Random Variables

Definitions

- [Sub-Gaussian] Let X be a random variable. Then X is <u>sub-Gaussian</u> if it satisfies any of the following equivalent properties:
    - [Tails] $\mathbb{P}[|X| \geq t] \leq 2e^{-\frac{t^2}{\kappa_1^2}}\ \forall t \geq 0$
    - [Moments] $(\mathbb{E}[|X|^p])^{\frac{1}{p}} \leq \kappa_2\sqrt{p}\ \forall p \geq 1$
    - [MGF Bounded over an Interval] $\mathbb{E}\left[e^{\lambda^2 X^2}\right] \leq e^{\lambda^2\kappa_3^2}\ \forall \lambda: |\lambda| \leq \frac{1}{\kappa_3}$
    - [MGF Bounded at a Point] $\exists \kappa_4 > 0$ s.t. $\mathbb{E}\left[e^{\frac{X^2}{\kappa_4^2}}\right] \leq 2$
    - [Uniform MGF] If $\mathbb{E}[X] = 0$, then $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\kappa_5^2\lambda^2}\ \forall \lambda \in \mathbb{R}$
- [$\|\cdot\|_{\psi_2}$] Let X be a sub-Gaussian random variable. Define $\|X\|_{\psi_2} := \inf_{t\geq 0}\left\{\mathbb{E}\left[e^{\frac{X^2}{t^2}}\right] \leq 2\right\}$.
    - $\|\cdot\|_{\psi_2}$ is a norm on the space of sub-Gaussian random variables
    - X is sub-Gaussian $\Leftrightarrow \|X\|_{\psi_2} < \infty$
    - If $X \sim N(0,1)$, then $\|X\|_{\psi_2} = \frac{2\sqrt{2}}{\sqrt{3}}$
- [Sub-Exponential] A random variable X is <u>sub-exponential</u> if it satisfies any of the following equivalent properties:
    - [Tails] $\mathbb{P}[|X| \geq t] \leq 2e^{-\frac{t}{\kappa_1}}\ \forall t \geq 0$
    - [Moments] $\|X\|_p := (\mathbb{E}[|X|^p])^{\frac{1}{p}} \leq \kappa_2 p\ \forall p \geq 1$
    - [MGF Bounded over an Interval] $\mathbb{E}\left[e^{\lambda|X|}\right] \leq e^{\lambda\kappa_3}\ \forall \lambda: 0 \leq \lambda \leq \frac{1}{\kappa_3}$
    - [MGF Bounded at a Point] $\exists \kappa_4 > 0$ s.t. $\mathbb{E}\left[e^{\frac{|X|}{\kappa_4}}\right] \leq 2$
    - [Uniform MGF] If $\mathbb{E}[X] = 0$, then $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\kappa_5^2\lambda^2}\ \forall \lambda: |\lambda| \leq \frac{1}{\kappa_5}$
- [$\|\cdot\|_{\psi_1}$] Let X be a sub-exponential random variable. Define $\|X\|_{\psi_1} := \inf_{t\geq 0}\left\{\mathbb{E}\left[e^{\frac{|X|}{t}}\right] \leq 2\right\}$.
- [Bernstein Condition] Let X have mean μ and variance σ^2. Then X satisfies the <u>Bernstein condition with parameter b</u> if $|\mathbb{E}[(X-\mu)^k]| \leq \frac{1}{2}(k!)\sigma^2 b^{k-2}\ \forall k \geq 2$.

Tools

- [Markov] Let $X \geq 0$ be a nonnegative random variable. Then $\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$.
- [Chebyshev] Let X be a random variable. Then $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathrm{Var}[X]}{t^2}$.
- [Generalised Markov] $\mathbb{P}[|X| \geq t] \leq \inf_{p\geq 0}\frac{\mathbb{E}[|X|^p]}{t^p}$
- [Chernoff] $\mathbb{P}[X \geq t] \leq \inf_{\lambda\geq 0}\frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}}$

Normal Distribution Bounds

- [Normal Tail] Let $Z \sim N(0,1)$. Then for $t \geq 1$, $\frac{1}{\sqrt{2\pi}}\left(\frac{1}{t} - \frac{1}{t^3}\right)e^{-\frac{t^2}{2}} \leq \mathbb{P}[Z \geq t] \leq \frac{1}{\sqrt{2\pi}}\frac{1}{t}e^{-\frac{t^2}{2}}$
    - [Weaker Tail] $\mathbb{P}[Z \geq t] \leq \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$
- [Truncated Normal] Let $Z \sim N(0,1)$. Then $\forall t \geq 1$, $\mathbb{E}\left[Z^2\mathbb{1}_{\{Z\geq t\}}\right] = \frac{1}{\sqrt{2\pi}}te^{-\frac{t^2}{2}} + \mathbb{P}[Z \geq t] \leq \frac{1}{\sqrt{2\pi}}\left(t + \frac{1}{t}\right)e^{-\frac{t^2}{2}}$

Bernoulli and Binomial Random Variables

- [2.3.1] Let $S = X_1 + \cdots + X_n$ where $X_i \sim \mathrm{Bernoulli}(\mu_i)$. Let $\mu = \sum_{i=1}^n \mu_i$. Then, for $t > \mu$, $\mathbb{P}[S > t] \le e^{-\mu}\left(\frac{e\mu}{t}\right)^t$ and for $t < \mu$, $\mathbb{P}[S < t] \le e^{-\mu}\left(\frac{e\mu}{t}\right)^t$.
  - *Prove by Chernoff, then apply* $1 + x \le e^x$
  - Let $S \sim \mathrm{Binomial}(n, p)$, then $\mathbb{P}[S > t] \le e^{-np}\left(\frac{enp}{t}\right)^t$
- [HW1 P4] Let $S \sim \mathrm{Binomial}(n, p)$, then:
  - $\mathbb{P}[S \ge n(p+t)] \le e^{-n\left((p+t)\log\left(\frac{p+t}{p}\right) + (1-p-t)\log\left(\frac{1-p-t}{1-p}\right)\right)}$
  - $\mathbb{P}[S \ge (1+\delta)np] \le \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{np}$
- [Hoeffding] Let $S \sim \mathrm{Binomial}(n, p)$, then $\mathbb{P}[S \ge np + t] \le e^{-\frac{2t^2}{n}}$

Bounded Random Variables

- Let X be a zero-mean random variable s.t. $X \in [a, b]$ a.s. Then $\mathbb{E}\left[e^{\lambda X}\right] \le e^{\frac{\lambda^2 (b-a)^2}{8}}$.

