i|$, so $\forall u \ge 0$,

$$\mathbb{P}(|f(\varepsilon_1, \ldots, \varepsilon_n) - \mathbb{E}[f(\varepsilon_1, \ldots, \varepsilon_n)]| \ge u) \le 2e^{-\frac{u^2}{2\sigma^2}},$$

where $\sigma^2 = \sum_{i=1}^n \sup_{t \in T} t_i^2$.

Later in the course, we will improve this to $\sigma^2 = \sup_{t \in T} \sum_{i=1}^n t_i^2 = \sup_{t \in T} ||t||_2^2$.

Example 1.17 (U-statistics)**.** Let $X_1, \ldots, X_n$ be iid, $h : \mathbb{R}^2 \mapsto \mathbb{R}$ with $||h||_\infty \le B$, $h(x, y) = h(y, x)$. Consider

$$f(X_1, \ldots, X_n) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} h(X_i, X_j).$$

Then

$$||D_i f||_\infty \le \sup_{x, z, z'} \frac{1}{\binom{n}{2}} \sum_{j : j \ne i} |h(z, x_j) - h(z', x_j)| \le \frac{(n-1) \cdot 2B}{\binom{n}{2}} = 4B/n.$$

So $\mathbb{P}(|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \ge t) \le 2e^{-\frac{nt^2}{8B^2}}$.

If the kernel is non-degenerate, then we have CLT $\mathcal{O}(\frac{1}{\sqrt{n}})$. If the kernel is instead degenerate, then we have a smaller fluctuation $\mathcal{O}(\frac{1}{n})$.

2 Sub-exponential Variables, Random Vectors in High Dimensions

Readings: §2.7-2.8, 3.1-3.2 in [Ver18], §2.1-2.2 in [Wai19].

Example 2.1 (Erdős-Rényi Graph)**.** On graph $G = \mathcal{G}(n,p)$ be an Erdős-Rényi graph, i.e., $\mathbb{P}(i \sim j) = p$ independently for all $i \neq j \in \{1,\dots,n\}$. Let $f(G)$ be chromatic number, i.e., the minimal number of colors to color vertices so that no adjacent vertices have the same color. Let

$$\begin{aligned} X_1 &= \text{all edges from } 1 \to \{2,3,\dots,n\} \\ X_2 &= \text{all edges from } 2 \to \{3,\dots,n\} \\ &\vdots \\ X_{n-1} &= \text{edge from } n-1 \to \{n\} \end{aligned}$$

Fixing all but X_i, $f(G)$ is smallest when $X_i = (0,\dots,0)$ and largest when $X_i = (1,\dots,1)$, and thus $||D_i f||_\infty \le 1$ because we just twist one edge. This observation implies

$$\mathbb{P}(|f - \mathbb{E}[f]| \ge t) \le 2\exp\left(-\frac{t^2}{n-1}\right).$$

For any $p \in (0,1)$, $f(G) = \mathbb{E}[f(G)] + \mathcal{O}_{\mathbb{P}}(\sqrt{n})$. Also, it is known that $\mathbb{E}[f(G)] \asymp \frac{n}{\log n} \gg \sqrt{n}$.

Recall that a mean-zero random variable X is σ^2-subgaussian if $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}] \le e^{\frac{1}{2}\lambda^2\sigma^2}, \forall \lambda \in \mathbb{R}$, and when $X_1,\dots,X_n$ are independent, mean-zero, σ^2-subgaussian, by Hoeffding's inequality 1.7, $\forall t \ge 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le \exp\left(-\frac{t^2}{2n\sigma^2}\right).$$

What if X_i's have heavier tails?

- By CLT, for $t \asymp \sqrt{n}$, $\mathbb{P}(\sum_{i=1}^n X_i \ge t) \approx e^{-\mathcal{O}(t^2/n)}$.
- However, for $t \gg \sqrt{n}$, a lower bound is

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) &\ge \mathbb{P}\left((X_1 \ge 1.01t \text{ and } \left|\sum_{i=2}^n X_i\right| \le 0.01t\right) \\ &= \mathbb{P}(X_1 \ge 1.01t) \cdot \underbrace{\mathbb{P}\left(\left|\sum_{i=2}^n X_i\right| \le 0.01t\right)}_{\approx 1 \text{ for } t \gg \sqrt{n}} \\ &\approx \mathbb{P}(X_1 \ge 1.01t). \end{aligned}$$

  This needs not decay exponentially in t^2 if X_1 is <u>not</u> subgaussian. Hence, we would expect to see two types of bounds for $\sum_{i=1}^n X_i$ for moderate v.s. large t.

2.1 Subexponential Random Variables

Example 2.2. $X \sim$ Exponential(θ). Then its pdf is $f_\theta(x) = \theta e^{-\theta x}\mathbb{1}_{x\ge 0}$ and its expectation $\mathbb{E}[X] = \frac{1}{\theta}$.

$$\psi(\lambda) = \log \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] = \begin{cases} \log\frac{\theta}{\theta-\lambda} - \frac{\lambda}{\theta} & \text{if } \lambda < \theta \\ \infty & \text{if } \lambda \ge \theta \end{cases}$$

For the former case $\lambda < \theta$, the expression is approximately $\frac{\lambda^2}{2\theta^2}$ for small $|\lambda|$.

Definition 2.3 (subexponential)**.** A mean-zero random variable X is (σ^2, b)-subexponential if

$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}] \le \frac{\lambda^2\sigma^2}{2} \quad \text{for all } |\lambda| \le \frac{1}{b}.$$

Example 2.4. The following are some examples of subexponential random variables:

- $X \sim \text{Exponential}(\theta) - \frac{1}{\theta}$. $\psi(\lambda) = \log \frac{\theta}{\theta-\lambda} - \frac{\lambda}{\theta} \le \frac{\lambda^2}{\theta^2}$ for $|\lambda| \le \frac{\theta}{2}$, so X is $(\frac{2}{\theta^2}, \frac{2}{\theta})$-subexponential.
- $X \sim \chi_1^2 - 1$. $\psi(\lambda) = -\frac{1}{2}\log(1-2\lambda) - \lambda \le 2\lambda^2$ for $|\lambda| \le \frac{1}{4}$, so X is $(4,4)$-subexponential.
- $X \sim \chi_n^2 - n$. $\psi(\lambda) = -\frac{n}{2}\log(1-2\lambda) - n\lambda \le 2n\lambda^2$ for $|\lambda| \le \frac{1}{4}$, so X is $(4n,4)$-subexponential.

Proposition 2.5. If X is mean-zero and (σ^2, b)-subexponential, then

$$\mathbb{P}(|X| \ge t) \le 2\exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{\sigma^2}, \frac{t}{\sigma}\right\}\right), \quad \forall t \ge 0.$$

Proof. Let $t \ge 0$. $\forall \lambda \in [0, \frac{1}{b}], \mathbb{P}(X \ge t) \le e^{-\lambda t}\mathbb{E}[e^{\lambda X}] \le e^{-\lambda t + \frac{\lambda^2\sigma^2}{2}}$.

- If $t < \frac{\sigma^2}{b}$, pick $\lambda = \frac{t}{\sigma^2}$. Then $\mathbb{P}(X \ge t) \le e^{-\frac{t^2}{2\sigma^2}}$.
- If $t \ge \frac{\sigma^2}{b}$, pick $\lambda = \frac{1}{b}$. Then $\mathbb{P}(X \ge t) \le e^{-\frac{t}{b} + \frac{\sigma^2}{2b^2}} \le e^{-\frac{t}{2b}}$.

The left tail is the same by choosing $\lambda \in [-\frac{1}{b}, 0]$. □

Example 2.6. Suppose $X = Z_1^2 + \ldots + Z_n^2 \sim \chi_n^2$. Then

$$\mathbb{P}(|X - n| \ge t) \le 2e^{-\frac{1}{2}\min(\frac{t^2}{4n}, \frac{t}{4})}, \quad \forall t \ge 0.$$

So X has Gaussian tail for $t \le n$ and exponential tail for $t > n$.

Theorem 2.7 (Bernstein's Inequality)**.** If $X_1, \ldots, X_n$ are independent, $\mathbb{E}[X_i] = 0$, X_i is (σ_i^2, b_i)-subexponential, then $\sum_{i=1}^n X_i$ is $\left(\sum_{i=1}^n \sigma_i^2, \max_{1\le i\le n} b_i\right)$-subexponential. Thus, $\forall t \ge 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le \exp\left\{-\frac{1}{2}\min\left(\frac{t^2}{\sum_{i=1}^n \sigma_i^2}, \frac{t}{\max_{1\le i\le n} b_i}\right)\right\}.$$

Proof. $\forall |\lambda| \le 1/\max_{i=1}^n b_i$, $\mathbb{E}[e^{\lambda \sum_{i=1}^n X_i}] \le \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}] \le \exp\left\{\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2\right\}$. □

Lemma 2.8 (Bernstein)**.** If $\mathbb{E}[X] = 0$, $\text{Var}(X) \le \sigma^2$, and $|X| \le b$ with probability 1, then

$$\psi(\lambda) \le \frac{\lambda^2\sigma^2}{2(1 - |\lambda|b/3)}, \quad \forall |\lambda| < \frac{3}{b}.$$

In particular, X is $(2\sigma^2, \frac{2b}{3})$-subexponential.

Proof. Let $|\lambda| < \frac{3}{b}$.

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}[X^k] &\le 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k\ge 3} \frac{\lambda\sigma^2}{1\cdot 2\cdot 3^{k-2}}(|\lambda|b)^{k-2} \\
&= 1 + \frac{\lambda^2\sigma^2}{2}\cdot\frac{1}{1 - |\lambda|b/3} \quad \text{since } |\lambda| < 3/b \\
&\le \exp\left\{\frac{\lambda^2\sigma^2}{2(1-|\lambda|b/3)}\right\} \quad \text{since } 1 + x \le e^x,\ \forall x \in \mathbb{R}.
\end{aligned}$$

Thus, $\psi(\lambda) \le \frac{\lambda^2\sigma^2}{2(1-|\lambda|b/3)}$. Further bound by $\lambda^2\sigma^2$ for $|\lambda| \le \frac{3}{2b}$. □

Corollary 2.9. If $X_1, \dots, X_n$ are independent with $\mathbb{E}[X_i] = 0$, $\mathrm{Var}(X_i) \le \sigma^2$, and $|X_i| \le b$, then $\forall t \ge 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \ge t\right) \le 2\exp\left(-\frac{t^2/2}{n\sigma^2 + bt/3}\right).$$

Proof. By Lemma 2.8, X_i is $(2\sigma^2, \frac{2b}{3})$-subexponential. Hence, by Bernstein Inequality (Theorem 2.7),

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \ge t\right) \le 2\exp\left\{-\min\left(\frac{t^2}{4n\sigma^2}, \frac{3t}{4b}\right)\right\}, \quad \forall t \ge 0.$$

To improve this bound, we use $\psi(\lambda) \le \frac{\lambda^2\sigma^2}{2(1-|\lambda|b/3)}, \forall|\lambda| < 3/b$ from Lemma 2.8 and pick $\lambda = \frac{t}{n\sigma^2+bt/3}$:

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le e^{-\lambda t + n\psi(\lambda)} \le e^{-\lambda t + \frac{\lambda^2 n\sigma^2}{2(1-\lambda b/3)}} = \exp\left(-\frac{t^2/2}{n\sigma^2 + bt/3}\right), \quad \forall t \ge 0.$$

Same for the lower tail. □

Proposition 2.10. The following are equivalent:

(a) $\exists K_1 > 0$ such that $\forall t \ge 0$, $\mathbb{P}(|X| \ge t) \le 2e^{-t/K_1}$.

(b) $\exists K_2 > 0$ such that $\forall p \ge 1$, $||X||_{L^p} = (\mathbb{E}[|X|^p])^{\frac{1}{p}} \le K_2 p$.

(c) $\exists K_3 > 0$ such that $\mathbb{E}[e^{|X|/K_3}] \le 2$.

(d) If $\mathbb{E}[X] = 0$, then these are also equivalent to $\exists K_4 > 0$ such that $\forall|\lambda| \le \frac{1}{K_4}$, $\psi(\lambda) \le K_4^2\lambda^2$.

Proof. □

Definition 2.11 (Subexponential norm)**.** The subexponential norm of a subexponential random variable X is

$$||X||_{\psi_1} := \inf\{K > 0 \mid \mathbb{E}\exp(|X|/K) \le 2\}.$$

Under this notation, we have an equivalent form of Bernstein bound:

Theorem 2.12. If $X_1, \dots, X_n$ are independent, $\mathbb{E}X_i = 0$, and $||X_i||_{\psi_1} \le K$, then for a constant $c > 0$, $\forall t \ge 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \ge t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{nK^2}, \frac{t}{K}\right\}\right).$$

If $t > K$ then we get to exponential decay.

Proposition 2.13. $||X - \mathbb{E}X||_{\psi_1} \lesssim ||X||_{\psi_1}$.

Proposition 2.14. $||X^2||_{\psi_1} = ||X||_{\psi_2}^2$.

Proposition 2.15. $||XY||_{\psi_1} \le ||X||_{\psi_2}||Y||_{\psi_2}$.

Proof. Let $||X||_{\psi_2} = K$ and $||Y||_{\psi_2} = L$. Then

$$\mathbb{E}\left[\exp\left(\frac{|XY|}{KL}\right)\right] \le \mathbb{E}\left[\exp\left(\frac{1}{2}\left(\frac{X^2}{K^2} + \frac{Y^2}{L^2}\right)\right)\right] \le \frac{1}{2}\mathbb{E}\left[\exp\left(\frac{X^2}{K^2}\right)\right] + \frac{1}{2}\mathbb{E}\left[\exp\left(\frac{Y^2}{L^2}\right)\right] \le 2.$$

□

2.2 Random Vectors in High Dimensions

Proposition 2.16. Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ with independent entries. $\mathbb{E}X_i = 0$, $\mathrm{Var}(X_i) = 1$, and each X_i is σ^2-subgaussian. Then for a universal constant $c > 0$,

(a) $\mathbb{P}\left(\left|||X||_2 - \sqrt{n}\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\sigma^4}\right)$.

(b) If X' is an independent copy of X, then

$$\mathbb{P}\left(\frac{X^\top X'}{||X||_2||X'||_2} \geq t\right) \leq 2\left(\exp\left(-\frac{cn}{\sigma^4}\min(t^2, t)\right) + \exp\left(-\frac{cn}{\sigma^4}\right)\right)$$

Note that when $\sigma^2 \asymp 1$, we have $||X||_2 = \sqrt{n} \pm \mathcal{O}_{\mathbb{P}}(1)$ and $\frac{|X^\top X'|}{||X||_2||X'||_2} = \mathcal{O}_{\mathbb{P}}(\frac{1}{\sqrt{n}})$. So it seems like X' and X are on a sphere with radius $\sqrt{n}$ and are orthogonal.

Proof. Since $\psi(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$, $\sigma^2/2 \geq \lim_{\lambda\to 0^+}\psi(\lambda)/\lambda^2 = \frac{\log(1+\lambda^2/2+\cdots)}{\lambda^2} = 1/2$, so $\sigma^2 \geq 1$. Then,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{n}||X||_2^2 - 1\right| \geq u\right) &= \mathbb{P}\left(\left|\sum_{i=1}^n (X_i^2 - 1)\right| \geq nu\right) \\ &\leq \exp\left(-\frac{cn}{\sigma^4}\min\{u, u^2\}\right).\end{aligned}$$

We have the following fact: if $z \geq 0$, $|z-1| \geq \delta$, then $|z^2 - 1| \geq \max(\delta, \delta^2)$. Hence,

$$\mathbb{P}\left(\left|\frac{1}{n}||X||_2 - 1\right| \geq \delta\right) \leq \mathbb{P}\left(\left|\frac{1}{n}||X||_2^2 - 1\right| \geq \max(\delta, \delta^2)\right) \leq 2\exp\left(-\frac{cn}{\sigma^4}\delta^2\right).$$

Take $\delta = t/\sqrt{n}$, we have (a). For part (b), for any $u \geq 0$,

$$\mathbb{P}\left(\frac{1}{n}|X^\top X'| \geq u\right) = \mathbb{P}\left(\left|\sum_{i=1}^n X_i X_i'\right| \geq nu\right) \leq 2\exp\left(-c\min\left(\frac{n^2u^2}{n\sigma^2}, \frac{nu}{\sigma^2}\right)\right) = 2\exp\left(-\frac{cn}{\sigma^2}\min(u^2, u)\right).$$

Note that $||X_iX_i'||_{\psi_1} \leq ||X_i||_{\psi_2}||X_i'||_{\psi_2} \leq C\sigma^2$, so

$$\begin{aligned}&\mathbb{P}\left(\frac{|X^\top X'|}{||X||_{\psi_2}||X'||_{\psi_2}} \geq t\right) \\ \leq&\mathbb{P}\left(\frac{1}{n}|X^\top X'| \geq t/4\right) + \mathbb{P}(||X||_2 \leq \sqrt{n}/2) + \mathbb{P}(||X'||_2 \leq \sqrt{n}/2) \\ \leq&2\exp\left(-\frac{c'n}{\sigma^4}\min(t, t^2)\right) + 2\exp\left(-\frac{cn}{\sigma^4}\right).\end{aligned}$$

□

Theorem 2.17 (Johnson-Lindenstrauss)**.** Let $\mathscr{A} \subseteq \mathbb{R}^m$. $|\mathscr{A}| = n$. There exists a constant $C > 0$ such that $\forall\varepsilon > 0$, if $n > \frac{C}{\varepsilon^2}\log N$, there exists a linear map $\mathscr{P} : \mathbb{R}^m \mapsto \mathbb{R}^n$ such that

$$(1-\varepsilon)||a - a'||_2 \leq ||\mathscr{P}a - \mathscr{P}a'||_2 \leq (1+\varepsilon)||a - a'||_2, \quad \forall a, a' \in \mathscr{A}.$$

Proof. The proof is done via a probabilistic method. Let Z_{ij} be iid with $\mathbb{E}Z_{ij} = 0$, $\mathrm{Var}(Z_{ij}) = 1$, and Z_{ij} is σ^2-subgaussian for some constant $\sigma^2 > 0$. Take $\mathscr{P} = \frac{1}{\sqrt{n}}Z \in \mathbb{R}^{n\times m}$. Fix an element

$$u \in \left\{\frac{a - a'}{||a - a'||_2} \mid a, a' \in \mathscr{A}\right\}$$

and consider $X = Zu$. Then for each i, we have $X_i = Z_{i1}u_1 + \ldots + Z_{im}u_m$ satisfy $\mathbb{E}X_i = 0$, $\mathrm{Var}(X_i) = 1$, aand X_i is $\sigma^2 u_1^2 + \ldots + \sigma^2 u_m^2 = \sigma^2$-subgaussian by Hoeffding's inequality . Also, $X_1, \ldots, X_n$ are independent. Each u corresponds to two elements $a, a' \in \mathscr{A}$, and

$$\begin{aligned}\mathbb{P}(|||\mathscr{P}u||_2 - 1| \geq \varepsilon) &= \mathbb{P}\left(\left|\frac{1}{\sqrt{n}}||Zu||_2 - 1\right| \geq \varepsilon\right) \\ &= \mathbb{P}\left(\left|\frac{1}{\sqrt{n}}||X||_2 - 1\right| \geq \varepsilon\right) \\ &\leq 2\exp\left(-\frac{cn\varepsilon^2}{\sigma^4}\right).\end{aligned}$$

These are the cases with "bad" behaviors. How many of these pairs? At most $\binom{N}{2}$. Hence,

$$\begin{aligned}&\mathbb{P}((1-\varepsilon)||a-a'||_2 \leq ||\mathscr{P}a - \mathscr{P}a'||_2 \leq (1+\varepsilon)||a-a'||_2) \\ \geq &1 - \binom{N}{2}\cdot 2\exp\left(-\frac{cn\varepsilon^2}{\sigma^4}\right) \geq 1 - \exp\left(-\frac{cn\varepsilon^2}{\sigma^4} + 2\log N\right).\end{aligned}$$

We only need n for which this probability is positive to ensure existence, which gives the desired bound for n. □

2.3 Sums of Heavy-tailed Random Variables

For X with tails heavier than exponential, its MGF may not exist. Two common approaches for tail bounds:

(1) Truncation

(2) Moment inequalities

To illustrate (1), we have the following theorem:

Theorem 2.18. Suppose $X_1, \ldots, X_n$ are independent, $\mathbb{E}X_i = 0$, $\mathbb{P}(|X_i| \geq t] \leq 2\exp\left(-\left(\frac{t}{K}\right)^\alpha\right)$ for all $t \geq 0$, some $K > 0$, and $\alpha \in (0,1)$. Then there exist $C, c(\alpha) > 0$ such that

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq C\exp\left(-c(\alpha)\min\left(\frac{t^2}{K^2 n}, \left(\frac{t}{K}\right)^\alpha\right)\right).$$

[For more general statements, see [BMDLP23]].

Proof. Let $Z_i = X_i \mathbb{1}_{X_i \leq L}$. Then for the right tail, $\forall t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \mathbb{P}\left(\sum_{i=1}^n Z_i \geq t\right) + \mathbb{P}(X_i > L \text{ for some } i \in \{1,\ldots,n\}) \leq \mathbb{P}\left(\sum_{i=1}^n Z_i \geq t\right) + 2ne^{-(L/K)^\alpha}.$$

For the moment generating function,

$$\begin{aligned}\mathbb{E}[e^{\lambda Z_i}] &= 1 + \lambda \underbrace{\mathbb{E}[Z_i]}_{\leq \mathbb{E}[X_i]=0} + \frac{\lambda^2}{2}\mathbb{E}[Z_i^2 e^{\tilde{\lambda} Z_i}] \quad \text{for some (random) } \tilde{\lambda} \in (0,\lambda) \\ &\leq 1 + \frac{\lambda^2}{2}\mathbb{E}[Z_i^2 \mathbb{1}_{Z_i \leq 0} + Z_i^2 e^{\lambda Z_i}\mathbb{1}_{Z_i > 0}].\end{aligned}$$

We now need to control these two terms.

$$\begin{aligned}
\mathbb{E}[Z_i^2 \mathbb{1}_{Z_i \le 0}] \le \mathbb{E}[Z_i^2] \le \mathbb{E}[X_i^2] &= \int_0^\infty \mathbb{P}(X_i^2 \ge u)\mathrm{d}u \\
&= \int_0^\infty \mathbb{P}(|X_i| \ge t] 2t \mathrm{d}t \\
&\le \int_0^\infty 4t e^{-(t/K)^\alpha} \mathrm{d}t \le C(\alpha) K^2. \\
\mathbb{E}[Z_i^2 e^{\lambda Z_i} \mathbb{1}_{Z_i > 0}] &= \int_0^\infty \mathbb{P}(X_i^2 e^{\lambda X_i} \mathbb{1}_{0 < X_i \le L} \ge u)\mathrm{d}u \\
&= \int_0^\infty \mathbb{P}(L \ge X_i \ge t)(2t + \lambda t^2) e^{\lambda t} \mathrm{d}t \quad \text{change of variables: } u = t^2 e^{\lambda t} \\
&\le \int_0^L 2 e^{-(t/K)^\alpha} (2t + \lambda t^2) e^{\lambda t} \mathrm{d}t \\
&\text{(Take } \lambda \le \frac{1}{2L}(L/K)^\alpha \le \frac{1}{2t}(t/K)^\alpha, \ \forall t \in (0, L) \text{)} \\
&\le \int_0^\infty 2 \left(2t + \frac{t}{2}\left(\frac{t}{K}\right)^\alpha\right) e^{-\frac{1}{2}(t/K)^\alpha} \mathrm{d}t \le C(\alpha) K^2.
\end{aligned}$$

Now, we have $\mathbb{E}[e^{\lambda Z_i}] \le 1 + C(\alpha) K^2 \lambda^2 \le e^{C(\alpha) K^2 \lambda^2}$ for some constant $C(\alpha)$. By Chernoff bound,

$$\mathbb{P}\left(\sum_{i=1}^n Z_i \ge t\right) \le e^{-\lambda t} \prod_{i=1}^n e^{\lambda Z_i} \le e^{-\lambda t + C(\alpha) K^2 n \lambda^2}, \quad \forall \lambda \in [0, \frac{1}{2L}(L/K)^\alpha].$$

Hence, $\forall L > 0$, $\lambda \in [0, \frac{1}{2L}(L/K)^\alpha]$, we have

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le 2e^{-\lambda t + C(\alpha) K^2 n \lambda^2} + 2ne^{-(L/K)^\alpha}.$$

Pick $L = t$ and optimize over $\lambda \in [0, \frac{1}{2t}(t/K)^\alpha]$, we have

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le 2e^{-\frac{1}{4}\min\left(\frac{t^2}{C(\alpha)K^2 n}, \left(\frac{t}{K}\right)^\alpha\right)} + 2ne^{-\left(\frac{t}{K}\right)^\alpha}.$$

- If $\left(\frac{t}{K}\right)^\alpha \ge 2\log(2n)$, then $2ne^{-\left(\frac{t}{K}\right)^\alpha} = e^{-\left(\frac{t}{K}\right)^\alpha + \log(2n)} < e^{-\frac{1}{2}\left(\frac{t}{K}\right)^\alpha}$.
- If $\left(\frac{t}{K}\right)^\alpha \le 2\log(2n)$, then choose $C(\alpha)$ such that $e^{-\frac{1}{4}\cdot\frac{t^2}{C(\alpha)K^2 n}} \ge 1$.

The above cases imply that

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le Ce^{-c(\alpha)\min\left(\frac{t^2}{K^2 n}, \left(\frac{t}{K}\right)^\alpha\right)}.$$

The lower tail is similar. □

3 Hanson-Wright Inequality, Decoupling and Symmetrization, U-statistics

Readings: §6.1-6.4 in [Ver18].

For this lecture, we consider the concentration property of $X^\top AX = \sum_{i,j} a_{ij} X_i X_j$. This is known as "order 2 chaos". More generally, we would consider $f(X) = \sum_{i,j} h(X_i, X_j)$.

3.1 Hanson-Wright Inequality

Theorem 3.1 (Hanson-Wright)**.** Let $X = (X_1, \dots, X_n)^\top$ be independent entries with $\mathbb{E}X_i = 0$ and each X_i is σ^2-subgaussian. Then for universal constants $C, c > 0$, $\forall A \in \mathbb{R}^{n\times n}$, we have $\forall t \geq 0$,

$$\mathbb{P}(|X^\top AX - \mathbb{E}[X^\top AX]| \geq t) \leq C \exp\left(-c \min\left(\frac{t^2}{\sigma^2 ||A||_F^2}, \frac{t}{\sigma^2 ||A||_{\text{op}}}\right)\right).$$

For diagonal matrix A, $X^\top AX = \sum_{i=1}^n a_i X_i^2$, then for each i, $||a_i X_i^2||_{\psi_1} \leq |a_i|\sigma^2$, which implies that $\sum_{i=1}^n a_i (X_i^2 - \mathbb{E}[X_i^2])$ is $(\underbrace{\sum_{i=1}^n a_i^2}_{||A||_F^2}\sigma^4, \underbrace{\max_{1\leq i\leq n} |a_i|}_{||A||_{\text{op}}}\sigma^2)$-subexponential.

Example 3.2. Let $X = BZ$, where B is fixed and Z_i's are iid with $\mathbb{E}[Z_i] = 0$, $\text{Var}(Z_i) = 1$, and σ^2-subgaussian. For $B = I_n$, $||X||_2 = \sqrt{n} + \mathcal{O}_{\mathbb{P}}(1)$.

What is the behavior of $||X||_2$ for general B?

Note that $||X||_2^2 = Z^\top B^\top BZ$, so $\mathbb{E}[||X||_2^2] = \mathbb{E}[\sum_{i,j=1}^n (B^\top B)_{ij} Z_i Z_j] = \text{Trace}(B^\top B) = ||B||_F^2$. Also note that $\sigma^2 \geq \text{Var}(Z_i) = 1$, $||B^\top B||_{\text{op}} = ||B||_{\text{op}}^2$, and $||B^\top B||_F^2 \leq ||B||_{\text{op}}^2 ||B||_F^2$, so $\forall u \geq 0$,

$$\begin{aligned}
\mathbb{P}(|||X||_2^2 - ||B||_F^2| \geq u) &\leq C \exp\left(-c \min\left(\frac{u^2}{\sigma^4 ||B^\top B||_F^2}, \frac{u}{\sigma^2 ||B^\top B||_{\text{op}}}\right)\right) \\
&\leq C \exp\left(-\frac{c}{\sigma^4 ||B||_{\text{op}}^2} \min\left(\frac{u^2}{||B||_F^2}, u\right)\right) \\
\implies \mathbb{P}\left(\left|\frac{1}{||B||_F^2} ||X||_2^2 - 1\right| \geq s\right) &\leq C \exp\left(-\frac{c||B||_F^2}{\sigma^4 ||B||_{\text{op}}^2} \min(s^2, s)\right).
\end{aligned}$$

By a fact in lecture 2, we have

$$\mathbb{P}\left(\left|\frac{1}{||B||_F} ||X||_2 - 1\right| \geq \delta\right) \leq \mathbb{P}\left(\left|\frac{1}{||B||_F^2} ||X||_2^2 - 1\right| \geq \max(\delta, \delta^2)\right) \leq C \exp\left(-\frac{c||B||_F^2}{\sigma^4 ||B||_{\text{op}}^2} \delta^2\right),$$

which implies that

$$\mathbb{P}\left(|||X||_2 - ||B||_F| \geq t\right) \leq C \exp\left(-\frac{ct^2}{\sigma^4 ||B||_{\text{op}}^2}\right).$$

So, $||X||_2 = ||B||_F + o_{\mathbb{P}}(||B||_{\text{op}})$.

Proof of Hanson-Wright Inequality 3.1. To start with, $X^\top AX - \mathbb{E}[X^\top AX] = \sum_{i=1}^n a_{ii}(X_i^2 - \mathbb{E}[X_i^2]) + \sum_{i\neq j} a_{ij} X_i X_j$. The first term is $(\sum_{i=1}^n a_{ii}^2 \sigma^4, \max_{1\leq i\leq n} |a_{ii}|\sigma^2)$-subexponential. By Bernstein inequality,

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_{ii}(X_i^2 - \mathbb{E}[X_i^2])\right| \geq t/2\right) \leq C \exp\left(-c \min\left(\frac{t^2}{\sigma^4 ||A||_F^2}, \frac{t}{\sigma^2 ||A||_{\text{op}}}\right)\right).$$

It remains to bound $\mathbb{P}(|\sum_{i\neq j} a_{ij}X_iX_j| \geq t/2)$. Let $S = \sum_{i\neq j} a_{ij}X_iX_j = X^\top \bar{A} X$, where $\bar{A}$ is A with zero diagonal entries. By Chernoff bound, the right-tail $\mathbb{P}(S \geq t/2) \leq e^{-\frac{\lambda t}{2}}\mathbb{E}[e^{\lambda S}]$.

(i) Doucoupling: $\mathbb{E}[e^{\lambda S}] \leq \mathbb{E}[e^{4\lambda X^\top \bar{A}\tilde{X}}]$, where $\tilde{X}$ is an independent copy of X.

Proof. Let $\delta_1, \ldots, \delta_n \overset{iid}{\sim} \text{Bernoulli}(\frac{1}{2})$. Then

$$S = 4\mathbb{E}_\delta\left[\sum_{i\neq j} a_{ij}\delta_i(1-\delta_i)X_iX_j\right] = 4\mathbb{E}_{\mathcal{I}}\left[\sum_{i\in\mathcal{I}, j\notin\mathcal{I}} a_{ij}X_iX_j\right],$$

where $\mathcal{I} = \{i \mid \delta_i = 1\}$. Hence, we can bound the MGF of S by Jensen's inequality:

$$\mathbb{E}[e^{\lambda S}] = \mathbb{E}_X\left\{\exp\left(4\lambda\mathbb{E}_{\mathcal{I}}\left[\sum_{i\in\mathcal{I}, j\notin\mathcal{I}} a_{ij}X_iX_j\right]\right)\right\} \leq \mathbb{E}_{\mathcal{I}}\mathbb{E}_X\left[\exp\left(4\lambda\sum_{i\in\mathcal{I}, j\notin\mathcal{I}} a_{ij}X_iX_j\right)\right].$$

For fixed $\mathcal{I}$, let $\mathcal{X} = \{X_i \mid i \in \mathcal{I}\} \cup \{\tilde{X}_j \mid j \notin \mathcal{I}\}$. Then

$$\begin{aligned}\mathbb{E}\left[\exp(4\lambda X^\top A\tilde{X})\right] &= \mathbb{E}\left[\exp\left(4\lambda\sum_{i\in\mathcal{I}, j\notin\mathcal{I}} a_{ij}X_iX_j\right)\mathbb{E}\left[\exp\left(4\lambda\sum_{i\notin\mathcal{I} \text{ or } i,j\in\mathcal{I}} a_{ij}X_iX_j\right) \mid \mathcal{X}\right]\right] \\ &\geq \mathbb{E}\left[\exp\left(4\lambda\sum_{i\in\mathcal{I}, j\notin\mathcal{I}} a_{ij}X_iX_j\right)\right]\end{aligned}$$

□

(ii) From subgaussian to gaussian:

$$\mathbb{E}[\exp(4\lambda X^\top \bar{A}\tilde{X})] \leq \mathbb{E}[\exp(4\lambda\sigma^2 G^\top \bar{A}\tilde{G})],$$

where $G = (g_1, \ldots, g_n)^\top \sim \mathcal{N}(0, I_n)$. Same for $\tilde{G}$.

Proof. For any $v \in \mathbb{R}^n$, by Hoeffding inequality 1.7, $v^\top X = \sum_{i=1}^n v_iX_i$ is $\sum_{i=1}^n v_i^2\sigma^2 = \sigma^2||v||_2^2$-subgaussian, i.e.,

$$E[e^{\lambda v^\top X}] \leq \exp\left(\frac{\lambda^2\sigma^2||v||_2^2}{2}\right) = \mathbb{E}[e^{\lambda\sigma v^\top G}].$$

Thus,

$$\mathbb{E}[e^{4\lambda X^\top \bar{A}X}] = \mathbb{E}_{\tilde{X}}\left[\mathbb{E}_X\left[e^{4\lambda X^\top \bar{A}\tilde{X}}\right]\right] \leq \mathbb{E}_{\tilde{X}}\left[\mathbb{E}_G\left[e^{4\lambda\sigma G^\top \bar{A}\tilde{X}}\right]\right] = \mathbb{E}_G\left[\mathbb{E}_{\tilde{X}}\left[e^{4\lambda\sigma G^\top \bar{A}\tilde{X}}\right]\right] = \mathbb{E}\left[e^{4\lambda\sigma^2 G^\top \bar{A}\tilde{G}}\right]$$

□

(iii) Rotational Invariance. Let $\bar{A} = UDV^\top$ be the singular value decomposition, where $U, V \in \mathbb{R}^{n\times n}$ are orthogonal. Since $G, \tilde{G} \sim \mathcal{N}(0, I_n)$, $\text{Law}(G, \tilde{G}) = \text{Law}(U^\top G, V^\top \tilde{G})$, so

$$\mathbb{E}[\exp(4\lambda\sigma^2 G^\top A\tilde{G})] = \mathbb{E}[\exp(4\lambda\sigma^2 (U^\top G)^\top D(V^\top \tilde{G})] = \mathbb{E}\left[\exp\left(4\lambda\sigma^2\sum_{i=1}^n d_ig_i\tilde{g}_i\right)\right].$$

For each i, $||d_i g_i \tilde{g}_i||_{\psi_1} \le d_i ||g_i||_{\psi_2} ||\tilde{g}_i||_{\psi_2} \le C d_i$ since g_i and $\tilde{g}_i$ are gaussian. Thus,

$$4\lambda\sigma^2 \sum_{i=1}^n d_i g_i \tilde{g}_i \text{ is } (C\lambda^2\sigma^4 \underbrace{\sum_{i=1}^n d_i^2}_{||A||_F^2}, C\lambda\sigma^2 \underbrace{\max_{1\le i\le n} d_i}_{||A||_{\text{op}}})\text{-subexponential,}$$

which implies $\mathbb{E}\left[\exp(4\lambda\sigma^2 G^\top A\tilde{G})\right] \le \exp(C\lambda^2\sigma^4||A||_F^2)$ for $\lambda \le 1/(C\sigma^2||A||_{\text{op}})$. Hence,

$$\mathbb{P}(S \ge t/2) \le \exp(-\lambda t/2 + C\lambda^2\sigma^4||A||_F^2).$$

$$\text{Pick } \lambda \asymp \begin{cases} \frac{t}{\sigma^4||A||_F^2} & \text{if } \frac{t}{\sigma^4||A||_F^2} \lesssim \frac{1}{\sigma^2||A||_{\text{op}}} \\ \frac{1}{\sigma^2||A||_{\text{op}}} & \text{otherwise} \end{cases}$$

□

3.2 Symmetrization and Decoupling

Definition 3.3. ε is a Rademacher random variable if $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$.

Lemma 3.4 (Rademacher Symmetrization)**.** Let $X_1, \dots, X_n$ independent with $\mathbb{E}[X_i] = 0$. Let $F : [0,\infty) \mapsto \mathbb{R}$ be increasing and convex. Let $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim}$ Rademacher, and independent of $X_1, \dots, X_n$. Then,

$$\mathbb{E}F\left(\frac{1}{2}\left|\sum_{i=1}^n \varepsilon_i X_i\right|\right) \le \mathbb{E}F\left(\left|\sum_{i=1}^n X_i\right|\right) \le \mathbb{E}F\left(2\left|\sum_{i=1}^n \varepsilon_i X_i\right|\right)$$

Note that F is increasing and convex implies that $x \mapsto F(|x|)$ is convex.