Hoeffding and Bernstein

- [Hoeffding (Rademacher)] Let $X_1, \dots, X_n$ be independent Rademacher random variables and $a \in \mathbb{R}^n$. Then, for $t > 0$, $\mathbb{P}[\sum_{i=1}^n a_i X_i \ge t] \le e^{-\frac{t^2}{2\|a\|_2^2}}$
  - *Prove by Chernoff's inequality*
  - $\mathbb{P}[\sum_{i=1}^n X_i \ge t] \le e^{-\frac{t^2}{2n}}$
  - $\mathbb{P}[|\sum_{i=1}^n a_i X_i| \ge t] \le 2e^{-\frac{t^2}{2\|a\|_2^2}}$
- [Hoeffding (Bounded)] Let $X_1, \dots, X_n$ be independent and $X_i \in [a_i, b_i]$ almost surely. Then, for $t > 0$, $\mathbb{P}[\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \ge t] \le e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$
- [Hoeffding (Sub-Gaussian)] Let $X_1, \dots, X_n$ be independent and zero-mean sub-Gaussian random variables. Then:
  - $\sum_{i=1}^n X_i$ is a sub-Gaussian random variable.
  - For $t > 0$, $\mathbb{P}[|\sum_{i=1}^n X_i| \ge t] \le 2e^{-\frac{ct^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}}$
  - Let $K = \max_{1 \le i \le n} \|X_i\|_{\psi_2}$ and $a \in \mathbb{R}^n$. Then, $\mathbb{P}[|\sum_{i=1}^n a_i X_i| \ge t] \le 2e^{-\frac{ct^2}{K^2\|a\|_2^2}}$
- [Khintchine] Let $X_1, \dots, X_n$ be independent and zero-mean sub-Gaussian random variables with unit variances and $K = \max_{1 \le i \le n} \|X_i\|_{\psi_2}$. Then, for $p \in [2, \infty)$, $\|a\|_2 \le \|\sum_{i=1}^n a_i X_i\|_p \le CK\sqrt{p}\|a\|_2$.
  - *Direct application of Hoeffding*
- [Bernstein] Let $X_1, \dots, X_n$ be independent, zero-mean sub-exponential random variables and $a \in \mathbb{R}^n$. Then, for $t \ge 0$, $\mathbb{P}[|\sum_{i=1}^n a_i X_i| \ge t] \le 2e^{-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)}$ where $K = \max_{1 \le i \le n} \|X_i\|_{\psi_1}$ and c is an absolute constant.
  - For $t \ge 0$, $\mathbb{P}[|\sum_{i=1}^n X_i| \ge t] \le 2e^{-c\min\left(\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_{1 \le i \le n} \|X_i\|_{\psi_1}}\right)}$
    - *Prove by sub-exponential characterisation*
  - For $t \ge 0$, $\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \ge t\right] \le 2e^{-cn\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)}$ where $K = \max_{1 \le i \le n} \|X_i\|_{\psi_1}$
- [Bernstein (Bounded)] Let $X_1, \dots, X_n$ be independent, zero-mean, bounded random variables. Then, for $t \ge 0$, $\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \ge t\right] \le 2e^{-\frac{t^2}{2\left(\sigma^2 + \frac{Kt}{3}\right)}}$ where $\sigma^2 = \sum_{i=1}^n \mathbb{E}[X_i^2]$.

- [Bernstein (Bernoulli)] Let $X_1, \dots, X_n \sim \text{Bernoulli}(p) - p$. Then w.p. $1-\delta$, $\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \le \sqrt{\frac{2p(1-p)\log\left(\frac{2}{\delta}\right)}{n}} + \frac{2\log\left(\frac{2}{\delta}\right)}{3n}$
- [Bernstein (Useful)] Let $X_1, \dots, X_n$ be bounded in an interval of length B and having variance σ^2, then w.p. $1-\delta$, $\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \le \sqrt{\frac{2\sigma^2\log\left(\frac{2}{\delta}\right)}{n}} + \frac{B\log\left(\frac{2}{\delta}\right)}{3n}$
- [Bounded Difference] Let $X_1, \dots, X_n$ be independent random variables and $f: \mathbb{R}^n \to \mathbb{R}$ be measurable. Suppose $|f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x_i', \dots, x_n)| \le c_i$, then $\mathbb{P}[f(X) - \mathbb{E}[f(X)] \ge t] \le e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}$.
- [HDS 2.10] Let X satisfy Bernstein condition with parameter b. Then X is sub-exponential and $\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}}$ for all $|\lambda| < \frac{1}{b}$. Moreover, $\mathbb{P}[|X-\mu| \ge t] \le 2e^{-\frac{t^2}{2(\sigma^2+bt)}}$ $\forall t \ge 0$

Known Results

- [2.3.1 Chernoff's Inequality for Binomial] Let $X_1, \dots, X_n$ be independent with $X_i \sim \text{Bernoulli}(p_i)$. Let $S_n = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[S_n]$. Then for $t > \mu$, $\mathbb{P}[S_n \ge t] \le e^{-\mu}\left(\frac{e\mu}{t}\right)^t$ and for $t < \mu$, $\mathbb{P}[S_n \le t] \le e^{-\mu}\left(\frac{e\mu}{t}\right)^t$
  - *Prove by Chernoff's inequality*

Propositions

- [Sub-Gaussianity in $\|\cdot\|_{\psi_2}$]
  - [Tails] $\mathbb{P}[|X| \ge t] \le 2e^{-\frac{ct^2}{\|X\|_{\psi_2}^2}}$ $\forall t \ge 0$
  - [Moments] $\|X\|_p \coloneqq (\mathbb{E}[|X|^p])^{\frac{1}{p}} \le C\|X\|_{\psi_2}\sqrt{p}$ $\forall p \ge 1$
  - [MGF (Bounded)] $\exists \kappa_4$ s.t. $M_{X^2}\left(\frac{1}{\|X\|_{\psi_2}^2}\right) \coloneqq \mathbb{E}\left[e^{\frac{X^2}{\|X\|_{\psi_2}^2}}\right] \le 2$
  - If $\mathbb{E}[X] = 0$, then $\mathbb{E}\left[e^{\lambda X}\right] \le e^{C\lambda^2\|X\|_{\psi_2}^2}$ $\forall \lambda \in \mathbb{R}$
- [Properties of $\|\cdot\|_{\psi_2}$]
  - $\|\cdot\|_{\psi_2}$ is a norm
  - $\|X\|_{\psi_2} \le C\|X\|_\infty$ where $C = \frac{1}{\sqrt{\ln 2}}$
  - Let $Z \sim N(0, \sigma^2)$, then $\|Z\|_{\psi_2} \le C\sigma$
  - [2.6.1] Let $X_1, \dots, X_n$ be independent, zero-mean sub-Gaussians. Then $\sum_{i=1}^n X_i$ is also sub-Gaussian. Furthermore, $\|\sum_{i=1}^n X_i\|_{\psi_2}^2 \le C\sum_{i=1}^n \|X_i\|_{\psi_2}^2$
- [Sub-Exponentiality in $\|\cdot\|_{\psi_1}$]
  - $\mathbb{P}[|X| \ge t] \ge 2e^{-\frac{ct}{\|X\|_{\psi_1}}}$
  - $(\mathbb{E}[|X|^p])^{\frac{1}{p}} \le cp\|X\|_{\psi_1}$
  - If $\mathbb{E}[X] = 0$, then $\mathbb{E}\left[e^{\lambda X}\right] \le e^{c_1\lambda^2\|X\|_{\psi_1}^2}$ for $|\lambda| \le \frac{c_2}{\|X\|_{\psi_1}}$
  - $\mathbb{E}\left[e^{\frac{|X|}{\|X\|_{\psi_1}}}\right] \le 2$
- [Properties of $\|\cdot\|_{\psi_1}$]
  - $\|\cdot\|_{\psi_1}$ is a norm
  - [2.7.6] A random variable X is sub-Gaussian if and only if X^2 is sub-exponential.
    - $\|X\|_{\psi_2}^2 = \|X^2\|_{\psi_1}$
  - [2.7.7] Let X, Y be sub-Gaussian random variables. Then XY is sub-exponential with $\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$.