Proof of Lemma 3.4. Let $\tilde{X}_i$ be an independent copy of X_i for each i. On one hand, we have

$$\begin{aligned}
\mathbb{E}F\left(\left|\sum_{i=1}^n X_i\right|\right) &= \mathbb{E}F\left(\left|\mathbb{E}_{\tilde{X}} \sum_{i=1}^n (X_i - \tilde{X}_i)\right|\right) \\
&\le \mathbb{E}F\left(\left|\sum_{i=1}^n (X_i - \tilde{X}_i)\right|\right) \quad \text{by Jensen's inequality} \\
&= \mathbb{E}F\left(\left|\sum_{i=1}^n \varepsilon_i (X_i - \tilde{X}_i)\right|\right) \\
&\le \mathbb{E}F\left(\left|\sum_{i=1}^n \varepsilon_i X_i\right| + \left|\sum_{i=1}^n \varepsilon_i \tilde{X}_i\right|\right) \quad \text{by monotonicity} \\
&\le \frac{1}{2}\mathbb{E}F\left(2\left|\sum_{i=1}^n \varepsilon_i X_i\right|\right) + \frac{1}{2}\mathbb{E}F\left(2\left|\sum_{i=1}^n \varepsilon_i \tilde{X}_i\right|\right) \quad \text{by convexity} \\
&= \mathbb{E}F\left(2\left|\sum_{i=1}^n \varepsilon_i X_i\right|\right)
\end{aligned}$$

Similarly, by Jensen's inequality and monotonicity and convexity of F, we have

$$\begin{aligned}
\mathbb{E}F\left(\frac{1}{2}\left|\sum_{i=1}^n \varepsilon_i X_i\right|\right) &= \mathbb{E}_{\varepsilon,X}F\left(\frac{1}{2}\left|\mathbb{E}_{\tilde{X}}\sum_{i=1}^n \varepsilon_i(X_i-\tilde{X}_i)\right|\right)\\
&\le \mathbb{E}F\left(\frac{1}{2}\left|\sum_{i=1}^n \varepsilon_i(X_i-\tilde{X}_i)\right|\right)\\
&= \mathbb{E}F\left(\frac{1}{2}\left|\sum_{i=1}^n (X_i-\tilde{X}_i)\right|\right)\\
&\le \mathbb{E}F\left(\frac{1}{2}\left|\sum_{i=1}^n X_i\right| + \frac{1}{2}\left|\sum_{i=1}^n \tilde{X}_i\right|\right) =\le \mathbb{E}F\left(\left|\sum_{i=1}^n X_i\right|\right)
\end{aligned}$$

□

Lemma 3.5 (Decoupling, [dlP92])**.** Let $X_1,\dots,X_n$ be independent, $h_{ij}:\mathbb{R}^2\mapsto\mathbb{R}$, $F:[0,\infty)\mapsto\mathbb{R}$ convex and increasing. Let $\tilde{X}_1,\dots,\tilde{X}_n$ be independent copies of $X_1,\dots,X_n$. Then

$$\mathbb{E}F\left(\left|\sum_{i\neq j} h_{ij}(X_i,X_j)\right|\right) \le \mathbb{E}F\left(8\left|\sum_{i\neq j} h_{ij}(X_i,\tilde{X}_j)\right|\right).$$

If futher $h_{ij}=h_{ji}$ and $h_{ij}(x,y)=h_{ji}(y,x)$ then

$$\mathbb{E}F\left(\left|\sum_{i\neq j} h_{ij}(X_i,X_j)\right|\right) \ge \mathbb{E}F\left(\frac{1}{4}\left|\sum_{i\neq j} h_{ij}(X_i,\tilde{X}_j)\right|\right).$$

Proof of Lemma 3.5. For the upper bound we have

$$\begin{aligned}
\mathbb{E}F\left(\left|\sum_{i\neq j} h_{ij}(X_i,X_j)\right|\right) &= \mathbb{E}_X F\left(\left|\mathbb{E}_{\tilde{X}}\sum_{i\neq j} h_{ij}(X_i,X_j)\right|\right)\\
&\le \mathbb{E}_X F\left(|\mathbb{E}_{\tilde{X}}\mathrm{I}| + |\mathbb{E}_{\tilde{X}}\mathrm{II}| + |\mathbb{E}_{\tilde{X}}\mathrm{III}| + |\mathbb{E}_{\tilde{X}}\mathrm{IV}|\right) \quad \text{by monotonicity}\\
&\le \frac{1}{2}\mathbb{E}_X F(2|\mathbb{E}_{\tilde{X}}\mathrm{I}|) + \frac{1}{6}\left(\mathbb{E}_X F(6|\mathbb{E}_{\tilde{X}}\mathrm{II}|) + \mathbb{E}_X F(6|\mathbb{E}_{\tilde{X}}\mathrm{III}|) + \mathbb{E}_X F(6|\mathbb{E}_{\tilde{X}}\mathrm{IV}|)\right),
\end{aligned}$$

where

- $\mathrm{I} = \sum_{i\neq j}[h_{ij}(X_i,X_j) + h_{ij}(X_i,\tilde{X}_j) + h_{ij}(\tilde{X}_i,X_j) + h_{ij}(\tilde{X}_i,\tilde{X}_j)]$
- $\mathrm{II} = \sum_{i\neq j} h_{ij}(X_i,\tilde{X}_j)$, $\mathrm{III} = \sum_{i\neq j} h_{ij}(\tilde{X}_i,X_j)$, $\mathrm{IV} = \sum_{i\neq j} h_{ij}(\tilde{X}_i,\tilde{X}_j)$

We first deal with the latter three terms.

- $\mathbb{E}_X F(6|\mathbb{E}_{\tilde{X}}\mathrm{II}|) \le \mathbb{E}F(6|\mathrm{II}|)$ by Jensen's inequality.
- $\mathbb{E}_X F(6|\mathbb{E}_{\tilde{X}}\mathrm{III}|) \le \mathbb{E}F(6|\mathrm{III}|) = \mathbb{E}F(6|\mathrm{II}|)$.
- $F(6|\mathbb{E}_{\tilde{X}}\mathrm{IV}| = F(6|\sum_{i\neq j}\underbrace{\mathbb{E}h(\tilde{X}_i,\tilde{X}_j)}_{=\mathbb{E}h(X_i,\tilde{X}_j)}|) \le \mathbb{E}F(6|\mathrm{II}|)$ by Jensen's inequality.

Now let $Z_i = \{X_i, \tilde{X}_i\}$ and $\mathscr{Z} = (Z_1, \ldots, Z_n)$. Then conditioning on $\mathscr{Z}$, the first term

$$\mathrm{I} = 4 \sum_{i \neq j} \mathbb{E}[h_{ij}(X_i, \tilde{X}_j) \mid \mathscr{Z}],$$

and then by Jensen's inequality,

$$\mathbb{E}_X F(2|\mathbb{E}_{\tilde{X}}\mathrm{I}|) \leq \mathbb{E}F(2|\mathrm{I}|) = \mathbb{E}F\left(8\left|\sum_{i \neq j} \mathbb{E}[h_{ij}(X_i, \tilde{X}_j) \mid \mathscr{Z}]\right|\right) \leq \mathbb{E}F(8|\mathrm{II}|),$$

which implies (by previous steps and monotonicity)

$$\mathbb{E}F\left(\left|\sum_{i \neq j} h_{ij}(X_i, X_j)\right|\right) \leq \frac{1}{2}\mathbb{E}F(8|\mathrm{II}|) + \frac{1}{2}\mathbb{E}F(6|\mathrm{II}|) \leq \mathbb{E}F\left(8\left|\sum_{i \neq j} h_{ij}(X_i, \tilde{X}_j)\right|\right).$$

As for the lower bound, we have

$$\begin{aligned}
\mathbb{E}F\left(\left|\sum_{i \neq j} h_{ij}(X_i, \tilde{X}_j)\right|\right) &= \mathbb{E}F\left(\frac{1}{2}\left|\sum_{i \neq j}(h_{ij}(X_i, \tilde{X}_j) + h_{ij}(\tilde{X}_i, X_j)\right|\right) \quad \text{since } h_{ij}(x,y) = h_{ji}(y,x) \\
&\leq \mathbb{E}F\left(\frac{1}{2}|\mathrm{I}| + \frac{1}{2}\left|\sum_{i \neq j} h_{ij}(X_i, X_j)\right| + \frac{1}{2}\left|\sum_{i \neq j} h_{ij}(\tilde{X}_i, \tilde{X}_j)\right|\right) \\
&\leq \frac{1}{2}\mathbb{E}F(|\mathrm{I}|) + \frac{1}{4}\mathbb{E}F\left(2\left|\sum_{i \neq j} h_{ij}(X_i, X_j)\right|\right) + \frac{1}{4}\mathbb{E}F\left(2\left|\sum_{i \neq j} h_{ij}(\tilde{X}_i, \tilde{X}_j)\right|\right) \quad \text{convexity} \\
&= \frac{1}{2}\mathbb{E}F\left(4\left|\sum_{i \neq j} \mathbb{E}[h_{ij}(X_i, X_j) \mid \mathscr{Z}]\right|\right) + \frac{1}{2}\mathbb{E}F\left(2\left|\sum_{i \neq j} h_{ij}(X_i, X_j)\right|\right) \\
&\leq \frac{1}{2}\mathbb{E}F\left(4\left|\sum_{i \neq j} \mathbb{E}[h_{ij}(X_i, X_j) \mid \mathscr{Z}]\right|\right) + \frac{1}{2}\mathbb{E}F\left(2\left|\sum_{i \neq j} h_{ij}(X_i, X_j)\right|\right) \quad \text{Jensen} \\
&\leq \mathbb{E}F\left(4\left|\sum_{i \neq j} \mathbb{E}[h_{ij}(X_i, X_j)]\right|\right)
\end{aligned}$$

□

3.3 U-Statistics

Let $X_1, \ldots, X_n$ be iid. Let $U = \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j)$ with $||h||_\infty \leq B$ and $h(x,y) = h(y,x)$. Suppose $\mathbb{E}[U] = \mathbb{E}[h(X_i, X_j)] = 0$.

By asymptotic theory (e.g., in *Asymptotic Statistics* by Van der Vaart [Vaa98], §12):

- Let $h^{(1)}(x) = \mathbb{E}h(x, X_j)$. If $U = \frac{2}{n}\sum_{i=1}^n h^{(1)}(X_i) + \mathcal{O}_{\mathbb{P}}(\frac{1}{n})$, then

$$\sqrt{n}U \overset{\mathscr{D}}{\rightsquigarrow} \mathcal{N}(0, 4\mathbb{E}h^{(1)}(X_i)^2).$$

- If $h^{(1)}(x) = 0$ for all $x \in \mathbb{R}$, then h is degenerate. As $n \to \infty$,

$$nU \overset{\mathscr{D}}{\rightsquigarrow} \sum_{k=1}^{\infty} \lambda_k (Z_k^2 - 1), \quad \{Z_k\}_{k=1}^{\infty} \overset{iid}{\sim} \mathcal{N}(0,1), \lambda_k \geq 0.$$

From bounded differences inequality of Lecture 1:

$$\mathbb{P}(|U| \geq t) \leq 2\exp\left(-\frac{nt^2}{8B^2}\right).$$

Thus, $U = \mathcal{O}_{\mathbb{P}}(\frac{1}{\sqrt{n}})$. This is the correct order only if h is non-degenerate.

Theorem 3.6 (Arcones, Giné'93 [AG93])**.** Let $X_1, \ldots, X_n$ be iid. Suppose $||h||_\infty \leq B$, $h(x,y) = h(y,x)$, $\mathbb{E}h(X_i, X_j) = 0$, $\mathbb{E}h(x, X_j) = 0$ for all $x \in \mathbb{R}$ (degenerate). Then for some $C, c > 0$, $U = \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j)$ satisfies

$$\mathbb{P}(|U| \geq t) \leq C\exp\left(-\frac{cnt}{B}\right).$$

Proof. Let $F : [0,\infty) \mapsto \mathbb{R}$ be convex and increasing. Let $\varepsilon_i, \tilde{\varepsilon}_i$ be iid Rademacher, $\tilde{X}_i$ be independent copy of X_i. Then

$$\begin{aligned}
\mathbb{E}F\left(\left|\sum_{i\neq j} h(X_i, X_j)\right|\right) &\leq \mathbb{E}F\left(8\left|\sum_{i\neq j} h(X_i, \tilde{X}_j)\right|\right) \quad \text{by decoupling Lemma} \\
&= \mathbb{E}F\left[8\left|\sum_{i=1}^n \underbrace{\sum_{j:j\neq i} h(X_i, \tilde{X}_j)}_{\text{independent, mean-0}}\right|\right] \quad \text{conditioned on } \tilde{X}_1, \ldots, \tilde{X}_n \\
&\leq \mathbb{E}F\left(16\left|\sum_{i=1}^n \varepsilon_i \sum_{j:j\neq i} h(X_i, \tilde{X}_j)\right|\right) \quad \text{by symmetrization} \\
&= \mathbb{E}F\left[16\left|\sum_{j=1}^n \underbrace{\sum_{i:i\neq j} \varepsilon_i h(X_i, \tilde{X}_j)}_{\text{independent, mean-0}}\right|\right] \quad \text{conditioned on } \varepsilon_1, \ldots, \varepsilon_n, \tilde{X}_1, \ldots, \tilde{X}_n \\
&\leq \mathbb{E}F\left(32\left|\sum_{i\neq j} \varepsilon_i \tilde{\varepsilon}_j h(X_i, \tilde{X}_j)\right|\right) \quad \text{symmetrization} \\
&\leq \mathbb{E}F\left(128\left|\sum_{i\neq j} \varepsilon_i \varepsilon_j h(X_i, X_j)\right|\right) \quad \text{reverse decoupling}
\end{aligned}$$

Consider $F(x) = \exp(\lambda x/2)$: by applying MGF bound from proof of Hanson-Wright Inequality,

$$\begin{aligned}
\mathbb{E}e^{\lambda|\sum_{i<j} h(X_i,X_j)|} &\leq \mathbb{E}e^{128\lambda|\sum_{i<j} \varepsilon_i\varepsilon_j h(X_i,X_j)|} \\
&\leq \mathbb{E}_X\mathbb{E}_\varepsilon\left[e^{128\lambda\sum_{i<j} \varepsilon_i\varepsilon_j h(X_i,X_j)} + e^{-128\lambda\sum_{i<j} \varepsilon_i\varepsilon_j h(X_i,X_j)}\right] \\
&\leq \mathbb{E}[2e^{c\lambda^2||H||_F^2}], \quad \forall|\lambda| \leq \frac{1}{c||H||_{\text{op}}}, H \in \mathbb{R}^{n\times n}, H_{ij} = H_{ji} = h(X_i, X_j), H_{ii} \equiv 0 \\
&\leq 2e^{c\lambda^2 n^2 B^2}, \quad \forall|\lambda| \leq \frac{1}{cnB}.
\end{aligned}$$

Then, choose $\lambda \asymp 1/(nB)$, we have: $\mathbb{P}\left(\frac{1}{\binom{n}{2}}\left|\sum_{i<j} h(X_i, X_j)\right| \geq t\right) \leq e^{-\lambda\binom{n}{2}t} \cdot 2e^{C\lambda^2 n^2 B^2} \leq C'e^{-\frac{cnt}{B}}$. □

4 Matrix Concentration Inequalities

Readings: §5.4-5.6 in [Ver18], §8 in [Tro15].

Theorem 4.1 (Matrix Bernstein)**.** If $X_1, \dots, X_n$ are $d \times d$ independent symmetric matrices with $\mathbb{E}[X_i] = 0$, $||X_i||_{\text{op}} \le b$ with probability 1. Set $v = \left\|\sum_{i=1}^n \mathbb{E}[X_i^2]\right\|_{\text{op}}$ as the "matrix variance". Then

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\|_{\text{op}} \ge t\right) \le 2d \exp\left(-\frac{t^2/2}{v + bt/3}\right).$$

Remark 4.2.

1. v is the matrix variance, representing "maximal variance of $S = \sum_{i=1}^n X_i$" in any direction:
$$v = \left\|\sum_{i=1}^n \mathbb{E}X_i^2\right\|_{\text{op}} = ||\mathbb{E}S^2||_{\text{op}}^2 = \sup_{||u||_2=1} u^\top \mathbb{E}S^2 u = \sup_{||u||=1} \mathbb{E}\|Su\|_2^2.$$
2. This shows that $\left\|\sum_{i=1}^n \mathbb{E}X_i^2\right\|_{\text{op}} = \mathcal{O}_{\mathbb{P}}\left(\sqrt{v \log d} + b \log d\right)$. The $\log d$ factors may not be sharp, but this bound is good enough for many applications.
3. There are versions that relax $\|X_i\|_{\text{op}} \le b$ to moment-type assumptions, see [Tro12].

Example 4.3 (Covariance Estimation)**.** Let $X_1, \dots, X_n \in \mathbb{R}^d$ be iid with $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i X_i^\top = \Sigma$. Suppose that $\|X_i\|_2^2 \le Cd\|\Sigma\|_{\text{op}}$ with probability 1. Examples of this could be (with $\Sigma = I_d$ and $C = 1$):

- $X_i \sim \text{Unif}(\pm\sqrt{d}e_1, \dots, \pm\sqrt{d}e_d)$.
- $X_i \sim \text{Unif}(\sqrt{d}\mathbb{S}^{d-1})$ (sphere of radius $\sqrt{d}$).

We estimate Σ by $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$. How large is $\|\hat{\Sigma} - \Sigma\|_{\text{op}}$? Note that

$$\|\hat{\Sigma} - \Sigma\|_{\text{op}} = \left\|\sum_{i=1}^n \frac{1}{n}(X_i X_i^\top - \Sigma)\right\|_{\text{op}} := \left\|\sum_{i=1}^n Y_i\right\|_{\text{op}}.$$

- $\mathbb{E}Y_i = 0$
- $\|Y_i\|_{\text{op}} \le \frac{1}{n}\left(\|X_i X_i^\top\|_{\text{op}} + \|\Sigma\|_{\text{op}}\right) = \frac{1}{n}\left(\|X_i\|_2^2 + \|\Sigma\|_{\text{op}}\right) \le \frac{C'd}{n}\|\Sigma\|_{\text{op}}$.
- We also have
$$\begin{aligned}
\mathbb{E}Y_i^2 &= \frac{1}{n^2}\left(\mathbb{E}(X_i X_i^\top)^2 - \mathbb{E}X_i X_i^\top \Sigma - \mathbb{E}\Sigma X_i X_i^\top + \Sigma^2\right) \\
&= \frac{1}{n^2}\left(\mathbb{E}(X_i X_i^\top)^2 - \Sigma^2\right) \\
&\preceq \frac{1}{n^2}\mathbb{E}(X_i X_i^\top)^2 = \frac{1}{n^2}\mathbb{E}\|X_i\|_2^2 X_i X_i^\top \\
&\preceq \frac{1}{n^2} Cd\|\Sigma\|_{\text{op}}\Sigma
\end{aligned}$$
Together with $\mathbb{E}Y_i^2 \succeq 0$, we have $v := \left\|\sum_{i=1}^n \mathbb{E}Y_i^2\right\|_{\text{op}} \le \frac{Cd}{n}\|\Sigma\|_{\text{op}}^2$.

By matrix Bernstein:

$$\|\hat{\Sigma} - \Sigma\|_{\text{op}} = \mathcal{O}_{\mathbb{P}}\left(\sqrt{v \log d} + b \log d\right) = \mathcal{O}_{\mathbb{P}}\left(\left(\sqrt{\frac{d \log d}{n}} + \frac{d \log d}{n}\right)\|\Sigma\|_{\text{op}}\right).$$

We postpone the proof of matrix Bernstein to develop the tools we need.

4.1 Lieb's Concavity Theorem

For $X = U \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} U^\top \in \mathbb{R}^{n\times n}$, define $f(X)$ by functional calculus: $f(X) = U \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{bmatrix} U^\top$.

Idea: apply moment generating function approach to matrices.

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}\left(\sum_{i=1}^n X_i\right) \geq t\right) &\leq e^{-\lambda t}\mathbb{E}e^{\lambda\cdot\lambda_{\max}(\sum_{i=1}^n X_i)} \quad \forall \lambda \geq 0\\ &= e^{-\lambda t}\mathbb{E}\lambda_{\max}\left(e^{\lambda\sum_{i=1}^n X_i}\right)\\ &\leq e^{-\lambda t}\operatorname{Tr}\mathbb{E}e^{\lambda\sum_{i=1}^n X_i}\end{aligned}$$

Issue: for matrices, $\operatorname{Tr} e^{\lambda\sum_{i=1}^n X_i}$ is not necessarily no larger than $\operatorname{Tr} e^{\lambda X_1}\times\cdots\times e^{\lambda X_n}$.

Theorem 4.4 (Lieb)**.** For any $H \in \mathbb{R}^{d\times d}$ symmetric, $A \mapsto \operatorname{Tr} e^{H+\log A}$ is concave over $A \succ 0$.

With this, we have

$$\begin{aligned}\mathbb{E}\operatorname{Tr} e^{\lambda\sum_{i=1}^n X_i} &= \mathbb{E}\operatorname{Tr} e^{\lambda\sum_{i=1}^{n-1} X_i + \log e^{\lambda X_n}}\\ &\leq \mathbb{E}\operatorname{Tr} e^{\lambda\sum_{i=1}^{n-1} X_i + \log \mathbb{E}e^{\lambda X_n}}\\ &\leq \ldots \leq \operatorname{Tr} e^{\sum_{i=1}^n \log \mathbb{E}e^{\lambda X_i}}\end{aligned}$$

Further analyzing $\mathbb{E}e^{\lambda X_i}$ will show matrix Bernstein.

Definition 4.5. Let $A, H \in \mathbb{R}^{d\times d}$ be symmetric and $A, H \succ 0$. The matrix relative entropy is

$$D(A\|H) = \operatorname{Tr}[A(\log A - \log H) - (A - H)].$$

Lemma 4.6 (Matrix relative entropy)**.** $(A, H) \mapsto D(A\|H)$ is non-negative and convex over $\{(A, H) : A, H \succ 0\}$.

Remark 4.7. This is elementary restricting to diagonal A, H:

$$D(A\|H) = \sum_{i=1}^d [a_i \log\frac{a_i}{h_i} - (a_i - h_i)].$$

Let $f(x) = x - 1 - \log x$. We have $af(\frac{h}{a}) = a\log\frac{a}{h} - (a - h)$.

- $f(x) \geq 0$ on $(0,\infty)$. This implies $af(\frac{h}{a}) \geq 0$. That is, $D(A\|H) \geq 0$.
- f is convex on $(0,\infty)$. This implies that $af(\frac{a}{h})$ is convex in (a, h). Note that the Hessian is

$$\begin{bmatrix} \frac{h^2}{a^2} & -\frac{h}{a} \\ -\frac{h}{a} & 1 \end{bmatrix} \cdot \frac{1}{a} f''\left(\frac{h}{a}\right) \succeq 0,$$

  so $D(A\|H)$ is convex in (A, H).

We will extend this argument below to non-diagonal A, H.

Proof of Lieb's Theorem 4.4. For any $M, T \succ 0$,

$$D(T\|M) = \operatorname{Tr}[T(\log T - \log M) - (T - M)] \geq 0,$$

which implies that $\mathrm{Tr}M \geq \mathrm{Tr}[T\log M - T\log T + T]$. Equality holds for $T = M$. So

$$\mathrm{Tr}M = \sup_{T\succ 0} \mathrm{Tr}[T\log M - T\log T + T],$$

which implies that

$$\begin{aligned}\mathrm{Tr}\, e^{H+\log A} &= \sup_{T\succ 0} \mathrm{Tr}[T(H+\log A) - T\log T + T] \\ &= \sup_{T\succ 0} \underbrace{\mathrm{Tr}\,TH + \mathrm{Tr}\,A - D(T\|A)}_{\text{concave in } (T,A)}.\end{aligned}$$

Hence, $A \mapsto \mathrm{Tr}\, e^{H+\log A}$ is concave. □

4.2 Analysis of Matrix Relative Entropy

Definition 4.8. A function $f : \mathcal{I} \mapsto \mathbb{R}$ on an interval $\mathcal{I} \subseteq \mathbb{R}$ is

$$\begin{cases}\text{operator monotone increasing} \\ \text{operator convex} \\ \text{trace monotone increasing}\end{cases}$$

if, for all $d \geq 1$ and $A, B \in \mathbb{R}^{d\times d}$ symmetric with all eigenvalues in $\mathcal{I}$,

$$\begin{cases}A \preceq B \Longrightarrow f(A) \preceq f(B) \\ f(\lambda A + (1-\lambda)B) \preceq \lambda f(A) + (1-\lambda)f(B), \quad \forall \lambda \in [0,1] \\ A \preceq B \Longrightarrow \mathrm{Tr}\, f(A) \leq \mathrm{Tr}\, f(B)\end{cases}$$

Trace monotonicity is the simplest condition:

Proposition 4.9. If $f : \mathcal{I} \mapsto \mathbb{R}$ is monotone increasing, then it is also trace monotone increasing.

Proof. Let $\lambda_1(A) \geq \ldots \lambda_d(A)$ be eigenvalues of A. If $A \preceq B$, then $\lambda_i(A) \leq \lambda_i(B)$ for all $1 \leq i \leq d$. This is by Courant-Fischer Theorem, where

$$\lambda_i(A) = \max_{V\subseteq\mathbb{R}^d, \dim V = i} \min_{u\in V, \|u\|_2=1} u^\top A u \leq \max_V \min_u u^\top B u = \lambda_i(B).$$

Thus, $\mathrm{Tr}\, f(A) = \sum_{i=1}^d f(\lambda_i(A)) \leq \sum_{i=1}^d f(\lambda_i(B)) = \mathrm{Tr}\, f(B)$. □

However, it is **not** true that f increasing/convex implies that f is operator increasing/convex.

Example 4.10. Let $A = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$, $B = \begin{bmatrix}2 & 1\\ 1 & 1\end{bmatrix}$, $f(x) = x^2$ on $[0,\infty)$. Then $A \preceq B$ but

$$A^2 = \begin{bmatrix}2 & 2\\ 2 & 2\end{bmatrix} \not\preceq \begin{bmatrix}6 & 3\\ 3 & 2\end{bmatrix} = B^2 \Longrightarrow x^2 \text{ is not operator increasing.}$$

Let $A = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$, $B = \begin{bmatrix}3 & 1\\ 1 & 1\end{bmatrix}$, $f(x) = x^3$ on $[0,\infty)$. Then $A \preceq B$ but we can check

$$\left(\frac{A+B}{2}\right)^3 \not\preceq \frac{A^3+B^3}{2} \Longrightarrow x^3 \text{ is not operator convex.}$$

Proposition 4.11. (a) Fix any $u \geq 0$. $a \mapsto (a+u)^{-1}$ is operator monotone decreasing and operator convex on $(0, \infty)$.

(b) $a \mapsto \log a$ is operator monotone increasing and operator concave on $(0, \infty)$.

Proof. (a) Suppose $B \succeq A \succ 0$. Then $B + uI \succeq A + uI$. Note that if $B \succeq A$ then $M^\top BM \succeq M^\top AM$, which implies

$$\begin{aligned}
&\Longrightarrow I \succeq (B+uI)^{-1/2}(A+uI)(B+uI)^{-1/2} \\
&\Longrightarrow I \preceq [(B+uI)^{-1/2}(A+uI)(B+uI)^{-1/2}]^{-1} = (B+uI)^{1/2}(A+uI)^{-1}(B+uI)^{1/2} \\
&\Longrightarrow (A+uI)^{-1} \succeq (B+uI)^{-1}
\end{aligned}$$

This shows monotonicity. Now consider any $A, B \succ 0$ and $\lambda \in [0,1]$. Note that if $A \succ 0$, then $\begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \succ 0$ if and only if $C - B^\top A^{-1} B \succeq 0$. Then

$$\begin{aligned}
0 &\preceq \lambda \underbrace{\begin{bmatrix} A+uI & I \\ I & (A+uI)^{-1} \end{bmatrix}}_{\succeq 0} + (1-\lambda) \underbrace{\begin{bmatrix} B+uI & I \\ I & (B+uI)^{-1} \end{bmatrix}}_{\succeq 0} \\
&= \begin{bmatrix} \lambda A + (1-\lambda)B + uI & I \\ I & \lambda(A+uI)^{-1} + (1-\lambda)(B+uI)^{-1} \end{bmatrix},
\end{aligned}$$

which implies $\lambda(A+uI)^{-1} + (1-\lambda)(B+uI)^{-1} - (\lambda A + (1-\lambda)B + uI)^{-1} \succeq 0$. This shows convexity.

(b) Note that

$$\int_0^\infty \left(\frac{1}{1+u} - \frac{1}{a+u} \right) \mathrm{d}u = \lim_{L \to \infty} \log(1+u)|_0^L - \log(a+u)|_0^L = \log a + \lim_{L \to \infty} \log \frac{1+L}{a+L} = \log a.$$

If $A \preceq B$, then

$$\log A = \int_0^\infty \left[(1+u)^{-1} I - (A+uI)^{-1} \right] \mathrm{d}u \overset{\text{by(a)}}{\preceq} \int_0^\infty \left[(1+u)^{-1} I - (B+uI)^{-1} \right] \mathrm{d}u = \log B.$$

For any $A, B \succ 0$, $\lambda \in [0,1]$,

$$\begin{aligned}
\lambda \log A + (1-\lambda) \log B &= \int_0^\infty \left(\lambda \left[(1+u)^{-1} I - (A+uI)^{-1} \right] + (1-\lambda) \left[(1+u)^{-1} I - (B+uI)^{-1} \right] \right) \mathrm{d}u \\
&= \int_0^\infty \left((1+u)^{-1} I - \left[\lambda (A+uI)^{-1} + (1-\lambda)(B+uI)^{-1} \right] \right) \mathrm{d}u \\
&\overset{\text{by(a)}}{\succeq} \int_0^\infty \left[(1+u)^{-1} I - (\lambda A + (1-\lambda)B + uI)^{-1} \right] \mathrm{d}u = \log(\lambda A + (1-\lambda)B).
\end{aligned}$$

□

Lemma 4.12 (Operator Jensen)**.** Suppose $f : \mathcal{I} \mapsto \mathbb{R}$ is operator convex, $A_1 \in \mathbb{R}^{d_1 \times d_1}$, $A_2 \in \mathbb{R}^{d_2 \times d_2}$, $K_1 \in \mathbb{R}^{d_1 \times d}$, $K_2 \in \mathbb{R}^{d_2 \times d}$ such that $K_1^\top K_1 + K_2^\top K_2 = I$. Then

$$f(K_1^\top A_1 K_1 + K_2^\top A_2 K_2) \preceq K_1^\top f(A_1) K_1 + K_2^\top f(A_2) K_2.$$

Proof. Let $A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$ and $U = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}$. Note that $\begin{bmatrix} K_1 \\ K_2 \end{bmatrix}$ has orthonormal columns, and pick L_1, L_2 such that $Q = \begin{bmatrix} K_1 & L_1 \\ K_2 & L_2 \end{bmatrix}$ is orthogonal. Then

- $Q^\top AQ = \begin{bmatrix} K_1^\top A_1 K_1 + K_2^\top A_2 K_2 & \star \\ \star & \star \end{bmatrix}$
- $\frac{1}{2}\begin{bmatrix} T & B \\ B^\top & M \end{bmatrix} + \frac{1}{2}U^\top \begin{bmatrix} T & B \\ B^\top & M \end{bmatrix} U = \begin{bmatrix} T & 0 \\ 0 & M \end{bmatrix}$ for any T, B, M.

So, by block-diagonal we have

$$f(K_1^\top A_1 K_1 + K_2^\top A_2 K_2) = f([Q^\top AQ]_{11}) = f\left(\left[\frac{1}{2}Q^\top AQ + \frac{1}{2}U^\top Q^\top AQU\right]_{11}\right) = \left[f\left(\frac{1}{2}Q^\top AQ + \frac{1}{2}U^\top Q^\top AQU\right)\right]_{11}.$$

By operator convexity and that Q, QU are orthogonal,

$$f\left(\frac{1}{2}Q^\top AQ + \frac{1}{2}U^\top Q^\top AQU\right) \preceq \frac{1}{2}f(Q^\top AQ) + \frac{1}{2}f(U^\top Q^\top AQU) = \frac{1}{2}Q^\top f(A)Q + \frac{1}{2}U^\top Q^\top f(A)QU,$$

which implies

$$f(K_1^\top A_1 K_1 + K_2^\top A_2 K_2) \preceq \left[\frac{1}{2}Q^\top f(A)Q + \frac{1}{2}U^\top Q^\top f(A)QU\right]_{11} = [Q^\top f(A)Q]_{11} = K_1^\top f(A_1)K_1 + K_2^\top f(A_2)K_2.$$

□

Proof of Matrix Entropy Lemma 4.6. Let $A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1d}B \\ \vdots & & \vdots \\ a_{d1}B & \cdots & a_{dd}B \end{bmatrix} \in \mathbb{R}^{d^2 \times d^2}$ be the Kronecker product. $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. If $A = \sum_i \lambda_i u_i u_i^\top$, $B = \sum_j \mu_j v_j v_j^\top$ are the eigen-decompositions,

$$A \otimes B = \sum_{i,j} \lambda_i \mu_j (u_i \otimes v_j)(u_i \otimes v_j)^\top \Longrightarrow \log(A \otimes B) = \log A \otimes I + I \otimes \log B.$$

Let $f(x) = x - 1 - \log x$ be non-negative and operator convex on $(0, \infty)$ and $\varphi : \mathbb{R}^{d^2 \times d^2} \mapsto \mathbb{R}$ with

$$\varphi(M) = \mathrm{Vec}(I_d)^\top M \mathrm{Vec}(I_d), \quad \mathrm{Vec}(I_d) \in \mathbb{R}^{d^2}.$$

Specifically, we have

$$\varphi(A \otimes B) = \sum_{i,j=1}^{d} (A \otimes B)_{(i,i),(j,j)} = \sum_{i,j=1}^{d} A_{ij} B_{ij} = \mathrm{Tr}(AB^\top).$$

This implies that (since both $A, H, \log A, \log H$ are symmetric)

$$\begin{aligned} D(A\|H) &= \mathrm{Tr}(A(\log A - \log H) - (A - H)) \\ &= \mathrm{Tr}((A \log A)I^\top) - \mathrm{Tr}(A \log H^\top) - \mathrm{Tr}(AI^\top) + \mathrm{Tr}(IH^\top) \\ &= \varphi(A \log A \otimes I) - \varphi(A \otimes \log H) - \varphi(A \otimes I) + \varphi(I \otimes H) \\ &= \varphi(A \log A \otimes I - A \otimes \log H - (A \otimes I) + (I \otimes H)) \\ &= \varphi((A \otimes I)[A^{-1} \otimes H - I \otimes I - \underbrace{\log(A^{-1} \otimes H)}_{-\log A \otimes I + I \otimes \log H}]) \\ &= \varphi\left((A \otimes I) f(A^{-1} \otimes H)\right) \end{aligned}$$

Since $A \otimes I$ and $I \otimes H$ commute,

$$(A \otimes I) f(A^{-1} \otimes H) = (A \otimes I)^{1/2} f\left((A \otimes I)^{-1/2}(I \otimes H)(A \otimes I)^{-1/2}\right)(A \otimes I)^{1/2} \succeq 0$$

because f is nonnegative, which implies $D(A\|H) = \varphi((A \otimes I) f(A^{-1} \otimes H)) \geq 0$. Now consider any $A_1, A_2, H_1, H_2 \succ 0$, $\lambda \in [0,1]$. Set $A = \lambda A_1 + (1-\lambda)A_2, H = \lambda H_1 + (1-\lambda)H_2$, $K_1 = \sqrt{\lambda} A_1^{1/2} A^{-1/2}, K_2 = \sqrt{1-\lambda} A_2^{1/2} A^{-1/2}$, we have $K_1^\top K_1 + K_2^\top K_2 = I$. This implies

$$\begin{aligned}
A^{1/2} f(A^{-1/2} H A^{-1/2}) A^{1/2} &= A^{1/2} f\left(\lambda A^{-1/2} H_1 A^{-1/2} + (1-\lambda) A^{-1/2} H_2 A^{-1/2}\right) A^{1/2} \\
&= A^{1/2} f\left(K_1^\top A_1^{-1/2} H_1 A_1^{-1/2} K_1 + K_2^\top A_2^{-1/2} H_2 A_2^{-1/2} K_2\right) A^{1/2} \\
&\overset{\text{operator Jensen}}{\preceq} A^{1/2}\left[K_1^\top f(A_1^{-1/2} H_1 A_1^{-1/2}) K_1 + K_2^\top f(A_2^{-1/2} H_2 A_2^{-1/2}) K_2\right] A^{1/2} \\
&= \lambda A_1^{1/2} f(A_1^{-1/2} H_1 A_1^{-1/2}) A_1^{1/2} + (1-\lambda) A_2^{1/2} f(A_2^{-1/2} H_2 A_2^{-1/2}) A_2^{1/2}.
\end{aligned}$$

Apply this with $A \otimes I$ and $I \otimes H$ in place of A and H:

$$(A \otimes I) f(A^{-1} H) \preceq \lambda (A_1 \otimes I) f(A_1^{-1} \otimes H_1) + (1-\lambda)(A_2 \otimes I) f(A_2^{-1} \otimes H_2),$$

which implies

$$D(A\|H) \leq \lambda D(A_1\|H_1) + (1-\lambda) D(A_2\|H_2).$$

□

Proof of Matrix Bernstein Inequality 4.1. Let $S = \sum_{i=1}^n X_i, \mathbb{E}X_i = 0, \|X_i\|_{\text{op}} \leq b$. For any $\lambda \geq 0$, we have

$$\begin{aligned}
\mathbb{P}[\lambda_{\max}(S) \geq t] &\leq e^{-\lambda t} \mathbb{E}\, e^{\lambda \cdot \lambda_{\max}(S)} \\
&\leq e^{-\lambda t} \operatorname{Tr} \mathbb{E}\, e^{\lambda S} \\
&\leq e^{-\lambda t} \operatorname{Tr} \exp\left(\sum_{i=1}^n \log \mathbb{E}\, e^{\lambda X_i}\right) \quad \text{by Lieb's Theorem}
\end{aligned}$$

One can check that

$$e^{\lambda x} \leq 1 + \lambda x + \frac{\lambda^2 x^2}{1 - \lambda x/3} \leq 1 + \lambda x + \underbrace{\frac{\lambda^2}{1 - \lambda b/3}}_{:= g(\lambda)} x^2, \quad \forall |x| \leq b, \lambda \in [0, 3/b).$$

This implies that $\mathbb{E}\, e^{\lambda X_i} \preceq I + g(\lambda)\mathbb{E}X_i^2, \forall \lambda \in [0, 3/b)$. By operator monotonicity of log,

$$\log \mathbb{E}\, e^{\lambda X_i} \preceq \log(I + g(\lambda)\mathbb{E}X_i^2) \preceq g(\lambda)\mathbb{E}X_i^2.$$

By trace monotonicity of exp, let $v = \|\sum_{i=1}^n \mathbb{E}X_i^2\|_{\text{op}}$,

$$\mathbb{P}[\lambda_{\max}(S) \geq t] \leq e^{-\lambda t} \operatorname{Tr} e^{g(\lambda) \sum_{i=1}^n \mathbb{E}X_i^2} \leq d e^{-\lambda t} e^{g(\lambda) v} = d \exp\left(-\frac{t^2/2}{v + bt/3}\right)$$

for $\lambda = \frac{t}{v + bt/3}$. Similarly, $\mathbb{P}[\lambda_{\min}(S) \leq -t] \leq d \exp\left(-\frac{t^2/2}{v+bt/3}\right)$. □

5 Efron-Stein Inequality, Poincaré Inequalities, Tensorization of Entropy

Readings: §3.1-3.5, 3.7, 4.8-4.9, 4.13 in [BLM13].

$X_1, \dots, X_n$ are independent random variables. $f(X_1, \dots, X_n)$ is a "general" function. How to bound $\mathrm{Var}(f(X_1, \dots, X_n))$?

Theorem 5.1 (Efron-Stein)**.** If $X_1, \dots, X_n$ are independent, $Z = f(X_1, \dots, X_n)$ such that $\mathsf{Var}(Z) < \infty$, then $\mathsf{Var}(Z) \le \sum_{i=1}^n \mathbb{E}[\mathsf{Var}^{(i)}(Z)]$, where

$$\mathsf{Var}^{(i)}(Z) = \mathsf{Var}(Z \mid \underbrace{X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n}_{X^{(i)}}).$$

This is a "tensorization property" of variance, and reduces bounds for $\mathsf{Var}Z$ to variance bounds for univariate functions.

Proof. Define $M_i := \mathbb{E}[Z \mid X_1, \dots, X_i]$ and $\Delta_i := M_i - M_{i-1}$. Then $Z - \mathbb{E}Z = \sum_{i=1}^n \Delta_i$. Note that for $i > j$,

$$\mathbb{E}[\Delta_i \Delta_j] = \mathbb{E}[\Delta_j \mathbb{E}[\Delta_i \mid X_i, \dots, X_j]] = \mathbb{E}[\Delta_j (\underbrace{\mathbb{E}[M_i \mid X_1, \dots, X_j]}_{=M_j} - M_j)] = 0,$$

which implies that $\mathsf{Var}Z = \mathbb{E}[(Z - \mathbb{E}[Z])^2] = \sum_{i=1}^n \mathbb{E}[\Delta_i^2]$. Then apply

$$\Delta_i^2 = \Big(\mathbb{E}[Z - \mathbb{E}^{(i)}Z \mid X_1, \dots, X_i]\Big)^2 \le \mathbb{E}[(Z - \mathbb{E}^{(i)}Z)^2 \mid X_1, \dots, X_i]$$

we obtain

$$\mathsf{Var}(Z) \le \sum_{i=1}^n \mathbb{E}[(Z - \mathbb{E}^{(i)}Z)^2] = \sum_{i=1}^n \mathbb{E}\mathsf{Var}^{(i)}(Z).$$

□

Equivalent forms: set $v = \sum_{i=1}^n \mathbb{E}\mathsf{Var}^{(i)}(Z)$.

1. Apply $\mathsf{Var}Y = \frac{1}{2}\mathbb{E}(Y - Y')^2$ for Y' being an indepedent copy of Y. Then

$$v = \frac{1}{2}\sum_{i=1}^n \mathbb{E}[(Z - Z_i')^2], \quad Z_i' = f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n),$$

   where X_i' is an independent copy of X_i.

2. By symmetry, $Z - Z_i' \overset{\mathscr{D}}{=} Z_i' - Z$, so $v = \sum_{i=1}^n \mathbb{E}(Z - Z_i')_+^2 = \sum_{i=1}^n \mathbb{E}(Z - Z_i')_-^2$.

Example 5.2. Suppose f has bounded differences property $\|D_i f\|_\infty < \infty$. Then

$$\mathsf{Var}(Z) \le \sum_{n=1}^n \|D_i f\|_\infty^2.$$

Example 5.3. Recall Rademacher complexity: $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} \mathrm{Rademacher}(\pm 1)$. For $T \subseteq \mathbb{R}^n$,

$$Z = f(\varepsilon_1, \dots, \varepsilon_n) = \sup_{t \in T} \varepsilon^\top t.$$

Here $\|D_i f\|_\infty = 2\sup_{t\in T} |t_i|$, which implies that $\sum_{i=1}^n \|D_i f\|_\infty^2 = 4\sum_{i=1}^n \sup_{t\in T} |t_i|^2$. Let $t^* = t^*(\varepsilon)$ be the point where the supremum is attained for ε.

$$Z = \varepsilon^\top t^*, \quad Z_i' := (\varepsilon_1, \dots, \varepsilon_{i-1}, \varepsilon_i', \varepsilon_{i+1}, \dots \varepsilon_n)^\top t^*(\varepsilon).$$

This implies

$$(Z - Z_i)_+ \le \left((\varepsilon_i - \varepsilon_i')t_i^*\right)_+ \le \mathbb{1}_{\varepsilon_i \neq \varepsilon_i'} \cdot 2|t_i^*|.$$

Hence,

$$\mathsf{Var}(Z) \le \frac{1}{2}\sum_{i=1}^n \mathbb{E}[(Z - Z')_+^2] \le \frac{1}{2}\cdot 4\sum_{i=1}^n \mathbb{E}[|t_i(\varepsilon)|^2] = 2\sup_{t\in T} \|t^*(\varepsilon)\|^2.$$

Example 5.4 (First Passage Percolation)**.** Let G be a graph with n edges with iid weights $X_i \ge 0$ for $i \in [n]$. $\mathbb{E}[X_i^2] = \sigma^2$. Fixing $u, v \in G$, let

$$Z = \inf_{\text{path } P:\, u \to v} \sum_{e_i \in P} X_i.$$

Let P^* be the path where the infimum is attained for X.

$$(Z - Z_i')_- \le (X_i' - X_i)_+ \mathbb{1}_{e_i \in P^*} \le X_i' \cdot \mathbb{1}_{e_i \in P^*},$$

which implies that

$$\mathsf{Var}(Z) \le \sum_{i=1}^n (Z - Z_i')_-^2 \le \sigma^2 \mathbb{E}[\text{length of } P^*].$$

If, for example, $X_i \in [a, b]$ with probability 1, then the length of P^* is upper bounded by $\frac{b}{a} d_G(u, v)$, while $\mathbb{E}Z \ge a \cdot d_G(u, v)$.

Example 5.5 (Configuration Functions)**.** We call a property Π <u>hereditory</u> if, for any $k \ge 1$ and sequence $(x_1, \dots, x_k)$,

$$(x_1, \dots, x_n) \text{ satisfies } \Pi \implies \text{ any subsequence of } (x_1, \dots, x_n) \text{ satisfies } \Pi.$$

For example,

- $x_1, \dots, x_k$ are distinct.
- $x_1, \dots, x_k$ are increasing.
- $x_1, \dots, x_k$ are <u>shattered</u> by a collection of sets $\mathscr{A}$ if $\forall S \subseteq \{x_1, \dots, x_n\}$, $\exists A \in \mathscr{A}$ such that

$$S = \{x_1, \dots, x_n\} \cap A.$$

Let $Z = f(X_1, \dots, X_n)$ be length of longest subsequence $(X_1, \dots, X_k)$ that satisfies Π.

(i) # distinct values

(ii) length of longest increasing subsequence

(iii) VC-dimension of $\mathscr{A}$ restricted to $\{X_1, \dots, X_k\}$.

Any such f is self-bounding:

1. $0 \le f(X_1, \dots, X_n) - f(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \le 1$.
2. $\sum_{i=1}^n \{f(X_1, \dots, X_n) - f(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)\} \le f(X_1, \dots, X_n)$.

Since $\mathsf{Var}^{(i)}(Z) \le \mathbb{E}^{(i)}[(f(X) - f(X^{(i)})^2]$, where $X^{(i)} = X \setminus X_i$, we have

$$\mathsf{Var}(Z) \le \mathbb{E}\left[\sum_{i=1}^n \left(f(X) - f(X^{(i)})\right)^2\right] \le \mathbb{E}\left[\sum_{i=1}^n f(X) - f(X^{(i)})\right] \le \mathbb{E}f(X) = \mathbb{E}Z.$$

5.1 Poincaré Inequalities

For certain distributions $X_1, \dots, X_n$ and certain functions f,

$$\mathsf{Var}(f(X_1, \dots, X_n)) \le C\mathbb{E}\left[\text{“squared gradeint of } f\text{”}\right].$$

Theorem 5.6. Let $X_1, \dots, X_n \overset{iid}{\sim}$ Rademacher, $f : \{\pm 1\}^n \mapsto \mathbb{R}$. Define its discrete gradient as

$$\mathrm{grad} f(x_1, \dots, x_n) = [D_i f(x_1, \dots, x_n)]_{i=1}^n,$$

where $D_i f(x_1, \dots, x_n) = f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, -x_i, \dots, x_n)$. Then

$$\mathsf{Var}(f(X_1, \dots, X_n)) \le \frac{1}{4}\mathbb{E}\left\|\mathrm{grad} f(X_1, \dots, X_n)\right\|^2.$$

Proof. In one dimension, $n = 1$:

$$\mathsf{Var}(f(X)) = \left(\frac{f(1) - f(-1)}{2}\right)^2 = \frac{1}{4}\mathbb{E}\left(\mathrm{grad} f(X)\right)^2.$$

For general n, by tensorization (Theorem 5.1),

$$\mathsf{Var}(f(X_1, \dots, X_n) \le \sum_{i=1}^n \mathbb{E}[\underbrace{\mathsf{Var}^{(i)}(f(X_1, \dots, X_n))}_{=\frac{1}{4}\mathbb{E}^{(i)}[D_i f(X_1, \dots, X_n)^2]}] = \frac{1}{4}\mathbb{E}\left\|\mathrm{grad} f(X_1, \dots, X_n)\right\|^2$$

□

Example 5.7. Let $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim}$ Rademacher and $f(\varepsilon) = \sup_{t \in T} \varepsilon^\top t$. By previous example, we have $D_i f(\varepsilon)_+ \le 2|t_i^*(\varepsilon)|$. Then

$$D_i f(\varepsilon)_+ = D_i f(\varepsilon_1, \dots, -\varepsilon_i, \dots, \varepsilon_n) \overset{\mathscr{D}}{=} D_i f(\varepsilon)_- \Longrightarrow \mathbb{E} D_i f(\varepsilon)^2 \le 8 t_i^*(\varepsilon)^2.$$

Therefore, with Theorem 5.6, we have

$$\mathsf{Var} f(\varepsilon) \le \frac{1}{4}\mathbb{E}\|\mathrm{grad} f(\varepsilon)\|^2 \le 2\mathbb{E}\|t^*(\varepsilon)\|^2 \le 2\sup_{t \in T}\|t\|^2.$$

Theorem 5.8 (Gaussian Poincaré Inequality)**.** Let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(0, 1)$ and $f : \mathbb{R}^n \mapsto \mathbb{R}$ weakly differentiable. Then $\mathsf{Var} f(X_1, \dots, X_n) \le \mathbb{E}\|\nabla f(X_1, \dots, X_n)\|^2$.

Proof. In one dimension ($n = 1$), assume that $f : \mathbb{R} \mapsto \mathbb{R}$ has bounded support, twice continuously differentiable, so $\|f\|, \|f'\|, \|f''\| \le K$ for some K. Introduce $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim}$ Rademacher, $S_m := \frac{1}{\sqrt{m}}\sum_{i=1}^m \varepsilon_i$. By previous Poincaré inequality on hypercube,

$$\begin{aligned}
\mathsf{Var}(f(S_m)) &\le \sum_{i=1}^m \frac{1}{4}\mathbb{E}[D_i f(S_m)^2] \\
&= \sum_{i=1}^m \frac{1}{4}\mathbb{E}\left(f(S_m) - f(S_m - \frac{2\varepsilon_i}{\sqrt{m}})\right)^2 \\
&\le \frac{m}{4}\mathbb{E}\left(\frac{2}{\sqrt{m}}|f'(S_m)| + \frac{2K}{m}\right)^2 = \mathbb{E}[|f'(S_m)|^2] + \mathcal{O}\left(\frac{1}{\sqrt{m}}\right).
\end{aligned}$$

Since $S_m \overset{\mathscr{D}}{\rightsquigarrow} X \sim \mathcal{N}(0,1)$, as $m \to \infty$,

$$\mathsf{Var}(f(X)) = \lim_{m\to\infty} \mathsf{Var}(f(S_m)) \le \lim_{m\to\infty} \mathbb{E}[|f'(S_m)|^2] + \mathcal{O}\left(\frac{1}{\sqrt{m}}\right) = \mathbb{E}[|f'(X)|^2].$$

For general weakly differentiable $f : \mathbb{R} \mapsto \mathbb{R}$: apply to smooth, compactly supported $\{f_n\}$ such that $\lim_{n\to\infty} \mathsf{Var}(f_n(Z)) = \mathsf{Var}(f(Z))$, then $\lim_{n\to\infty} \mathbb{E} f_n'(Z)^2 = \mathbb{E} f'(Z)^2$. For general n, by tensorization,

$$\mathsf{Var} f(X_1, \dots, X_n) \le \sum_{i=1}^n \mathsf{Var}^{(i)}(f(X_1, \dots, X_n)) \le \mathbb{E}\|\nabla f(X_1, \dots, X_n)\|^2,$$

where for the first term, $\mathsf{Var}^{(i)}(f(X_1, \dots, X_n)) \le \mathbb{E}^{(i)}[(\partial_i f(X_1, \dots, X_n))^2]$ by result for $n = 1$. □

Corollary 5.9. If f is L-Lipschitz, i.e., $|f(x) - f(x')| \le L\|x - x'\|_2$, then $\mathsf{Var}(f(X_1, \dots, X_n)) \le L^2$.

Note that the result of this lemma is dimension-free, i.e., there is no explicit dependence on n.

Example 5.10 (Gaussian Complexity)**.** Let $g \sim \mathcal{N}(0, I), g \in \mathbb{R}^n$. $f(g) = \sup_{t\in T} g^\top t$, where $T \subseteq \mathbb{R}^n$. For fixed $t \in T$ and any $g, g' \in \mathbb{R}^n$, $|g^\top t - g'^\top t| \le \|g - g'\|_2 \|t\|_2$, so $g \mapsto g^\top t$ is $\|t\|_2$-Lipschitz. Then

$$|f(g) - f(g')| = \left|\sup_{t\in T} g^\top t - \sup_{t\in T} g'^\top t\right| \le \sup_{t\in T} |g^\top t - g'^\top t| \le \|g - g'\|_2 \sup_{t\in T} \|t\|_2,$$

which implies f is $\sup_{t\in T} \|t\|_2$-Lipschitz, and so $\mathsf{Var}(f(g)) \le \sup_{t\in T} \|t\|_2^2$.

Theorem 5.11 (Convex Poincaré Inequality)**.** Let $X_1, \dots, X_n$ are indepedent. $X_i \in [0,1]$ with probability 1. $f : [0,1]^n \mapsto \mathbb{R}$ is weakly differentiable and separately convex in each variable. Then

$$\mathsf{Var}(f(X_1, \dots, X_n)) \le \mathbb{E}\|\nabla f(X_1, \dots, X_n)\|^2.$$

Proof. For $n = 1$: let $x^* = \arg\min_{x\in[0,1]} f(x)$.

$$\mathsf{Var}(f(X)) \le \mathbb{E}[(f(X) - f(x^*))^2] \overset{\text{convexity}}{\le} \mathbb{E}[f'(X)^2 (X - x^*)^2] \le \mathbb{E}[f'(X)^2],$$

where the first inequality holds because $\mathbb{E}[f(X)]$ is the minimizer of $\mathbb{E}[(f(X) - c)^2]$ over all c and the second inequality holds because $0 \le f(X) - f(x^*) \le f'(X)(X - x^*)$.

For general $f : \mathbb{R}^n \mapsto \mathbb{R}$,

$$\mathsf{Var}(f(X_{1:n})) \le \sum_{i=1}^n \mathbb{E}[\mathsf{Var}^{(i)}(f(X_{1:n}))] \le \sum_{i=1}^n \mathbb{E}^{(i)}[(\partial_i f(X_{1:n}))^2] = \mathbb{E}[\|\nabla f(X_{1:n})\|^2].$$

□

Example 5.12. Let $X \in \mathbb{R}^{m\times n}$ have independent entries $X_{ij} \in [a, b]$. Let

$$\sigma_{\max}(X) = \sup_{u,v:\|u\|_2=\|v\|_2=1} u^\top X v.$$

- $|u^\top X v - u^\top X' v| \le \|u\|_2 \cdot \|X - X'\|_{\text{op}} \cdot \|v\|_2 \le \|X - X'\|_F$, which implies $X \mapsto u^\top X v$ is 1-Lipschitz, so $\sigma_{\max}(X)$ is 1-Lipschitz.
- $\sigma_{\max}(X)$ is supremum of linear functions, hence it is convex.

Then $\sigma_{\max}(X)$ is a convex, $(b - a)$-Lipschitz function of $\{Z_{ij}\}$ where $Z_{ij} = \frac{X_{ij} - a}{b - a} \in [0, 1]$. By previous Corollary, $\mathsf{Var}(\sigma_{\max}(X)) \le (b - a)^2$.

5.2 Tensorization of Entropy

Definition 5.13. For a nonnegative random variable Z with $\mathbb{E}[Z \ln Z] < \infty$, its *entropy* is

$$\mathrm{Ent} Z = \mathbb{E}[Z \ln Z] - \mathbb{E}[Z] - \ln \mathbb{E}[Z] \geq 0.$$

Interpretation: Suppose $Z \geq 0$ is defined on $(\Omega, \mathbb{P})$. If $\mathbb{E}Z = 1$, then we can interpret $Z(\omega) = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\omega)$ of $\mathbb{Q}$ on Ω with respect to $\mathbb{P}$. Then

$$\mathrm{Ent}(Z) = \mathbb{E}[Z \ln Z] = \int \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \ln \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \mathrm{d}\mathbb{P}(\omega) = \int \ln \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \mathrm{d}\mathbb{Q}(\omega) = \mathsf{KL}(\mathbb{Q} \| \mathbb{P}).$$

For $\mathbb{E}Z = c \neq 1$, let $Z = c\tilde{Z}$ so $\mathbb{E}\tilde{Z} = 1$. Then

$$\mathrm{Ent} Z = \mathbb{E}[c\tilde{Z} \ln(c\tilde{Z})] - \mathbb{E}[c\tilde{Z}] \ln \mathbb{E}[c\tilde{Z}] = c \mathrm{Ent} \tilde{Z},$$

so $\mathrm{Ent}(\cdot)$ is a homogeneous extension of KL divergence.

Compare with variance: $\mathsf{Var} Z = \mathbb{E}Z^2 - (\mathbb{E}Z)^2$. If $\mathbb{E}Z = 1$, then $Z = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\omega)$ and

$$\mathsf{Var} Z = \int \left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\omega) \right)^2 \mathrm{d}\mathbb{P}(\omega) - 1 = \chi^2(\mathbb{Q} \| \mathbb{P}),$$

so $\mathsf{Var}(\cdot)$ is a homogeneous extension of χ^2-divergence.

Theorem 5.14. If $X_1, \ldots, X_n$ are independent, $Z = f(X_1, \ldots, X_n)$ nonnegative with $\mathbb{E}[Z \ln Z] < \infty$, then $\mathrm{Ent} Z \leq \sum_{i=1}^n \mathbb{E}[\mathrm{Ent}^{(i)}(Z)]$, where

$$\mathrm{Ent}^{(i)}(Z) = \mathrm{Ent}(Z \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) = \mathbb{E}^{(i)}[Z \ln Z] - \mathbb{E}^{(i)}[Z] \ln \mathbb{E}^{(i)}[Z]$$

Lemma 5.15 (Duality)**.** Let $Z \geq 0$ be a random variable on $(\Omega, \mathbb{P})$ and $\mathbb{E}[Z \ln Z] < \infty$. Then

$$\mathrm{Ent} Z = \sup_{U : \Omega \mapsto [-\infty, \infty), \mathbb{E} e^U = 1} \mathbb{E}[UZ]$$

Proof. By homogeneity, assume $\mathbb{E}Z = 1$. Consider any $U : \Omega \mapsto [-\infty, \infty)$ where $\mathbb{E}[e^U] = 1$. Let $\mathbb{Q}$ be such that $e^U = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$.

$$1 = \mathbb{E}_{\mathbb{P}}[Z] = \int Z \mathrm{d}\mathbb{P} = \int Z \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} \mathrm{d}\mathbb{Q} = \int e^{-U} Z \mathrm{d}\mathbb{Q} = \mathbb{E}_{\mathbb{Q}}[e^{-U} Z].$$

Then, we have

$$\begin{aligned} \mathrm{Ent}_{\mathbb{P}}(Z) = \mathbb{P}[Z \ln Z] = \mathbb{E}_{\mathbb{Q}}[e^{-U} Z \ln Z] &= \mathbb{E}_{\mathbb{Q}}[e^{-U} Z \ln e^{-U} Z] + \mathbb{E}_{\mathbb{Q}}[e^{-U} U Z] \\ &= \mathrm{Ent}_{\mathbb{Q}}(e^{-U} Z) + \mathbb{E}_{\mathbb{P}}[UZ], \end{aligned}$$

so $\mathrm{Ent}_{\mathbb{P}}(Z) \geq \sup_{U : \Omega \mapsto [-\infty, \infty), \mathbb{E}[e^U] = 1} \mathbb{E}_{\mathbb{P}}[UZ]$. Equality holds when $e^U = Z$, i.e., $U = \ln Z$. □

Remark 5.16. The supremum can be restricted to $U = f(Z)$ since equality holds for such U.

Proof of Theorem 5.14. Let $Z = f(X_1, \ldots, X_n)$ and $M_i := \mathbb{E}[Z \mid X_{1:i}]$.

$$\mathrm{Ent}(Z) = \mathbb{E}[Z \ln Z] - \mathbb{E}[Z] \ln \mathbb{E}[Z] = \mathbb{E}[Z(\ln Z - \ln \mathbb{E}[Z])] = \mathbb{E}\left[\sum_{i=1}^n Z(\ln M_i - \ln M_{i-1}) \right].$$

By duality, $\mathrm{Ent}^{(i)}(Z) \geq \mathbb{E}^{(i)}[Z(\ln M_i - \log \mathbb{E}^{(i)}[M_i])] = \mathbb{E}^{(i)}[Z(\ln M_i - M_{i-1})]$, where for each i, let $U = \ln M_i - \ln M_{i-1} = \ln M_i - \ln \mathbb{E}^{(i)}[M_i]$, so $\mathbb{E}^{(i)}[e^U] = \mathbb{E}^{(i)}\left[\frac{M_i}{\mathbb{E}^{(i)}[M_i]} \right] = 1$. This implies that

$$\mathrm{Ent} Z \leq \mathbb{E} \sum_{i=1}^n \mathrm{Ent}^{(i)} Z.$$

□

6 Entropy Method, Log-Sobolev Inequalities, Concentration of Gaussian Measure

Readings: §5.1-5.5, 6.3-6.7 in [BLM13].

Let $X_1, \dots, X_n$ be independent and $Z = f(X_1, \dots, X_n)$.

Recall: For $Y \geq 0$, $\mathsf{Ent}\, Y = \mathbb{E}Y \log Y - \mathbb{E}Y \log \mathbb{E}Y$. We also showed that $\mathsf{Ent}\, Z \leq \sum_{i=1}^n \mathbb{E}[\mathsf{Ent}^{(i)} Z]$.

Entropy method: exponential tail bounds for Z by building entropy of $Y = e^{\lambda Z}$ via $\mathbb{E}Y$.

Lemma 6.1 (Herbst's argument)**.** Suppose for some $\sigma^2 > 0$, we have

$$\mathsf{Ent}(e^{\lambda Z}) \leq \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda Z}], \qquad \forall \lambda > 0\,.$$

Then $\mathbb{E}[e^{\lambda(Z - \mathbb{E}Z)}] \leq e^{\lambda^2 \sigma^2/2}$, $\forall \lambda > 0$, and

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}, \qquad t \geq 0\,.$$

Proof. Let $F(\lambda) = \mathbb{E}[e^{\lambda Z}]$ be the MGF of Z and $F'(\lambda) = \mathbb{E}[Ze^{\lambda Z}]$. Then,

$$\begin{aligned}\mathsf{Ent}(e^{\lambda Z}) &= \mathbb{E}[\lambda Z e^{\lambda Z}] - \mathbb{E}[e^{\lambda Z}] \log \mathbb{E}[e^{\lambda Z}] \\ &= \lambda F'(\lambda) - F(\lambda) \log F(\lambda) \leq \frac{\lambda^2 \sigma^2}{2} F(\lambda)\,,\end{aligned}$$

which implies

$$\frac{\sigma^2}{2} \geq \frac{1}{\lambda} \cdot \frac{F'(\lambda)}{F(\lambda)} - \frac{1}{\lambda^2} \log F(\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda} \left(\frac{1}{\lambda} \log F(\lambda) \right).$$

By L'Hôpital,

$$\lim_{\lambda \to 0^+} \frac{\log F(\lambda)}{\lambda} = \lim_{\lambda \to 0^+} \frac{F'(\lambda)}{F(\lambda)} = \mathbb{E}[Z]\,,$$

so, by integrating on both sides, we obtain

$$\frac{1}{\lambda} \log F(\lambda) - \mathbb{E}[Z] = \int_0^\lambda \frac{\mathrm{d}}{\mathrm{d}u} \left[\frac{1}{u} \log F(u) \right] \mathrm{d}u \leq \int_0^\lambda \frac{\sigma^2}{2} \mathrm{d}u = \frac{\lambda \sigma^2}{2}\,,$$

which implies

$$F(\lambda) \leq \exp\left(\lambda \mathbb{E}[Z] + \frac{\lambda^2 \sigma^2}{2} \right) \Longrightarrow \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq e^{\frac{\lambda^2 \sigma^2}{2}}\,.$$

□

6.1 Log-Sobolev inequalities

By tensorization of entropy, for $Z = f(X_1, \dots, X_n)$ with independent X_i's,

$$\mathsf{Ent}(e^{\lambda Z}) \leq \sum_{i=1}^n \mathbb{E}[\mathsf{Ent}^{(i)}(e^{\lambda Z})]\,.$$

LSI: for certain distributions of $X_1, \dots, X_n$ and functions f,

$$\mathsf{Ent}[f(X_1, \dots, X_n)^2] \leq C \cdot \mathbb{E}\,[\text{“squared gradient” of } f]\,.$$

Theorem 6.2. Let $X_1, \dots, X_n \overset{iid}{\sim} \mathsf{Rademacher}$, $f : \{\pm 1\}^n \mapsto \mathbb{R}$,

$$\operatorname{grad} f(x) = \{D_i f(x)\}_{i=1}^n = \{f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, -x_i, \dots, x_n)\}_{i=1}^n .$$

Then $\mathsf{Ent}[f(X_1, \dots, X_n)^2] \le \frac{1}{2}\mathbb{E}\|\operatorname{grad} f(X_1, \dots, X_n)\|^2$.

Proof. When $n = 1$, $f : \{\pm 1\} \mapsto \mathbb{R}$. Let $a = f(1)$ and $b = f(-1)$.

$$\begin{aligned}
\mathsf{Ent}(f(X)^2) &= \mathbb{E}[f(X)^2 \log f(X)^2] - \mathbb{E}[f(X)^2] \log \mathbb{E}[f(X)^2] \\
&= \frac{a^2 \log a^2 + b^2 \log b^2}{2} - \frac{a^2 + b^2}{2} \log \frac{a^2 + b^2}{2} \\
\mathbb{E}[(\operatorname{grad} f(X))^2] &= \frac{1}{2}((a-b)^2 + (b-a)^2) = (a-b)^2
\end{aligned}$$

It suffices to show that, WLOG $a \ge b \ge 0$,

$$\frac{a^2 \log a^2 + b^2 \log b^2}{2} - \frac{a^2 + b^2}{2} \log \frac{a^2 + b^2}{2} - \frac{1}{2}(a-b)^2 \le 0 .$$

If we fix $b \ge 0$, then define the above as a function of a, denoted as $h(a)$. Observe that

- $h(b) = b^2 \log b^2 - b^2 \log b^2 - \frac{1}{2}(b-b)^2 = 0$.
- $h'(a)|_{a=b} = 0$.
- $h''(a) \le 0$.

These imply that for $a \ge b$, $h(a) \le 0$, as desired. Thus, we have shown $n = 1$ case.

For general n, by tensorization,

$$\mathsf{Ent}(f(X_1, \dots, X_n)^2) \le \sum_{i=1}^n \mathbb{E}[\underbrace{\mathsf{Ent}^{(i)}(f(X_1, \dots, X_n)^2)}_{\le \frac{1}{2}\mathbb{E}^{(i)}[D_i f(X_1, \dots, X_n)^2]}] \le \frac{1}{2}\mathbb{E}\|\operatorname{grad} f\|_2^2 .$$

□

Theorem 6.3. Let $X_1, \dots, X_n \overset{iid}{\sim} \mathsf{Rademacher}$ and $Z = f(X_1, \dots, X_n)$. Suppose $\forall (x_1, \dots, x_n) \in \{\pm 1\}^n$,

$$\sum_{i=1}^n (D_i f(x_1, \dots, x_n))_+^2 \le \sigma^2 .$$

Then $\mathbb{P}(Z - \mathbb{E}[Z] \ge t) \le \exp(-t^2/\sigma^2)$.

Corollary 6.4. If $\forall (x_1, \dots, x_n) \in \{\pm 1\}^n$, we have

$$\|\operatorname{grad}(f(x_1, \dots, x_n))\|_2^2 = \sum_{i=1}^n (D_i f)^2 \le \sigma^2 ,$$

then $\mathbb{P}(|Z - \mathbb{E}[Z]| \ge t) \le 2\exp(-t^2/\sigma^2)$.

Proof. Let $g(X_1, \dots, X_n) = e^{\frac{\lambda}{2} f(X_1, \dots, X_n)}$, where $\lambda > 0$.

$$\begin{aligned}
\mathsf{Ent}(e^{\lambda Z}) &= \mathsf{Ent}\, g(X_1, \dots, X_n)^2 \\
&\le \frac{1}{2}\mathbb{E}[\|\operatorname{grad}(g(X_1, \dots, X_n))\|_2^2] \\
&= \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[\left(e^{\frac{\lambda}{2} f(X_1, \dots, X_i, \dots, X_n)} - e^{\frac{\lambda}{2} f(X_1, \dots, -X_i, \dots, X_n)}\right)^2\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\left(e^{\frac{\lambda}{2} f(X_1, \dots, X_i, \dots, X_n)} - e^{\frac{\lambda}{2} f(X_1, \dots, -X_i, \dots, X_n)}\right)_+^2\right].
\end{aligned}$$

By convexity, for $z \ge y$, $e^z - e^y \le (z - y)e^z$. So,

$$\begin{aligned}
&\left(e^{\frac{\lambda}{2} f(X_1, \dots, X_i, \dots, X_n} - e^{\frac{\lambda}{2} f(X_1, \dots, -X_i, \dots, X_n}\right)_+ \\
\le &\frac{\lambda}{2}(f(X_1, \dots, X_i, \dots, X_n) - f(X_1, \dots, -X_i, \dots, X_n))_+ e^{\frac{\lambda}{2} f(X_1, \dots, X_i, \dots, X_n)} \\
= &\frac{\lambda}{2} D_i f(X_{1:n})_+ e^{\frac{\lambda}{2} Z}.
\end{aligned}$$

Substitute into the above, we have

$$\begin{aligned}
\mathsf{Ent}(e^{\lambda Z}) &\le \sum_{i=1}^n \mathbb{E}\left[\frac{\lambda^2}{4} D_i f(X_{1:n})_+^2 e^{\lambda Z}\right] \\
&= \frac{\lambda^2}{4}\mathbb{E}\left[e^{\lambda Z}\sum_{i=1}^n D_i f(X_{1:n})_+^2\right] \le \frac{\lambda^2 \sigma^2}{4}\mathbb{E}[e^{\lambda Z}].
\end{aligned}$$

Therefore, by Herbst's argument with $\sigma^2/2$, we have

$$\mathbb{P}(Z - \mathbb{E}[Z] \ge t) \le \exp(-t^2/\sigma^2).$$

□

Example 6.5. Let $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} \mathsf{Rademacher}(\{\pm 1\})$ and $Z = f(\varepsilon) = \sup_{t \in T} \varepsilon^\top t$, where $T \subseteq \mathbb{R}^n$. Recall

$$D_i f(\varepsilon)_+ \le |t_i^*(\varepsilon)| \Longrightarrow \sum_{i=1}^n D_i f(\varepsilon)_+^2 \le 4\|t^*(\varepsilon)\|_2^2 \le 4 \sup_{t \in T} \|t\|_2^2.$$

Then,

$$\mathbb{P}(Z \ge \mathbb{E}[Z] + u) \le \exp\left(-\frac{u^2}{4 \sup_{t \in T} \|t\|_2^2}\right).$$

The lower tail will be covered in the next lecture.

Corollary 6.6. Let $X_1, \dots, X_n \overset{iid}{\sim} \mathsf{Rademacher}$ and $Z = f(X_1, \dots, X_n)$. If for all $(x_1, \dots, x_n) \in \{\pm 1\}^n$, $\|\operatorname{grad} f(x_1, \dots, x_n)\|_2^2 \le \sigma^2$, then

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \ge t) \le 2e^{-t^2/\sigma^2}.$$

Proof. In this case, we have

$$\sum_{i=1}^n (D_i f(x_1, \dots, x_n))_+^2, \sum_{i=1}^n (D_i f(x_1, \dots, x_n))_-^2 \le \sigma^2.$$

Apply previous result to both f and $-f$. □

Theorem 6.7 (Gaussian LSI)**.** Let $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(0,1)$, where $f : \mathbb{R}^n \mapsto \mathbb{R}$ weakly differentiable. Then

$$\mathsf{Ent}(f(X_1, \dots, X_n)^2) \le 2\mathbb{E}[\|\nabla f(X_1, \dots, X_n)\|_2^2] \,.$$

Proof. For $n = 1$, assume $f : \mathbb{R} \mapsto \mathbb{R}$ has compact support and twice continuously differentiable. Introduce $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} \mathsf{Rademacher}$ and define $S_m = \frac{\varepsilon_1 + \dots + \varepsilon_m}{\sqrt{m}}$. By previous LSI on hypercube:

$$\mathsf{Ent}(f(S_m)^2) \le \frac{1}{2}\mathbb{E}[\|\operatorname{grad} f(S_m)\|_2^2] \,.$$

Recall that $\lim_{m\to\infty} \mathbb{E}[\|\operatorname{grad} f(S_m)\|_2^2] = 4\mathbb{E}[f'(X)^2]$ from proof of Poincaré inequality, Lecture 5. Then

$$\mathsf{Ent}(f(X)^2) = \lim_{m\to\infty} \mathsf{Ent}(f(S_m)^2) \le 2\mathbb{E}[f'(X)^2] \,.$$

Approximate general weakly differentiable $f : \mathbb{R} \mapsto \mathbb{R}$ by sequence of smooth, compactly supported functions. Then in higher dimensions, $f : \mathbb{R}^n \mapsto \mathbb{R}$, by tensorization,

$$\mathsf{Ent}(f(X_1, \dots, X_n)^2) \le \sum_{i=1}^n \mathbb{E}[\underbrace{\mathsf{Ent}^{(i)}(f(X_1, \dots, X_n)^2)}_{\le 2\mathbb{E}^{(i)}[(\partial_i f)^2]}] \le 2\mathbb{E}[\|\nabla f(X_1, \dots, X_n)\|_2^2] \,.$$

□

Theorem 6.8 (Tsirelson-Ibragimov-Sudakov inequality)**.** If $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(0,1)$ and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L-Lipschitz, $Z = f(X_1, \dots, X_n)$, then

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \ge t) \le 2\exp\left(-\frac{t^2}{2L^2}\right), \qquad \forall t \ge 0 \,.$$

Remark 6.9. This result is dimension-free.

Proof. For any L-Lipschitz function $f : \mathbb{R}^n \mapsto \mathbb{R}$ that is weakly differentiable, we have $\|\nabla f(x)\|_2 \le L$ a.e. Then, $\forall \lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathsf{Ent}(e^{\lambda Z}) &= \mathsf{Ent}(e^{\lambda f(X_1,\dots,X_n)} \\
&\le 2\mathbb{E}\left[\left\|\nabla\left(e^{\frac{\lambda}{2}f(X_1,\dots,X_n)}\right)\right\|_2^2\right] \quad \text{by Gaussian LSI} \\
&= 2\mathbb{E}\left[\left\|\frac{\lambda}{2}e^{\frac{\lambda}{2}f(X_1,\dots,X_n)}\nabla f(X_1, \dots, X_n)\right\|_2^2\right] \\
&= \frac{\lambda^2}{2}\mathbb{E}[e^{\lambda Z}\|\nabla f(X_1, \dots, X_n)\|_2^2] \\
&\le \frac{\lambda^2 L^2}{2}\mathbb{E}[e^{\lambda Z}] \,.
\end{aligned}$$

By Herbst's argument applied to both f and $-f$,

$$\mathbb{E}[e^{\lambda Z}] \le \exp\left(\frac{\lambda^2 L^2}{2}\right), \qquad \lambda \in \mathbb{R} \,,$$

i.e., Z is L^2-subgaussian. □

Example 6.10. Let $Y \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^d$. Consider

$$\|Y\|_p = \left(\sum_{i=1}^d |Y_i|^p \right)^{1/p}, \qquad p \geq 1 .$$

We can write $Y = \Sigma^{1/2} X, X \sim \mathcal{N}(0, I_d)$. Set $f(X) = \|\Sigma^{1/2} X\|_p$. For any $x, x' \in \mathbb{R}^d$, we have

$$\begin{aligned} |f(x) - f(x')| &= \left| \|\Sigma^{1/2} x\|_p - \|\Sigma^{1/2} x'\|_p \right| \\ &\leq \|\Sigma^{1/2}(x - x')\|_p \leq \|\Sigma^{1/2}\|_{\ell_2 \to \ell_p} \|x - x'\|_2 . \end{aligned}$$

Let $L = \|\Sigma^{1/2}\|_{\ell_2 \to \ell_p}$. Then

$$\mathbb{P}(\|Y\|_p \geq \mathbb{E}[\|Y\|_p] + t) \leq \exp\left(-\frac{t^2}{2L^2} \right) .$$

Example 6.11. Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0, I_d) \in \mathbb{R}^d$, $\theta_* \in \mathbb{R}^d$ fixed. For each $i = 1, \ldots, n$,

$$y_i = x_i^\top \theta_* .$$

Fix $\theta \in \mathbb{R}^d$, loss function $\ell : \mathbb{R} \mapsto \mathbb{R}$ with $\|\ell'\|_\infty \leq K$. View

$$F(X) = \frac{1}{n} \sum_{i=1}^n \ell(y_i - x_i^\top \theta) = \frac{1}{n} \sum_{i=1}^n \ell(x_i^\top (\theta_* - \theta))$$

as a function of $X = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times d}$. We have

$$\|\nabla_{x_i} F(X)\|_2^2 = \left\| \frac{1}{n} \ell'(x_i^\top (\theta_* - \theta)) \cdot (\theta_* - \theta) \right\|_2^2 \leq \frac{K^2}{n^2} \|\theta_* - \theta\|_2^2 ,$$

which implies

$$\|\nabla F(X)\|_2^2 = \sum_{i=1}^n \|\nabla_{x_i} F(X)\|_2^2 \leq \frac{K^2}{n} \|\theta_* - \theta\|_2^2 .$$

So $F(X)$ is $\frac{K\|\theta_* - \theta\|_2}{\sqrt{n}}$-Lipschitz, which implies

$$\mathbb{P}(|F(X) - \mathbb{E}[F(X)]| \geq t) \leq \exp\left(-\frac{nt^2}{2K^2 \|\theta_* - \theta\|_2^2} \right) .$$

Remark 6.12. First, compared with the Poincaré inequality, i.e., $\mathsf{Var}(f(X)) \leq \mathbb{E}\|\nabla f(X_1, \ldots, X_n)\|_2^2$, the TIS inequality requires stronger uniform bound on $\|\nabla f\|_2$ rather than only in expectation, but gives stronger statement of exponential tail decay.