- [Centering Lemmas]
    - $\|X - \mathbb{E}[X]\|_2 \leq \|X\|_2$
    - Let X be a sub-Gaussian random variable. Then $X - \mathbb{E}[X]$ is sub-Gaussian with $\|X - \mathbb{E}[X]\|_{\psi_2} \leq C\|X\|_{\psi_2}$ where C is an absolute constant.
        - *Apply norm then Jensen*
    - Let X be a sub-exponential random variable. Then $X - \mathbb{E}[X]$ is sub-exponential with $\|X - \mathbb{E}[X]\|_{\psi_1} \leq C\|X\|_{\psi_1}$ where C is an absolute constant.
- [Sub-Gaussian Maxima Inequality] Let $X_1, \dots, X_n$ be sub-Gaussian random variables, not necessarily independent, s.t. $\mathbb{E}\left[e^{\lambda X_i}\right] \leq e^{\frac{\lambda^2 \sigma_i^2}{2}}$ $\forall \lambda \in \mathbb{R}$. Then $\mathbb{E}\left[\max_{1 \leq i \leq n} X_i\right] \leq \sqrt{2 \log n} \max_{1 \leq i \leq n} \sigma_i$
    - Add in optimisation parameter λ, then apply softmax technique

Orlicz Space

- [Orlicz Function] A function $\psi: [0, \infty) \to [0, \infty)$ is an Orlicz function if ψ is convex, increasing and satisfies $\psi(0) = 0$ and $\lim_{x \to \infty} \psi(x) = \infty$.
- [Orlicz Norm] Let X be a random variable. Then, the Orlicz norm of an Orlicz function ψ is $\|X\|_\psi := \inf_{t > 0}\left\{t : \mathbb{E}\left[\psi\left(\frac{|X|}{t}\right)\right] \leq 1\right\}$.
- [Orlicz Space] Let ψ be an Orlicz function. Then, the Orlicz space $L_\psi := \left\{X : \|X\|_\psi < \infty\right\}$.
    - L_ψ is complete i.e. a Banach space.
    - $L^\infty \subset L_{\psi_2} \subset L^p \ \forall p \in [1, \infty)$
- [Examples]
    - If $\psi(x) = x^p$, then $L_\psi = L_p$
    - $\|\cdot\|_{\psi_2}$ corresponds to $\psi(x) = e^{x^2} - 1$
    - $\|\cdot\|_{\psi_1}$ corresponds to $\psi(x) = e^x - 1$

Random Vectors

Definitions

• [Second Moment Matrix] Let $X \in \mathbb{R}^n$ be a random vector. Then the second moment matrix is: $\Sigma[X] = \mathbb{E}[XX^T] = \sum_{i=1}^n s_i u_i u_i^T$
• [Isotropic] A random vector $X \in \mathbb{R}^d$ is isotropic if $\Sigma[X] = \mathbb{E}[XX^T] = \mathbb{I}_d$.
  ○ $X \in \mathbb{R}^n$ is isotropic if and only if $\mathbb{E}[\langle X, v\rangle^2] = \|v\|_2^2 \ \forall v \in \mathbb{R}^n$
  ○ $X \in \mathbb{R}^n$ is isotropic if and only all one-dimensional marginals of X has unit variance i.e. $\mathbb{E}[\langle X, v\rangle^2] = 1$ for $\forall v: \|v\|_2 = 1$
  ○ If $X \in \mathbb{R}^n$ is isotropic, then $\mathbb{E}[\|X\|_2^2] = n$
  ○ If $X, Y \in \mathbb{R}^n$ are isotropic, then $\mathbb{E}[\langle X, Y\rangle^2] = n$
• [Spherical Distribution] $X \sim \text{Uniform}(\sqrt{n}\mathbb{S}^{n-1})$
  ○ Isotropic, but coordinates of X are not independent
• [Rotation Invariance] Let U be an orthogonal matrix and $v \sim N(0, \mathbb{I}_d)$. Then $Uv \sim N(0, \mathbb{I}_d)$.
• [Sub-Gaussian] Let $X \in \mathbb{R}^d$ be a random vector. Then X is sub-Gaussian if $\langle X, v\rangle$ is a sub-Gaussian random variable $\forall v \in \mathbb{R}^d$.
  ○ i.e. projection of X onto any direction yields a sub-Gaussian random variable
• [Sub-Gaussian Norm] Denote $\|X\|_{\psi_2} = \sup_{v \in \mathbb{S}^{d-1}} \|\langle X, v\rangle\|_{\psi_2}$
• [Sub-Exponential Vector] Let $X \in \mathbb{R}^d$ be a random vector, with not necessarily independent coordinates. Then X is sub-exponential if $\|\langle X, v\rangle\|_{\psi_1} \le C\|\langle X, v\rangle\|_2 \ \forall v \in S^{d-1}$ for some absolute constant C.

Tools (Donsker-Varadhan)

• [Donsker-Varadhan Variational Formula] Let $f: \mathcal{X}, \Theta \to \mathbb{R}$ and π be a fixed distribution on $\Theta \subset \mathbb{R}^d$. Then, with probability $1-\delta$, simultaneously for all measures ρ on Θ with $\text{KL}(\rho\|\pi) < \infty$, $\mathbb{E}_{\theta\sim\rho}[f(X,\theta)] \le \mathbb{E}_{\theta\sim\rho}\left[\log\left(\mathbb{E}_X\left[e^{f(X,\theta)}\right]\right)\right] + \text{KL}(\rho\|\pi) + \log\left(\frac{1}{\delta}\right)$.
  ○ $\mathbb{E}_{\theta\sim\rho}\left[\log \mathbb{E}_X\left[e^{f(X,\theta)}\right]\right]$ is some constant (no randomness)
  ○ $KL(\rho, \pi)$ is the price for uniformity (i.e. depends on the specific distribution ρ)
  ○ In essence, choose some nice measure π s.t. all relevant ρ are s.t. $\rho \ll \pi$
    ▪ $\pi \sim N\left(0, \frac{1}{\beta}\mathbb{I}_d\right)$
    ▪ $\rho_v \sim N\left(v, \frac{1}{\beta}\mathbb{I}_d\right)$
    ▪ $\text{KL}(\rho_v\|\pi) = \frac{\beta}{2}$
• [Donsker-Varadhan Variational Formula] Let $\Theta \subset \mathbb{R}^d$ be a parameter space and π be a measure supported on Θ. Let $h: \Theta \to \mathbb{R}$ be a fixed function. Then: $$\mathbb{E}_{\theta\sim\pi}\left[e^{h(\theta)}\right] = \sup_{\rho: KL(\rho\|\pi)<\infty}\left\{e^{\mathbb{E}_{\theta\sim\rho}[h(\theta)] - \text{KL}(\rho\|\pi)}\right\}$$
  ○ i.e. allows the change of measure from π to ρ, incurring a $\text{KL}(\rho\|\pi)$ penalty.
  ○ The supremum is achieved by taking $\rho = \pi'$ defined as $\pi'(\theta) = \frac{e^{h(\theta)}\pi(\theta)}{\mathbb{E}_{\theta\sim\pi}[e^{h(\theta)}]}$
  ○ $\log\left(\mathbb{E}_{\theta\sim\pi}\left[e^{h(\theta)}\right]\right) = \sup_\rho\{\mathbb{E}_{\theta\sim\rho}[h(\theta)] - KL(\rho\|\pi)\}$