Second, the LSI $\mathsf{Ent}(f(X)^2) \leq 2\mathbb{E}\|\nabla f(X_1, \ldots, X_n)\|_2^2$ implies Poincaré: set $f = 1 + \varepsilon g$ with g bounded. Then $\mathbb{E}\|\nabla f\|_2^2$ and

$$\mathsf{Ent}(f^2) = \mathbb{E}[f^2 \log f^2] - \mathbb{E}[f^2] \log \mathbb{E}[f^2] = 2\varepsilon^2 \mathsf{Var}(g) + \mathcal{O}(\varepsilon^3) .$$

Hence, the LSI implies

$$2\varepsilon^2 \mathsf{Var}(g) + \mathcal{O}(\varepsilon^3) \leq 2\varepsilon^2 \mathbb{E}\|\nabla g\|_2^2 .$$

Let $\varepsilon \to 0^+$, we have

$$\mathsf{Var}(g) \leq \mathbb{E}\|\nabla g\|_2^2,$$

which is the Poincaré inequality.

Third, Poincaré and LSI for many other distrbutions could be established via analysis of continuous-time Markov chains: see §2 and §3, [vH14].

6.2 Further applications of the entropy method

Lemma 6.13. Let $X_1, \dots, X_n$ be independent, $Z = f(X_1, \dots, X_n)$, $Z_i = f^{(i)}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ for functions $f^{(1)}, \dots, f^{(n)}$ such that $Z \geq Z_1, \dots, Z_n$ with probability 1. Then

$$\mathsf{Ent}(e^{\lambda Z}) \leq \sum_{i=1}^{n} \mathbb{E}\left[\frac{\lambda^2}{2}(Z - Z_i)^2 e^{\lambda Z}\right], \qquad \lambda \geq 0\,.$$

Proof. By simple calculus, for $Y \geq 0$,

$$\mathsf{Ent}(Y) = \inf_{u>0} \mathbb{E}[Y \log Y - Y \log u - Y + u]$$

with inf attained at $u^* = \mathbb{E}[Y]$. Apply with $Y = e^{\lambda Z}$ and $u = e^{\lambda Z_i}$,

$$\begin{aligned}
\mathsf{Ent}(e^{\lambda Z}) &\leq \sum_{i=1}^{n} \mathbb{E}[\mathsf{Ent}^{(i)}(e^{\lambda Z})] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\mathbb{E}^{(i)}\left[e^{\lambda Z} \log e^{\lambda Z} - e^{\lambda Z} \log e^{\lambda Z_i} - e^{\lambda Z} + e^{\lambda Z_i}\right] \quad \text{taking } U = e^{\lambda Z_i} \\
&= \sum_{i=1}^{n} \mathbb{E}[e^{\lambda Z} \underbrace{(\lambda Z - \lambda Z_i - 1 + e^{\lambda(Z_i - Z)})}_{\text{apply } t-1+e^{-t} \leq \frac{t^2}{2},\ \forall t \geq 0}]
\end{aligned}$$

□

Theorem 6.14. Suppose

$$\sup_{x_1,\dots,x_n} \sum_{i=1}^{n} (f(x_1, \dots, x_n) - \inf_{x_i'} f(x_1, \dots, x_i', \dots, x_n))_+^2 \leq \sigma^2\,.$$

Let $X_1, \dots, X_n$ be independent and $Z = f(X_1, \dots, X_n)$. Then

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right), \qquad t \geq 0\,.$$

Proof. For each $i = 1, \dots, n$, take

$$Z_i := \inf_{X_i'} f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n)\,.$$

Hence, $Z \geq Z_1, \dots, Z_n$ and $\sum_{i=1}^n (Z - Z_i)^2 \leq \sigma^2$, which implis

$$\mathsf{Ent}(e^{\lambda Z}) \leq \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda Z}]\,.$$

By Herbst's argument, $\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp(-\frac{t^2}{2\sigma^2})$. □

Compare with Efron-Stein:

$$\mathsf{Var}(Z) \leq \mathbb{E} \sum_{i=1}^{n} (f(X_1, \dots, X_n) - f(X_1, \dots, X_i', \dots, X_n))_+^2\,.$$

Theorem 6.15. Let $X_1, \dots, X_n$ be independent and $X_i \in [0,1]$. $f : [0,1]^n \mapsto \mathbb{R}$ be separately convex, L-Lipschitz, $Z = f(X_1, \dots, X_n)$. Then

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{t^2}{2L^2}\right), \qquad \forall t \geq 0\,.$$

Proof. Let $x_i^* = \arg\min_{x_i' \in [0,1]} f(x_1, \dots, x_i', \dots, x_n)$. Then by convexity in x_i,

$$\Delta_i := f(x_1, \dots, x_n) - \inf_{x_i'} f(x_1, \dots, x_i', \dots, x_n) \leq \underbrace{|x_i - x_i^*|}_{\leq 1} \cdot |\partial_i f(x_1, \dots, x_n)|\,.$$

Therefore, we have

$$\sum_{i=1}^n \Delta_i^2 \leq \|\nabla f(x_1, \dots, x_n)\|_2^2 \leq L^2\,,$$

and the result follows from previous theorem. □

Compare this with convex Poincaré:

$$\mathsf{Var}(Z) \leq \mathbb{E}\|\nabla f(X_1, \dots, X_n)\|_2^2\,.$$

Example 6.16. Let $X \in \mathbb{R}^{m \times n}$ have independent entries. $X_{ij} \in [a,b]$. Recall that $\sigma_{\max}(X)$ is convex, $(b-a)$-Lipschitz function of $\{Z_{ij}\} = \left\{\frac{X_{ij}-a}{b-a}\right\} \in [0,1]$. Thus,

$$\mathbb{P}(\sigma_{\max}(X) \geq \mathbb{E}[\sigma_{\max}(X)] + t) \leq \exp\left(-\frac{t^2}{2(b-a)^2}\right).$$

Note that this provides only an upper tail bound. If f is jointly convex, we will see different method for also getting lower tail bound in the next lecture.

Example 6.17. Let $Z = f(X_1, \dots, X_n)$ be the longest subsequence satisfying a hereditary property. Recall from Lecture 5

$$0 \leq f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \leq 1$$
$$\sum_{i=1}^n [f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)] \leq f(x_1, \dots, x_n)$$

Set $Z_i = f(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. Then $Z \geq Z_i$, and

$$\mathsf{Ent}(e^{\lambda Z}) \leq \sum_{i=1}^n \mathbb{E}\left[\frac{\lambda^2}{2}(Z - Z_i)^2 e^{\lambda Z}\right] \leq \frac{\lambda^2}{2}\mathbb{E}[Z e^{\lambda Z}]\,.$$

7 Transportation Method, Transport Inequalities, Convex Lipschitz Concentration

Readings: §4.10-4.11, §8.1-8.6 in [BLM13].

Recall duality of entropy: for any $Z \geq 0$ with $\mathbb{E}[Z \log Z] < \infty$, we have

$$\mathsf{Ent}(Z) = \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z] = \sup_{U:\Omega \mapsto [-\infty,\infty),\ \mathbb{E}[e^U]=1} \mathbb{E}[UZ]\,.$$

Theorem 7.1 (Gibbs variational principle)**.** For any random variable U on $(\Omega, \mathbb{P})$ with $\mathbb{E}_{\mathbb{P}}[e^U] < \infty$,

$$\log \mathbb{E}_{\mathbb{P}}[e^U] = \sup_{\mathbb{Q} \ll \mathbb{P}} \{\mathbb{E}_{\mathbb{Q}}[U] - D_{KL}(\mathbb{Q}\|\mathbb{P})\}\,.$$

Proof. Consider any $\mathbb{Q} \ll \mathbb{P}$ and set $Z = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$, $V = \log \frac{e^U}{\mathbb{E}_{\mathbb{P}}[e^U]}$ for any U such that $\mathbb{E}_{\mathbb{P}}[e^U] < \infty$. Then we have $\mathbb{E}_{\mathbb{P}}[U] = 1$ and $\mathbb{E}_{\mathbb{P}}[e^V] = 1$, so $\log \mathbb{E}_{\mathbb{P}}[Z] = 0$ and by duality,

$$D_{KL}(\mathbb{Q}\|\mathbb{P}) = \mathbb{E}_{\mathbb{Q}}\left[\log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right] = \mathsf{Ent}_{\mathbb{P}}(Z) \geq \mathbb{E}_{\mathbb{P}}[VZ] = \mathbb{E}_{\mathbb{P}}\left[\log \frac{e^U}{\mathbb{E}_{\mathbb{P}}[e^U]} Z\right] = \mathbb{E}_{\mathbb{P}}[UZ] - \log \mathbb{E}_{\mathbb{P}}[e^U]\,,$$

which implies

$$\log \mathbb{E}_{\mathbb{P}}[e^U] \geq \mathbb{E}_{\mathbb{P}}[UZ] - D_{KL}(\mathbb{Q}\|\mathbb{P})\,.$$

Take sup over all $\mathbb{Q} \ll \mathbb{P}$ on the right hand side. Equality hold when $e^V = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$. □

Theorem 7.2 (Transportation method)**.** Let Z be a random variable on $(\Omega, \mathbb{P})$. The folloiwng are equivalent:

(a) $\forall \lambda \geq 0$, we have $\log \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq \frac{\lambda^2 \sigma^2}{2}$.

(b) For any $\mathbb{Q} \ll \mathbb{P}$, we have

$$\mathbb{E}_{\mathbb{Q}}[Z] - \mathbb{E}_{\mathbb{P}}[Z] \leq \sqrt{2\sigma^2 D_{KL}(\mathbb{Q}\|\mathbb{P})}\,.$$

Proof. For any $\mathbb{Q} \ll \mathbb{P}$, by Gibbs variational principle, we have

$$\log \mathbb{E}_{\mathbb{P}}[e^{\lambda(Z - \mathbb{E}_{\mathbb{P}}[Z])}] = \sup_{\mathbb{Q} \ll \mathbb{P}} \{\mathbb{E}_{\mathbb{Q}}[\lambda(Z - \mathbb{E}_{\mathbb{P}}[Z])] - D_{KL}(\mathbb{Q}\|\mathbb{P})\}\,.$$

Then we have

$$\begin{aligned}
\text{(a)} &\iff \forall \mathbb{Q} \ll \mathbb{P},\ \lambda \mathbb{E}_{\mathbb{Q}}[Z] - \lambda \mathbb{E}_{\mathbb{P}}[Z] - D_{KL}(\mathbb{Q}\|\mathbb{P}) \leq \frac{\lambda^2 \sigma^2}{2} \\
&\iff \forall \mathbb{Q} \ll \mathbb{P},\ D_{KL}(\mathbb{Q}\|\mathbb{P}) \geq \sup_{\lambda \geq 0} \{\lambda(\mathbb{E}_{\mathbb{Q}}[Z] - \mathbb{E}_{\mathbb{P}}[Z]) - \lambda^2 \sigma^2/2\}
\end{aligned}$$

For the sup problem, if $\mathbb{E}_{\mathbb{Q}}[Z] \geq \mathbb{E}_{\mathbb{P}}[Z]$, then with $\lambda^* = \frac{\mathbb{E}_{\mathbb{Q}}[Z] - \mathbb{E}_{\mathbb{P}}[Z]}{\sigma^2}$, we have $D_{KL}(\mathbb{Q}\|\mathbb{P}) \geq \frac{(\mathbb{E}_{\mathbb{Q}}[Z] - \mathbb{E}_{\mathbb{P}}[Z])^2}{2\sigma^2}$; if $\mathbb{E}_{\mathbb{Q}}[Z] < \mathbb{E}_{\mathbb{P}}[Z]$, then the supremum is 0. Thus, the above equivalence becomes

$$\begin{aligned}
\text{(a)} &\iff \forall \mathbb{Q} \ll \mathbb{P},\ D_{KL}(\mathbb{Q}\|\mathbb{P}) \geq \frac{(\mathbb{E}_{\mathbb{Q}}[Z] - \mathbb{E}_{\mathbb{P}}[Z])_+^2}{2\sigma^2} \\
&\iff \forall \mathbb{Q} \ll \mathbb{P},\ \mathbb{E}_{\mathbb{Q}}[Z] - \mathbb{E}_{\mathbb{P}}[Z] \leq \sqrt{2\sigma^2 D_{KL}(\mathbb{Q}\|\mathbb{P})}\,.
\end{aligned}$$

Hence, $\forall (X_1, \ldots, X_n) \in \mathcal{X}^n$ with joint law $\mathbb{P}$ and $Z = f(X_1, \ldots, X_n)$, then

$$\mathbb{E}_{\mathbb{P}}[e^{\lambda(Z - \mathbb{E}_{\mathbb{P}}[Z])}] \leq \exp(\lambda^2 \sigma^2/2) \qquad \forall \lambda \geq 0$$

if and only if, for any $(Y_1, \ldots, Y_n) \in \mathcal{X}^n$ with joint law $\mathbb{Q}$,

$$\mathbb{E}[f(Y_1, \ldots, Y_n)] - \mathbb{E}[f(X_1, \ldots, X_n)] \leq \sqrt{2\sigma^2 D_{KL}(\mathbb{Q}\|\mathbb{P})}\,. \qquad (\star)$$

□

Goal: bound left-hand side of ($\star$) by coupling $\mathbb{P}$ and $\mathbb{Q}$, and use bounded difference / Lipschitz properties of f.

7.1 Bounded differences revisited

Theorem 7.3 (Bounded differences)**.** Let $X_1, \dots, X_n$ be independent, $Z = f(X_1, \dots, X_n)$.

$$\|D_i f\|_\infty = \sup_{x_1,\dots,x_n,x_i'} |f(x_1,\dots,x_n) - f(x_1,\dots,x_i',\dots,x_n)| \,.$$

Then $Z - \mathbb{E}[Z]$ is $\frac{1}{4}\sum_{i=1}^n \|D_i f\|_\infty^2$-subgaussian.

[We proved this in Lecture 1 using the martingale method]

Under any coupling of $(X_1,\dots,X_n) \sim \mathbb{P}$ with $(Y_1,\dots,Y_n) \sim \mathbb{Q}$, we have

$$\begin{aligned}\mathbb{E}[f(Y_1,\dots,Y_n) - f(X_1,\dots,X_n)] &\le \mathbb{E}\sum_{i=1}^n \|D_i f\|_\infty \mathbb{1}_{\{X_i \neq Y_i\}} \\ &\le \sum_{i=1}^n \|D_i f\|_\infty \mathbb{P}(X_i \neq Y_i)\,.\end{aligned}$$

Lemma 7.4 (Pinsker's inequality)**.** Let $\mathbb{Q} \ll \mathbb{P}$ and

$$\mathsf{TV}(\mathbb{P},\mathbb{Q}) = \inf_{(X,Y)\sim \text{couplings}(\mathbb{P},\mathbb{Q})} \mathbb{P}(X \neq Y)\,.$$

Then $\mathsf{TV}(\mathbb{P},\mathbb{Q})^2 \le \frac{1}{2} D_{KL}(\mathbb{Q}\|\mathbb{P})$.

Proof. Let p and q be the densities of $\mathbb{P}$ and $\mathbb{Q}$ with respect to an underlying measure μ. Let $\eta(x) = \min(p(x), q(x))$, $\tilde{p}(x) = p(x) - \eta(x)$, $\tilde{q}(x) = q(x) - \eta(x)$. Consider the coupling of (X, Y) such that:

- with probability $\int \eta \mathrm{d}\mu$, let $X = Y \sim \frac{\eta}{\int \eta \mathrm{d}\mu}$.
- with probability $1 - \int \eta \mathrm{d}\mu = \int \tilde{p}\mathrm{d}\mu = \int \tilde{q}\mathrm{d}\mu$, let $X \sim \frac{\tilde{p}}{\int \tilde{p}\mathrm{d}\mu}$ independent of $Y \sim \frac{\tilde{q}}{\int \tilde{q}\mathrm{d}\mu}$. Note that in this case $X \neq Y$ because $\tilde{p}$ and $\tilde{q}$ have disjoint support.

The marginal density of X is then

$$\frac{\eta}{\int \eta \mathrm{d}\mu} \cdot \int \eta \mathrm{d}\mu + \frac{\tilde{p}}{\int \tilde{p}\mathrm{d}\mu} \cdot \left(1 - \int \eta \mathrm{d}\mu\right) = \eta + \frac{\tilde{p}}{\int \tilde{p}\mathrm{d}\mu} \cdot \int \tilde{p}\mathrm{d}\mu = \eta + \tilde{p} = p\,.$$

Similarly, the marginal density of Y is q. Thus,

$$\mathbb{P}(X \neq Y) = 1 - \int \eta \mathrm{d}\mu = \frac{1}{2}\int (p-\eta)\mathrm{d}\mu + \frac{1}{2}\int (q-\eta)\mathrm{d}\mu = \frac{1}{2}\int |p-q|\mathrm{d}\mu\,,$$

so

$$\begin{aligned}\mathsf{TV}(\mathbb{P},\mathbb{Q}) &\le \frac{1}{2}\int |p-q|\mathrm{d}\mu \quad \text{in fact, this is equality b/c we use the optimal coupling} \\ &= \frac{1}{2}\mathbb{E}_{\mathbb{P}}[|q/p - 1|] \\ &\le \mathbb{E}_{\mathbb{P}}\sqrt{\left(\frac{4}{3}\cdot\frac{2}{3}\cdot\frac{q}{p}\right)\left(\frac{q}{p}\log\frac{q}{p} - \frac{q}{p} + 1\right)} \quad \text{by } (x-1)^2 \le (\frac{4}{3} + \frac{2}{3}x)(x\log x - x + 1) \\ &\le \frac{1}{2}\sqrt{\mathbb{E}_{\mathbb{P}}\left[\frac{4}{3} + \frac{2}{3}\cdot\frac{q}{p}\right]}\sqrt{\mathbb{E}_{\mathbb{P}}\left[\frac{q}{p}\log\frac{q}{p} - \frac{q}{p} + 1\right]} \quad \text{by Cauchy-Schwarz} \\ &= \sqrt{\frac{1}{2}D_{KL}(\mathbb{Q}\|\mathbb{P})}\,.\end{aligned}$$

□

Theorem 7.5 (Marton)**.** Let $\mathbb{P} = \bigotimes_{i=1}^n \mathbb{P}_i$ be a product distribution on $\mathbb{R}^n$ such that for some $w : \mathbb{R}^2 \mapsto [0,\infty)$, convex $\phi : [0,\infty) \mapsto [0,\infty)$, and all $i \in \{1,\dots,n\}$ and $\mathbb{Q}_i \ll \mathbb{P}_i$,

$$\inf_{(X_i,Y_i)\sim\text{couplings}(\mathbb{P}_i,\mathbb{Q}_i)} \phi(\mathbb{E}[w(X_i,Y_i)]) \le D_{KL}(\mathbb{Q}_i\|\mathbb{P}_i)\,.$$

Then, for all $\mathbb{Q} \ll \mathbb{P}$,

$$\inf_{(X,Y)\sim\text{couplings}(\mathbb{P},\mathbb{Q})} \sum_{i=1}^n \phi(\mathbb{E}[w(X_i,Y_i)]) \le D_{KL}(\mathbb{Q}\|\mathbb{P})\,.$$

Proof. The proof is done by induction on n. The base case $n = 1$ is the given condition. Suppose the result holds for $n-1$, and consider $\mathbb{Q} \ll \mathbb{P} = \bigotimes_{i=1}^n \mathbb{P}_i$. For $Y = (Y_1,\dots,Y_n) \sim \mathbb{Q}$, let $\mathbb{Q}^{(n-1)}$ be the marginal law of $(Y_1,\dots,Y_{n-1})$ and $\mathbb{Q}_n(\cdot \mid Y_1,\dots,Y_{n-1})$ be the conditional law of Y_n.

By chain rule for KL divergence,

$$D_{KL}(\mathbb{Q}\|\mathbb{P}) = D_{KL}(\mathbb{Q}^{(n-1)}\|\bigotimes_{i=1}^{n-1}\mathbb{P}_i) + \mathbb{E}_{(Y_1,\dots,Y_{n-1})\sim\mathbb{Q}^{(n-1)}}[D_{KL}(\mathbb{Q}(\cdot \mid Y_1,\dots,Y_{n-1})\|\mathbb{P}_n]\,.$$

Consider the coupling (X,Y) of $(\mathbb{P},\mathbb{Q})$ such that

- $(X_1,\dots,X_{n-1})$ and $(Y_1,\dots,Y_{n-1})$ is from the coupling of $\bigotimes_{i=1}^{n-1}\mathbb{P}_i$ and $\mathbb{Q}^{(n-1)}$ that minimizes the objective $\sum_{i=1}^{n-1}\phi(\mathbb{E}[w(X_i,Y_i)])$.
- X_n and $Y_n \mid Y_1,\dots,Y_{n-1}$ follows the coupling of $\mathbb{P}_n$ and $\mathbb{Q}_n(\cdot \mid Y_1,\dots,Y_{n-1})$ that minimizes the objective $\phi(\mathbb{E}[w(X_n,Y_n) \mid Y_1,\dots,Y_{n-1}])$.

Under this construction, together with the chain rule, we have

$$\begin{aligned}
D_{KL}(\mathbb{Q}\|\mathbb{P}) &\ge \underbrace{\sum_{i=1}^{n-1}\phi(\mathbb{E}[w(X_i,Y_i)])}_{\text{induction hypothesis}} + \underbrace{\mathbb{E}_{\mathbb{Q}^{(n-1)}}[\phi(\mathbb{E}[w(X_n,Y_n) \mid Y_1,\dots,Y_{n-1}])]}_{\text{base case } n=1} \\
&\overset{\text{Jensen}}{\ge} \sum_{i=1}^{n-1}\phi(\mathbb{E}[w(X_i,Y_i)]) + \phi(\mathbb{E}[w(X_n,Y_n)]) \quad \text{since } \phi \text{ is convex} \\
&\ge \inf_{(X,Y)\sim\text{couplings}(\mathbb{P},\mathbb{Q})} \sum_{i=1}^n \phi(\mathbb{E}[w(X_i,Y_i)])\,.
\end{aligned}$$

□

With these tools, we can provide an alternative proof of the bounded difference theorem:

Proof. Recall that for any coupling (X,Y) of $(\mathbb{P},\mathbb{Q})$,

$$\mathbb{E}f(Y_1,\dots,Y_n) - \mathbb{E}f(X_1,\dots,X_n) \le \sum_{i=1}^n \|D_i f\|_\infty \cdot \mathbb{P}(X_i \ne Y_i) \le \left(\sum_{i=1}^n \|D_i f\|_\infty^2\right)^{1/2}\left(\sum_{i=1}^n \mathbb{P}(X_i \ne Y_i)^2\right)^{1/2}\,.$$

By Pinsker's inequality, $\forall \mathbb{Q}_i \ll \mathbb{P}_i$, there exists a coupling (X_i,Y_i) of $(\mathbb{P}_i,\mathbb{Q}_i)$ such that

$$\mathbb{P}(X_i \ne Y_i)^2 \le \frac{1}{2} D_{KL}(\mathbb{Q}_i\|\mathbb{P}_i)\,.$$

Thus, by Marton's tensorization theorem with $w(x,y) = \mathbb{1}_{x\neq y}$ and $\phi(x) = 2x^2$, invoking Pinsker's inequality, there exists a coupling (X,Y) of $(\mathbb{P},\mathbb{Q})$ such that

$$\sum_{i=1}^n \mathbb{P}(X_i \neq Y_i)^2 \leq \frac{1}{2} D_{KL}(\mathbb{Q}\|\mathbb{P})\,,$$

which, by substituting to above, implies that

$$\mathbb{E}f(Y_1,\ldots,Y_n) - \mathbb{E}f(X_1,\ldots,X_n) \leq \sqrt{\left(\frac{1}{2}\sum_{i=1}^n \|D_i f\|_\infty^2\right) \cdot D_{KL}(\mathbb{Q}\|\mathbb{P})}\,.$$

Let $Z = f(X_1,\ldots,X_n)$. By Theorem 7.2, we have: $\forall \lambda \geq 0$,

$$\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \leq \exp(\lambda^2\sigma^2/2)$$

for $\sigma^2 = \frac{1}{4}\sum_{i=1}^n \|D_i f\|_\infty^2$. For $\lambda \leq 0$, apply to $-f$. In particular, we know that

$$\mathbb{E}[-f(Y_1,\ldots,Y_n)] - \mathbb{E}[-f(X_1,\ldots,X_n)] \leq \sum_{i=1}^n \|D_i f\|_\infty \mathbb{P}(X_i \neq Y_i)\,,$$

which implies that

$$\mathbb{E}_{\mathbb{P}}[e^{(-\lambda)(-Z-\mathbb{E}_{\mathbb{P}}[-Z])}] \leq \exp((-\lambda)^2\sigma^2/2)\,, \qquad \forall(-\lambda) \geq 0\,.$$

Note that the left-hand side is equal to $\mathbb{E}_{\mathbb{P}}[e^{\lambda(Z-\mathbb{E}_{\mathbb{P}}[Z])}]$, so we finish the proof. □

7.2 Gaussian concentration revisited

Theorem 7.6 (Tsirelson-Ibragimov-Sudakov)**.** Let $X_1,\ldots,X_n \overset{iid}{\sim} \mathcal{N}(0,1)$. $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L-Lipschitz. $Z = f(X_1,\ldots,X_n)$. Then $Z - \mathbb{E}[Z]$ is L^2-subgaussian.

Lemma 7.7 (Stein)**.** Let $Z \sim \mathcal{N}(0,1)$ whose density is $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$. Suppose that $f : \mathbb{R} \mapsto \mathbb{R}$ is weakly differentiable such that $\mathbb{E}[|Zf(Z)|] < \infty$ and $\lim_{|z|\to\infty} f(z)\phi(z) = 0$. Then $\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)]$.

Proof. Apply $\phi'(z) = -z\phi(z)$ and integration-by-parts,

$$\begin{aligned}\int_a^b f'(z)\phi(z)\mathrm{d}z &= f(z)\phi(z)|_a^b - \int_a^b f(z)\phi'(z)\mathrm{d}z \\ &= f(b)\phi(b) - f(a)\phi(a) + \int_a^b zf(z)\phi(z)\mathrm{d}z\,.\end{aligned}$$

Take $a \to -\infty$ and $b \to \infty$ on both sides, we have $\mathbb{E}[f'(Z)] = \mathbb{E}[Zf(Z)]$. □

Lemma 7.8 (T_2 inequality)**.** Let $\mathbb{P} = \mathcal{N}(0,1)$ and $\mathbb{Q} \ll \mathbb{P}$. Then

$$\min_{(X,Y)\sim\text{couplings}(\mathbb{P},\mathbb{Q})} \mathbb{E}(X-Y)^2 \leq 2D_{KL}(\mathbb{Q}\|\mathbb{P})\,.$$

Note that the left-hand side is the squared Wasserstein-2 distance $W_2^2(\mathbb{P},\mathbb{Q})$.

Proof. Let $\mathbb{P}$ and $\mathbb{Q}$ be the CDFs and $p \equiv \phi$ and q be densities. Suppose first that $\forall x \in \mathbb{R}$, $\varepsilon \le \frac{q(x)}{p(x)} \le \frac{1}{\varepsilon}$, so $\mathbb{Q}$ is strictly increasing.

Consider the coupling $X \sim \mathcal{N}(0,1)$ and $Y = T(X) := \mathbb{Q}^{-1}(\mathbb{P}(X))$. By the chain rule,

$$T'(x) = \frac{p(x)}{q(\mathbb{Q}^{-1}(\mathbb{P}(x)))} = \frac{p(x)}{q(T(x))} = \frac{p(x)}{q(y)}\,.$$

Hence, we have

$$\begin{aligned} D_{KL}(\mathbb{Q}\|\mathbb{P}) = \mathbb{E}_{Y\sim\mathbb{Q}} \log \frac{q(Y)}{p(Y)} &= \mathbb{E} \log \left(\frac{1}{p(Y)} \cdot \frac{p(X)}{T'(X)} \right) \\ &= \mathbb{E}[-X^2/2 + Y^2/2 - \log T'(X)] \\ &\ge \mathbb{E}[-X^2/2 + Y^2/2 + 1 - T'(X)] \end{aligned}$$

On the other hand, for t large enough, we have

$$\mathbb{Q}(t) = \int_{-\infty}^{t} q(y)\mathrm{d}y \le \frac{1}{\varepsilon}\int_{0}^{t} p(y)\mathrm{d}y = \frac{1}{\varepsilon}\mathbb{P}(t) \le \mathbb{P}(t/2)\,.$$

This implies that $T(x) = \mathbb{Q}^{-1}(\mathbb{P}(x)) \le 2x$ for all large x. Similarly, $T(-x) \ge -2x$ for all large x. Thus,

$$\lim_{|x|\to\infty} T(x)\phi(x) = 0\,, \qquad \mathbb{E}[|XT(X)|] = \mathbb{E}[|XY|] < \infty\,.$$

By Stein's lemma,

$$\begin{aligned} D_{KL}(\mathbb{Q}\|\mathbb{P}) &\ge \mathbb{E}[-X^2/2 + Y^2/2 + 1 - T'(X)] \\ &= \mathbb{E}[-X^2/2 + Y^2/2 + 1 - XT(X)] \\ &= \frac{1}{2}\mathbb{E}[(X-Y)^2] \quad \text{since } Y = T(X), \mathbb{E}[X^2] = \mathbb{E}[Y^2] = 1\,. \end{aligned}$$

For general $q(x)$, let $w^\varepsilon(x) = \max\{\varepsilon, \min(\frac{1}{\varepsilon}, \frac{q(x)}{p(x)})\}$, then we let

$$q^\varepsilon(x) = \frac{w^\varepsilon(x)p(x)}{Z_\varepsilon}, \qquad Z_\varepsilon = \int w^\varepsilon(x)p(x)\mathrm{d}x\,,$$

and define $Y^\varepsilon \sim \mathbb{Q}^\varepsilon$ whose density is q^ε. Now by definition, $w^\varepsilon(x) \le \varepsilon + \frac{q(x)}{p(x)}$. As $\varepsilon \to 0^+$, by dominated convergence:

- $Z^\varepsilon \to 1, q^\varepsilon(x) \to q(x), Y^\varepsilon \overset{\mathscr{D}}{\rightsquigarrow} Y$.
- $D_{KL}(\mathbb{Q}^\varepsilon\|\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\frac{q^\varepsilon}{p}\log\frac{q^\varepsilon}{p}] \to \mathbb{E}_{\mathbb{P}}[\frac{q}{p}\log\frac{q}{p}] = D_{KL}(\mathbb{Q}\|\mathbb{P})$.

Therefore,

$$\mathbb{E}[(X-Y)^2] \le \liminf_{\varepsilon\to 0^+} \mathbb{E}[(X-Y^\varepsilon)^2] \le \lim_{\varepsilon\to 0^+} 2D_{KL}(\mathbb{Q}^\varepsilon\|\mathbb{P}) = 2D_{KL}(\mathbb{Q}\|\mathbb{P})\,.$$

□

Corollary 7.9. Let $\mathbb{P} = \mathcal{N}(0, I)$ on $\mathbb{R}^n$ and $\mathbb{Q} \ll \mathbb{P}$. Then

$$\min_{(X,Y)\sim\text{couplings}(\mathbb{P},\mathbb{Q})} \mathbb{E}[\|X-Y\|_2^2] \le 2D_{KL}(\mathbb{Q}\|\mathbb{P})\,.$$

Proof. Apply Marton's tensorization theorem with $w(x,y) = (x-y)^2$ and $\phi(x) = x/2$. □

With these tools, we can provide an alternative proof of the Tsirelson-Ibragimov-Sudakov theorem:

Proof. For any $\mathbb{Q} \ll \mathbb{P}$, there exists a coupling (X, Y) of $(\mathbb{P}, \mathbb{Q})$ such that

$$\begin{aligned}\mathbb{E}[f(Y)] - \mathbb{E}[f(X)] &\le L\mathbb{E}[\|X - Y\|_2] \\ &\le \sqrt{L^2\mathbb{E}[\|X - Y\|_2^2]} \\ &\le \sqrt{2L^2 D_{KL}(\mathbb{Q}\|\mathbb{P})}\,.\end{aligned}$$

Apply this to both f and $-f$. Then let $Z = f(X)$ and the above are equivalent to $\forall\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_{\mathbb{P}}[e^{\lambda(Z - \mathbb{E}_{\mathbb{P}}[Z])}] \le \exp(\lambda^2 L^2/2)\,.$$

Thus, $Z - \mathbb{E}_{\mathbb{P}}[Z]$ is L^2-subgaussian by definition. □

7.3 Convex Lipschitz concentration

Theorem 7.10 (Talagrand)**.** Let $X_1, \dots, X_n$ independent and $X_i \in [0, 1]$. If $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L-Lipschitz and convex, $Z = f(X_1, \dots, X_n)$, then $\mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \le \exp(\lambda^2 L^2/2), \forall\lambda \in \mathbb{R}$.

Lemma 7.11. For any $\mathbb{Q} \ll \mathbb{P}$,

$$\min_{(X,Y)\sim\text{couplings}(\mathbb{P},\mathbb{Q})} \mathbb{E}[\mathbb{P}(X \ne Y \mid X)^2] + \mathbb{E}[\mathbb{P}(X \ne Y \mid Y)^2] \le 2D_{KL}(\mathbb{Q}\|\mathbb{P})\,.$$

Proof. Let $\eta(x) = \min(p(x), q(x)), \tilde{p}(x) = p(x) - \eta(x), \tilde{q}(x) = q(x) - \eta(x)$. Recall the TV coupling of (X, Y):

- with probability $\int \eta \mathrm{d}\mu$, let $X = Y \sim \frac{\eta}{\int \eta \mathrm{d}\mu}$.
- with probability $1 - \int \eta \mathrm{d}\mu = \int \tilde{p}\mathrm{d}\mu = \int \tilde{q}\mathrm{d}\mu$, let $X \sim \frac{\tilde{p}}{\int \tilde{p}\mathrm{d}\mu}$ independent of $Y \sim \frac{\tilde{q}}{\int \tilde{q}\mathrm{d}\mu}$. Note that in this case $X \ne Y$ because $\tilde{p}$ and $\tilde{q}$ have disjoint support.

Then conditioned on $X = x$ or $Y = y$, we have

$$\begin{aligned}\mathbb{P}(X \ne Y \mid X = x) &= \frac{\tilde{p}(x)}{p(x)} = 1 - \frac{\eta(x)}{p(x)} = \left(1 - \frac{q(x)}{p(x)}\right)_+ \\ \mathbb{P}(X \ne Y \mid Y = y) &= \frac{\tilde{q}(y)}{q(y)} = 1 - \frac{\eta(y)}{q(y)} = \left(1 - \frac{p(y)}{q(y)}\right)_+ .\end{aligned}$$

Hence, we have

$$\begin{aligned}&\min_{(X,Y)\sim\text{couplings}(\mathbb{P},\mathbb{Q})} \mathbb{E}[\mathbb{P}(X \ne Y \mid X)^2] + \mathbb{E}[\mathbb{P}(X \ne Y \mid Y)^2] \\ \le&\mathbb{E}_{\mathbb{P}}\left(1 - \frac{q}{p}\right)_+^2 + \mathbb{E}_{\mathbb{Q}}\left(1 - \frac{p}{q}\right)_+^2 \quad \text{[In fact, equality holds for this coupling.]} \\ =&\mathbb{E}_{\mathbb{P}}\left[\left(1 - \frac{q}{p}\right)_+^2 + \frac{q}{p}\left(1 - \frac{p}{q}\right)_+^2\right] \\ \le&2\mathbb{E}_{\mathbb{P}}\left[\frac{q}{p}\log\frac{q}{p} - \frac{q}{p} + 1\right] \\ =&2D_{KL}(\mathbb{Q}\|\mathbb{P})\,,\end{aligned}$$

where in the last inequality we used $(1 - x)_+^2 + x(1 - \frac{1}{x})_+^2 \le 2(x\log x - x + 1)$. □

Theorem 7.12 (Marton)**.** Let $\mathbb{P} = \bigotimes_{i=1}^n \mathbb{P}_i$ be a product distribution on $\mathbb{R}^n$ such that for some $w : \mathbb{R}^2 \mapsto [0,\infty)$, convex $\phi : [0,\infty) \mapsto [0,\infty)$, and all $i \in \{1,\dots,n\}$ and $\mathbb{Q}_i \ll \mathbb{P}_i$,

$$\min_{(X_i,Y_i)\sim\text{couplings}(\mathbb{P}_i,\mathbb{Q}_i)} \mathbb{E}[\phi(\mathbb{E}[w(X_i,Y_i) \mid X_i]) + \phi(\mathbb{E}[w(X_i,Y_i) \mid Y_i])] \le D_{KL}(\mathbb{Q}_i \| \mathbb{P}_i)\,.$$

Then, for all $\mathbb{Q} \ll \mathbb{P}$,

$$\min_{(X,Y)\sim\text{coupling}(\mathbb{P},\mathbb{Q})} \sum_{i=1}^n \mathbb{E}[\phi(\mathbb{E}[w(X_i,Y_i) \mid X]) + \phi(\mathbb{E}[w(X_i,Y_i) \mid Y])] \le D_{KL}(\mathbb{Q} \| \mathbb{P})\,.$$

Proof. Induction on n, analogous to the preceeding tensorization theorem. □

Proof of Talagrand's theorem. By convexity of f,

$$f(y) - f(x) \le \langle \nabla f(y), y - x\rangle = \sum_{i=1}^n \partial_i f(y)(y_i - x_i) \le \sum_{i=1}^n |\partial_i f(y)| \mathbb{1}_{\{x_i \ne y_i\}}\,.$$

for any $\mathbb{Q} \ll \mathbb{P}$, there exists a coupling of $X = (X_1,\dots,X_n) \sim \mathbb{P}$ and $Y = (Y_1,\dots,Y_n) \sim \mathbb{Q}$ such that

$$\begin{aligned}
\mathbb{E}[f(Y_1,\dots,Y_n)] - \mathbb{E}[f(X_1,\dots,X_n)] &\le \sum_{i=1}^n \mathbb{E}[|\partial_i f(Y)| \cdot \mathbb{1}_{\{X_i \ne Y_i\}}] \\
&= \sum_{i=1}^n \mathbb{E}[|\partial_i f(Y) \mathbb{E}[\mathbb{1}_{\{X_i \ne Y_i\}} \mid Y]] \\
&\le \left(\mathbb{E}\sum_{i=1}^n |\partial_i f(Y)|^2\right)^{1/2} \left(\mathbb{E}\sum_{i=1}^n \mathbb{P}(X_i \ne Y_i \mid Y)^2\right)^{1/2} \\
&\le L\left(\mathbb{E}\sum_{i=1}^n \mathbb{P}(X_i \ne Y_i \mid Y)^2\right)^{1/2} \\
&\le L\sqrt{2D_{KL}(\mathbb{Q}\|\mathbb{P})} \quad \text{by Marton's tensorization}\,.
\end{aligned}$$

Similarly, $f(x) - f(y) \le \langle \nabla f(x), x - y\rangle$, so by conditioning on X in the intemediate steps,

$$\mathbb{E}[f(X_1,\dots,X_n)] - \mathbb{E}[f(Y_1,\dots,Y_n)] \le L\sqrt{2D_{KL}(\mathbb{Q}\|\mathbb{P})}\,.$$

Finally, apply transportation method to both f and $-f$, we finish the proof. □

Example 7.13. Let $X \in \mathbb{R}^{m\times n}$ have independent entries $X_{ij} \in [0,1]$. Recall that $\sigma_{\max}(X)$ is a convex, 1-Lipschitz function of X. Then

$$\mathbb{P}(|\sigma_{\max}(X) - \mathbb{E}[\sigma_{\max}(X)]| \ge t) \le 2\exp(-t^2/2) \qquad t \ge 0\,.$$

8 Maximal Inequalities, Covering Nets, Norms of Random Matrices

Readings: §5.1-5.2 in [vH14], §4.1-4.4, 4.6 in [Ver18]

Setting: we have a random process $\{X_t\}_{t\in T}$. A few examples:

- $\{g^\top t\}_{t\in T}$, where $g \sim \mathcal{N}(0, I)$ is a gaussian random vector.
- $\{u^\top X v\}_{u\in\mathbb{S}^{n-1}, v\in\mathbb{S}^{m-1}}$, where $X \in \mathbb{R}^{n\times m}$ is a random matrix.
- $\{\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_1)]\}_{f\in\mathscr{F}}$, where $\mathscr{F}$ is some function class and $X_1, \dots, X_n$ are iid.

The goals are two-holded:

- Sharp upper bounds for $\mathbb{E}[\sup_{t\in T} X_t]$.
- Sharp tail bounds for $\mathbb{P}(\sup_{t\in T} X_t \ge u)$.

These two goals are often equivalent if $\sup_{t\in T} X_t$ concentrates around its mean.

Principle: If $\{X_t\}_{t\in T}$ is "sufficiently continuous", then the size of $\sup_{t\in T} X_t$ is controlled by "complexity" of T.

Basic idea: If T has finite cardinality and $X_t \ge 0, \forall t \in T$, then $\mathbb{E}\sup_{t\in T} X_t \le \sum_{t\in T}\mathbb{E}[X_t]$.

Example 8.1. For any random variables $X_1, \dots, X_n$, we have

$$\mathbb{E}\sup_{1\le i\le n} X_i \le \sum_{i=1}^n \mathbb{E}[|X_i|] \le n\cdot \sup_{1\le i\le n}\mathbb{E}[|X_i|]$$
$$\mathbb{E}\sup_{1\le i\le n} X_i \le \left(\mathbb{E}\sup_{1\le i\le n}|X_i|^p\right)^{1/p} \le n^{1/p}\sup_{1\le i\le n}(\mathbb{E}[|X_i|^p])^{1/p}$$

Lemma 8.2 (Maximal inequality)**.** Suppose $|T|$ is finite and $\log\mathbb{E}[e^{\lambda X_t}] \le \psi(\lambda)$ for all $\lambda \ge 0$ and $t \in T$, where ψ is convex and $\psi(0) = \psi'(0) = 0$. Set $\psi^*(t) := \sup_{\lambda\ge 0}\{\lambda t - \psi(\lambda)\}$ be the Fenchel dual of ψ. Then

$$\mathbb{E}\sup_{t\in T} X_t \le \psi^{*-1}(\log|T|)\,.$$

Corollary 8.3. If $\mathbb{E}[e^{\lambda X_t}] \le \lambda^2\sigma^2/2$ for all $\lambda \ge 0$ and $t \in T$, then

$$\mathbb{E}\sup_{t\in T} X_t \le \sqrt{2\sigma^2\log|T|}\,.$$

Proof. Apply the maximal inequality lemma with $\psi(\lambda) = \lambda^2\sigma^2/2$ and $\psi^*(t) = t^2/(2\sigma^2)$ for $t \ge 0$. □

Proof of the maximal inequality lemma. By Jensen inequality, $\forall\lambda > 0$, we have

$$\begin{aligned}
\mathbb{E}\sup_{t\in T} X_t &\le \frac{1}{\lambda}\log\mathbb{E}[e^{\lambda\sup_{t\in T} X_t}] \\
&= \frac{1}{\lambda}\log\mathbb{E}\sup_{t\in T} e^{\lambda X_t} \\
&\le \frac{1}{\lambda}\log\sum_{t\in T}\mathbb{E}[e^{\lambda X_t}] \\
&\le \frac{1}{\lambda}\log\left(|T|\cdot\mathbb{E}\sup_{t\in T} e^{\lambda X_t}\right) \\
&\le \frac{1}{\lambda}(\log|T| + \psi(\lambda)) \quad \text{by given condition,}
\end{aligned}$$

which implies $\log|T| \geq \lambda \mathbb{E}\sup_{t\in T} X_t - \psi(\lambda)$. Taking sup on both sides for $\lambda \geq 0$, we have

$$\log|T| \geq \psi^*(\mathbb{E}\sup_{t\in T} X_t)\,.$$

Since ψ^* is convex and increasing, we have

$$\mathbb{E}\sup_{t\in T} X_t \leq \psi^{*-1}(\log|T|)\,.$$

□

Remark 8.4. This is closely related to the union bound:

$$\begin{aligned}
&\mathbb{P}(\sup_{t\in T} X_t \geq u) \leq \sum_{t\in T}\mathbb{P}(X_t \geq u) \leq |T|e^{-\lambda u+\psi(\lambda)}, \quad \forall\lambda\geq 0\\
\Longrightarrow&\mathbb{P}(\sup_{t\in T} X_t \geq u) \leq |T|e^{-\psi^*(u)} = e^{\log|T|-\psi^*(u)}\\
\Longrightarrow&\mathbb{P}(\sup_{t\in T} X_t \geq \psi^{*-1}(\log|T|+s)) \leq e^{-s}\,.
\end{aligned}$$

Integrating this tail bound gives

$$\begin{aligned}
\psi^*(\mathbb{E}\sup_{t\in T} X_t) &\overset{\text{Jensen}}{\leq} \mathbb{E}\left[\psi^*(\sup_{t\in T} X_t)\right]\\
&= \int_0^\infty \mathbb{P}(\psi^*(\sup_{t\in T} X_t)\geq x)\mathrm{d}x\\
&= \int_0^{\log|T|} \mathbb{P}(\psi^*(\sup_{t\in T} X_t)\geq x)\mathrm{d}x + \int_0^\infty \mathbb{P}(\psi^*(\sup_{t\in T} X_t)\geq \log|T|+s)\mathrm{d}s\\
&\leq \log|T| + \int_0^\infty e^{-s}\mathrm{d}s = \log|T|+1\,,
\end{aligned}$$

so $\mathbb{E}\sup_{t\in T} X_t \leq \psi^{*-1}(\log|T|+1)$.

Intuition: if $\{X_t\}_{t\in T}$ are independent and $\sum_{t\in T}\mathbb{P}(X_t\geq u)$ is small, then

$$\mathbb{P}(\sup_{t\in T} X_t \geq u) = 1-\prod_{t\in T}(1-\mathbb{P}(X_t\geq u)) \approx 1-\prod_{t\in T} e^{-\mathbb{P}(X_t\geq u)} \approx \sum_{t\in T}\mathbb{P}(X_t\geq u)\,,$$

so the union bound is tight. Loose by a factor of $|T|$ if $X_t = X$ for all $t\in T$.

8.1 Covering nets

Idea: trade off dependence of $\{X_t\}_{t\in T}$ with the continuity $X_t \approx X_s$ for $t\approx s$.

Definition 8.5. Let (T,d) be a metric space. $\mathcal{N}\subseteq T$ is an *ε-net* of T if, for all $t\in T$, there exists $\pi(t)\in\mathcal{N}$ such that $\mathrm{d}(t,\pi(t))\leq\varepsilon$. The *covering number* is

$$N(T,\mathrm{d},\varepsilon) = \inf\{|\mathcal{N}| : \mathcal{N} \text{ is an } \varepsilon\text{-net of } T\}\,.$$

Proposition 8.6. If $\{X_t\}_{t\in T}$ is L-Lipschitz, i.e,, there exists (possibly random) $L\geq 0$ for which $|X_t - X_s| \leq L\mathrm{d}(X_t,X_s)$ for all $s,t\in T$ **and** $\log\mathbb{E}[e^{\lambda X_t}] \leq \frac{\lambda^2\sigma^2}{2}$, $\forall\lambda\geq 0$ and $t\in T$, then

$$\mathbb{E}\sup_{t\in T} X_t \leq \inf_{\varepsilon>0}\left\{\varepsilon\mathbb{E}[L] + \sqrt{2\sigma^2\log N(T,\mathrm{d},\varepsilon)}\right\}$$

Proof. For any ε-net $\mathcal{N}$,

$$\begin{aligned}\sup_{t\in T} X_t &= \sup_{t\in T} (X_t - X_{\pi(t)} + X_{\pi(t)}) \\ &\le L \cdot \mathrm{d}(t,\pi(t)) + \sup_{s\in\mathcal{N}} X_s\,.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\sup_{t\in T} X_t &\le \mathbb{E}[L] \underbrace{\mathrm{d}(t,\pi(t))}_{\le\varepsilon \text{ by definition of } \mathcal{N}} + \underbrace{\mathbb{E}\sup_{s\in\mathcal{N}} X_s}_{\le\sqrt{2\sigma^2\log|\mathcal{N}|} \text{ by previous Corollary}} \\ &\le \varepsilon\mathbb{E}[L] + \sqrt{2\sigma^2 \log N(T,\mathrm{d},\varepsilon)}\,.\end{aligned}$$

Finally, take inf over $\varepsilon > 0$ and all ε-nets $\mathcal{N}$. □

Remark 8.7. Sometimes $T \subseteq S$ (a larger space), and it is simpler to construct an ε-net of T with points in S. Let $N^{\mathrm{ext}}(T,\mathrm{d},\varepsilon) = \inf\{|\mathcal{N}| : \mathcal{N} \subseteq S \text{ is } \varepsilon\text{-net of } T\}$. Then

$$N^{\mathrm{ext}}(T,\mathrm{d},\varepsilon) \le N(T,\mathrm{d},\varepsilon) \le N^{\mathrm{ext}}(T,\mathrm{d},\varepsilon/2)\,.$$

Example 8.8. Let $Z_1,\dots,Z_n$ be iid random variables on $[0,1]$ and

$$\mathscr{F} = \{f : [0,1] \mapsto \mathbb{R}, 1-\text{Lipchitz}, f(0) = 0\}\,.$$

Let $X_f = \frac{1}{n}\sum_{i=1}^n (f(Z_i) - \mathbb{E}[f(Z_i)])$ and $W = \sup_{f\in\mathscr{F}} X_f$. This is the 1-Wasserstein distance between $\frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ and law of Z_i.

- $|X_f - X_g| \le \frac{1}{n}\sum_{i=1}^n |f(Z_i) - g(Z_i) - \mathbb{E}[f(Z_i)] + \mathbb{E}[g(Z_i)]| \le 2\|f-g\|_\infty$. Thus, $f \mapsto X_f$ is 2-Lipschitz with respect to $\mathrm{d}(f,g) = \|f-g\|_\infty$.
- By 1-Lipschitzness and $f(0) = 0$, for each i, $f(Z_i) \in [-1,1]$. Then, by Hoeffding's Lemma (Lemma 1.8) and Hoeffding's inequality (Theorem 1.7),

$$\log\mathbb{E}[e^{\lambda X_f}] \le \frac{\lambda^2}{2n}, \qquad \lambda \ge 0\,.$$

- Let $\mathcal{N}^{\mathrm{ext}}$ be the set of piecewise constant functions on $[0,1]$ with $f(0) = 0$, jumps at $\varepsilon, 2\varepsilon, 3\varepsilon, \dots$, values in $\cdots, -2\varepsilon, -\varepsilon, 0, \varepsilon, 2\varepsilon, \cdots$, where value changes by at most ε at each jump. For any $f \in \mathscr{F}$, if $f(k\varepsilon) \in [j\varepsilon, (j+1)\varepsilon)$, then $\forall x \in [k\varepsilon, (k+1)\varepsilon)$, we have

  $$|f(x) - j\varepsilon| \le |f(x) - f(k\varepsilon)| + |f(k\varepsilon) - j\varepsilon| \le \varepsilon + \varepsilon = 2\varepsilon\,,$$

  where the first is bounded by Lipschitzness and the second because $f(k\varepsilon) \in [j\varepsilon, (j+1)\varepsilon)$. Also,

  $$f((k+1)\varepsilon) \in [(j-1)\varepsilon, (j+1)\varepsilon) = [(j-1)\varepsilon, j\varepsilon) \cup [j\varepsilon, (j+1)\varepsilon) \cup [(j+1)\varepsilon, (j+2)\varepsilon)\,.$$

  Thus, there exists $\pi(f) \in \mathcal{N}^{\mathrm{ext}}$ such that $\|f - \pi(f)\|_\infty \le 2\varepsilon$, where we round the value of f at the left endpoint of each interval down to the nearest $j\varepsilon$, so $\mathcal{N}^{\mathrm{ext}}$ is a 2ε-net for $\mathscr{F}$ with respect to $\mathrm{d}(f,g) = \|f-g\|_\infty$. Therefore, $|\mathcal{N}^{\mathrm{ext}}| \le 3^{1/\varepsilon}$, which implies

  $$N(\mathscr{F}, \|\cdot\|_\infty, 4\varepsilon) \le N^{\mathrm{ext}}(\mathscr{F}, \|\cdot\|_\infty, 2\varepsilon) \le 3^{1/\varepsilon}\,,$$

  so $\mathbb{E}[W] \le \inf_{\varepsilon>0}\left\{2\varepsilon + \sqrt{\frac{2}{n}\log 3^{4/\varepsilon}} \le Cn^{-1/3}\right\}$.

In later lectures, we will:

(i) Improve this bound to $Cn^{-1/2}$, using the typical size

$$|X_f - X_g| \lesssim \frac{1}{\sqrt{n}}\|f-g\|_\infty \qquad \text{w.h.p.}$$

(ii) Extend to non-Lipschitz classes $\mathscr{F}$ where $N(\mathscr{F}, \|\cdot\|_\infty, \varepsilon)$ is infinite, by using other metrics for $\mathscr{F}$.

8.2 Norm of random matrices

Let $X \in \mathbb{R}^{n\times m}$. The operator norm is

$$\|X\|_{\mathrm{op}} = \sup_{u\in\mathbb{B}^n, v\in\mathbb{B}^m} u^\top X v\,, \qquad \mathbb{B}^n := \{u\in\mathbb{R}^n : \|u\|_2 \le 1\}\,.$$

Then for any $u, u' \in \mathbb{B}^n$ and $v, v' \in \mathbb{B}^m$,

$$\begin{aligned} |X_{u,v} - X_{u',v'}| &= |u^\top X v - u'^\top X v'| \\ &\le |u^\top X (v-v')| + |(u-u')^\top X v'| \\ &\le \|X\|_{\mathrm{op}}(\|v-v'\|_2 + \|u-u'\|_2)\,. \end{aligned}$$

This implies that $\{X_{u,v}\}_{u,v}$ is $\|X\|_{\mathrm{op}}$-Lipschitz w.r.t. the metric $\mathrm{d}((u,v),(u',v')) = \|u-u'\|_2 + \|v-v'\|_2$.

To bound $N(\mathbb{B}^n \times \mathbb{B}^m, \mathrm{d}, \varepsilon)$, we need the following tools.

Definition 8.9. Let (T, d) be a metric space. $\mathscr{D} \subseteq T$ is an *ε-packing* of T if $\mathrm{d}(s,t) > \varepsilon$ for all $s \ne t \in \mathscr{D}$. The *packing number* is

$$D(T, \mathrm{d}, \varepsilon) = \sup\{|\mathscr{D}| : \mathscr{D} \text{ is } \varepsilon\text{-packing of } T\}\,.$$

Proposition 8.10. For any $\varepsilon > 0$, we have

$$D(T, \mathrm{d}, 2\varepsilon) \le N(T, \mathrm{d}, \varepsilon) \le D(T, \mathrm{d}, \varepsilon)\,.$$

Proof. Let $\mathscr{D}$ be any 2ε-packing and $\mathscr{N}$ be any ε-net. For any $t \in \mathscr{D} \subseteq T$, $\exists \pi(t) \in \mathscr{N}$ such that $\mathrm{d}(t, \pi(t)) \le \varepsilon$ by the definition of ε-net. Furthermore, for distinct $s, t \in \mathscr{D}$, we have $\pi(t) \ne \pi(s)$ because

$$\mathrm{d}(\pi(t), \pi(s)) \ge \mathrm{d}(t,s) - \mathrm{d}(t, \pi(t)) - \mathrm{d}(s, \pi(s)) > 2\varepsilon - \varepsilon - \varepsilon = 0\,.$$

This implies that $|\mathscr{D}| \le |\mathscr{N}|$, because we already need $|\mathscr{D}|$ points in $\mathscr{N}$ to cover $\mathscr{D}$, and we potentially need more points in $\mathscr{N}$ in order to cover $T \setminus \mathscr{D}$. Hence, $D(T, \mathrm{d}, 2\varepsilon) \le N(T, \mathrm{d}, \varepsilon)$.

For the other part, let $\mathscr{D}$ be any *maximal* ε-packing, i.e., $\mathscr{D} \cup \{t\}$ is not an ε-packing for any $t \in T \setminus \mathscr{D}$. Then, for any $t \in T \setminus \mathscr{D}$, $\exists s \in \mathscr{D}$ such that $\mathrm{d}(t,s) \le \varepsilon$. So, $\mathscr{D}$ is also an ε-net, implying that $N(T, \mathrm{d}, \varepsilon) \le |\mathscr{D}| \le D(T, \mathrm{d}, \varepsilon)$. □

Proposition 8.11. Let $\mathbb{B}^n = \{u \in \mathbb{R}^n : \|u\|_2 \le 1\}$. then for any $\varepsilon \in (0,1)$,

$$\left(\frac{1}{\varepsilon}\right)^n \le N(\mathbb{B}^n, \|\cdot\|_2, \varepsilon) \le \left(\frac{2}{\varepsilon} + 1\right)^n < \left(\frac{3}{\varepsilon}\right)^n\,.$$

Proof. Let $\mathsf{vol}(\cdot)$ be the volume in $\mathbb{R}^n$. If $\mathscr{N}$ is an ε-net, then $\mathbb{B}^n = \mathbb{B}(0,1) \subseteq \bigcup_{t\in\mathscr{N}} \mathbb{B}(t,\varepsilon)$.

$$\implies \mathsf{vol}(\mathbb{B}(0,1)) \le \sum_{t\in\mathscr{N}} \mathsf{vol}(\mathbb{B}(t,\varepsilon)) = |\mathscr{N}|\varepsilon^n \,\mathsf{vol}(\mathbb{B}(0,1)) \implies |\mathscr{N}| \ge \left(\frac{1}{\varepsilon}\right)^n\,.$$

Hence, $N(\mathbb{B}^n, \|\cdot\|_2, \varepsilon) \geq (\frac{1}{\varepsilon})^n$. On the other hand, if $\mathscr{D}$ is an ε-packing, then $\{\mathbb{B}(t, \varepsilon/2) : t \in \mathscr{D}\}$ are disjoint balls contained in $\mathbb{B}(0, 1+\varepsilon/2)$. This implies that

$$\mathsf{vol}(\mathbb{B}(0, 1+\varepsilon/2)) \geq \sum_{t\in\mathscr{D}} \mathsf{vol}(\mathbb{B}(t, \varepsilon/2)) \Longrightarrow \left(1+\frac{\varepsilon}{2}\right)^n \mathsf{vol}(\mathbb{B}(0,1)) \geq |\mathscr{D}| \left(\frac{\varepsilon}{2}\right)^n \mathsf{vol}(\mathbb{B}(0,1)),$$

so $|\mathscr{D}| \leq (\frac{2}{\varepsilon}+1)^n$. Thus, $N(\mathbb{B}, \|\cdot\|_2, \varepsilon) \leq D(\mathbb{B}, \|\cdot\|_2, \varepsilon) \leq (\frac{2}{\varepsilon}+1)^n$. □

Theorem 8.12. Suppose $X \in \mathbb{R}^{n\times m}$ has independent, mean-zero, σ^2-subgaussian entries. Then $\mathbb{E}\|X\|_{\mathrm{op}} \leq C\sigma(\sqrt{n}+\sqrt{m})$ for a universal constant $C > 0$.

Proof. Recall that $\|X\|_{\mathrm{op}} = \sup_{u\in\mathbb{B}^n, v\in\mathbb{B}^m} X_{u,v} = \sup_{u\in\mathbb{B}^n, v\in\mathbb{B}^m} u^\top X v$, where $X_{u,v}$ is $\|X\|_{\mathrm{op}}$-Lipschitz w.r.t. $\mathrm{d}((u,v),(u',v')) = \|u-u'\|_2 + \|v-v'\|_2$. If $\mathscr{N}_u$ is an ε-net of $\mathbb{B}^n$ and $\mathscr{N}_v$ be an ε-net of $\mathbb{B}^m$ w.r.t. $\|\cdot\|_2$, then $\mathscr{N}_u \times \mathscr{N}_v$ is a 2ε-net of $\mathbb{B}^n \times \mathbb{B}^m$ in d, which implies

$$N(\mathbb{B}^n \times \mathbb{B}^m, \mathrm{d}, 2\varepsilon) \leq N(\mathbb{B}^n, \|\cdot\|_2, \varepsilon) \cdot N(\mathbb{B}^m, \|\cdot\|_2, \varepsilon) \leq \left(\frac{3}{\varepsilon}\right)^{n+m}, \qquad \forall \varepsilon > 0\,.$$

For each $(u,v) \in \mathbb{B}^n \times \mathbb{B}^m$, $u^\top X v$ is σ^2-subgaussian by Hoeffding's inequality (Theorem 1.7), so

$$\mathbb{E}\|X\|_{\mathrm{op}} \leq \varepsilon \underbrace{\mathbb{E}\|X\|_{\mathrm{op}}}_{\text{Lipschitz constant of } X_{u,v}} + \sqrt{2\sigma^2 \log \underbrace{(6/\varepsilon)^{n+m}}_{|N(\mathbb{B}^n\times\mathbb{B}^m, \mathrm{d}, \varepsilon)| \leq (3/(\varepsilon/2))^{n+m}}}$$

Pick $\varepsilon = \frac{1}{2}$, we have $\mathbb{E}\|X\|_{\mathrm{op}} \leq C\sigma(\sqrt{n}+\sqrt{m})$ for some universal constant $C > 0$. □

Remark 8.13. $\|X\|_{\mathrm{op}}$ is a convex, 1-Lipschitz function of its entries. If the entries are bounded, e.g., $X_{ij} \in [-C\sigma, C\sigma]$, or satisfy an LSI such as $\mathsf{Ent}(f(X_{ij})^2) \leq C\sigma^2 \mathbb{E}[f'(X_{ij})^2]$, then by previous lectures, we have

$$\mathbb{P}(\|X\|_{\mathrm{op}} \geq \mathbb{E}\|X\|_{\mathrm{op}} + \sigma t) \leq \exp(-ct^2), \qquad t > 0,$$

for some constant $c > 0$.

Alternatively, assuming only that X_{ij} are σ^2-subgaussian, the above argument may be adjusted into a union bound: $\forall t \geq 0$,

$$\begin{aligned}
&\mathbb{P}(\|X\|_{\mathrm{op}} \geq 2\varepsilon\|X\|_{\mathrm{op}} + C\sigma(\sqrt{n}+\sqrt{m}) + \sigma t)\\
\leq &\mathbb{P}(\exists (u,v) \in \mathscr{N}_u \times \mathscr{N}_v : X_{u,v} \geq C\sigma(\sqrt{n}+\sqrt{m}) + \sigma t)\\
\leq &\left(\frac{6}{\varepsilon}\right)^{n+m} \exp\left(-\frac{1}{2}(C(\sqrt{n}+\sqrt{m}) + t)^2\right)\\
\leq &\left(\frac{6}{\varepsilon}\right)^{n+m} \exp\left(-\frac{C^2}{2}(n+m) - \frac{t^2}{2}\right).
\end{aligned}$$

Take $\varepsilon = \frac{1}{4}$, the above implies that $\exists \tilde{C} > 0$ such that

$$\mathbb{P}(\|X\|_{\mathrm{op}} \geq \tilde{C}\sigma(\sqrt{n}+\sqrt{m}) + 2\sigma t) \leq e^{-t^2/2}.$$

9 Chaining, Dudley's Inequality, Moduli of Continuity

Readings: §5.3-5.4 in [vH14], §8.1-8.2 in [Ver18].

Recall that $\{X_t\}_{t\in T}$, $\mathcal{N} \subseteq T$ is an ε-convering of (T, d) if $\forall t \in T$, $\exists \pi(t) \in \mathcal{N}$ such that $\mathrm{d}(t, \pi(t)) \le \varepsilon$.

General strategy last time:

$$\sup_{t\in T} X_t \le \underbrace{\sup_{t\in T} X_{\pi(t)}}_{\text{(I)}} + \underbrace{\sup_{t\in T} (X_t - X_{\pi(t)})}_{\text{(II)}}$$

(I) If each X_t is σ^2-subgaussian, then by maximal inequality,

$$\mathbb{E} \sup_{t\in T} X_{\pi(t)} = \mathbb{E} \sup_{s\in \mathcal{N}} X_s \le \sqrt{2\sigma^2 \log |\mathcal{N}|}\,.$$

(II) If $|X_t - X_s| \le L\mathrm{d}(t, s)$ for all $s, t \in T$, then

$$\mathbb{E} \sup_{t\in T} (X_t - X_{\pi(t)}) \le \varepsilon \cdot \mathbb{E}[L]\,.$$

Using a "worst-case" Lipschitz bound for (II) can be conservative, and often we want a bound via the typical tail behavior of $\{X_t - X_{\pi(t)}\}$.

Definition 9.1. A mean-zero process $\{X_t\}_{t\in T}$ is a σ^2-*subgaussian process* w.r.t. metric d if

$$\mathbb{E}[e^{\lambda(X_t - X_s)}] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\mathrm{d}(t, s)^2\right) \qquad \forall s, t \in T,\ \forall \lambda \in \mathbb{R}\,.$$

9.1 Chaining and Dudley's inequality

The idea of chaining:

- Approximate $\sup_{t\in T} (X_t - X_{\pi(t)})$ by a further (finer) ε'-net $\mathcal{N}'$ so that for any $t \in T$, $\exists \pi'(t) \in \mathcal{N}'$ satisfying $\mathrm{d}(t, \pi'(t)) \le \varepsilon'$. Then

  $$\sup_{t\in T} (X_t - X_{\pi(t)}) \le \sup_{t\in T} (X_{\pi'(t)} - X_{\pi(t)}) + \sup_{t\in T} (X_t - X_{\pi'(t)})\,.$$

  For the former term, we bound it via subgaussianity of $X_{\pi'(t)} - X_{\pi(t)}$.

- Recursively apply this to the latter term.

Definition 9.2. $\{X_t\}_{t\in T}$ is *separable* if there is a countable subset $T_o \subseteq T$ such that almost surely, for all $t \in T$ there exists $\{t_k\}_{k=1}^{\infty}$ such that

$$X_{t_k} \to X_t, \qquad \mathrm{d}(t_k, t) \to 0 \qquad \text{as } k \to \infty\,.$$

In particular $\sup_{t\in T} X_t = \sup_{t\in T_o} X_t$.

Theorem 9.3 (Dudley's inequality)**.** If $\{X_t\}_{t\in T}$ is mean-zero, separable, σ^2-subgaussian on (T, d), then for a finite constant $C > 0$,

$$\mathbb{E} \sup_{t\in T} X_t \le C\sigma \int_0^{\infty} \sqrt{\log N(T, \mathrm{d}, \varepsilon)}\, \mathrm{d}\varepsilon\,.$$

Note that we usually work with bounded T, so the upper bound of the integral is just $\mathsf{diam}(T) = \max_{s,t\in T} \mathrm{d}(s, t)$. When $\varepsilon > \mathsf{diam}(T)$, $N(T, \mathrm{d}, \varepsilon) = 1$ and so $\log N(T, \mathrm{d}, \varepsilon) = 0$.

Proof. Suppose first that T is finite. Let $\varepsilon_k = 2^{-k}$ for $k \in \mathbb{Z}$ and $\mathcal{N}_k$ be the smallest ε_k-net of T. Let $\kappa, K \in \mathbb{Z}$ such that $\mathcal{N}_\kappa$ is a single point $\{t_o\}$ and $\mathcal{N}_K = T$. These two numbers κ and K exist because $|T| < \infty$ and $\mathsf{diam}(T) < \infty$. Let $\pi_k(t) \in \mathcal{N}_k$ be a point such that $\mathrm{d}(t, \pi_k(t)) \le \varepsilon_k = 2^{-k}$. Then by telescoping the difference, we have

$$X_t - X_{t_o} = \sum_{k=\kappa+1}^{K} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right).$$

Note that

$$\mathrm{d}(\pi_k(t), \pi_{k-1}(t)) \le \mathrm{d}(\pi_k(t), t) + \mathrm{d}(t, \pi_{k-1}(t)) \le 2^{-k} + 2^{-(k-1)} = 3 \cdot 2^{-k}.$$

By σ^2-subgaussianity of $\{X_t\}_{t\in T}$ on (T, d), we have $\forall \lambda \ge 0$,

$$\mathbb{E}[e^{\lambda(X_{\pi_k(t)} - X_{\pi_{k-1}(t)})}] \le \frac{\lambda^2\sigma^2}{2}\mathrm{d}(\pi_k(t), \pi_{k-1}(t))^2 \le \frac{\lambda^2\sigma^2(3\cdot 2^{-k})^2}{2}.$$

Hence, we have

$$\begin{aligned}
\mathbb{E}\sup_{t\in T} X_t &= \mathbb{E}[\sup_{t\in T}(X_t - X_{t_o})] \quad \text{since } \mathbb{E}[X_t] = 0, \forall t \in T \\
&= \sum_{k=\kappa+1}^{K} \mathbb{E}[\sup_{t\in T} \underbrace{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}}_{\sigma^2(3\cdot 2^{-k})^2\text{-subgaussian}}] \\
&\le \sum_{k=\kappa+1}^{K} \sqrt{2\sigma^2(3\cdot 2^{-k})^2 \log\left(|\mathcal{N}_k| \cdot |\mathcal{N}_{k-1}|\right)} \quad \text{by Corollary 8.3} \\
&\le \sum_{k=\kappa+1}^{K} 6\sigma \cdot 2^{-k}\sqrt{\log|\mathcal{N}_k|} \quad \text{since } |\mathcal{N}_k| \ge |\mathcal{N}_{k-1}| \\
&\le \sum_{k=\kappa+1}^{K} 12\sigma \int_{2^{-(k+1)}}^{2^{-k}} \sqrt{\log N(T, \mathrm{d}, \varepsilon_k)}\,\mathrm{d}\varepsilon \\
&\le \sum_{k=\kappa+1}^{K} 12\sigma \int_{2^{-(k+1)}}^{2^{-k}} \sqrt{\log N(T, \mathrm{d}, \varepsilon)}\,\mathrm{d}\varepsilon \\
&\le 12\sigma \int_0^\infty \sqrt{\log N(T, \mathrm{d}, \varepsilon)}\,\mathrm{d}\varepsilon.
\end{aligned}$$

If T has infinite cardinality, then let $T_o = \{t_1, t_2, t_3, \ldots\}$ be a countable set from separability. Then,

$$\begin{aligned}
\mathbb{E}\sup_{t\in T} X_t &= \mathbb{E}\sup_{t\in T_o} X_t \\
&= \mathbb{E}\sup_{t\in T_o} (X_t - X_{t_o}) \\
&= \mathbb{E}\lim_{k\to\infty} \sup_{t\in\{t_1,\ldots,t_k\}} (X_t - X_1) \\
&\overset{\text{MCT}}{=} \lim_{k\to\infty} \mathbb{E} \sup_{t\in\{t_1,\ldots,t_k\}} (X_t - X_{t_o}),
\end{aligned}$$

and then apply the result for each finite set $\{t_1, \ldots, t_k\}$. □

Example 9.4. $Z_1, \ldots, Z_n$ are iid on $[0,1]$. Let $\mathscr{F} = \{f : [0,1] \mapsto \mathbb{R}, 1\text{-Lipschitz}, f(0) = 0\}$. Also let $X_f = \frac{1}{n}\sum_{i=1}^n (f(Z_i) - \mathbb{E}[f(Z_i)])$ and $W = \sup_{f\in\mathscr{F}} X_f$. From lecture 8 we know that $N(\mathscr{F}, \|\cdot\|_\infty, \varepsilon) \le C^{1/\varepsilon}$,

$|X_f - X_g| \le 2\|f-g\|_\infty$, and that $\log \mathbb{E}[e^{\lambda X_f}] \le \frac{\lambda^2}{2n}$, i.e., X_f is $\frac{1}{n}$-subgaussian for any $f \in \mathscr{F}$, which implies

$$\mathbb{E}[W] \le \inf_{\varepsilon > 0} \left\{ 2\varepsilon + \sqrt{\frac{2}{n} \log N(\mathscr{F}, \|\cdot\|_\infty, \varepsilon)} \right\} \asymp n^{-1/3} .$$

However, this is not optimal. Instead, we can apply the idea of chaining to this problem.

By Hoeffding's inequality,

$$\mathbb{E}[e^{\lambda(X_f - X_g)}] \le \exp\left(\frac{\lambda^2}{2n}\|f-g\|_\infty^2\right) = \exp\left(\frac{\lambda^2}{2n}\mathrm{d}(f,g)^2\right) .$$

Hence, with ℓ_∞-norm, $\{X_f\}_{f\in\mathscr{F}}$ is mean-zero, $\frac{1}{n}$-subgaussian on $(\mathscr{F}, \|\cdot\|_\infty)$. Therefore, we can apply Dudley's inequality (Theorem 9.3):

$$\begin{aligned}\mathbb{E}[W] &\le \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\mathscr{F}, \|\cdot\|_\infty, \varepsilon)}\, \mathrm{d}\varepsilon \\ &\le \frac{C'}{\sqrt{n}} \int_0^2 \sqrt{1/\varepsilon}\, \mathrm{d}\varepsilon \asymp n^{-1/2} .\end{aligned}$$

This is the optimal bound up to a constant factor.

Example 9.5 (Gaussian width)**.** Let $T \subseteq \mathbb{R}^n$ be bounded and $g \sim \mathcal{N}(0, I)$. Let $X_t = g^\top t$ for $t \in T$, and the Gaussian width of T is defined as $w(T) := \mathbb{E}\sup_{t\in T} X_t$. This is a geometric measure of how large T looks in a random Gaussian direction. By the Gaussian MGF, we have

$$\mathbb{E}[e^{\lambda(X_t - X_s)}] = \mathbb{E}[e^{\lambda g^\top (t-s)}] = \exp\left(\frac{\lambda^2}{2}\|t-s\|_2^2\right) .$$

Hence, $\{X_t\}_{t\in T}$ is 1-subgaussian w.r.t. $\mathrm{d}(t,s) = \|t-s\|_2$. By Dudley's inequality,

$$w(T) = \mathbb{E}\sup_{t\in T} X_t \le C \int_0^\infty \sqrt{\log N(T, \|\cdot\|_2, \varepsilon)}\, \mathrm{d}\varepsilon .$$

Consider, for example, $T := \mathbb{B}^n = \{x \in \mathbb{R}^n \mid \|x\|_2 \le 1\}$. Then by proposition in lecture 8, we have: $\forall \varepsilon \in (0,1)$, $N(\mathbb{B}^n, \|\cdot\|_2, \varepsilon) \le (\frac{3}{\varepsilon})^n$, and $= 1$ when $\varepsilon \ge 1$. Hence,

$$w(\mathbb{B}^n) = \mathbb{E}\|g\|_2 \le C \int_0^1 \sqrt{n \log(3/\varepsilon)}\, \mathrm{d}\varepsilon \le C'\sqrt{n} .$$

Remark 9.6. The same proof of Dudley's inequality shows that

$$\mathbb{E}\sup_{t\in T} X_t \le C \int_\delta^\infty \sqrt{\log N(T, \mathrm{d}, \varepsilon)}\, \mathrm{d}\varepsilon + 2\mathbb{E} \sup_{s,t\in T : \mathrm{d}(s,t)\le\delta} |X_s - X_t|$$

by running chaining down to the scale $2^{-K} \approx \delta$. This is useful when $\sqrt{\log N(T, \mathrm{d}, \varepsilon)}$ is not integrable at 0.