Tools (Gaussian Concentrations)

• Let ϕ be a convex function. Then: $$\mathbb{E}_{X\sim N(0,\mathbb{I}_d)}[\phi(f(X) - \mathbb{E}[f(X)])] \le \mathbb{E}_{X,Y\sim N(0,\mathbb{I}_d), X\perp Y}\left[\phi\left(\frac{\pi}{2}\langle\nabla f(X), Y\rangle\right)\right]$$
  ○ *Prove by interpolation* $Z_k(\theta) = X_k \sin\theta + Y_k\cos\theta$ *for* $\theta \in \left[0, \frac{\pi}{2}\right]$ *and Jensen's*
• [Gaussian Concentration] Let $f: \mathbb{R}^n \to \mathbb{R}$ be a L-Lipschitz function and $X \sim N(0, \mathbb{I}_d)$. Then:
  ○ $f(X)$ is $\frac{\pi L}{2}$-sub-Gaussian
  ○ $\mathbb{P}[f(X) - \mathbb{E}[f(X)] \ge t] \le e^{-\frac{t^2}{2L^2}}$

- $\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}}$
- $\|\cdot\|_2$ is 1-Lipschitz.

- [3.1.1 Corollary] Let $G \sim N(0, \mathbb{I}_d)$. Then $\left\|\|G\|_2 - \sqrt{d}\right\|_{\psi_2} \leq C$.
  - i.e. $\|G\|_2 \approx \sqrt{d}$ with high probability, as expected

Propositions

- [Sub-Gaussian Concentration of Norm] Let X be a sub-Gaussian random vector s.t. $\mathbb{E}[X] = 0$, $\mathbb{E}[XX^T] = \Sigma$, $\mathbb{E}\left[e^{\lambda\langle X,v\rangle}\right] \leq e^{\frac{\lambda^2 v^T \Sigma v}{2}}$. Then, with probability $1 - \delta$, $\|X\|_2 \leq \sqrt{\mathrm{tr}(\Sigma)} + \sqrt{2\lambda_{\max}(\Sigma)\log\left(\frac{1}{\delta}\right)}$
  - *Prove by Donsker-Varadhan and lemmas*
  - Let $X_1, \ldots, X_n \sim N_d(\mu, \Sigma)$. Then $\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right\|_2 \leq \sqrt{\mathrm{tr}\left(\frac{\Sigma}{n}\right)} + \sqrt{\frac{2\,\lambda_{\max}(\Sigma)\log\left(\frac{1}{\delta}\right)}{n}}$
- [Sub-Exponential Concentration of Norm] Let X be a sub-exponential random vector s.t. $\mathbb{E}[X] = 0$, $\mathbb{E}[XX^T] = \Sigma$. Then with probability $1 - \delta$, $\|X\|_2 \leq C\left(\sqrt{\mathrm{tr}(\Sigma)\log\left(\frac{1}{\delta}\right)} + \log\left(\frac{1}{\delta}\right)\sqrt{\lambda_{\max}(\Sigma)}\right)$ where C is some absolute constant.
  - *Prove by Donsker-Varadhan (similar to sub-Gaussian covariance)*
- [Concentration of Norm] Let $X \in \mathbb{R}^d$ be a random vector with independent, sub-Gaussian coordinates X_i satisfying $\mathbb{E}[X_i^2] = 1$. Let $K = \max_{1 \leq i \leq d} \|X_i\|_{\psi_2}$. Then $\left\|\|X\|_2 - \sqrt{d}\right\|_{\psi_2} \leq CK^2$, where C is an absolute constant.
  - $\mathbb{P}\left[\left|\|X\|_2 - \sqrt{d}\right| \geq t\right] \leq 2e^{-c\frac{t^2}{K^4}}\ \forall t \geq 0$
  - *Prove by Bernstein on X_i^2*
- [3.2.3] Let $X \in \mathbb{R}^n$ be a random vector. Then X is isotropic if and only if $\mathbb{E}[\langle X, x\rangle^2] = \|x\|_2^2$ $\forall x \in \mathbb{R}^n$.
  - X is isotropic if and only if all one-dimensional marginal distribution of X have unit variance.
- [3.2.4] Let $X, Y \in \mathbb{R}^n$ be two independent isotropic random vectors. Then $\mathbb{E}[\langle X, Y\rangle^2] = n$
- [3.4.2] Let $X \in \mathbb{R}^n$ with independent, zero-mean, sub-Gaussian coordinates. Then X is sub-Gaussian with $\|X\|_{\psi_2} \leq C \max_{1 \leq i \leq n} \|X_i\|_{\psi_2}$
- [3.4.6] Let $X \sim \mathrm{Uniform}\left(\sqrt{n}\mathbb{S}^{n-1}\right)$. Then X is sub-Gaussian and $\|X\|_{\psi_2} \leq C$, where C is an absolute constant (independent of n).

Random Matrices

Definitions

- [Operator Norm] Let $A \in \mathbb{R}^{m\times n}$, then $\|A\| \coloneqq \max_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|Av\|}{\|v\|} = \max_{v\in S^{n-1}} \|Av\|$
  - Equivalently, $\|A\| = \max_{u\in\mathbb{R}^m, v\in\mathbb{R}^n:\|u\|=\|v\|=1} u^T Av = \max_{u\in S^{m-1}, v\in S^{n-1}} u^T Av$
  - Equivalently, $\|A\| = \sigma_1(A)$
- [Frobenius Norm] $\|A\|_F = \left(\sum_{i,j}^{m,n} A_{ij}^2\right)^{\frac{1}{2}}$
- [ϵ-Cover] Let $K \subset \mathbb{R}^d$, then an ϵ-cover w.r.t. distance ρ is a subset $N_\epsilon \subset K$ s.t. $\forall x \in K$, $\exists x_0 \in N_\epsilon$ s.t. $\rho(x, x_0) \leq \epsilon$
- [ϵ-Separated] Let $K \subset \mathbb{R}^d$ equipped with distance ρ, then an ϵ-separated set P_ϵ is s.t. $\forall x_1 \neq x_2 \in P_\epsilon$, $\rho(x_1, x_2) > \epsilon$.
- [Covering Number] The covering number $\mathcal{N}(K, \rho, \epsilon)$ of a set K equipped with a distance function ρ is the smallest cardinality of an ϵ-cover.
- [Packing Number] The packing number $\mathcal{P}(K, \rho, \epsilon)$ of a set K equipped with a distance function ρ is the largest cardinality of an ϵ-separated set.
- [Effective Rank] Let Σ be a covariance matrix. Then the effective rank of Σ is $r(\Sigma) = \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}$.
- [Spectral Mapping Theorem] Let $f: I \to \mathbb{R}$ be a function on $I \subset \mathbb{R}$ and A be a Hermitian matrix whose eigenvalues are contained in I. If λ is an eigenvalue of A, then $f(\lambda)$ is an eigenvalue of $f(A)$.
- Let $f: \mathbb{R} \to \mathbb{R}$ be a function. Then:
  - [Operator Monotone] $f(A) \preccurlyeq f(B)$ whenever $A \preccurlyeq B$
  - [Operator Concave] $\lambda f(A) + (1-\lambda) f(B) \preccurlyeq f(\lambda A + (1-\lambda)B)$ $\forall \lambda \in [0,1]$ $\forall A, B$
- Let X be a symmetric random matrix. Then:
  - [Moment Generating Function] $M_X(\lambda) \coloneqq \mathbb{E}\left[e^{\lambda X}\right]$
  - [Cumulant Generating Function] $\Xi_X(\theta) \coloneqq \log \mathbb{E}\left[e^{\theta X}\right]$