Theorem 9.7 (Tail inequality)**.** If $\{X_t\}_{t\in T}$ is separable, mean-zero, and σ^2-subgaussian on (T, d), then for universal constants $C, C' > 0$, any $t_o \in T$, and any $u \ge 0$,

$$\mathbb{P}\left[\sup_{t\in T} X_t > X_{t_o} + C'\sigma\left(\int_0^{\mathsf{diam}(T)} \sqrt{\log N(T, \mathrm{d}, \varepsilon)}\, \mathrm{d}\varepsilon + \mathsf{diam}(T)\cdot u\right)\right] \le Ce^{-u^2/2} .$$

Proof. Suppose T has finite cardinality. Let $\varepsilon_k = 2^{-k}$ and $\mathcal{N}_k$ be the smallest ε_k-net. Also, let $\mathcal{N}_\kappa = \{t_o\}$ and $\mathcal{N}_K = T$ as before, and

$$\sup_{t\in T}(X_t - X_{t_o}) = \sum_{k=\kappa+1}^{K}\Big(\underbrace{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}}_{\sigma^2(3\cdot 2^{-k})^2\text{-subgaussian}}\Big).$$

By subgaussianity and union bound, $\forall z \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}(\sup_{t\in T}(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) > 3\sigma\cdot 2^{-k}(\sqrt{2\log|\mathcal{N}_k||\mathcal{N}_{k-1}|} + z))\\
\le&|\mathcal{N}_k||\mathcal{N}_{k-1}|\sup_{t\in T}\mathbb{P}(X_{\pi_k(t)} - X_{\pi_{k-1}(t)} > 3\sigma\cdot 2^{-k}\sqrt{2\log|\mathcal{N}_k||\mathcal{N}_{k-1}|} + z)\\
\le&|\mathcal{N}_k||\mathcal{N}_{k-1}|\exp\left(-\frac{1}{2}(\sqrt{2\log|\mathcal{N}_k||\mathcal{N}_{k-1}|} + z)^2\right)\quad\text{by subgaussianity}\\
\le&e^{-z^2/2}.
\end{aligned}$$

Now for any $z_{\kappa+1},\dots,z_K$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in T}(X_t - X_{t_o}) \ge \sum_{k=\kappa+1}^{K} 3\sigma\cdot 2^{-k}(\sqrt{2\log|\mathcal{N}_k||\mathcal{N}_{k-1}|} + z_k)\right)\\
\overset{\text{union bound}}{\le}&\sum_{k=\kappa+1}^{K}\mathbb{P}\left(\sup_{t\in T}(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \ge 3\sigma\cdot 2^{-k}(\sqrt{2\log|\mathcal{N}_k||\mathcal{N}_{k-1}|} + z_k)\right)\\
\le&\sum_{k=\kappa+1}^{K}\exp(-z_k^2/2).
\end{aligned}$$

Now, set $z_k = u + \sqrt{k-\kappa}$. We have

$$\begin{aligned}
&\sum_{k=\kappa+1}^{K} 3\sigma^2\cdot 2^{-k}(\sqrt{2\log|\mathcal{N}_k||\mathcal{N}_{k-1}|} + z_k)\\
\le&\sum_{k=\kappa+1}^{K} 6\sigma\cdot 2^{-k}\sqrt{\log|\mathcal{N}_k|} + \sum_{k=\kappa+1}^{K} 3\sigma\cdot 2^{-k}u + \sum_{k=\kappa+1}^{K} 3\sigma\cdot 2^{-k}\sqrt{k-\kappa}\\
\le&12\sigma\int_0^{\mathsf{diam}(T)}\sqrt{\log N(T,\mathrm{d},\varepsilon)}\,\mathrm{d}\varepsilon + C\sigma 2^{-\kappa}u + C\sigma 2^{-\kappa}\quad\text{by previous proof and geometric series}\\
\le&C'\sigma\left(\int_0^{\mathsf{diam}(T)}\sqrt{\log N(T,\mathrm{d},\varepsilon)}\,\mathrm{d}\varepsilon + \mathsf{diam}(T)\cdot u\right).
\end{aligned}$$

The last inequality holds because of the following argument. Choose $\kappa \in \mathbb{Z}$ to be the largest integer such that $2^{-\kappa} \ge \mathsf{diam}(T)$. Then $2^{-(\kappa+1)} \le \mathsf{diam}(T) \iff 2^{-\kappa} \le 2\,\mathsf{diam}(T)$. Also note that $N(T,\mathrm{d},2^{-(\kappa+2)}) \ge 2$, so $2^{-\kappa} \le C\sum_{k=\kappa+1}^{K} 2^{-k}\sqrt{\log|\mathcal{N}_k|}$. Hence, the third term can be absorbed into the first. Moreover,

$$\sum_{k=\kappa+1}^{K} e^{-z_k^2/2} \le \sum_{k=\kappa+1}^{K} e^{-u^2/2-(k-\kappa)/2} \le Ce^{-u^2/2}.$$

Therefore,

$$\mathbb{P}\left[\sup_{t\in T}(X_t - X_{t_o}) > C'\sigma\left(\int_0^{\mathsf{diam}(T)}\sqrt{\log N(T,\mathrm{d},\varepsilon)} + \mathsf{diam}(T)\cdot u\right)\right] \le Ce^{-u^2/2}.$$

If T has infinite cardinality, let $T_o = \{t_1, t_2, \ldots\}$ be the countable set from separability, and apply

$$\begin{aligned}\mathbb{P}(\sup_{t\in T}\,(X_t - X_{t_o}) > x) &= \mathbb{P}(\sup_{t\in T_o}\,(X_t - X_{t_o}) > x)\\ &= \mathbb{P}\left(\bigcup_{k\geq 1}\left\{\sup_{t\in\{t_1,\ldots,t_k\}}\,(X_t - X_{t_o}) > x\right\}\right)\\ &\overset{\text{MCT}}{=} \lim_{k\to\infty}\mathbb{P}\left(\sup_{t\in\{t_1,\ldots,t_k\}}\,(X_t - X_{t_o}) > x\right) \leq Ce^{-u^2}\,.\end{aligned}$$

□

9.2 Modulus of continuity

Definition 9.8 (Modulus of continuity)**.** A function $\omega : (0,\infty) \mapsto (0,\infty)$ for which $\sup_{s,t\in T}\frac{|X_t - X_s|}{\omega(\mathrm{d}(t,s))} < \infty$ a.s. is called a *modulus-of-continuity* for $\{X_t\}_{t\in T}$.

Theorem 9.9 (Modulus of continuity)**.** Suppose $\{X_t\}_{t\in T}$ is mean-zero, separable, σ^2-subgaussian on (T,d) with $N(T,\mathrm{d},\varepsilon) \geq (c_0/e)^q$ for some $c_0, q > 0$, $\forall\varepsilon > 0$. Set

$$\omega(\delta) = \int_0^\delta \sqrt{\log N(T,\mathrm{d},\varepsilon)}\,\mathrm{d}\varepsilon\,.$$

Then, for constants $C, C', c > 0$ depending on c_0, q, and $\mathsf{diam}(T)$,

$$\mathbb{P}\left(\sup_{s,t\in T}\frac{|X_t - X_s|}{\omega(\mathrm{d}(t,s))} \geq C\sigma(1+y)\right) \leq C'e^{-cy^2}, \qquad \forall y \geq 0\,.$$

In particular, this sup is finite with probability 1, so $\omega(\cdot)$ is a modulus of continuity.

Example 9.10. Let $\{B_t\}_{t\in[0,1]}$ be 1-dimensional Brownian motion. Then

$$\mathbb{E}[e^{\lambda(B_t - B_s)}] = \exp\left(\frac{\lambda^2}{2}|t-s|\right),$$

which implies that $\{B_t\}$ is subgaussian w.r.t. $\mathrm{d}(t,s) = \sqrt{|t-s|}$. Hence, $\forall\varepsilon \in (0,1)$,

$$N([0,1], \sqrt{|\cdot|}, \varepsilon) \leq \frac{1}{\varepsilon^2}\,.$$

Applying the theorem above on modulus of continuity,

$$\omega(\delta) = \int_0^\delta \sqrt{\log(1/\varepsilon)^2}\,\mathrm{d}\varepsilon \asymp \int_0^\delta \sqrt{\log(1/\varepsilon)}\,\mathrm{d}\varepsilon \asymp \delta\sqrt{\log\frac{1}{\delta}}\,.$$

So, $\sqrt{|t-s|\log\frac{1}{|t-s|}}$ is a modulus-of-continuity for $\{B_t\}_{t\in[0,1]}$.

Proof of Theorem 9.9. Slice the product space $T\times T$ by values of $\mathrm{d}(t,s)$ for $(t,s)\in T\times T$.

Set $\alpha_k = 2^{-k}\,\mathsf{diam}(T)$ and $\mathscr{A}_k = \{(t,s)\in T\times T \mid \mathrm{d}(t,s)\leq\alpha_k\}$.

$$\begin{aligned}\mathbb{P}\left(\sup_{s,t\in T}\frac{|X_t - X_s|}{\omega(\mathrm{d}(t,s))} \geq x\right) &= \mathbb{P}\left(\sup_{k\geq 0}\sup_{s,t\in T:\mathrm{d}(t,s)\in[\alpha_{k+1},\alpha_k]}\frac{|X_t - X_s|}{\omega(\mathrm{d}(t,s))} \geq x\right)\\ &\leq \sum_{k=0}^{\infty}\mathbb{P}\left(\sup_{s,t\in T:\mathrm{d}(t,s)\in[\alpha_{k+1},\alpha_k]}\frac{|X_t - X_s|}{\omega(\mathrm{d}(t,s))} \geq x\right)\\ &\leq \sum_{k=0}^{\infty}\mathbb{P}\left(\sup_{s,t\in T:\mathrm{d}(t,s)\in[\alpha_{k+1},\alpha_k]}|X_t - X_s| \geq x\cdot\omega(\alpha_{k+1})\right)\end{aligned}$$

Apply chaining to $X_{t,s} = X_t - X_s$ for $(t,s) \in \mathscr{A}_k$:

$$\begin{aligned}\mathbb{E}[e^{\lambda(X_{t,s}-X_{u,v})}] &= \mathbb{E}[e^{\lambda(X_t-X_u)} \cdot e^{\lambda(X_v-X_s)}] \\ &\le \sqrt{\mathbb{E}[e^{2\lambda(X_t-X_u)}]\mathbb{E}[e^{2\lambda(X_v-X_s)}]} \\ &\le \sqrt{e^{4\lambda^2\sigma^2(X_t-X_u)^2/2} e^{4\lambda^2\sigma^2(X_v-X_s)^2/2}} \\ &= \exp(\lambda^2\sigma^2 \mathrm{d}(t,u)^2 + \lambda^2\sigma^2\mathrm{d}(s,v)^2)\,.\end{aligned}$$

On the other hand, if $(s,t),(u,v) \in \mathscr{A}_k$, then

$$\begin{aligned}\mathbb{E}[e^{\lambda(X_{t,s}-X_{u,v})}] &= \mathbb{E}[e^{\lambda(X_t-X_s)+\lambda(X_u-X_v)}] \\ &\le \mathbb{E}[e^{\lambda^2\sigma^2\mathrm{d}(t,s)^2+\lambda^2\sigma^2\mathrm{d}(u,v)^2}] \\ &\le \exp(2\lambda^2\sigma^2\alpha_k^2)\,.\end{aligned}$$

Let $\tilde{\mathrm{d}}((t,s),(u,v)) = \min(\sqrt{2\mathrm{d}(t,u)^2 + 2\mathrm{d}(v,s)^2}, 2\alpha_k)$. Then

$$\mathbb{E}[e^{\lambda(X_{t,s}-X_{u,v})}] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\tilde{\mathrm{d}}((t,s),(u,v))^2\right),$$

so $X_{t,s}$ is σ^2-subgaussian w.r.t. $\tilde{\mathrm{d}}$. Equip $\mathscr{A}_k$ with $\tilde{\mathrm{d}}$. Apply the tail version of Dudley's inequality (Theorem 9.7) with $X_{t,t} = 0$ and $\mathsf{diam}(\mathscr{A}_k) \le 4\alpha_k$,

$$\mathbb{P}\left[\sup_{(t,s)\in\mathscr{A}_k} X_{t,s} \ge C'\sigma\left(\int_0^{4\alpha_k}\sqrt{\log N(\mathscr{A}_k,\tilde{\mathrm{d}},\varepsilon)}\,\mathrm{d}\varepsilon + 4\alpha_k u\right)\right] \le Ce^{-u^2/2} \qquad \cdots(\star)$$

If $\mathscr{N}$ is an ε-net of (T,d), then $\mathscr{N}\times\mathscr{N}$ is a 2ε-net of $(T\times T,\tilde{\mathrm{d}})$ because

$$\tilde{\mathrm{d}}((t,s),(\pi(t),\pi(s))) \le \sqrt{2\varepsilon^2+2\varepsilon^2} \le 2\varepsilon\,.$$

Hence, $N(\mathscr{A}_k,\tilde{\mathrm{d}},4\varepsilon) \le N^{\mathrm{ext}}(\mathscr{A}_k,\tilde{\mathrm{d}},2\varepsilon) \le N(T,\mathrm{d},\varepsilon)^2$, so $(\star)$ becomes

$$\mathbb{P}\left[\sup_{(t,s)\in\mathscr{A}_k} X_{t,s} \ge C''\sigma\left(\int_0^{\alpha_k}\sqrt{\log N(T,\mathrm{d},\varepsilon)}\,\mathrm{d}\varepsilon + \alpha_k u\right)\right] \le Ce^{-u^2/2}\,.$$

Let $u = \frac{y}{\alpha_k}\int_0^{\alpha_k}\sqrt{\log N(T,\mathrm{d},\varepsilon)}\,\mathrm{d}\varepsilon = \frac{y}{\alpha_k}\omega(\alpha_k)$ by definition. Thus,

$$\mathbb{P}\left(\sup_{s,t\in\mathscr{A}_k} X_{t,s} \ge C''\sigma(1+y)\omega(\alpha_k)\right) \le Ce^{-u^2/2}\,.$$

- $\omega(\alpha_k) = \int_0^{\alpha_{k+1}} + \int_{\alpha_{k+1}}^{\alpha_k}\sqrt{\log N(T,\mathrm{d},\varepsilon)}\,\mathrm{d}\varepsilon \le 2\omega(\alpha_{k+1})$.
- $u \ge y\sqrt{\log N(T,\mathrm{d},\alpha_k)} \ge y\sqrt{\log(\frac{c_0}{\alpha_k})^q}$, and $u \ge \frac{y}{2}\sqrt{\log N(T,\mathrm{d},\frac{\alpha_k}{2})} \ge \frac{y\sqrt{\log 2}}{2}$ since $\frac{\alpha_k}{2} \le \frac{\mathsf{diam}(T)}{2}$.

These two observations imply that

$$\begin{aligned}&\mathbb{P}\left(\sup_{s,t\in\mathscr{A}_k} X_t - X_s \ge 2C''\sigma(1+y)\omega(\alpha_{k+1})\right) \le Ce^{-\frac{y^2}{2}\max(\log(\frac{c_0}{\alpha_k})^q,\frac{\log 2}{4})} = C\min(2^{-\frac{y^2}{8}},(c_0/\alpha_k)^{-\frac{qy^2}{2}}) \\ \Longrightarrow &\mathbb{P}\left(\sup_{s,t\in T}\frac{X_t-X_s}{\omega(\mathrm{d}(t,s))} \ge 2C''\sigma(1+y)\right) \le \sum_{k=0}^{\infty} C\min(2^{-\frac{y^2}{8}},(2^{-k}\,\mathsf{diam}(T)/c_0)^{\frac{qy^2}{2}}) \le C'e^{-cy^2}\,.\end{aligned}$$

□

10 Empirical Processes, VC Dimension

Readings: §7.1-7.2 in [vH14], §8.3-8.4 in [Ver18].

10.1 Empirical process

$X_1, \dots, X_n$ are iid random variables on $\mathcal{X}$. $\mathcal{F}$ is a class of functions $f : \mathcal{X} \mapsto \mathbb{R}$. The associated *empirical process* is

$$Z_f = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_i)], \quad \forall f \in \mathcal{F}.$$

Example 10.1. Let $\mathcal{X} = [0,1]$ and $\mathcal{F} = \{f : [0,1] \mapsto \mathbb{R} \mid f \text{ is } L - \text{Lipschitz}, f(0) = 0\}$. Then $W = \sup_{f\in\mathcal{F}} Z_f$ is 1-Wasserstein distance between $\frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ and law of Z.

Example 10.2. (Glivenko-Cantelli ULLN) Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} \mid t \in \mathbb{R}\}$. Define

$$\begin{aligned} Z_f &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le t\}} - \mathbb{P}(X_1 \le t) \\ &= F_n(t) - F(t) \xrightarrow{\mathbb{P}} 0 \end{aligned}$$

as $n \to \infty$ for each fixed $f \in \mathcal{F}$ by LLN, where F_n is the empirical CDF and F is the CDF.

By Glivenko-Cantelli's ULLN, $\sup_{f\in\mathcal{F}} |Z_f| \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$.

Question: how to bound the expectation / tail of $\sup_{f\in\mathcal{F}} |Z_f|$?

Example 10.3 (Classification risk)**.** Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be iid, where $X_i \in \mathcal{X}$ and $Y_i \in \{0,1\}$. $\mathcal{F}$ is the class of classifiers that maps $\mathcal{X}$ to $\{0,1\}$.

- Empirical risk: $R_n(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{f(X_i)\neq Y_i\}}$.
- True (test) risk: $R(f) = \mathbb{P}(f(X) \neq Y) = \mathbb{P}(f(X_i) \neq Y_i), \forall i$.

Question: how to bound $\sup_{f\in\mathcal{F}} |R_n(f) - R(f)|$?

Note: these function classes are not continuous, do not have finite covers in $\|\cdot\|_\infty$. For example, for any finite set $\mathcal{N} \subseteq \mathcal{F} = \{\mathbb{1}_{(-\infty,t]} \mid t \in \mathbb{R}\}$, there exists $f \in \mathcal{F}$ such that $\min_{g\in\mathcal{N}} \|f - g\|_\infty = 1$, which implies that for any $\varepsilon < 1$, $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = \infty$.

Lemma 10.4 (Rademacher symmetrization)**.** For any independent random variables $X_1, \dots, X_n$ and function class $\mathcal{F}$,

$$\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_1)] \le 2\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(X_i)\,,$$

where $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} \mathsf{Rademacher}(\pm 1)$ independent of $X_1, \dots, X_n$.

Proof. Let $X_1', \dots, X_n'$ be independent copies of $X_1, \dots, X_n$.

$$\begin{aligned} \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}[f(X_i)]) &= \mathbb{E}_X \sup_{f\in\mathcal{F}} \mathbb{E}_{X'} \sum_{i=1}^{n} (f(X_i) - f(X_i')) \\ &= \mathbb{E}_{X,X'} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} (f(X_i) - f(X_i')) \\ &= \mathbb{E}_{X,X',\varepsilon} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i (f(X_i) - f(X_i')) \\ &\le \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(X_i) + \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} (-\varepsilon_i) f(X_i') = 2\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \varepsilon_i f(X_i)\,, \end{aligned}$$

where we know that $\{f(X_i) - f(X_i')\}_{f\in\mathscr{F}}$ and $\{\varepsilon_i(f(X_i) - f(X_i'))\}_{f\in\mathscr{F}}$ have the same distribution. □

The idea is to bound $\mathbb{E}_\varepsilon \sup_{f\in\mathscr{F}} \frac{1}{n}\varepsilon_i f(X_i) := \mathbb{E}_\varepsilon \sup_{f\in\mathscr{F}} \tilde{Z}_f$ conditioned on any fixed $(X_1,\dots,X_n)\in\mathscr{X}^n$. Note that conditioned on $(X_1,\dots,X_n)$,

$$\tilde{Z}_f - \tilde{Z}_g = \frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - g(X_i))$$

is $\frac{1}{n^2}\sum_{i=1}^n (f(X_i)-g(X_i))^2$-subgaussian, i.e., $\{\tilde{Z}_f\}_{f\in\mathscr{F}}$ is $\frac{1}{n}$-subgaussian w.r.t.

$$\mathrm{d}(f,g) = \|f-g\|_{\mathbb{L}^2(\mathbb{P}_n)} = \sqrt{\frac{1}{n}\sum_{i=1}^n (f(X_i)-g(X_i))^2}\,.$$

By Dudley's theorem,

$$\mathbb{E}_{\varepsilon,X} \sup_{f\in\mathscr{F}} \tilde{Z}_f \lesssim \frac{1}{\sqrt{n}}\mathbb{E}_X \int_0^\infty \sqrt{\log N(\mathscr{F}, \|\cdot\|_{\mathbb{L}^2(\mathbb{P}_n)}, \varepsilon)}\,\mathrm{d}\varepsilon\,. \tag{10.1}$$

Covering $\mathscr{F}$ at only n points $X_1,\dots,X_n$ is easier than covering on all $\mathscr{X}$. For example, for $\mathscr{F} = \{\mathbb{1}_{(-\infty,t]} \mid t\in\mathbb{R}\}$, then $\exists \mathscr{N}\subseteq\mathscr{F}$ of $n+1$ functions such that $\forall f\in\mathscr{F}$, $\exists \pi(f)\in\mathscr{N}$ such that $\|f-\pi(f)\|_{\mathbb{L}^2(\mathbb{P}_n)} = 0$. This is because although there are infinitely many functions in $\mathscr{F}$, on a fixed sample of size n, it induces only $n+1$ distinct labelings. To be specific, let $f_t(x) := \mathbb{1}_{\{x\le t\}}$, we can choose

$$\mathscr{N} = \{f_{X_{(1)}},\dots,f_{X_{(n)}}\}\cup\{f_{X_{(0)}}\}\,,\qquad X_{(0)} := X_{(1)} - 1\,.$$

For each $f_{t_o}\in\mathscr{F}$, if we can find k for which $X_{(k)}\le t_o < X_{(k+1)}$, then f_{t_o} agrees with $f_{X_{(k)}}$ on all sample points, so $\|f_{t_o} - f_{X_{(k)}}\|_{\mathbb{L}^2(\mathbb{P}_n)} = 0$; otherwise we either have $t_o < X_{(0)}$ or $t_o \ge X_{(n)}$. In the former case, $f_{t_o}\equiv f_{X_{(0)}}$ on all sample points; in the latter case, $f_{t_o}\equiv f_{X_{(n)}}$ on all sample points. Hence, $n+1$ functions suffice to cover $\mathscr{F}$.

In general, for any class $\mathscr{F}: \mathscr{X}\mapsto\{0,1\}$, there are 2^n possibilities for $(f(X_1),\dots,f(X_n))$ for any $f\in\mathscr{F}$, so there exists a net $\mathscr{N}^{\mathrm{ext}}$ of 2^n functions so that $\forall f\in\mathscr{F}$, $\exists\pi(f)\in\mathscr{N}^{\mathrm{ext}}$ such that $\|f-\pi(f)\|_{\mathbb{L}^2(\mathbb{P}_n)} = 0$.

10.2 VC dimension

Definition 10.5 (VC dimension)**.** Let $\mathscr{F}$ be a class of functions $f:\mathscr{X}\mapsto\{0,1\}$. A set $\Lambda\subseteq\mathscr{X}$ is *shattered* by $\{f|_\Lambda \mid f\in\mathscr{F}\}$ contains all $2^{|\Lambda|}$ possible functions $\Lambda\mapsto\{0,1\}$. The *VC dimension* is

$$\mathsf{VC}(\mathscr{F}) := \sup_{\Lambda\subseteq\mathscr{X} \text{ shattered by } \mathscr{F}} |\Lambda|\,.$$

Equivalently, let $\mathscr{A}$ be a class of subsets of $\mathscr{X}$. $\Lambda\subseteq\mathscr{X}$ is shattered by $\mathscr{A}$ if $\forall S\subseteq\Lambda$, there exists $A\in\mathscr{A}$ such that $\Lambda\cap A = S$.

Example 10.6. $\mathscr{F} = \{\mathbb{1}_{(-\infty,t]} \mid t\in\mathbb{R}\}$ on $\mathbb{R}$. Any singleton $\{x\}$ is shattered. Any two points $\{x,y\}$ are not shattered. WLOG, $x<y$, then we cannot have $f(x)=1$ and $f(y)=0$, so $\mathsf{VC}(\mathscr{F}) = 1$.

Example 10.7. $\mathscr{F} = \{\mathbb{1}_{[a,b]} \mid a,b\in\mathbb{R}\}$ on $\mathbb{R}$. Any two points $\{x,y\}$ are shattered. Any three points $\{x,y,z\}$ are not shattered. WLOG, we let $x<y<z$. We cannot have an $f\in\mathscr{F}$ such that $f(x)=1, f(y)=0, f(z)=1$, so $\mathsf{VC}(\mathscr{F}) = 2$.

Example 10.8. $\mathscr{F}$ = {indicators of closed half planes} on $\mathbb{R}^2$. Three points in general position are shattered, any 4 points are not shattered, so $\mathsf{VC}(\mathscr{F}) = 3$.

Example 10.9. $\mathcal{X} = \{x_1, x_2, x_3\}$. Represent $f : \mathcal{X} \mapsto \{0,1\}$ by binary strings. $\mathcal{F} = \{001, 010, 100, 111\}$ (number of $x \in \mathcal{X}$ where $f(x) = 1$ is odd). The set $\{x_1, x_2\}$ is shattered. The set $\{x_1, x_2, x_3\}$ is not shattered, so $\mathsf{VC}(\mathcal{F}) = 2$.

Theorem 10.10. Let $\mathcal{F} : \mathcal{X} \mapsto \{0,1\}$ and $\mathbb{P}$ be any probability measure on $\mathcal{X}$. Define $\|f - g\|^2_{\mathbb{L}^2(\mathbb{P})} = \mathbb{E}_{\mathbb{P}}(f(X) - g(X))^2$. Then for a universal constant $C > 0$ and any $\varepsilon > 0$,

$$N(\mathcal{F}, \|\cdot\|_{\mathbb{L}^2(\mathbb{P})}, \varepsilon) \le \left(\frac{C}{\varepsilon}\right)^{C \cdot \mathsf{VC}(\mathcal{F})}.$$

Lemma 10.11 (Pajor)**.** Suppose $\mathcal{X}$ is a finite set, and $\mathcal{F}$ is any class of functions $f : \mathcal{X} \mapsto \{0,1\}$. Then

$$|\mathcal{F}| \le \text{number of subsets of } \mathcal{X} \text{ shattered by } \mathcal{F} \text{ (including } \emptyset)$$

Proof. We prove by induction on $|\mathcal{X}|$. Base case: $\mathcal{X} = \{x\}$. $\emptyset$ is always shattered, so right hand side ≥ 1. If $|\mathcal{F}| = 2$ i.e., $\mathcal{F}$ contains both an f for which $f(x) = 0$ and another with $f(x) = 1$, then $\{x\}$ also shattered, so both LHS and RHS are 2. The lemma holds.

Inductive step: let $\mathcal{X} = \mathcal{X}_0 \cup \{x\}$ and suppose that the lemma holds for $\mathcal{X}_0$. Let $S(\mathcal{F})$ be the number of sets shattered by $\mathcal{F}$. Write

$$\mathcal{F} = \underbrace{\{f \in \mathcal{F} \mid f(x) = 0\}}_{\mathcal{F}_0} \cup \underbrace{\{f \in \mathcal{F} \mid f(x) = 1\}}_{\mathcal{F}_1},$$

so $|\mathcal{F}| = |\mathcal{F}_0| + |\mathcal{F}_1| \le S(\mathcal{F}_0) + S(\mathcal{F}_1)$ by induction hypothesis applied to $\mathcal{X}_0$. Consider two cases:

- If $\Lambda \subseteq \mathcal{X}_0$ is shattered by exactly one of $\mathcal{F}_0$ and $\mathcal{F}_1$, then trivially Λ is shattered by $\mathcal{F}$. So whenever we have such Λ, it counts towards both $S(\mathcal{F}_0) + S(\mathcal{F}_1)$ and $S(\mathcal{F})$.
- If $\Lambda \subseteq \mathcal{X}_0$ is shattered by both $\mathcal{F}_0$ and $\mathcal{F}_1$, then Λ is shattered by $\mathcal{F}$ trivially. In addition, $\Lambda \cup \{x\}$ is also shattered by $\mathcal{F}$. This is because for labelings with $x \mapsto 0$, we use the functions in $\mathcal{F}_0 \subseteq \mathcal{F}$ to shattered $\Lambda \cup \{x\}$, whereas for labelings with $x \mapsto 1$, we use functions in $\mathcal{F}_1 \subseteq \mathcal{F}$.

Combining these two cases, we find that every subset $\Lambda \subseteq \mathcal{X}_0$ counted towards $S(\mathcal{F}_0) + S(\mathcal{F}_1)$ corresponds to a distinct subset to be shattered by $\mathcal{F}$, which implies $S(\mathcal{F}) \ge S(\mathcal{F}_0) + S(\mathcal{F}_1)$. □

Corollary 10.12 (Sauer-Shelah)**.** If $|\mathcal{X}| = m$ and $\mathsf{VC}(\mathcal{F}) = d$, then $|\mathcal{F}| \le \sum_{j=0}^{d} \binom{m}{j} \le (\frac{em}{d})^d$.

Proof. By Pajor's Lemma,

$$\begin{aligned} |\mathcal{F}| &\le [\text{number of shattered subsets of } \mathcal{X}] \\ &\le [\text{number of subsets of size at most } \mathsf{VC}(\mathcal{F}) = d] = \sum_{j=0}^{d} \binom{m}{j}. \end{aligned}$$

The second inequality is elementary. □

Proof of Theorem 10.10. By packing number upper bound for covering number,

$$N(\mathcal{F}, \|\cdot\|_{\mathbb{L}^2(\mathbb{P})}, \varepsilon) \le D(\mathcal{F}, \|\cdot\|_{\mathbb{L}^2(\mathbb{P})}, \varepsilon).$$

Let $\mathcal{D}$ be an ε-packing with $|\mathcal{D}| = D(\mathcal{F}, \|\cdot\|_{\mathbb{L}^2(\mathbb{P})}, \varepsilon)$. Then for $f, g \in \mathcal{D}$ with $f \ne g$,

$$\varepsilon^2 < \|f - g\|^2_{\mathbb{L}^2(\mathbb{P})} = \mathbb{E}_{\mathbb{P}}(f(X) - g(X))^2.$$

The idea is to apply a probabilistic method. Sample m points $X_1, \dots, X_m \overset{iid}{\sim} \mathbb{P}$. By Hoeffding's inequality, since $|f(X) - g(X)| \le 1$,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}(f(X_i)-g(X_i))^2 - \mathbb{E}_{\mathbb{P}}(f(X)-g(X))^2\right| \ge \frac{\varepsilon^2}{2}\right] \le 2e^{-cm\varepsilon^4}$$

for some universal constant $c > 0$. By union bound,

$$\mathbb{P}\left(\exists f \ne g \in \mathscr{D} : \left|\frac{1}{m}\sum_{i=1}^{m}(f(X_i)-g(X_i))^2 - \mathbb{E}_{\mathbb{P}}(f(X)-g(X))^2\right| \ge \frac{\varepsilon^2}{2}\right) \le \frac{|\mathscr{D}|(|\mathscr{D}|-1)}{2}\cdot 2e^{-cm\varepsilon^4}.$$

Hence, for the complementary event,

$$\mathbb{P}\left(\forall f \ne g \in \mathscr{D} : \left|\frac{1}{m}\sum_{i=1}^{m}(f(X_i)-g(X_i))^2 - \mathbb{E}_{\mathbb{P}}(f(X)-g(X))^2\right| < \frac{\varepsilon^2}{2}\right) \ge 1 - |\mathscr{D}|^2 e^{-cm\varepsilon^4}.$$

We are to select sample size m for which the probability on the RHS is bounded away from 0, e.g., we can choose $m = C\varepsilon^{-4}\log|\mathscr{D}|$. This shows in particular that with nontrivial probability over $(X_1, \dots, X_m) \in \mathscr{X}$, every pair $f \ne g \in \mathscr{D}$ satisfies

$$\left|\frac{1}{m}\sum_{i=1}^{m}(f(X_i)-g(X_i))^2 - \mathbb{E}_{\mathbb{P}}(f(X)-g(X))^2\right| < \frac{\varepsilon^2}{2},$$

which, combined with $\|f-g\|_{\mathbb{L}^2(\mathbb{P})} \ge \varepsilon$, implies

$$\frac{1}{m}\sum_{i=1}^{m}(f(X_i)-g(X_i))^2 \ge \mathbb{E}_{\mathbb{P}}(f(X)-g(X))^2 - \left|\frac{1}{m}\sum_{i=1}^{m}(f(X_i)-g(X_i))^2 - \mathbb{E}_{\mathbb{P}}(f(X)-g(X))^2\right| \ge \varepsilon^2 - \frac{\varepsilon^2}{2} = \frac{\varepsilon^2}{2}.$$

This means that the functions $f \in \mathscr{D}$ restricted to this set of realized values $\{X_1, \dots, X_m\}$ are all distinct. Let d_m be the VC-dimension of $\mathscr{D}$ on $\{X_1, \dots, X_m\}$. By the previous corollary, we have

$$|\mathscr{D}| \le \left(\frac{em}{d_m}\right)^{d_m} = \left(\frac{C\varepsilon^{-4}\log|\mathscr{D}|}{d_m}\right)^{d_m}.$$

Apply $\frac{\log|\mathscr{D}|}{2d_m} = \log|\mathscr{D}|^{\frac{1}{2d_m}} \le |\mathscr{D}|^{\frac{1}{2d_m}}$:

$$|\mathscr{D}| \le \left(\frac{C\varepsilon^{-4}\log|\mathscr{D}|}{d_m}\right)^{d_m} \le (2C\varepsilon^{-4})^{d_m}\sqrt{|\mathscr{D}|} \Longrightarrow |\mathscr{D}| \le \left(\frac{C'}{\varepsilon}\right)^{C'd_m}$$

for some $C' > 0$ large enough, i.e., $C' \ge 8$ and $C'^{C'} \ge 4C^2$. Finally, note that

$$\begin{aligned} d_m &\le \text{VC dimension of } \mathscr{F} \text{ restricted to } \{X_1, \dots, X_m\} \\ &\le \text{VC dimension of } \mathscr{F} \text{ on } \mathscr{X} = \mathsf{VC}(\mathscr{F}). \end{aligned}$$

□

Theorem 10.13. Let $X_1, \dots, X_n \in \mathscr{X}$ be iid, $\mathscr{F}$ a class of functions $\mathscr{X} \mapsto \{0,1\}$. Then, for a universal constant $C > 0$,

$$\mathbb{E}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \mathbb{E}[f(X_i)])\right| \le C\sqrt{\frac{\mathsf{VC}(\mathscr{F})}{n}}.$$

Proof. Let $\bar{\mathscr{F}} = \{1 - f \mid f \in \mathscr{F}\}$. Then with universal constants $C, C', C'' > 0$,

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)] \right| &= \mathbb{E} \sup_{f \in \mathscr{F} \cup \bar{\mathscr{F}}} \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)] \\
&\le 2\mathbb{E}_{\varepsilon, X} \left[\sup_{f \in \mathscr{F} \cup \bar{\mathscr{F}}} \frac{1}{n} \sum_{i=1}^n f(X_i) \right] \quad \text{by Lemma 10.4} \\
&\le \frac{C}{\sqrt{n}} \mathbb{E}_X \int_0^1 \sqrt{\log N(\mathscr{F} \cup \bar{\mathscr{F}}, \|\cdot\|_{\mathbb{L}^2(\mathbb{P}_n)}, \varepsilon)} \, \mathrm{d}\varepsilon \quad \text{by Eq (10.1)} \\
&\le \frac{C}{\sqrt{n}} \mathbb{E}_X \int_0^1 \sqrt{\log(2N(\mathscr{F}, \|\cdot\|_{\mathbb{L}^2(\mathbb{P}_n)}, \varepsilon)} \, \mathrm{d}\varepsilon \\
&\le \frac{C}{\sqrt{n}} \mathbb{E}_X \int_0^1 \sqrt{C' \cdot \mathsf{VC}(\mathscr{F}) \log \frac{C'}{\varepsilon}} \, \mathrm{d}\varepsilon \quad \text{by Theorem 10.10} \\
&\le C'' \sqrt{\frac{\mathsf{VC}(\mathscr{F})}{n}} \qquad \square
\end{aligned}$$

Example 10.14 (Excess risk in classification)**.** $\{(X_i, Y_i)\}_{i=1}^n$ are iid and $Y_i \in \{0, 1\}$.

- Optimal estimator in $\mathscr{F}$: $f_* = \arg\min_{f \in \mathscr{F}} R(f) = \mathbb{P}(f(X_i) \neq Y_i)$.
- Empirical risk minimizer: $\hat{f} = \arg\min_{f \in \mathscr{F}} R_n(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{f(X_i) \neq Y_i\}}$.

How large is the *excess risk* $R(\hat{f}) - R(f_*)$?

$$\begin{aligned}
R(\hat{f}) &\le R_n(\hat{f}) + \sup_{f \in \mathscr{F}} |R_n(f) - R(f)| \\
&\le R_n(f_*) + \sup_{f \in \mathscr{F}} |R_n(f) - R(f)| \\
&\le R(f_*) + 2 \sup_{f \in \mathscr{F}} |R_n(f) - R(f)|
\end{aligned}$$

Let $\mathscr{L} = \{\ell(x, y) = \mathbb{1}_{\{f(x) = y\}} \mid f \in \mathscr{F}\}$. Then with a universal constant $C > 0$,

$$\mathbb{E} \sup_{f \in \mathscr{F}} |R_n(f) - R(f)| = \mathbb{E} \sup_{\ell \in \mathscr{L}} \left| \frac{1}{n} \sum_{i=1}^n \ell(X_i, Y_i) - \mathbb{E}\ell(X_i, Y_i) \right| \le \mathbb{E} \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\mathscr{L}, \|\cdot\|_{\mathbb{L}^2(\mathbb{P}_n)}, \varepsilon)} \, \mathrm{d}\varepsilon \,.$$

If $\mathscr{N}$ is an ε-net of $\mathscr{F}$, i.e., $\forall f \in \mathscr{F}$, $\exists \pi(f) \in \mathscr{N}$ such that

$$\frac{1}{n} \sum_{i=1}^n (f(X_i) - \pi(f)(X_i))^2 = \sum_{i=1}^n \mathbb{1}\{f(X_i) \neq \pi(f)(X_i)\} \le \varepsilon^2 \,,$$

then $\{\mathbb{1}\{g(x) \neq y\} : g \in \mathscr{N}\}$ is an ε-net of $\mathscr{L}$ because

$$\frac{1}{n} \sum_{i=1}^n (\mathbb{1}\{f(X_i) \neq Y_i\} - \mathbb{1}\{\pi(f)(X_i) \neq Y_i\})^2 = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{f(X_i) \neq \pi(f)(X_i)\} \le \varepsilon^2 \,.$$

By Theorem 10.13, with universal constants $C', C'' > 0$,

$$\begin{aligned}
&\implies \mathbb{E} \sup_{f \in \mathscr{F}} |R_n(f) - R(f)| \le \mathbb{E} \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\mathscr{L}, \|\cdot\|_{\mathbb{L}^2(\mathbb{P}_n)}, \varepsilon)} \, \mathrm{d}\varepsilon \le C' \sqrt{\frac{\mathsf{VC}(\mathscr{F})}{n}} \\
&\implies \mathbb{E}[R(\hat{f}) - R(f_*)] \le C'' \sqrt{\frac{\mathsf{VC}(\mathscr{F})}{n}} \,.
\end{aligned}$$

11 Gaussian Processes, Gaussian Comparison Inequalities

Readings: §6.1-6.2 in [vH14], §7.1-7.4 in [Ver18].

Definition 11.1. $\{X_t\}_{t\in T}$ is a Gaussian process if for any finite subset $T_0 \subseteq T$, $\{X_t\}_{t\in T_0}$ has multivariate Gaussian law.

We will assume throughout that $\{X_t\}$ has mean-zero. Then $\Sigma(\cdot,\cdot)$ is sufficient to specify the law of $\{X_t\}_{t\in T_0}$ for any finite $T_0 \subseteq T$.

Goal: upper and lower bounds of $\sup_{t\in T} X_t$ using additional Gaussian comparison inequalities.

11.1 Gaussian Comparison Inequalities

Theorem 11.2 (Sudakov-Fernique)**.** If $\{X_t\}_{t\in T}$ and $\{Y_t\}_{t\in T}$ are both mean-zero, separable Gaussian processes, such that $\mathbb{E}(X_t - X_s)^2 \le \mathbb{E}(Y_t - Y_s)^2$ for all $s,t \in T$. Then

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t.$$

Theorem 11.3 (Slepian)**.** If, in addition, $\mathbb{E}X_t^2 = \mathbb{E}Y_t^2$ for all $t \in T$, then $\forall \tau \in \mathbb{R}$,

$$\mathbb{P}\left(\sup_{t\in T} X_t \ge \tau\right) \le \mathbb{P}\left(\sup_{t\in T} Y_t \ge \tau\right).$$

Example 11.4. Let $X \in \mathbb{R}^{n\times m}$ be iid standard Gaussian entries. We want to bound

$$\mathbb{E}\left[||X||_{\mathrm{op}}\right] = \mathbb{E}\left[\sup_{t\in\mathbb{S}^{n-1}, u\in\mathbb{S}^{m-1}} t^\top X u\right].$$

Denote $X_{t,u} := t^\top X u$. We compare it with $Y_{t,u} := t^\top g + u^\top h$, where $g \sim \mathcal{N}(0, I_n)$ and $h \sim \mathcal{N}(0, I_m)$ are independent.

- For $\{X_{t,u}\}$,

$$\begin{aligned}
\mathbb{E}(X_{t,u} - X_{s,v})^2 &= \mathbb{E}(t^\top X u - s^\top X v)^2 \\
&= \mathbb{E}\left(\sum_{i,j} X_{ij}(t_i u_j - s_i v_j)\right)^2 \\
&= \sum_{i,j}(t_i u_j - s_i v_j)^2 \quad X_{ij}\text{'s are independent} \\
&= ||tu^\top - sv^\top||_F^2 = 2 - 2(s^\top t)(u^\top v)
\end{aligned}$$

- For $\{Y_{t,u}\}$,

$$\begin{aligned}
\mathbb{E}(Y_{t,u} - Y_{s,v})^2 &= \mathbb{E}((t-s)^\top g + (u-v)^\top h)^2 \\
&= \mathbb{E}((t-s)^\top g)^2 + \mathbb{E}((u-v)^\top h)^2 \\
&= ||t-s||^2 + ||u-v||^2 \\
&= 4 - 2s^\top t - 2u^\top v
\end{aligned}$$

When we take the difference, we have

$$\mathbb{E}(Y_{t,u} - Y_{s,v})^2 - \mathbb{E}(X_{t,u} - X_{s,v})^2 = 2(1 - u^\top v)(1 - s^\top t) \ge 0.$$

Hence, by Sudakov-Fernique inequality 11.2,

$$\begin{aligned}
\mathbb{E}||X||_{\text{op}} &= \mathbb{E}\left[\sup_{t\in\mathbb{S}^{n-1},u\in\mathbb{S}^{m-1}} t^\top Xu\right] \\
&\le \mathbb{E}\left[\sup_{t\in\mathbb{S}^{n-1},u\in\mathbb{S}^{m-1}} Y_{t,u}\right] \\
&= \mathbb{E}\left[\sup_{t\in\mathbb{S}^{n-1}} t^\top g\right] + \mathbb{E}\left[\sup_{u\in\mathbb{S}^{m-1}} u^\top h\right] \\
&= \mathbb{E}[||g||_2] + \mathbb{E}[||h||_2] \le \sqrt{\mathbb{E}||g||_2^2} + \sqrt{\mathbb{E}||h||_2^2} = \sqrt{n} + \sqrt{m}
\end{aligned}$$

We know from lecture 8: if X has iid 1-subgaussian entries, then the expected operator norm $\mathbb{E}||X||_{\text{op}} \lesssim \sqrt{n} + \sqrt{m}$.

To prove Sudakov-Fernique, we need the following lemmas on Gaussian integral by parts:

Lemma 11.5 (Stein)**.** If $X \sim \mathcal{N}(0,\Sigma)$ in $\mathbb{R}^n$, $f : \mathbb{R}^n \mapsto \mathbb{R}$ continuously differentiable, $\mathbb{E}|\partial_i f| < \infty$, and $\mathbb{E}|X_i f(X)| < \infty$ for any $i \in \{1,\dots,n\}$, then

$$\mathbb{E}[Xf(X)] = \Sigma\mathbb{E}[\nabla f(X)].$$

Note that when $n = 1$ and $\Sigma = 1$, we have the 1D Stein's lemma $\mathbb{E}Xf(X) = \mathbb{E}f'(X)$.

Proof. For $n = 1$. Suppose f is compactly supported. $X \sim \mathcal{N}(0,1)$.

$$\mathbb{E}[Xf(X)] = \int_{-\infty}^{\infty} xf(x)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathrm{d}x = \int_{-\infty}^{\infty} f'(x)\mathrm{d}x = \mathbb{E}[f'(X)].$$

For general f, we would take a (countable) sequence of compactly supported $f^{(k)}$ that converges to f pointwise. Then we let $k \to \infty$ and apply DCT.

For general n and f, we write $X = \Sigma^{\frac{1}{2}}Z$, where $Z \sim \mathcal{N}(0, I_n)$.

$$\begin{aligned}
\mathbb{E}[Xf(X)] &= \mathbb{E}\left[\Sigma^{\frac{1}{2}}Zf(\Sigma^{\frac{1}{2}}Z)\right] \\
&= \Sigma^{\frac{1}{2}}\left(\mathbb{E}[Z_i f(\Sigma^{\frac{1}{2}}Z)]\right)_{i=1}^n \\
&= \Sigma^{\frac{1}{2}}\left(\mathbb{E}[\partial_{Z_i} f(\Sigma^{\frac{1}{2}}Z)\right)_{i=1}^n \quad \text{by case for } n = 1 \\
&= \Sigma^{\frac{1}{2}}\left(\mathbb{E}[e_i^\top\Sigma^{\frac{1}{2}}\nabla f(\Sigma^{\frac{1}{2}}Z)]\right)_{i=1}^n \\
&= \Sigma\mathbb{E}[\nabla f(X)].
\end{aligned}$$

□

Lemma 11.6 (Gaussian Interpolation)**.** Let $X \sim \mathcal{N}(0,\Sigma^X)$ and $Y \in \mathcal{N}(0,\Sigma^Y)$ are independent random vectors in $\mathbb{R}^n$. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be twice continuously differentiable. Define

$$Z(u) := \sqrt{u}X + \sqrt{1-u}Y, \quad \forall u \in [0,1].$$

Then

$$\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}[f(Z(u))] = \frac{1}{2}\sum_{1\le i,j\le n}(\Sigma_{ij}^X - \Sigma_{ij}^Y)\mathbb{E}[\partial_{ij}^2 f(Z(u))].$$

Proof.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}[f(Z(u))] &= \sum_{i=1}^{n}\mathbb{E}\left[\partial_i f(Z(u))\frac{\mathrm{d}}{\mathrm{d}u}Z(u)\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[\partial_i f(Z(u))\left(\frac{1}{2\sqrt{u}}X_i - \frac{1}{2\sqrt{1-u}}Y_i\right)\right] \\
&= \sum_{i=1}^{n}\frac{1}{2\sqrt{u}}e_i^\top\mathbb{E}[X\partial_i f(Z(u))] - \sum_{i=1}^{n}\frac{1}{2\sqrt{1-u}}e_i^\top\mathbb{E}[Y\partial_i f(Z(u))] \\
&= \sum_{i=1}^{n}\frac{1}{2\sqrt{u}}e_i^\top\Sigma^X\mathbb{E}[\nabla_X\partial_i f(Z(u))] - \sum_{i=1}^{n}\frac{1}{2\sqrt{1-u}}e_i^\top\Sigma^Y\mathbb{E}[\nabla_Y\partial_i f(Z(u))] \quad \text{by Lemma 11.5} \\
&= \sum_{i=1}^{n}\frac{1}{2}e_i^\top\Sigma^X\left(\mathbb{E}[\partial_{ij}^2 f(Z(u))]\right)_{j=1}^{n} - \sum_{i=1}^{n}\frac{1}{2}e_i^\top\Sigma^Y\left(\mathbb{E}[\partial_{ij}^2 f(Z(u))]\right)_{j=1}^{n} \\
&= \frac{1}{2}\sum_{1\le i,j\le n}(\Sigma_{ij}^X - \Sigma_{ij}^Y)\mathbb{E}[\partial_{ij}^2 f(Z(u))].
\end{aligned}$$

□

Proof of Theorem 11.2. Suppose first that T is finite and $|T| = n$. Identify $\{X_t\}_{t\in T}$ as $X_1, \dots, X_n$ and $\{Y_t\}_{t\in T}$ as $Y_1, \dots, Y_n$. Fix any $\beta > 0$ and consider $f_\beta : \mathbb{R}^n \mapsto \mathbb{R}$ defined as:

$$f_\beta(z) = \frac{1}{\beta}\log\left(\sum_{i=1}^{n}e^{\beta z_i}\right).$$

By definition, we can observe that as $\beta \uparrow +\infty$, it picks out the maximum entry of z. Here,

$$\partial_i f_\beta(z) = \frac{e^{\beta z_i}}{\sum_{k=1}^{n}e^{\beta z_k}} := p_i(z), \quad \partial_{ij}^2 f_\beta(z) = \frac{\beta e^{\beta z_i}}{\sum_{k=1}^{n}e^{\beta z_k}}\delta_{ij} - \frac{\beta e^{\beta z_i}e^{\beta z_j}}{\left(\sum_{k=1}^{n}e^{\beta z_k}\right)^2} = \beta p_i(z)\delta_{ij} - \beta p_i(z)p_j(z).$$

By Lemma 11.6,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}[f_\beta(Z(u))] &= \frac{1}{2}\sum_{1\le i,j\le n}(\Sigma_{ij}^X - \Sigma_{ij}^Y)\mathbb{E}[\partial_{ij}^2 f_\beta(Z(u))] \\
&= \frac{\beta}{2}\sum_{i=1}^{n}(\Sigma_{ii}^X - \Sigma_{ii}^Y)\mathbb{E}[p_i(Z(u))(1 - p_i(Z(u)))] \\
&\quad - \frac{\beta}{2}\sum_{i\ne j}(\Sigma_{ij}^X - \Sigma_{ij}^Y)\mathbb{E}[p_i(Z(u))p_j(Z(u))] \\
&= \frac{\beta}{2}\sum_{i\ne j}(\Sigma_{ii}^X - \Sigma_{ii}^Y - \Sigma_{ij}^X + \Sigma_{ij}^Y)\mathbb{E}[p_i(Z(u))p_j(Z(u))] \quad \text{since } 1 - p_i = \sum_{j\ne i}p_j \\
&= \frac{\beta}{4}\sum_{i\ne j}(\Sigma_{ii}^X + \Sigma_{jj}^X - \Sigma_{ii}^Y - \Sigma_{ii}^Y - 2\Sigma_{ij}^X + 2\Sigma_{ij}^Y)\mathbb{E}[p_i(Z(u))p_j(Z(u))] \quad \text{by symmetrization} \\
&= \frac{\beta}{4}\sum_{i\ne j}\left(\underbrace{\mathbb{E}[(X_i - X_j)^2] - \mathbb{E}[(Y_i - Y_j)^2]}_{\le 0}\right)\mathbb{E}[\underbrace{p_i(Z(u))p_j(Z(u))}_{\ge 0}] \le 0, \quad \forall u \in (0,1),
\end{aligned}$$

which implies

$$\mathbb{E}[f_\beta(X)] = \mathbb{E}[f_\beta(Z(1))] \le \mathbb{E}[f_\beta(Z(0))] = \mathbb{E}[f_\beta(Y)]. \tag{11.1}$$

Note that by definition of f_β, we have:

$$f_\beta(z) = \frac{1}{\beta}\log\left(\sum_{i=1}^n e^{\beta z_i}\right) \ge \frac{1}{\beta}\log e^{\beta \max_{1\le i\le n} z_i} = \max_{1\le i\le n} z_i$$

$$f_\beta(z) \le \frac{1}{\beta}\log(n\cdot e^{\beta \max_{1\le i\le n} z_i} = \frac{\log n}{\beta} + \max_{1\le i\le n} z_i$$

Take $\beta \uparrow +\infty$ in Eq 11.1, we have

$$\mathbb{E}\left[\max_{1\le i\le n} X_i\right] \le \mathbb{E}\left[\max_{1\le i\le n} Y_i\right]. \tag{11.2}$$

For $|T| = \infty$. By separability, $\exists t_1, t_2, \ldots \in T$ such that

$$\sup_{t\in T} X_t = \lim_{k\to\infty} \max_{t\in\{t_1,\ldots,t_k\}} X_t,$$

and similarly for Y_t. Apply Eq 11.2 for each fixed k and take $k \to \infty$. □

Proof of Theorem 11.3. Suppose first that $|T| = n$ is finite and identify $\{X_t\}_{t\in T}$ as $X_1, \ldots, X_n$ and $\{Y_t\}_{t\in T}$ as $Y_1, \ldots, Y_n$. Approximate

$$\mathbb{1}_{\max_{1\le i\le n} Z_i < k} = \prod_{i=1}^n \mathbb{1}_{Z_i<k} \quad \text{by} \quad f_\beta(Z) = \prod_{i=1}^n h_\beta(Z_i),$$

where $h_\beta : \mathbb{R} \mapsto [0,1]$ is a decreasing smooth approximation to $\mathbb{1}_{Z_i<k}$ and $h_\beta(x) \to \mathbb{1}_{x<k}$ as $\beta \to \infty$. Here,

$$\partial^2_{ij} f_\beta(z) = h'_\beta(z_i)h'_\beta(z_j)\prod_{k\notin\{i,j\}} h_\beta(z_k) \ge 0, \quad \forall i \ne j.$$

Also, by $\mathbb{E}[(X_i - X_j)^2] \le \mathbb{E}[(Y_i - Y_j)^2]$ and $\mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2]$, we have

$$\mathbb{E}[X_iX_j] = \Sigma^X_{ij} \ge \Sigma^Y_{ij} = \mathbb{E}[Y_iY_j].$$

Now we return to the previous proof, where

$$\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}[f_\beta(Z(u))] = \frac{1}{2}\sum_{1\le i,j\le n}(\Sigma^X_{ij} - \Sigma^Y_{ij})\mathbb{E}[\partial^2_{ij} f_\beta(Z(u))],$$

For $1 \le i, j \le n$, we know that $\Sigma^X_{ij} - \Sigma^Y_{ij} \begin{cases} = 0 & i = j \\ \ge 0 & i \ne j \end{cases}$, so

$$\mathbb{E}[f_\beta(X)] = \mathbb{E}[f_\beta(Z(1))] \ge \mathbb{E}[f_\beta(Z(0))] = \mathbb{E}[f_\beta(Y)].$$

Let $\beta \uparrow +\infty$, we conclude that $\forall \tau \in \mathbb{R}$,

$$\mathbb{P}(\max_{1\le i\le n} X_i < \tau) \ge \mathbb{P}(\max_{1\le i\le n} Y_i < \tau) \Longleftrightarrow \mathbb{P}(\max_{1\le i\le n} X_i \ge \tau) \le \mathbb{P}(\max_{1\le i\le n} Y_i \ge \tau).$$

□

11.2 Gaussian Process Lower Bounds

Definition 11.7. The *canonical metric* (or natural distance) associated to a mean-zero Gaussian process $\{X_t\}_{t\in T}$ is

$$d(t,s) = \left(\mathbb{E}[(X_t - X_s)^2]\right)^{\frac{1}{2}}.$$

Remarks:

1. $\mathbb{E}[\exp(\lambda(X_t - X_s))] = \exp(\frac{\lambda^2}{2}\mathbb{E}[(X_t - X_s)^2]) = \exp(\frac{\lambda^2}{2}d(t,s)^2)$, so $\{X_t\}_{t\in T}$ is always 1-subgaussian in this metric.

2. Consider $X_t = g^\top t$ for $g \sim \mathcal{N}(0, I_n)$, where $T \subseteq \mathbb{R}^n$. Then

$$d(t,s)^2 = \mathbb{E}[(X_t - X_s)^2] = \mathbb{E}[(g^\top(t-s))^2] = ||t-s||_2^2.$$

   So in this example, $d(t,s) = ||t-s||_2$ is the Euclidean metric.

3. Let $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}$ and $\mathcal{F}$ be a class of functions with $\mathbb{E}_{X\sim\mathbb{P}}[f(X)] = 0, \forall f \in \mathcal{F}$. Define renormalized empirical process

$$Z_f = \frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i).$$

   As $n \to \infty$, the finite-dimensional marginals of $\{Z_f\}_{f\in\mathcal{F}}$ will converge in distribution to a Gaussian limit, with $\Sigma(f,g) = \mathbb{E}_{X\sim\mathbb{P}}[f(X)g(X)]$. Then

$$d(f,g)^2 = \mathbb{E}[(Z_f - Z_g)^2] = \mathbb{E}_{X\sim\mathbb{P}}[(f(X) - g(X))^2],$$

   i.e., $d(\cdot,\cdot)$ is the $L^2(\mathbb{P})$ metric on $\mathcal{F}$. By Dudley Theorem + 1.,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \lesssim \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\mathrm{d}\varepsilon.$$

Theorem 11.8 (Sudakov)**.** Let $\{X_t\}_{t\in T}$ be a mean-zero, separable Gaussian process, $d(t,s)$ its canonical metric. Then for a universal constant $c > 0$,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \geq c \sup_{\varepsilon>0}\ \varepsilon\sqrt{\log N(T,d,\varepsilon)}.$$

Proof. Let $\mathcal{D}$ be an ε-packing of (T,d). Then $\mathbb{E}[\sup_{t\in T} X_t] \geq \mathbb{E}[\sup_{t\in\mathcal{D}} X_t]$. Here, $\forall s \neq t \in \mathcal{D}$, we have

$$\mathbb{E}[(X_t - X_s)^2] = d(t,s)^2 \geq \varepsilon^2.$$

Consider $\{Y_t\}_{t\in\mathcal{D}}$ that consists of iid $\mathcal{N}(0, \varepsilon^2/2)$. Then $\mathbb{E}[(Y_t - Y_s)^2] = \varepsilon^2\delta_{st}$. By Theorem 11.2,

$$\mathbb{E}\left[\sup_{t\in\mathcal{D}} X_t\right] \geq \mathbb{E}\left[\sup_{t\in\mathcal{D}} Y_t\right] \geq c\varepsilon\sqrt{\log|\mathcal{D}|}.$$

Take $\mathcal{D}$ be the maximal ε-packing, we have

$$|\mathcal{D}| \geq D(T,d,\varepsilon) \geq N(T,d,\varepsilon).$$

Thus, we have $\mathbb{E}[\sup_{t\in T} X_t] \geq c\varepsilon\sqrt{\log N(T,d,\varepsilon)}$. We finish by taking supremum over $\varepsilon > 0$. □

Example 11.9. Consider $X_t = g^\top t$, where $g \sim \mathcal{N}(0, I_n)$, $T \subseteq \mathbb{R}^n$, and $d(t,s) = ||t-s||_2$.

1. Let $T = B^n$, the unit ball in $\mathbb{R}^n$. Recall that $(1/\varepsilon)^n \le N(t, d, \varepsilon) \le (3/\varepsilon)^n$, we have

$$\int_0^\infty \sqrt{\log N(t, d, \varepsilon)} \mathrm{d}\varepsilon \asymp \int_0^1 \sqrt{n \log \frac{1}{\varepsilon}} \mathrm{d}\varepsilon \asymp \sqrt{n}.$$

   Also,

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(T, d, \varepsilon)} \asymp \sup_{\varepsilon \in (0,1)} \varepsilon \sqrt{n \log \frac{1}{\varepsilon}} \asymp \sqrt{n}.$$

2. Consider $T = \left\{ \frac{e_k}{\sqrt{1 + \log k}} \right\}_{k=1}^n$. Here $M(n) = \int_0^\infty \sqrt{\log N(T, d, \varepsilon)} \mathrm{d}\varepsilon \to \infty$ as $n \to \infty$ (HW9). Heuristically,

$$N(T, d, \varepsilon) \approx |\{k : 1/\sqrt{\log k} \ge \varepsilon\}| = |\{k : \exp(1/\varepsilon^2) \ge k\}| \approx \exp(1/\varepsilon^2),$$

   so $\varepsilon \sqrt{\log N(T, d\varepsilon)} \approx \varepsilon \cdot \sqrt{1/\varepsilon^2} = 1$. Also by HW9, $\sup_{t \in T} X_t \lesssim 1$. Hence, Sudakov is tight here but Dudley is not.

Definition 11.10. A Gaussian process $\{X_t\}_{t \in T}$ is stationary is there is a group G acting on T s.t.

- $\forall g \in G$, $\{X_t\}_{t \in T} \overset{\text{law}}{=} \{X_{g \cdot t}\}_{t \in T}$.
- $\forall s, t \in T$, $\exists g \in G$ such that $g \cdot s = t$.

Example 11.11. Let $T = \{u \in \mathbb{R}^n \mid ||u||_2 = 1\}$ and $G = \{\text{all rotations of } T\}$. Then G is stationary if $\Sigma(t, s)$ depends only on spherical distance between t and s.

Theorem 11.12 (Fernique)**.** Let $\{X_t\}_{t \in T}$ be a stationary, mean-zero, and separable Gaussian process. Then for a universal constant $c > 0$,

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \ge c \int_0^\infty \sqrt{\log N(T, d, \varepsilon)} \mathrm{d}\varepsilon.$$

12 Generic Chaining, Majorizing Measures Theorem

Readings: §6.3-6.4 in [vH14], §8.5-8.6 in [Ver18].

Recap: Let $\{X_t\}_{t\in T}$ be a mean-zero, separable Gaussian process, and $\mathrm{d}(t,s) = \sqrt{\mathbb{E}(X_t - X_s)^2}$ be the canonical metric.

- If $\{X_t\}_{t\in T}$ is stationary, then by Fernique's theorem, $\mathbb{E}\sup_{t\in T} X_t \gtrsim \int_0^\infty \sqrt{\log N(T,\mathrm{d},\varepsilon)}\,\mathrm{d}\varepsilon$. Combined with Dudley's inequality, $\mathbb{E}\sup_{t\in T} X_t \asymp \int_0^\infty \sqrt{\log N(T,\mathrm{d},\varepsilon)}\,\mathrm{d}\varepsilon$.
- In general, we have

$$\sup_{\varepsilon>0} \varepsilon\sqrt{\log N(T,\mathrm{d},\varepsilon)} \overset{\text{Sudakov}}{\lesssim} \mathbb{E}\sup_{t\in T} X_t \overset{\text{Dudley}}{\lesssim} \int_0^\infty \sqrt{\log N(T,\mathrm{d},\varepsilon)}\,\mathrm{d}\varepsilon\,.$$

- An example where bounds do not match is $X_t = g^\top t, g \sim \mathcal{N}(0,I)$, and $t \in T = \{\frac{e_k}{\sqrt{1+\log k}}\}_{k=1}^n$.

Goal: tighter characterization of $\mathbb{E}\sup_{t\in T} X_t$ when T is "inhomogeneous", i.e., $\{X_t\}$ is non-stationary.

Definition 12.1 (Labeled net)**.** Fix $\alpha \in (0,1)$. A *labeled net* $(\mathscr{A},\ell)$ of (T,d) consists of

- An increasing sequence of partitions $\mathscr{A} = \{\mathscr{A}_k\}_{k\in\mathbb{Z}}$ such that $\mathscr{A}_\kappa = \{T\}$ for some $\kappa \in \mathbb{Z}$, and $\max_{A\in\mathscr{A}_k} \mathsf{diam}(A) \le 2\cdot\alpha^k, \forall k \in \mathbb{Z}$.
- An ordering $1,2,3,\ldots$ of the children of each $A \in \mathscr{A}$. Letting $S_1,\ldots,S_m$ be the children of A, we denote this ordering by $\ell(S_1) = 1,\ldots,\ell(S_m) = m$. See Fig 1.

Figure 1: Example of an ordering of the children of each $A \in \mathscr{A}$

Additionally, let $A_k(t)$ be the element of $\mathscr{A}_k$ containing t, and

$$\gamma(T) = \inf_{\text{labeled nets } (\mathscr{A},\ell)} \sup_{t\in T} \sum_{k\in\mathbb{Z}} \alpha^k \sqrt{\log \ell(A_k(t))}\,.$$

Theorem 12.2 (Generic chaining upper bound)**.** If $\{X_t\}_{t\in T}$ is separable, mean-zero, 1-subgaussian on (T,d), then $\forall\alpha \in (0,1)$, $\exists C > 0$ such that $\mathbb{E}\sup_{t\in T} X_t \le C\gamma(T)$.

Theorem 12.3 (Majorizing measures lower bound)**.** If, in addition, $\{X_t\}_{t\in T}$ is a Gaussian process and d is the canonical metric, then for some constants $c,\alpha > 0$, $\mathbb{E}\sup_{t\in T} X_t \ge c\gamma(T)$.

Corollary 12.4 (Talagrand's comparison inequality)**.** Let $\{X_t\}_{t\in T}, \{Y_t\}_{t\in T}$ be separable mean-zero processes. $\{Y_t\}_{t\in T}$ is Gaussian process with canonical metric d, $\{X_t\}_{t\in T}$ is subgaussian w.r.t. d, i.e.,

$$\log \mathbb{E}e^{\lambda(X_t - X_s)} \le \frac{\lambda^2}{2}\mathbb{E}(Y_t - Y_s)^2 \qquad \forall s,t \in T, \lambda \ge 0\,.$$

Then for a universal constant $C > 0$,

$$\mathbb{E}\sup_{t\in T} X_t \le C\,\mathbb{E}\sup_{t\in T} Y_t\,.$$

Proof. Follows immediately from the previous two theorems. Note that This extends Sudakov-Fernique's theorem to subgaussian processes, up to constant factor. □

12.1 Deferred proof of Fernique's Theorem 11.12

Lemma 12.5. If $\{X_t\}_{t\in T}$ is separable Gaussian process, then

$$\sup_{t\in T} X_t - \mathbb{E}\sup_{t\in T} X_t \text{ is } \sup_{t\in T}\mathsf{Var}(X_t)\text{-subgaussian.}$$

Proof. By separability, it suffices to consider $|T| = n$. Write $X = \Sigma^{1/2}Z$, where $Z \sim \mathcal{N}(0, I)$. $\forall i \in [n]$, $X_i = e_i^\top \Sigma^{1/2} Z$ is Lipschitz in Z with constant $\|e_i^\top \Sigma^{1/2}\|_2 = \sqrt{\Sigma_{ii}} = \sqrt{\mathsf{Var}(X_i)}$. Therefore,

$$\sup_{1\le i\le n} X_i \text{ is } \sqrt{\sup_{1\le i\le n}\mathsf{Var}(X_i)}\text{-Lipschitz.}$$

Apply Theorem 7.10, we finish the proof. □

Theorem 12.6 (Super-Sudakov, Theorem 6.11 [vH14])**.** Let $\{X_t\}_{t\in T}$ be a separable Gaussian process and let $\mathscr{D}$ be an ε-packing of (T, d), then we have

$$\mathbb{E}\sup_{t\in T} X_t \ge c\varepsilon\sqrt{\log|\mathscr{D}|} + \min_{s\in\mathscr{D}} \mathbb{E}\sup_{t\in\mathbb{B}(s,\alpha\varepsilon)} X_t\,,$$

where c and $\alpha < \frac{1}{2}$ are universal constants and $\mathbb{B}(s,\varepsilon) = \{t \in T \mid \mathrm{d}(t,s) \le \varepsilon\}$.

Proof of Theorem 11.12. The idea is to apply chaining and packing to construct multi-scale lower bound. Fix any $t_0 \in T$. Let $\mathbb{B}(t,\varepsilon) = \{s \in T \mid \mathrm{d}(s,t) \le \varepsilon\}$ and $\mathbb{B}(\varepsilon) = \mathbb{B}(t_0,\varepsilon)$. Fix $\alpha \in (0, \frac{1}{2})$ and consider $G_k := \mathbb{E}\sup_{t\in\mathbb{B}(\alpha^k)} X_t$. Let $\mathscr{D}_k$ be a maximal α^{k+2}-packing of $\mathbb{B}(\alpha^{k+1})$. Then we observe that $\{\mathbb{B}(s,\alpha^{k+3}) \mid s \in \mathscr{D}_k\}$ are disjoint balls contained in $\mathbb{B}(\alpha^k)$. See Fig 2.

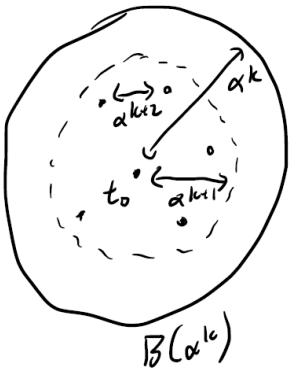

Figure 2: Illustration of our construction.

$$\begin{aligned}
G(k) = \mathbb{E}\sup_{t\in\mathbb{B}(\alpha^k)} X_t &\ge \mathbb{E}\sup_{s\in\mathscr{D}_k}\sup_{t\in\mathbb{B}(s,\alpha^{k+3})} X_t \\
&= \mathbb{E}\sup_{S\in\mathscr{D}_k} X_s + \sup_{t\in\mathbb{B}(s,\alpha^{k+3})}(X_t - X_s) \quad \text{define the latter as } Y_s \\
&= \mathbb{E}\sup_{s\in\mathscr{D}_k}(X_s + Y_s - \mathbb{E}[Y_s] + \mathbb{E}[Y_s])
\end{aligned}$$

By Super-Sudakov (Theorem 12.6) and $\mathbb{E}X_t \equiv 0$, we have $\mathbb{E}\sup_{s\in\mathscr{D}_k} X_s \ge c\alpha^{k+2}\sqrt{\log|\mathscr{D}_k|}$. Moreover, $\mathbb{E}Y_s = \mathbb{E}\sup_{t\in\mathbb{B}(s,\alpha^{k+3})} X_t = G(k+3)$. Finally, by zero-mean assumption, we have

$$\mathsf{Var}(X_t - X_s) = \mathrm{d}(t,s)^2 \le (\alpha^{k+3})^2,$$

so $Y_s - \mathbb{E}Y_s$ is $(\alpha^{k+3})^2$-subgaussian by previous lemma, $\forall s \in \mathscr{D}_k$, which, by maximal inequality, implies

$$\mathbb{E} \sup_{s \in \mathscr{D}_s} |Y_s - \mathbb{E}[Y_s]| \le C\alpha^{k+3}\sqrt{\log |\mathscr{D}_k|}\,.$$

Combining the above, for $\alpha \in (0, \frac{1}{2})$ small enough and some $c' > 0$,

$$\begin{aligned} G(k) &\ge c\alpha^{k+2}\sqrt{\log |\mathscr{D}_k|} - C\alpha^{k+3}\sqrt{\log |\mathscr{D}_k|} + G(k+3) \\ &= (c - C\alpha)\alpha^{k+2}\sqrt{\log |\mathscr{D}_k|} + G(k+3) \\ &\ge c'\alpha^{k+2}\sqrt{\log N(\mathbb{B}(\alpha^{k+1}), \mathrm{d}, \alpha^{k+2})} + G(k+3)\,. \end{aligned}$$

Summing over $k, k+1, k+2$ and iterating this bound,

$$G(k) + G(k+1) + G(k+2) \ge \sum_{j \ge k} c'\alpha^{j+2}\sqrt{\log N(\mathbb{B}(\alpha^{j+1}), \mathrm{d}, \alpha^{j+2})}\,.$$

Pick $\kappa \in \mathbb{Z}$ such that $\alpha^{\kappa+2} \ge \mathsf{diam}(T)$. This κ exists because otherwise $\mathbb{E}\sup_{t\in T} X_t = \infty$ by Sudakov lower bound, and so the theorem is trivial. Then we observe that

$$G(\kappa), G(\kappa+1), G(\kappa+2) = \mathbb{E}\sup_{t \in T} X_t, \text{ and } \log N(\mathbb{B}(\alpha^j), \mathrm{d}, \alpha^{j+1}) = 0, \forall j \le \kappa\,.$$

Substituting into the above, we have

$$\mathbb{E}\sup_{t \in T} X_t \ge \frac{c'}{3}\sum_{j \in \mathbb{Z}} \alpha^{j+1}\sqrt{\log N(\mathbb{B}(\alpha^j), \mathrm{d}, \alpha^{j+1})}\,.$$

Finally, apply $\sqrt{a-b} \le \sqrt{a} - \sqrt{b}$ for $a \ge b \ge 0$ and

$$N(T, \mathrm{d}, \alpha^{k+1}) \le N(T, \mathrm{d}, \alpha^k) \cdot N(\mathbb{B}(\alpha^k), \mathrm{d}, \alpha^{k+1})\,,$$

Therefore,

$$\begin{aligned} \mathbb{E}\sup_{t \in T} X_t &\ge \frac{c'}{3}\sum_{j \in \mathbb{Z}} \alpha^{j+1}\sqrt{\log \frac{N(T, \mathrm{d}, \alpha^{j+1})}{N(T, \mathrm{d}, \alpha^j)}} \\ &= \frac{c'}{3}\sum_{j \in \mathbb{Z}} \alpha^{j+1}\sqrt{\log N(T, \mathrm{d}, \alpha^{j+1})} - \frac{c'}{3}\sum_{j \in \mathbb{Z}} \alpha^{j+1}\sqrt{\log N(T, \mathrm{d}, \alpha^j)} \\ &\ge \frac{c'(1-\alpha)}{3}\sum_{j \in \mathbb{Z}} \alpha^j\sqrt{\log N(T, \mathrm{d}, \alpha^j)} \\ &= \frac{c'\alpha}{3}\sum_{j \in \mathbb{Z}} (\alpha^{j-1} - \alpha^j)\sqrt{\log N(T, \mathrm{d}, \alpha^j)} \\ &\ge \frac{c'\alpha}{3}\int_0^\infty \sqrt{\log N(T, \mathrm{d}, \varepsilon)}\,\mathrm{d}\varepsilon\,. \end{aligned}$$

□

12.2 Proof of majorizing measures lower bound

Denote $\mathbb{B}(t,\varepsilon) = \{s \in T \mid \mathrm{d}(s,t) \le \varepsilon\}$ and $G(A) = \mathbb{E}\sup_{t\in A} X_t$.

Lemma 12.7. Let $\{X_t\}_{t\in T}$ be a separable, mean-zero Gaussian process, d be the canonical metric. For some universal constants $\alpha, c > 0$, the following holds:

Let $\mathscr{D} = \{t_1, \ldots, t_m\}$ be any ε-packing of T, $\forall \varepsilon > 0$. Order $t_1, \ldots, t_m$ such that

$$G(\mathbb{B}(t_1, \alpha\varepsilon)) \ge \ldots \ge G(\mathbb{B}(t_m, \alpha\varepsilon))\,,$$

then $\mathbb{E}\sup_{t\in T} \ge \max_{1\le i\le m}\left\{c\varepsilon\sqrt{\log i} + G(\mathbb{B}(t_i,\alpha\varepsilon))\right\}$.

Proof. Since $\mathscr{D} \subseteq T$, we have

$$\begin{aligned}\mathbb{E}\sup_{t\in T} X_t &\ge \mathbb{E}\max_{1\le i\le n}\sup_{t\in\mathbb{B}(t_i,\alpha\varepsilon)} X_t \\ &= \mathbb{E}\max_{1\le i\le m} X_{t_i} + \mathbb{E}\max_{1\le i\le m}\underbrace{\sup_{s\in\mathbb{B}(t_i,\alpha\varepsilon)}(X_s - X_{t_i})}_{Y_i} = \mathbb{E}\max_{1\le i\le m}(X_{t_i} + Y - \mathbb{E}Y_i + \mathbb{E}Y_i)\,.\end{aligned}$$

For the three terms, we observe that

- $\mathbb{E}\max_{1\le i\le m} X_{t_i} \ge c\varepsilon\sqrt{\log m}$ by lower bound in Sudakov' theorem (Theorem 11.8).
- $\mathbb{E}\min_{1\le i\le m}(Y_i - \mathbb{E}[Y_i]) \ge -C\alpha\varepsilon\sqrt{\log m}$ since Y_i is $\alpha^2\varepsilon^2$-subgaussian.
- $\min_{1\le i\le m}\mathbb{E}Y_i \le \min_{1\le i\le m} G(\mathbb{B}(t_i,\alpha\varepsilon))$ by given condition.

Therefore, for α sufficiently small,

$$\mathbb{E}\sup_{t\in T} X_t \ge c'\varepsilon\sqrt{\log m} + G(\mathbb{B}(t_m,\alpha\varepsilon))\,.$$

The same argument applies to $\mathscr{D} = \{t_1,\ldots,t_i\}$ for any $1 \le i \le m$. We finish by taking max over i. □

Proof of Theorem 12.3. If $N(T,\mathrm{d},\varepsilon) = \infty$ for any $\varepsilon > 0$, then $\mathbb{E}\sup_{t\in T} X_t = \infty$ by the Sudakov lower bound. If $N(T,\mathrm{d},\varepsilon) < \infty$, $\forall\varepsilon > 0$, then $\mathsf{diam}(T) < \infty$. Construct the following labeled net $(\mathscr{A}, \ell)$:

Let $\kappa \in \mathbb{Z}$ be the smallest integer such that $2\cdot\alpha^\kappa \ge \mathsf{diam}(T)$. Take $\mathscr{A}_k = \{T\}$ for all $k \le \kappa$ and $\ell(T) = 1$. Given $A \in \mathscr{A}_k$, partition A into $S_1, \ldots, S_m \in \mathscr{A}_{k+1}$ in the following manner (see Fig 3):

- Choose $t_1 \in A$ maximizing $G(A \cap \mathbb{B}(t_1, \alpha^{k+2}))$. Set $S_1 = A \cap \mathbb{B}(t_1, \alpha^{k+1})$ and $\ell(S_1) = 1$.
- Choose $t_2 \in A \setminus S_1$ maximizing $G(A \cap \mathbb{B}(t_2, \alpha^{k+2}))$. Set $S_2 = (A \setminus S_1) \cap \mathbb{B}(t_2, \alpha^{k+1})$ and $\ell(S_2) = 2$.
- Continue in the above manner ...

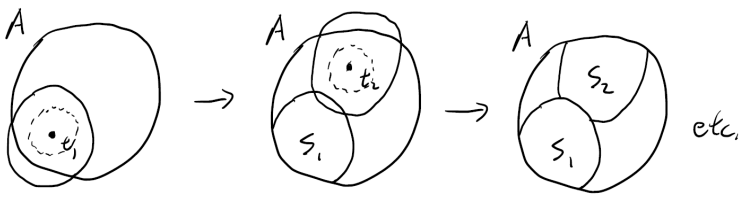

Figure 3: Illustration of our construction.

Here $\mathscr{D} = \{t_1, \ldots, t_m\}$ is an α^{k+1}-packing of A, so $m \le D(T, \mathrm{d}, \alpha^{k+1}) < \infty$ and

$$G(A \cap \mathbb{B}(t_1, \alpha^{k+2})) \ge \ldots \ge G(A \cap \mathbb{B}(t_m, \alpha^{k+2}))\,.$$

By the previous lemma, we have

$$G(A) \ge \max_{1 \le i \le m} \left\{ c\alpha^{k+1}\sqrt{\log \ell(S_i)} + G(A \cap \mathbb{B}(t_i, \alpha^{k+2})) \right\}\,.$$

We want to turn this relationship into a recursive bound. $\forall t \in S_i$, we have $A_k(t) = A$, $A_{k+1}(t) = S_i$,

$$\mathsf{diam}(A_{k+3}(t)) \le 2 \cdot \alpha^{k+3} < \alpha^{k+2}, \quad \forall \alpha \in (0, \frac{1}{2})\,,$$

which, by definition of t_i,

$$G(A_{k+3}(t)) \le G(A \cap \mathbb{B}(t, \alpha^{k+2})) \le G(A \cap \mathbb{B}(t_i, \alpha^{k+2}))\,.$$

Hence, $\forall k \in \mathbb{Z}$, $\forall t \in T$, we have

$$G(A_k(t)) \ge c\alpha^{k+1}\sqrt{\log \ell(A_{k+1}(t))} + G(A_{k+3}(t))\,.$$

Using the same idea as in the proof of the Fernique lower bound, $\forall t \in T$,

$$\begin{aligned}\mathbb{E}\sup_{t\in T} X_t &= \frac{1}{3}\left(G(A_{\kappa-2}(t)) + G(A_{\kappa-1}(t)) + G(A_\kappa(t))\right) \\ &\ge \frac{c}{3}\sum_{k=\kappa-1}^{\infty} \alpha^k \sqrt{\log \ell(A_k(t))} = \frac{c}{3}\sum_{k\in\mathbb{Z}} \alpha^k \sqrt{\log \ell(A_k(t))} \ge c'\gamma(T)\,.\end{aligned}$$

□

12.3 Proof of generic chaining upper bound

Proof of Theorem 12.2. Suppose first that $|T| < \infty$. Let $(\mathscr{A}, \ell)$ be any labeled net. Denote $\kappa, K \in \mathbb{Z}$ such that $\mathscr{A}_\kappa = \{T\}$ and $\mathscr{A}_K = \{\text{all single points of } T\}$. Take $\kappa \le k \le K$. Let $t_A \in A$ be an arbitrary point for each $A \in \mathscr{A}_k$ and $\pi_k(t) = t_{A_k(t)}$. Recall that $\forall t \in T$, $A_k(t)$ is the set in $\mathscr{A}_k$ to which t belongs and is unique, and then $t_{A_k(t)}$ is an arbitrary point from it. Further denote $t_0 = \pi_\kappa(t)$, which is an arbitrary point from T because $\mathscr{A}_\kappa = \{T\}$. See Fig 4.