Covering and Packing Numbers

- $\mathcal{P}(K, \rho, 2\epsilon) \leq \mathcal{N}(K, \rho, \epsilon) \leq \mathcal{P}(K, \rho, \epsilon)$
- $\left(\frac{1}{\epsilon}\right)^d \leq \mathcal{N}\left(B_2^d, \|\cdot\|_2, \epsilon\right) \leq \left(1 + \frac{2}{\epsilon}\right)^d$
- Let $A \in \mathbb{R}^{m\times n}$ and $\mathcal{M}, \mathcal{N}$ be nets of S^{m-1} and S^{n-1} respectively. Then $\sup_{u\in\mathcal{M}, v\in\mathcal{N}} u^T Av \leq \|A\| \leq \frac{1}{1-2\epsilon} \sup_{u\in\mathcal{M}, v\in\mathcal{N}} u^T Av$.

Matrix Calculus

- [Properties (Deterministic)]
  - Let $f, g: \mathbb{R} \to \mathbb{R}$ satisfy $f(x) \leq g(x)$ $\forall x \in [l, u]$. Suppose A is symmetric and eigenvalues of A all lie in $[l, u]$. Then $f(A) \preccurlyeq g(A)$.
  - $\|A\|_{\mathrm{op}} \leq \|A\|_2 \leq \sqrt{r}\|A\|_{\mathrm{op}}$
  - Let $C \succcurlyeq B \succcurlyeq 0$ and $A \succcurlyeq 0$, then $\mathrm{tr}(AB) \leq \mathrm{tr}(AC)$
  - Let f be a monotone function. Then $\mathrm{tr} \circ f$ is also monotone.
    - $A \succcurlyeq B$ implies $\mathrm{tr}\big(f(A)\big) \geq \mathrm{tr}\big(f(B)\big)$
  - $\|X\|_{\mathrm{op}} \leq t \Leftrightarrow -t\mathbb{I} \preccurlyeq X \preccurlyeq t\mathbb{I}$
  - Let $f: \mathbb{R} \to \mathbb{R}$ be increasing and X, Y be commuting matrices. Then $X \preccurlyeq Y \Rightarrow f(X) \preccurlyeq f(Y)$.
    - Commuting symmetric matrices are simultaneously diagonalisable by an orthogonal matrix.
  - Let $f, g: \mathbb{R} \to \mathbb{R}$ be two functions and $f(x) \leq g(x)$ $\forall x \in \mathbb{R}$ satisfying $|x| \leq K$. Then, $f(X) \preccurlyeq g(X)$ for $\|X\|_{\mathrm{op}} \leq K$.
  - Let $X \preccurlyeq Y$, then $\mathrm{tr}(X) \preccurlyeq \mathrm{tr}(Y)$.

- [Weyl Monotonicity] Let A, B be symmetric matrices. Let $\lambda_i(A)$ denote the ith largest eigenvalue of A.
  - $\lambda_{i+j-1}(A+B) \le \lambda_i(A) + \lambda_j(B) \le \lambda_{i+j-n}(A+B)$
  - If $A \preccurlyeq B$, then $\lambda_i(A) \le \lambda_i(B)\ \forall i$
- [Properties (Random)]
  - Let A, B be random matrices with $A \succcurlyeq B$. Then $\mathbb{E}[A] \succcurlyeq \mathbb{E}[B]$.
- [Exponentiation] $e^A := \mathbb{I} + \sum_{i=1}^{\infty} \frac{1}{i!} A^i$
  - [Golden-Thompson] $\mathrm{tr}(e^{A+B}) \le \mathrm{tr}(e^A e^B)$
  - If $A \preccurlyeq B$, then $\mathrm{tr}(e^A) \le \mathrm{tr}(e^B)$
  - [Warning!] $e^{X_1+X_2} \ne e^{X_1} e^{X_2}$ unless $X_1 X_2 = X_2 X_1$
- [Logarithm] The function $f(x) = \log x$ is operator concave
  - $\log\big((1-\lambda)A + \lambda B\big) \ge (1-\lambda)\log A + \lambda \log B$
  - $\log(e^A) = A$
- [Lieb] Let $H \in \mathbb{R}^{d\times d}$ be a fixed symmetric matrix. Then, $f: A \to \mathrm{tr}\big(e^{H+\log A}\big)$ is concave on the space of symmetric, positive definite matrices.
  - $\mathbb{E}\big[\mathrm{tr}\big(e^{\lambda \sum_{i=1}^n X_i}\big)\big] \le \mathrm{tr}\left(e^{\sum_{i=1}^n \log \mathbb{E}[e^{\lambda X_i}]}\right)$
  - $\mathbb{E}[\mathrm{tr}(e^{H+X})] \le \mathrm{tr}\big(e^{H+\log \mathbb{E}[e^X]}\big)$
- Let X be a $d \times d$ symmetric, zero-mean matrix s.t. $\|X\|_{\mathrm{op}} \le K$ a.s. Then $\mathbb{E}\big[e^{\lambda X}\big] \preccurlyeq e^{g(\lambda)\mathbb{E}[X^2]}$ where $g(\lambda) = \frac{\frac{\lambda^2}{2}}{1-\frac{|\lambda|K}{3}}$ provided $|\lambda| \le \frac{3}{K}$.