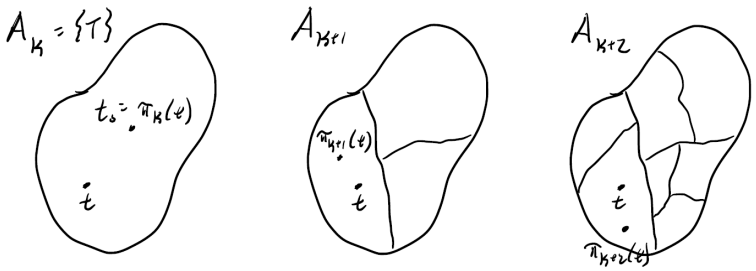

Figure 4: Illustration of our construction.

Idea: we use $\pi_k(t)$ to approach t. then $\pi_K(t) = t$ because $\{t\} \in \mathscr{A}_K$. To continue with this idea:

$$X_t - X_{t_0} = \sum_{k=\kappa+1}^{K} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \qquad \text{and} \qquad \mathrm{d}(\pi_k(t), \pi_{k-1}(t)) \le \mathsf{diam}(A_{k-1}(t)) \le 2\alpha^{k-1}\,.$$

Since $\{X_t\}_{t\in T}$ is 1-subgaussian w.r.t. d, we conclude that $X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$ is $(2\alpha^{k-1})^2$-subgaussian. Pick any labeling $U : \mathscr{A} \mapsto [1,\infty)$, $\forall x \geq 0$,

$$\mathbb{P}\left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \geq x\alpha^{k-1}\sqrt{\log U(A_k(t))}\right) \leq \exp\left(-\frac{x^2\alpha^{2k-2}\log U(A_k(t))}{2(2\alpha^{k-1})^2}\right) = U(A_k(t))^{-\frac{x^2}{8}}\,.$$

Define the event

$$\Omega = \left\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)} < x\alpha^{k-1}\sqrt{\log U(A_k(t))} \mid \forall k = \kappa+1,\ldots,K, t\in T\right\}\,.$$

Then $\mathbb{P}(\Omega^c) \leq \sum_{k=\kappa+1}^{K}\sum_{A\in\mathscr{A}_k} U(A)^{-\frac{x^2}{8}}$, and on Ω,

$$\sup_{t\in T}(X_t - X_{t_0}) \leq \frac{x}{\alpha}\sup_{t\in T}\sum_{k=\kappa+1}^{K}\alpha^k\sqrt{\log U(A_k(t))} \qquad \cdots(\star)\,.$$

Intuition: set $U(A) = \ell(A), \forall A$ to get $\gamma(T)$ in $(\star)$. However, this choice is too small, e.g., if $U \equiv 1$, then we don't have something small for the probabilistic bound on Ω^c because $1^{-\frac{x^2}{8}} = 1$. Instead, we note that for any sequence $u_{\kappa+1}, u_{\kappa+2}, \ldots$, setting $U_k = \prod_{j=\kappa+1}^{k} u_j$ and use $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, $U_\kappa = 1$:

$$\sum_{k=\kappa+1}^{\infty}\alpha^k\sqrt{\log U_k} \leq \sum_{k=\kappa+1}^{\infty}\alpha^k\sqrt{\log u_k} + \sum_{k=\kappa+2}^{\infty}\alpha^k\sqrt{\log U_{k-1}} = \sum_{k=\kappa+1}^{\infty}\alpha^k\sqrt{\log u_k} + \sum_{k=\kappa+1}^{\infty}\alpha^{k+1}\sqrt{\log U_k}\,,$$

which implies $\sum_{k=\kappa+1}^{\infty}\alpha^k\sqrt{\log U_k} \leq \frac{1}{1-\alpha}\sum_{k=\kappa+1}^{\infty}\alpha^k\sqrt{\log u_k}$. Now for each k, $\forall t \in A_k$, define

$$U(A_k) = \prod_{j=\kappa+1}^{k} 2\ell(A_j(t))$$

Then $(\star)$ becomes:

$$\sup_{t\in T}(X_t - X_{t_0}) \leq \frac{x}{\alpha(1-\alpha)}\sup_{t\in T}\sum_{k=\kappa+1}^{\infty}\alpha^k\sqrt{\log 2\ell(A_k(t))} \leq Cx\sup_{t\in T}\sum_{k=\kappa+1}^{\infty}\alpha^k\sqrt{\log \ell(A_k(t))}$$

This is because $\log \ell(A_{\kappa+1}) \geq \log 2$ for some $A_{\kappa+1} \in \mathscr{A}_{\kappa+1}$. Also, $\sum_{s\in\text{children}(A)}\frac{1}{\ell(s)^2} \leq \sum_{i=1}^{\infty}\frac{1}{i^2} < 2$, so

$$\sum_{A\in\mathscr{A}_k}\prod_{j=\kappa+1}^{k}\frac{1}{\ell(A_j(t_A))^2} = \sum_{A\in\mathscr{A}_k}\frac{1}{\ell(A)^2}\underbrace{\prod_{j=\kappa+1}^{k-1}\frac{1}{\ell(A_j(t_A))^2}}_{\text{same for all } A \text{ w/ same parent in } \mathscr{A}_{k-1}}$$
$$\leq 2\sum_{A\in\mathscr{A}_{k-1}}\prod_{j=\kappa+1}^{k-1}\frac{1}{\ell(A_j(t_A))^2} \leq \cdots \leq 2^{k-\kappa}\,.$$

Now we revisit the bound for $\mathbb{P}(\Omega^c)$ with $x \geq 4$ (so that $x^2/8 \geq 2$). By definition of U, for $U \in \mathscr{A}_k$,

$$U(A) = \prod_{j=\kappa+1}^{k} 2\ell(A_j(t_A)) = 2^{k-\kappa}\prod_{j=\kappa+1}^{k}\ell(A_j(t_A)) \Longrightarrow U(A)^{-\frac{x^2}{8}} = 2^{-\frac{x^2}{8}(k-\kappa)}\prod_{j=\kappa+1}^{k}\frac{1}{(\ell(A_j(t_A)))^{\frac{x^2}{8}}}\,.$$

By plugging the above into the bound for $\mathbb{P}(\Omega^c)$, we have:

$$\begin{aligned}
\mathbb{P}(\Omega^c) &\le \sum_{k=\kappa+1}^{\infty} \sum_{A\in\mathscr{A}_k} U(A)^{-\frac{x^2}{8}} \\
&= \sum_{k=\kappa+1}^{\infty} (2^{k-\kappa})^{-\frac{x^2}{8}} \sum_{A\in\mathscr{A}_k} \prod_{j=\kappa+1}^{k} \frac{1}{(\ell(A_j(t_A)))^{\frac{x^2}{8}}} \\
&\le \sum_{k=\kappa+1}^{\infty} 2^{-\frac{x^2}{8}(k-\kappa)} \cdot 2^{k-\kappa} \\
&\le \sum_{k=\kappa+1}^{\infty} 2^{-\frac{x^2}{16}(k-\kappa)} \le C \cdot 2^{-\frac{x^2}{16}} .
\end{aligned}$$

Thus

$$\mathbb{P}\left(\sup_{t\in T}(X_t - X_{t_0}) \ge C'x \cdot \sup_{t\in T} \sum_{k\in\mathbb{Z}} \alpha^k \sqrt{\log \ell(A_k(t))}\right) \le C \cdot 2^{-\frac{x^2}{16}} .$$

Integrate this tail bound, and use that $\mathbb{E}X_t \equiv 0$:

$$\begin{aligned}
\mathbb{E}\sup_{t\in T} X_t &= \mathbb{E}\sup_{t\in T}(X_t - X_{t_0}) \\
&= \int_0^{\infty} \mathbb{P}(\sup_{t\in T} X_t - X_{t_0} \ge y)\mathrm{d}y \\
&\le C'' \sup_{t\in T} \sum_{k\in\mathbb{Z}} \alpha^k \sqrt{\log \ell(A_k(t))} \quad \text{by change of variable.}
\end{aligned}$$

Recall the definition of $\gamma(T)$. Take inf over all labeled nets $(\mathscr{A}, \ell)$, we have $\mathbb{E}\sup_{t\in T} X_t \lesssim \gamma(T)$. For the case $|T| = \infty$, by separability, $\exists t_1, t_2, \ldots \in T$ such that

$$\mathbb{E}\sup_{t\in T} X_t = \lim_{k\to\infty} \mathbb{E} \sup_{t\in\{t_1,\ldots,t_k\}} X_t \le \limsup_{k\to\infty} C''\gamma(\{t_1, \ldots, t_k\}) .$$

If $T' \subseteq T$, any labeled net $(\mathscr{A}, \ell)$ of T restricted to a labeled net of T' implies $\gamma(T') \le \gamma(T)$, so $\mathbb{E}\sup_{t\in T} X_t \le C''\gamma(T)$. □

13 Matrix Deviations, Random Projections, Dvoretsky-Milman Theorem

Readings: §8.7, §9.1-9.2, §11.1-11.3 in [Ver18].

Recall Talagrand's comparison inequality (Corollary 12.4), which we restate here:

Theorem 13.1 (Talagrand's comparison inequality)**.** Let $\{X_t\}_{t\in T}, \{Y_t\}_{t\in T}$ be separable mean-zero processes. $\{Y_t\}_{t\in T}$ is Gaussian process with canonical metric d, $\{X_t\}_{t\in T}$ is subgaussian w.r.t. d, i.e.,

$$\log \mathbb{E} e^{\lambda(X_t - X_s)} \le \frac{\lambda^2}{2}\mathbb{E}(Y_t - Y_s)^2 \qquad \forall s,t \in T, \lambda \ge 0\,.$$

Then for a universal constant $C > 0$,

$$\mathbb{E}\sup_{t\in T} X_t \le C\,\mathbb{E}\sup_{t\in T} Y_t\,.$$

Starting from this theorem, we will derive several interesting results in high-dimensional probability, statistics, and geometry.

13.1 Chevet's inequality and matrix deviations

Theorem 13.2 (Chevet's inequality)**.** Let $X \in \mathbb{R}^{n\times m}$ have independent, mean-zero, σ^2-subgaussian entries. Then for any $S \subseteq \mathbb{R}^n$, $T \subseteq \mathbb{R}^m$,

$$\mathbb{E}\sup_{u\in S, v\in T} u^\top X v \le C\sigma(w(S)\,\mathsf{rad}(T) + w(T)\,\mathsf{rad}(S))\,,$$

where:

$$w(T) = \mathbb{E}_{g\sim\mathcal{N}(0,I)}\sup_{t\in T} g^\top t \quad \text{is the Gaussian width}$$
$$\mathsf{rad}(T) = \sup_{t\in T}\|t\|_2 \quad \text{is the radius}$$

Proof. Denote $X_{uv} := u^\top X v$. Then $X_{uv} - X_{wz} = \sum_{i,j} X_{ij}(u_i v_j - w_i z_j)$ is τ^2-subgaussian for

$$\begin{aligned}
\tau^2 &= \sigma^2\|uv^\top - wz^\top\|_F^2 \\
&\le \sigma\left(\|(u-w)v^\top\|_F + \|w(v-z)^\top\|_F\right)^2 \\
&= \sigma^2\left(\|u-w\|_2\cdot\|v\|_2 + \|v-z\|_2\cdot\|w\|_2\right)^2 \\
&\le 2\sigma^2(\mathsf{rad}(T)^2\|u-w\|_2^2 + \mathsf{rad}(S)^2\|v-z\|_2^2)
\end{aligned}$$

Consider the process $Y_{uv} = \sqrt{2\sigma^2}(g^\top u\cdot\mathsf{rad}(T) + h^\top v\cdot\mathsf{rad}(S))$, where $g\sim\mathcal{N}(0,I_n)$ and $h\sim\mathcal{N}(0,I_m)$ are independent. Then

$$\begin{aligned}
\mathbb{E}(Y_{uv} - Y_{wz})^2 &= \mathbb{E}\left[(\mathsf{rad}(T)(u-w)^\top g + \mathsf{rad}(S)(v-z)^\top h)^2\right] \\
&= 2\sigma^2(\mathsf{rad}(T)^2\|u-w\|_2^2 + \mathsf{rad}(S)^2\|v-z\|_2^2)\,.
\end{aligned}$$

Therefore, by Talagrand's comparison inequality (Theorem 13.1),

$$\begin{aligned}
\mathbb{E}\sup_{u\in S, v\in T} u^\top X v &\le C\cdot\mathbb{E}\sup_{u\in S, v\in T} Y_{uv} \\
&\le C'\sigma\mathbb{E}[\mathsf{rad}(T)\sup_{u\in S} u^\top g + \mathsf{rad}(S)\sup_{v\in T} v^\top h] \\
&= C'\sigma(\mathsf{rad}(T)w(S) + \mathsf{rad}(S)w(T))\,.
\end{aligned}$$

□

Example 13.3. If $S = \mathbb{B}^n$ (unit ball in ℓ_2), then $w(S) = \mathbb{E}[\|g\|_2] \le \sqrt{n}$ and $\mathsf{rad}(S) = 1$. This recovers $\mathbb{E}[\|X\|_{\mathrm{op}}] \le C(\sqrt{n} + \sqrt{m})$ from lecture 7.

Example 13.4. If $S = \frac{1}{\sqrt{n}}[-1,1]^n$ (rescaled unit ball in ℓ_∞), then $\mathsf{rad}(S) = 1$ and $w(S) = \frac{1}{\sqrt{n}}\mathbb{E}\|g\|_1 \asymp \sqrt{n}$. So we have

$$\mathbb{E} \sup_{u \in \frac{1}{\sqrt{n}}[-1,1]^n, v \in \frac{1}{\sqrt{n}}[-1,1]^n} u^\top X v \le C(\sqrt{n} + \sqrt{m}) .$$

This is the same bound as for $\mathbb{B}^n$.

Example 13.5. If $S = \{t \in \mathbb{R}^n \mid \|t\|_1 \le 1\}$ (unit ball in ℓ_1). Then $w(S) = \mathbb{E}\|g\|_\infty \asymp \sqrt{\log n}$ and $\mathsf{rad}(S) = 1$. So we have

$$\mathbb{E} \sup_{u,v : \|u\|_1, \|v\|_1 \le 1} u^\top X v \le C(\sqrt{n} + \sqrt{m}) .$$

Lemma 13.6 (Matrix deviation inequality)**.** Let $X \in \mathbb{R}^{n \times m}$ have independent, mean-zero, isotropic σ^2-subgaussian rows (i.e., $\mathbb{E}X_i X_i^\top] = I$ and $u^\top X_i$ is σ^2-subgaussian for all unit vectors $u \in \mathbb{R}^n$).Then for any $T \subseteq \mathbb{R}^m$ containing 0,

$$\mathbb{E} \sup_{u \in T} \left| \|Xu\|_2 - \mathbb{E}[\|Xu\|_2] \right| \le C\sigma^2 \cdot w(T) .$$

Proof. Define $X_u := \|Xu\|_2 - \mathbb{E}[\|Xu\|_2]$. We claim that

$$\|X_u - X_v\|_{\psi_2} \le C\sigma^2 \|u - v\|_2^2, \quad \forall u, v \in \mathbb{R}^n \tag{$\star$}$$

Now we assume that ($\star$) holds. Let $X_0 = 0$.

$$\begin{aligned} \mathbb{E} \sup_{u \in T} |X_u| &= \mathbb{E} \sup_{u \in T} [\max(X_u, 0) + \max(-X_u, 0)] \\ &\le \mathbb{E} \sup_{u \in T} X_u + \mathbb{E} \sup_{u \in T} -X_u \end{aligned}$$

Define $Y_u = u^\top g$ for $u \in T$, where $g \sim \mathcal{N}(0, I_m)$. Then $\mathbb{E}[(Y_u - Y_v)^2] = \mathbb{E}((u-v)^\top g)^2 = \|u - v\|_2^2$. By Talagrand's comparison inequality (Theorem 13.1),

$$\begin{aligned} \mathbb{E} \sup_{u \in T} X_u &\le C\sigma^2 \mathbb{E} \sup_{u \in T} Y_u \\ \mathbb{E} \sup_{u \in T} -X_u &\le C\sigma^2 \mathbb{E} \sup_{u \in T} -Y_u \\ \implies \mathbb{E} \sup_{u \in T} |X_u| &\le 2C\sigma^2 \mathbb{E} \sup_{u \in T} Y_u = 2C\sigma^2 \cdot w(T) \end{aligned}$$

It remains to show that ($\star$) holds. We divide into the following two cases:

(1) Suppose first that $\|u\|_2 = \|v\|_2 = 1$. Note that

$$\begin{aligned} \mathbb{P}\left(\frac{|\|Xu\|_2 - \|Xv\|_2|}{\|u - v\|_2} \ge s \right) = \mathbb{P}\left(|Z| := \left| \frac{\|Xu\|_2^2 - \|Xv\|_2^2}{\|u - v\|_2} \right| \ge s(\|Xu\|_2 + \|Xv\|_2) \right) \\ \le \underbrace{\mathbb{P}(|Z| \ge s\sqrt{n})}_{\textbf{(I)}} + \underbrace{\mathbb{P}(\|Xu\|_2 \le \sqrt{n}/2) + \mathbb{P}(\|Xv\|_2 \le \sqrt{n}/2)}_{\textbf{(II)}} \end{aligned}$$

- Xu has mean-zero, variance 1, σ^2-subgaussian entries. By Lecture 2 (Proposition 2.16),

$$\mathbb{P}(|\|Xu\|_2 - \sqrt{n}| \ge t) \le 2e^{-\frac{ct^2}{\sigma^4}} \quad \overset{t = \sqrt{n}/2}{\implies} \mathbb{P}(\|Xu\|_2 \le \sqrt{n}/2) \le 2e^{-\frac{cn}{2\sigma^4}} .$$

  Similarly for $\|Xv\|_2$. So, **(II)** $\le 4e^{-\frac{cn}{2\sigma^4}}$.

- By definition, we can decompose Z into $Z = \sum_{i=1}^n Z_i$ with $Z_i = \frac{(X_i^\top u)^2 - (X_i^\top v)^2}{\|u-v\|_2}$. Note that $\mathbb{E}Z_i = 0$, and write $Z_i = \frac{X_i^\top(u-v)\cdot X_i^\top(u+v)}{\|u-v\|_2}$, by Proposition 2.15, Z_i is sub-exponential with

$$\|Z_i\|_{\psi_1} \le \frac{1}{\|u-v\|_2}\|X_i^\top(u-v)\|_{\psi_2}\cdot\|X_i^\top(u+v)\|_{\psi_2} \le \frac{C_1\sigma\|u-v\|_2\cdot C_2\sigma\|u+v\|_2}{\|u-v\|_2} \le C'\sigma^2\,.$$

By Bernstein's inequality (Theorem 2.12),

$$\textbf{(I)} \le 2\exp\left(-c\min\left(\frac{(s\sqrt{n})^2}{n(C'\sigma^2)^2}, \frac{s\sqrt{n}}{C'\sigma^2}\right)\right) = 2\exp\left(-c'\min\left(\frac{s^2}{\sigma^4}, \frac{s\sqrt{n}}{\sigma^2}\right)\right).$$

Since $\sigma^2 \ge 1$, if $s \le 2\sqrt{n}$, then $\frac{s\sqrt{n}/\sigma^2}{s^2/\sigma^4} \ge \frac{\sigma^2\sqrt{n}}{s} \ge \frac{\sigma^2}{2} \ge \frac{1}{2}$. Hence,

$$\textbf{(I)} \le 2e^{-\frac{c's^2}{2\sigma^4}} \quad \Longrightarrow \textbf{(I)} + \textbf{(II)} \le 6e^{-\frac{c''s^2}{\sigma^4}}\,.$$

If $s > 2\sqrt{n}$, we directly study

$$\begin{aligned}\mathbb{P}\left(\frac{|\|Xu\|_2 - \|Xv\|_2|}{\|u-v\|_2} \ge s\right) &\le \mathbb{P}\left(\left\|X\frac{u-v}{\|u-v\|_2}\right\|_2 \ge s\right)\\ &\le \mathbb{P}\left(\left\|X\frac{u-v}{\|u-v\|_2}\right\|_2 - \sqrt{n} \ge \frac{s}{2}\right) \le 2e^{-\frac{\tilde{c}s^2}{\sigma^4}}\,.\end{aligned}$$

Therefore, $\frac{|\|Xu\|_2 - \|Xv\|_2|}{\|u-v\|_2}$ is $C\sigma^4$-subgaussian, i.e.,

$$\|X_u - X_v\|_{\psi_2} \le C\,\big\|\|Xu\|_2 - \|Xv\|_2\big\|_{\psi_2} \le C'\sigma^2\|u-v\|_2\,.$$

(2) For general $u, v \in \mathbb{R}^m$, assume that WLOG, $\|u\|_2 = 1$ and $\|v\|_2 \ge 1$. Let $\bar{v} = \frac{v}{\|v\|_2}$. Since $\|\cdot\|_{\psi_2}$ is indeed a norm (Definition 1.13),

$$\|X_u - X_v\|_{\psi_2} \le \|X_u - X_{\bar{v}}\|_{\psi_2} + \|X_{\bar{v}} - X_v\|_{\psi_2}\,.$$

For the first term, $\|X_u - X_{\bar{v}}\|_{\psi_2} \le C\sigma^2\|u-\bar{v}\|_2$ by case (1) above. Since

$$|X_{\bar{v}} - X_v| = |X_{\bar{v}} - \|v\|_2 X_{\bar{v}}| = \|\bar{v}-v\|_2|X_{\bar{v}}|\,,$$

we have $\|X_{\bar{v}} - X_v\|_{\psi_2} = \|\bar{v}-v\|_2\|X_{\bar{v}}\|_{\psi_2} \le C\sigma^2\|\bar{v}-v\|_2$. Then,

$$\begin{aligned}\|X_u - X_v\|_{\psi_2} &\le C\sigma^2(\|u-\bar{v}\|_2 + \|\bar{v}-v\|_2)\\ &\le C\sigma^2(\|u-v\|_2 + 2\|\bar{v}-v\|_2)\\ &\le 3C\sigma^2\|u-v\|_2\,,\end{aligned}$$

where the last inequality uses

$$\|\bar{v}-v\|_2 = \left\|\frac{\|u\|_2}{\|v\|_2}v - v\right\|_2 \le \left|\frac{\|u\|_2}{\|v\|_2} - 1\right|\|v\|_2 = |\|u\|_2 - \|v\|_2| \le \|u-v\|_2\,.$$

By combining cases (1) and (2), we finish the proof for ($\star$). □

Remark 13.7.

- We can replace $\mathbb{E}\|Xu\|_2$ by $\sqrt{n}\|u\|_2$. This is because by Jensen's inequality, we have $\mathbb{E}\|Xu\|_2 \le \sqrt{n}\|u\|_2$, which implies

$$\mathbb{P}(\|Xu\|_2 - \sqrt{n}\|u\|_2 > t\|u\|_2) \le \mathbb{P}(\|Xu\|_2 - \mathbb{E}\|Xu\|_2 > t\|u\|_2) \le 2e^{-\frac{ct^2}{\sigma^4}},$$

and further implies $\mathbb{E}\|Xu\|_2 \in \sqrt{n}\|u\|_2 \pm C\sigma^2\|u\|_2$ and since $g^\top u \sim \mathcal{N}(0, \|u\|_2^2)$, $\mathbb{E}|g^\top u| = \sqrt{\frac{2}{\pi}}\|u\|_2$,

$$\sup_{u\in T} C\sigma^2\|u\|_2 \le \sup_{u\in T} C'\sigma^2\mathbb{E}|g^\top u| \le C'\sigma^2\mathbb{E}\sup_{u\in T}|g^\top u| = C'\sigma^2 w(T).$$

- If $T \subseteq \mathbb{B}^m$ (unit ball in ℓ_2), then $w(T) \asymp \sqrt{m}$. This bounds the singular valuesof X.

$$\mathbb{E}\sup_{u\in\mathbb{R}^m:\|u\|_2=1}\left|\|Xu\|_2 - \sqrt{n}\right| = \mathbb{E}\sup_{u\in T}\left|\|Xu\|_2 - \sqrt{n}\|u\|_2\right| \le C\sigma^2\sqrt{m}.$$

This means that $s_{\min}(X), s_{\max}(X) \in \sqrt{n} \pm \mathcal{O}_{\mathbb{P}}(\sqrt{m})$.

Theorem 13.8 (Low-rank covariance estimation)**.** Let $X_1, \dots, X_n \in \mathbb{R}^m$ be iid with $\mathbb{E}X_i = 0$, $\mathsf{Cov}(X_i) = \Sigma \in \mathbb{R}^{m\times m}$, and $\Sigma^{-\frac{1}{2}}X_i$ is σ^2-subgaussian. Let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ be the sample covariance estimate of Σ. For any $\delta > 0$, there exists $C(\delta) > 0$ such that

$$\mathbb{P}\left[\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} > C(\delta)\sigma^4\left(\sqrt{\frac{r}{n}} + \frac{r}{n}\right)\|\Sigma\|_{\mathrm{op}}\right] \le \delta,$$

where $r = \mathsf{Tr}(\Sigma)/\|\Sigma\|_{\mathrm{op}}$ is the "stable rank" of Σ.

Proof. Let $Z_i = \Sigma^{-\frac{1}{2}}X_i$. Collect these Z_i's as rows of $Z = \begin{bmatrix} — & Z_1^\top & — \\ & \vdots & \\ — & Z_n^\top & — \end{bmatrix} \in \mathbb{R}^{n\times m}$. Then

$$\begin{aligned}
\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} &= \left\|\Sigma^{\frac{1}{2}}\left(\frac{1}{n}Z^\top Z - I_m\right)\Sigma^{\frac{1}{2}}\right\|_{\mathrm{op}} \\
&= \sup_{v\in\mathbb{B}^m}\left|v^\top\Sigma^{\frac{1}{2}}\left(\frac{1}{n}Z^\top Z - I_m\right)\Sigma^{\frac{1}{2}}v\right| \\
&= \sup_{u\in T}\left|u^\top\left(\frac{1}{n}Z^\top Z - I_m\right)u\right| \quad \text{let } T = \Sigma^{\frac{1}{2}}\mathbb{B}^m \\
&= \frac{1}{n}\sup_{u\in T}\left|\|Zu\|_2^2 - n\|u\|_2^2\right|.
\end{aligned}$$

By matrix deviation lemma and Markov's inequality,

$$\mathbb{P}\left(\sup_{u\in T}\left|\|Zu\|_2 - \sqrt{n}\|u\|_2\right| > \frac{C\sigma^2 w(T)}{\delta}\right) \le \frac{\mathbb{E}\sup_{u\in T}\left|\|Zu\|_2 - \sqrt{n}\|u\|_2\right|}{C\sigma^2 w(T)/\delta} = \delta,$$

so with probability $\ge 1-\delta$, $\forall u \in T$,

$$\left|\|Zu\|_2 - \sqrt{n}\|u\|_2\right| \le C(\delta)\sigma^2 w(T),$$

which further implies

$$\begin{aligned}
\frac{1}{n}\left|\|Zu\|_2^2 - n\|u\|_2^2\right| &= \frac{1}{n}\left|\|Zu\|_2 - \sqrt{n}\|u\|_2\right|\cdot\left|\|Zu\|_2 + \sqrt{n}\|u\|_2\right| \\
&\le \frac{1}{n}C(\delta)\sigma^2 w(T)\left(C(\delta)\sigma^2 w(T) + 2\sqrt{n}\|u\|_2\right).
\end{aligned}$$

Here, $w(T) = \mathbb{E}\sup_{u\in T} g^\top u = \mathbb{E}\|\Sigma^{\frac{1}{2}} g\|_2 \le \left(\mathbb{E}\|\Sigma^{\frac{1}{2}} g\|_2^2\right)^{\frac{1}{2}} = \sqrt{\mathsf{Tr}\,\Sigma}$ and $\|u\|_2 \le \|\Sigma^{1/2}\|_{\mathrm{op}} \le \sqrt{\|\Sigma\|_{\mathrm{op}}}$, so, still, with probability $\ge 1-\delta$,

$$\begin{aligned}\frac{1}{n}\left|\|Zu\|_2^2 - n\|u\|_2^2\right| &\le \frac{1}{n} C(\delta)\sigma^2\sqrt{\mathsf{Tr}\,\Sigma}\left(C(\delta)\sigma^2\sqrt{\mathsf{Tr}(\Sigma)} + 2\sqrt{n\|\Sigma\|_{\mathrm{op}}}\right)\\ &\le C'(\delta)\sigma^4\frac{\mathsf{Tr}\,\Sigma}{n} + C''(\delta)\sigma^2\sqrt{\frac{\|\Sigma\|_{\mathrm{op}}\,\mathsf{Tr}\,\Sigma}{n}}\\ &\le \tilde{C}(\delta)\sigma^4\left(\frac{r}{n} + \sqrt{\frac{r}{n}}\right)\|\Sigma\|_{\mathrm{op}}\,.\end{aligned}$$

□

13.2 Random projections

Let $X \in \mathbb{R}^{n\times m}$ have iid $\mathcal{N}(0,1)$ entries. Consider $P = \frac{1}{\sqrt{n}}X$ as a random projection from $\mathbb{R}^m$ to $\mathbb{R}^n$. For a given $T \subseteq \mathbb{R}^m$, what does the projected set PT look like?

Corollary 13.9. For any $\delta > 0$, with probability $\ge 1-\delta$,

$$\|Pu - Pv\|_2 \in \|u-v\|_2 \pm \frac{C(\delta)}{\sqrt{n}} w(T), \qquad \forall u, v \in T\,.$$

In particular, $\mathsf{diam}(PT) \le \mathsf{diam}(T) \pm \frac{C(\delta)}{\sqrt{n}} w(T)$.

Proof. Let $S = T - T = \{u - v \mid u, v \in T\}$. Then $w(S) = \mathbb{E}\sup_{u,v\in T} g^\top(u-v) \le W(T) + w(-T) = 2w(T)$.

By matrix deviation lemma and Markov inequality, let $\delta > 0$,

$$\begin{aligned}&\mathbb{P}\left(\sup_{u-v\in S}\left|\|P(u-v)\|_2 - \|u-v\|_2\right| > \frac{Cw(S)}{\delta\sqrt{n}}\right)\\ \le&\frac{\delta\sqrt{n}}{Cw(S)}\cdot\frac{1}{\sqrt{n}}\mathbb{E}\sup_{u-v\in S}\left|\|X(u-v)\|_2 - \sqrt{n}\|u-v\|_2\right|\\ \le&\frac{\delta\sqrt{n}}{Cw(S)}\cdot\frac{1}{\sqrt{n}}Cw(S) = \delta\,.\end{aligned}$$

Hence, with probability $\ge 1-\delta$,

$$\left|\|P(u-v)\|_2 - \|u-v\|_2\right| \le \frac{Cw(S)}{\delta\sqrt{n}} \le \frac{C(\delta)}{\sqrt{n}} w(T)\,.$$

This shows the first statement. For the second,

$$\mathsf{diam}(PT) = \sup_{u,v\in T}\|Pu - Pv\|_2 \le \sup_{u,v\in T}\|u-v\|_2 + \frac{C(\delta)}{\sqrt{n}}w(T) = \mathsf{diam}(T) + \frac{C(\delta)}{\sqrt{n}}w(T)\,.$$

□

Interpretation: $\mathsf{diam}(PT)$ has a phase transition. When $\mathsf{diam}(T) \gg \frac{1}{\sqrt{n}}w(T)$, i.e., $n \gg \frac{w(T)^2}{\mathsf{diam}(T)^2}$, we have $\mathsf{diam}(PT) \approx \mathsf{diam}(T)$, and P is a near-isometry:

$$\|Pu - Pv\|_2 = \|u-v\|_2 + o(\mathsf{diam}\,T) \qquad \forall u, v \in T\,.$$

Here $\frac{w(T)^2}{\mathsf{diam}(T)^2}$ is called the stable dimension of T.

If T is finite, then by the maximal of Gaussian (Exercise 2.5.10 in [Ver18]),

$$w(T) = \mathbb{E} \sup_{t \in T} g^\top t \le C\sqrt{\log |T|} \cdot \mathsf{diam}(T)\,,$$

so the above holds as long as $n \gg \log |T|$, recovering the condition of the Johnson-Lindenstrauss Theorem (Theorem 2.17) from lecture 2.

In the complementary regime $\mathsf{diam}(T) \ll \frac{1}{\sqrt{n}} w(T)$, the following theorem shows that PT looks instead like a Euclidean ball in $\mathbb{R}^n$.

Theorem 13.10 (Dvoretsky-Milman)**.** Suppose T contains 0, let $\mathsf{conv}(PT)$ be the convex hull of PT, and $\mathbb{B}_2^n$ be the unit ball in $\mathbb{R}^n$. Then with probability $\ge 1 - \delta$,

$$r_- \mathbb{B}_2^n \subseteq \mathsf{conv}(PT) \subseteq r_+ \mathbb{B}_2^n\,,$$

where $r_\pm = \frac{1}{\sqrt{n}} w(T) \pm C(\delta)\,\mathsf{diam}(T)$.

Proof. The claim that $r_- \mathbb{B}_2^n \subseteq \mathsf{conv}(PT) \subseteq r_+ \mathbb{B}_2^n$ is equivalent to:

$$r_- \le \sup_{x \in PT} x^\top u \le r_+\,, \qquad \forall u \in \mathbb{S}^{n-1}\,.$$

Consider

$$\begin{aligned} Z &:= \sup_{u \in \mathbb{S}^{n-1}} \left| \sup_{x \in PT} x^\top u - \mathbb{E} \sup_{x \in PT} x^\top u \right| \\ &= \frac{1}{\sqrt{n}} \sup_{u \in \mathbb{B}_2^n} \left| \sup_{y \in T} y^\top X^\top u - \mathbb{E} \sup_{y \in T} y^\top X^\top u \right| := \frac{1}{\sqrt{n}} \sup_{u \in \mathbb{B}_2^n} |X_u|\,. \end{aligned}$$

We claim that

$$\|X_u - X_v\|_{\psi_2} \le C \cdot \mathsf{diam}(T) \|u - v\|_2 \qquad \forall u, v \in \mathbb{B}^n \tag{▲}$$

Assume that (▲) holds. Let $Y_u = u^\top g$ for $u \in \mathbb{B}_2^n$ and $g \sim \mathcal{N}(0, I_n)$. Then by subgaussian comparison theorem (Theorem 13.1),

$$\begin{aligned} \mathbb{E} \sup_{u \in \mathbb{B}^n} |X_u| &\le \mathbb{E} \sup_{u \in \mathbb{B}^n} X_u + \mathbb{E} \sup_{u \in \mathbb{B}^n} -X_u \\ &\le 2C\,\mathsf{diam}(T) \mathbb{E} \sup_{u \in \mathbb{B}^n} Y_u \\ &\le C' \sqrt{n}\,\mathsf{diam}(T)\,. \end{aligned}$$

Then by the same argument as in the matrix deviation inequality (Theorem 13.6), with probability $\ge 1 - \delta$,

$$Z = \frac{1}{\sqrt{n}} \sup_{u \in \mathbb{B}^n} |X_u| \le C(\delta)\,\mathsf{diam}(T)\,.$$

For every $u \in \mathbb{S}^{n-1}$, we have

$$\mathbb{E} \sup_{x \in PT} x^\top u = \frac{1}{\sqrt{n}} \sup_{y \in T} y^\top X^\top u = \frac{1}{\sqrt{n}} w(T)\,.$$

Hence, $Z \le C(\delta)\,\mathsf{diam}(T)$ implies $r_- \le \sup_{x \in PT} x^\top u \le r_+$ as desired.

It remains to show (▲). Let $f(x) = \sup_{y \in T} y^\top x$. Note that $X \mapsto y^\top(a + X^\top b)$ is $\|y\|_2 \cdot \|b\|_2$-Lipschitz, so $X \mapsto f(a + X^\top b)$ is $\mathsf{diam}(T)\|b\|_2$-Lipschitz. The same holds for $X \mapsto f(a - X^\top b)$. By Gaussian concentration,

$$\|f(a + X^\top b) - \mathbb{E}f(a + X^\top b)\|_{\psi_2} \le C\,\mathsf{diam}(T) \cdot \|b\|_2$$
$$\|f(a - X^\top b) - \mathbb{E}f(a - X^\top b)\|_{\psi_2} \le C\,\mathsf{diam}(T) \cdot \|b\|_2\,.$$

Here $\mathbb{E}f(a + X^\top b) = \mathbb{E}f(a - X^\top b)$, so by triangular inequality,

$$\|f(a + X^\top b) - f(a - X^\top b)\|_{\psi_2} \le C'\,\mathsf{diam}(T) \cdot \|b\|_2\,.$$

Take $b = \frac{u-v}{2}$ and $a = X\frac{u+v}{2}$. Since X has mean-zero Gaussian entries, a is independent of Xb because $\langle u + v, u - v\rangle = 0$. This shows that

$$\|f(X^\top u) - f(X^\top v)\|_{\psi_2} \le C''\,\mathsf{diam}(T)\|u - v\|_2$$

conditioned on a, and hence also unconditionally, which is the claim (▲). □

References

[AG93] Miguel A Arcones and Evarist Giné. Limit theorems for u-processes. *The Annals of Probability*, pages 1494–1542, 1993.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013.

[BMDLP23] Milad Bakhshizadeh, Arian Maleki, and Victor H De La Pena. Sharp concentration results for heavy-tailed distributions. *Information and Inference: A Journal of the IMA*, 12(3):1655–1685, 2023.

[dlP92] Victor H de la Pena. Decoupling and khintchine's inequalities for u-statistics. *The Annals of Probability*, pages 1877–1892, 1992.

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[Vaa98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[vH14] Ramon van Handel. Probability in high dimension. *Princeton University*, 2014.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.