Matrix Toolkits

- [Matrix Laplace Transform] Let X be a symmetric random matrix. Then $\forall t \in \mathbb{R}$, $\mathbb{P}[\lambda_{\max}(X) \ge t] \le \inf_{\lambda>0} \frac{\mathbb{E}[\mathrm{tr}(e^{\lambda X})]}{e^{\lambda t}}$
  - i.e. can control extreme eigenvalues of X via the trace of MGF
  - *Prove by scalar Chernoff's method*
- [Matrix Chernoff] Let X be a symmetric random matrix. Then $\forall t \in \mathbb{R}$, $\mathbb{P}[\lambda_{\max}(\sum_{i=1}^n X_i) \ge t] \le \inf_{\lambda>0} e^{-\lambda t} \mathrm{tr}\left(e^{\sum_{i=1}^n \log \mathbb{E}[e^{\lambda X_i}]}\right)$
  - *Prove by Matrix Laplace + Lieb on the conditional expectation*
  - $\mathbb{E}\big[\|X\|_{\mathrm{op}}\big] \le \inf_{\lambda>0} \left\{\frac{1}{\lambda} \log\big(\mathbb{E}\big[\lambda_{\max}\big(e^{\lambda X}\big) + \lambda_{\max}\big(e^{-\lambda X}\big)\big]\big)\right\}$
    - *Prove by Jenson and that* $e^{\lambda \|X\|_{\mathrm{op}}} \le \lambda_{\max}\big(e^{\lambda X}\big) + \lambda_{\max}\big(e^{-\lambda X}\big)$
- [Matrix Bernstein] Let $(X_n)_n$ be a sequence of independent, random, zero-mean, symmetric matrices with $X_i \in \mathbb{R}^{d\times d}$ and $\mathbb{E}[X_i] = 0$, $\|X_i\|_{\mathrm{op}} \le K$ a.s. $\forall i$ and $\sigma^2 = \left\|\sum_{i=1}^n \mathbb{E}[X_i^2]\right\|_{\mathrm{op}}$. Then, for $t \ge 0$, $\mathbb{P}\left[\|\sum_{i=1}^n X_i\|_{\mathrm{op}} \ge t\right] \le 2d e^{\left(-\frac{\frac{t^2}{2}}{\sigma^2 + \frac{Kt}{3}}\right)}$.
  - $\mathbb{E}\left[\|\sum_{i=1}^n X_i\|_{\mathrm{op}}\right] \le \sqrt{2\log(2d)\,\sigma^2} + \frac{2}{3}\log(2d)\,K$
  - *Prove by applying algebraic result* $e^x \le 1 + x + \frac{1}{1-\frac{|x|}{3}} \frac{x^2}{2}$
- [Matrix Hoeffding] Let $\epsilon_1, \dots, \epsilon_n$ be independent Rademacher random variables and $A_1, \dots, A_n$ be symmetric $d \times d$ deterministic matrices. Then for $t \ge 0$ and $\sigma^2 = \left\|\sum_{i=1}^n A_i^2\right\|_{\mathrm{op}}$

$$\mathbb{P}\left[\left\|\sum_{i=1}^n \epsilon_i A_i\right\|_{\mathrm{op}} \ge t\right] \le 2d e^{-\frac{t^2}{2\sigma^2}}$$

- [Matrix Khintchine] Let $\epsilon_1, \dots, \epsilon_n$ be independent Rademacher random variables and $A_1, \dots, A_n$ be symmetric $d \times d$ deterministic matrices. Then, with $\sigma^2 = \left\|\sum_{i=1}^n A_i^2\right\|_{\mathrm{op}}$:

- $\mathbb{E}\left[\|\sum_{i=1}^{n} \epsilon_i A_i\|_{\text{op}}\right] \leq C\sqrt{\sigma^2 \log d}$

Results

- Let $X \in \mathbb{R}^{m\times n}$ with X_{ij} being independent sub-Gaussian elements s.t. $\mathbb{E}[X_{ij}] = 0$ and $K = \max_{i,j}\|X_{ij}\|_{\psi_2} < \infty$. Then $\|X\|_{\text{op}} \leq CK\left(\sqrt{n} + \sqrt{m} + \sqrt{\log\left(\frac{1}{\delta}\right)}\right)$ where C is an absolute constant.
  - *Prove by ϵ-net argument.*
- Let $X_1, \dots, X_n$ be sub-Gaussian, zero-mean, independent samples with covariance matrix Σ. Then $\left\|\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T - \Sigma\right\|_{\text{op}} \leq C\|\Sigma\|_{\text{op}}\left(\sqrt{\frac{r(\Sigma)}{n}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\right)$ with probability $1-\delta$ whenever $n \geq C_1\left(r(\Sigma) + \log\left(\frac{1}{\delta}\right)\right)$, where C is an absolute constant.

Applications

- [Unconstrained OLS]
- [Constrained OLS] $\mathbb{E}\left[\frac{1}{n}\left\|X\hat{\beta} - X\beta^*\right\|_2^2\right] \leq \min_{\beta\in K}\frac{1}{n}\|X\beta - X\beta^*\|_2^2 + \frac{4\sigma^2 d}{n}$

Vapnik-Chervonenkis Dimension

Definitions

• [Shatter Function] Let $\mathcal{A}$ be a collection of subsets of $\mathcal{X}$. Then: $$S_{\mathcal{A}}(n) = \max_{x_1,\dots,x_n \in \mathcal{X}} \left|\{(\mathbb{1}\{x_1 \in A\}, \dots, \mathbb{1}\{x_n \in A\}) : A \in \mathcal{A}\}\right|$$ • [VC Dimension of $\mathcal{A}$] The <u>VC dimension of $\mathcal{A}$</u> is the largest d s.t. $S_{\mathcal{A}}(d) = 2^d$
• [Shatter Function] Let $\mathcal{F}$ be a family of boolean functions i.e. $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$. Then: $$S_{\mathcal{F}}(n) = \max_{x_1,\dots,x_n \in \mathcal{X}} \left|\{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\}\right|$$ • [Shattered] Let $\mathcal{F}$ be a class of boolean functions from $\mathcal{X}$ to $\{0,1\}$. A subset $\Lambda \subset \mathcal{X}$ is <u>shattered</u> by $\mathcal{F}$ if $\forall g : \Lambda \to \{0,1\}$, $\exists f \in \mathcal{F}$ s.t. $f\|_{\Lambda} = g$.
• [Shatter Number] $N_{\mathcal{F}}(x_1, \dots, x_n) = \left|\{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\}\right|$
• [VC Dimension of $\mathcal{F}$] The <u>VC dimension of $\mathcal{F}$</u> is the largest cardinality of a subset $\Lambda \subset \Omega$ shattered by $\mathcal{F}$.
  ○ $\mathrm{VC}(\mathcal{F}) = \max\{n : \exists x_1, \dots, x_n \text{ s.t. } \left|\{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\}\right| = 2^n\}$
  ○ $\mathrm{VC}(\mathcal{F}) = \max\{n : \exists x_1, \dots, x_n \text{ s.t. } N_{\mathcal{F}}(x_1, \dots, x_n) = 2^n\}$
  ○ $\mathrm{VC}(\mathcal{F}) = \max\{n : S_{\mathcal{F}}(n) = 2^n\}$
  ○ i.e. VC dimension of $\mathcal{F}$ is the size of the largest shattered set

Theorems

• [Properties of Shatter Functions]
  ○ $S_{\mathcal{A}}(n) \le 2^n$
  ○ If $\|\mathcal{A}\| < \infty$, then $S_{\mathcal{A}}(n) \le \|\mathcal{A}\|$
• [Radon] Let there be $p+2$ points in $\mathbb{R}^p$. Then, exists a grouping of these points into two groups A, B disjoint s.t. their convex hulls intersect.
• [Lemma] Let $\mathcal{F}$ be a class of functions s.t. $\forall f \in \mathcal{F}$, $\|f(x)\| \le b$ a.s., then: $\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\}\right] \le b\sqrt{\frac{2\log(2S_{\mathcal{F}}(n))}{n}}$
  ○ Prove by maximum of sub-Gaussians applied to ϵ
• [VC Bound] Let $\mathcal{F}$ be a class of boolean functions with VC dimension d. Then, with probability $1-\delta$, $\left\|R(\hat{f}) - R(f^*)\right\| \le C\left(\sqrt{\frac{d\log\left(\frac{en}{d}\right)}{n}} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}}\right)$.
• [Pajor] Let $\mathcal{F}$ be a class of Boolean functions on a finite set Ω. Then: $$\|\mathcal{F}\| \le \|\{\Lambda \subset \Omega : \Lambda \text{ is shattered by } \mathcal{F}\}\|$$ • [Sauer-Shelah] Let $\mathcal{A}$ have VC dimension $d < \infty$. Then $S_{\mathcal{A}}(n) \le \sum_{i=1}^d \binom{n}{i} \le \left(\frac{en}{d}\right)^d$
  ○ Intuitively, a collection $\mathcal{A}$ with finite VC dimension has a shatter function that grows at most polynomially, instead of exponentially
  ○ Sets $\mathcal{A}$ with finite VC dimension satisfy uniform law of large numbers
  ○ *Prove by combinatorics*
  ○ [Corollary] Let $\mathcal{A}$ have VC dimension d. Then, with probability $1-\delta$, $$\sup_{A \in \mathcal{A}} \|\mathbb{P}_n[A] - \mathbb{P}[A]\| \le 4\sqrt{\frac{d\log\left(\frac{en}{d}\right)}{n}} + \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}$$ ▪ Prove by lemmas.
• [Dvoretzky-Kiefer-Wolfowitz] Let $F(x)$ be the true CDF and $F_n(x)$ be the empirical CDF. Then, with probability $1-\delta$, $\sup_{t \in \mathbb{R}} \|F_n(t) - F(t)\| \le C\left(\frac{1}{\sqrt{n}} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}}\right) \le C'\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}}$, where C and C' are absolute constants.
  ○ *Prove by symmetrisation, bounded difference, bound for Rademacher complexity with VC*

- [Warren] The VC dimension of a binary class induced by polynomials of d variables and power at most p is $\leq 2d\log(12p)$.
  - Let $\mathcal{F}$ be a function class where elements are of the form $f(x) = \mathbb{1}\{P(x) \geq 0\}$ where $P(x)$ is a polynomial of max degree p and of d variables.

Common VC Dimension Examples

- [Intervals] Let $\mathcal{F} = \left\{\mathbb{1}_{[a,b]}: a, b \in \mathbb{R}, a \leq b\right\}$. Then $\mathrm{VC}(\mathcal{F}) = 2$.
- [Half-Intervals] Let $\mathcal{F} = \left\{\mathbb{1}_{(-\infty,t]}: t \in \mathbb{R}\right\}$. Then $\mathrm{VC}(\mathcal{F}) = 1$.
- [Half-Spaces] The VC dimension of half-spaces in $\mathbb{R}^n$ is $n + 1$.

Metric Entropy

Key Idea
• Bound terms like: $\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\}\right]$ • [Dudley] The process $\left(\sum_{i=1}^n \epsilon_i f(x_i)\right)_{f\in\mathcal{F}}$ is a sub-Gaussian process indexed by f with metric $d(f,g)^2 = \sum_{i=1}^n \left(f(x_i)-g(x_i)\right)^2 = n\\|f-g\\|_{L_2(\hat{\mathcal{P}}_n)}^2$. ○ Any bound on $\mathcal{N}(\mathcal{F},\mathcal{P}_n,\epsilon)$ implies a bound on $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\}\right]$ ○ $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\sum_{i=1}^n \epsilon_i f(x_i)\right\}\right] \le 2\mathbb{E}\left[\sup_{\substack{f,g\in\mathcal{F}\\ d(f,g)\le\delta}} d(f,g)\right] + 16\int_{\frac{\delta}{4}}^{\frac{\mathrm{diam}_d(\mathcal{F})}{2}}\sqrt{\log\mathcal{N}(\mathcal{F},d,\epsilon)}\,\mathrm{d}\epsilon$ ○ $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\}\right] \le 2\delta + \frac{16}{\sqrt{n}}\int_{\frac{\delta}{4}}^{\frac{\mathrm{diam}_{L_2(\mathcal{P}_n)}(\mathcal{F})}{2}}\sqrt{\log\mathcal{N}(\mathcal{F},L_2(\mathcal{P}_n),\epsilon)}\,\mathrm{d}\epsilon$

Definitions
• [Gaussian Process] A stochastic process $(X_t)_{t\in T}$, where T is an index set, is a Gaussian process if for every finite set of indices $t_1,\dots,t_k\in T$, the distribution of $\left(X_{t_1},\dots,X_{t_k}\right)$ is multivariate Gaussian. • [Sub-Gaussian Process A] The process $(X_t)_{t\in\mathcal{T}}$ w.r.t. the metric $d(t,s)$ on $\mathcal{T}$ is <u>sub-Gaussian</u> if: ○ $(X_t)_{t\in\mathcal{T}}$ is zero-mean i.e. $\mathbb{E}[X_t]=0$ ○ $\forall s,t\in\mathcal{T}$, $\mathbb{E}\left[e^{\lambda(X_s-X_t)}\right]\le e^{\frac{\lambda^2 d(t,s)^2}{2}}$ • [Sub-Gaussian Process B] The process $(X_t)_{t\in\mathcal{T}}$ w.r.t. the metric $d(t,s)$ on $\mathcal{T}$ is <u>sub-Gaussian</u> if $\exists c>0$ absolute constant s.t. $\\|X_t-X_s\\|_{\psi_2}\le c\,d(t,s)$ • Let $\mathcal{T}\subset\mathbb{R}^d$. ○ [Gaussian Width] $\mathcal{W}(\mathcal{T}) = \mathbb{E}\left[\sup_{t\in\mathcal{T}}\langle g,t\rangle\right] = \mathbb{E}\left[\sup_{t\in\mathcal{T}}\sum_{i=1}^n g_i t_i\right]$ ○ [Rademacher Average] $\mathbb{E}\left[\sup_{t\in\mathcal{T}}\langle \epsilon,t\rangle\right] = \mathbb{E}\left[\sup_{t\in\mathcal{T}}\sum_{i=1}^n \epsilon_i t_i\right]$ ○ [Gaussian Complexity] $\mathbb{E}\left[\sup_{t\in\mathcal{T}}\lvert\langle g,t\rangle\rvert\right] = \mathbb{E}\left[\sup_{t\in\mathcal{T}}\left\lvert\sum_{i=1}^n g_i t_i\right\rvert\right]$ ○ [Rademacher Complexity] $\mathbb{E}\left[\sup_{t\in\mathcal{T}}\lvert\langle \epsilon,t\rangle\rvert\right] = \mathbb{E}\left[\sup_{t\in\mathcal{T}}\left\lvert\sum_{i=1}^n \epsilon_i t_i\right\rvert\right]$ • [Empirical Rademacher Complexity] The <u>empirical Rademacher complexity</u> of a class of functions $\mathcal{F}=\{f:\mathcal{X}\to\mathbb{R}\}$ is $R_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\lvert\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\rvert\right]$ • [Full Rademacher Complexity] The <u>full Rademacher complexity</u> of a class of functions $\mathcal{F}=\{f:\mathcal{X}\to\mathbb{R}\}$ is $R(\mathcal{F}) = \mathbb{E}_X\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\lvert\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\rvert\right]\right] = \mathbb{E}_X[R_n(\mathcal{F};X)]$ • [Function Norms] Let $\mathcal{P}$ be a distribution and $\mathcal{P}_n$ be the empirical distribution (by sampling $X_1,\dots,X_n\sim\mathcal{P}$). Then the following two are norms: ○ $\\|f\\|_{L^2(\mathcal{P})}^2 \coloneqq \mathbb{E}_{X\sim\mathcal{P}}[f^2(X)]$ ○ $\\|f\\|_{L^2(\mathcal{P}_n)}^2 \coloneqq \frac{1}{n}\sum_{i=1}^n\left(f(x_i)\right)^2$ • [Parametric Class of Functions] A family of functions $\mathcal{F}$ is a <u>parametric class of functions</u> if $\sup_{\mathcal{P}_n}\mathcal{N}\left(\mathcal{F},\\|\cdot\\|_{L_2(\mathcal{P}_n)},\epsilon\right)\le\left(\frac{C}{\epsilon}\right)^p$ where C is an absolute constant. ○ p can be thought of as dimension ($\uparrow p \Rightarrow$ more complex the class $\mathcal{F}$ is)

- [Nonparametric Class] A family of functions $\mathcal{F}$ is nonparametric if $\sup_{\mathcal{P}_n} \log\left(\mathcal{N}\left(\mathcal{F}, \|\cdot\|_{L_2(\mathcal{P}_n)}, \epsilon\right)\right) \lessapprox \left(\frac{C}{\epsilon}\right)^p$ for some p
  - $\log \mathcal{N}\left(\mathcal{F}, \|\cdot\|_{L_2(\mathcal{P}_n)}, \epsilon\right) \leq C'\epsilon^{-p}\ \forall \mathcal{P}_n$
- Let $(X, Y) \sim \mathcal{P}$ and $f: \mathcal{X} \to \mathbb{R}$. Let $l(f(X), Y)$ be a loss function. Then:
  - [Population Risk] The population risk is $R(f) = \mathbb{E}[l(f(X), Y)]$
  - [Empirical Risk] The empirical risk is $R_n(f) = \frac{1}{n}\sum_{i=1}^n l(f(X_i), Y_i)$
  - [Excess Risk] The excess risk is $\hat{\xi} = R(\hat{f}) - \inf_{f\in\mathcal{F}} R(f)$ where $\hat{f} = \arg\min_{f\in\mathcal{F}} R_n(f)$
  - [Generalisation Error] The generalisation error is $\left|R_n(\hat{f}) - R(\hat{f})\right|$.
- [Bracket] Let $\mathcal{F}$ be a function class, $\mathcal{P}$ be a fixed distribution over $\mathcal{X}$ and $L_q(\mathcal{P})$ be a norm. A bracket is a pair of functions $l, u: \mathcal{X} \to \mathbb{R}$, not necessarily in $\mathcal{F}$, s.t. $[l, u] = \{f \in \mathcal{F}: l(x) \leq f(x) \leq u(x)\ \forall x \in \mathcal{X}\}$
- [ϵ-Bracket] Let $\mathcal{P}$ be a distribution over $\mathcal{X}$. Then $[l, u]$ is an ϵ-bracket if $\|u - l\|_{L_q(\mathcal{P})} \leq \epsilon$
  - $\|u - l\|_{L_q(\mathcal{P})} = \left(\mathbb{E}_{X\sim\mathcal{P}}\left[\left(u(X) - l(X)\right)^q\right]\right)^{\frac{1}{q}}$
- [Bracketing Entropy] $\mathcal{N}_{[\,]}\left(\mathcal{F}, \|\cdot\|_{L_q(\mathcal{P})}, \epsilon\right)$ is the minimum number of ϵ-brackets to cover $\mathcal{F}$

Gaussian Processes

- [Stein's Lemma]
  - Let $X \sim N(\mu, \sigma^2)$ and $f: \mathbb{R} \to \mathbb{R}$ differentiable and $\mathbb{E}[|f'(X)|] < \infty$, then $\mathbb{E}[(X - \mu)f(X)] = \sigma^2\mathbb{E}[f'(X)]$
  - Let $X \sim N_d(\mu, \sigma^2\mathbb{I}_d)$ and $f: \mathbb{R}^d \to \mathbb{R}^d$ differentiable and $\mathbb{E}[\|\mathrm{D}f(X)\|_F] < \infty$, then $\mathbb{E}[(X - \mu)^T f(X)] = \sigma^2\mathbb{E}\left[\mathrm{tr}\left(\mathrm{D}f(X)\right)\right] = \sigma^2\sum_{i=1}^d \mathbb{E}\left[\frac{\partial f_i}{\partial x_i}(X)\right]$
- [Slepian Gaussian Comparison] Let $(X_t)_t$ and $(Y_t)_t$ be two zero-mean Gaussian processes s.t. $\mathbb{E}[X_t^2] = \mathbb{E}[Y_t^2]$ and $\forall t, s \in T$, $\mathbb{E}[(X_t - X_s)^2] \leq \mathbb{E}[(Y_t - Y_s)^2]$. Then:
  - $\forall t \in T$, $\mathbb{P}\left[\sup_{t\in T} X_t \geq t\right] \leq \mathbb{P}\left[\sup_{t\in T} Y_t \geq t\right]$
  - $\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq \mathbb{E}\left[\sup_{t\in T} Y_t\right]$
- [Sudakov-Fernique] Let $(X_t)_t$ and $(Y_t)_t$ be two Gaussian processes with zero means s.t. $\forall t, s \in T$, $\mathbb{E}[(X_t - X_s)^2] \leq \mathbb{E}[(Y_t - Y_s)^2]$. Then $\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq \mathbb{E}\left[\sup_{t\in T} Y_t\right]$.
  - It is the same as Slepian Gaussian Comparison with one assumption dropped
  - Approximate supremum with softmax
- [Sudakov Minoration] Let $(X_t)_t$ be a zero-mean Gaussian process. Define $d(t, s) \coloneqq \sqrt{\mathbb{E}[(X_t - X_s)^2]}$. Then, $\mathbb{E}\left[\sup_{t\in T} X_t\right] \geq C\epsilon\sqrt{\log \mathcal{N}(T, d, \epsilon)}$.
- [Gaussian Contraction] Let $\phi_1, \ldots, \phi_d: \mathbb{R} \to \mathbb{R}$ be L-Lipschitz functions. Then: $\mathbb{E}\left[\sup_{t\in\mathcal{T}} \sum_{i=1}^d g_i\phi_i(t_i)\right] \leq L\mathbb{E}\left[\sup_{t\in\mathcal{T}} \sum_{i=1}^d g_i t_i\right] = L\mathcal{W}(\mathcal{T})$

Symmetrisation Lemmas

- Let $\epsilon_1, \ldots, \epsilon_n$ be independent Rademacher random variables with $x_1, \ldots, x_n \sim \mathcal{P}$. Then:
  - $\mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[f(x)] - \frac{1}{n}\sum_{i=1}^n f(x_i)\right\}\right] \leq 2\mathbb{E}_X\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\}\right]\right]$
  - $\mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}[f(x)]\right\}\right] \leq 2\mathbb{E}_X\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\}\right]\right]$
  - $\mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}[f(x)] - \frac{1}{n}\sum_{i=1}^n f(x_i)\right|\right] \leq 2\mathbb{E}_X\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right|\right]\right]$

Theorems

- [Dudley Integral] Let $(X_t)_{t\in\mathcal{T}}$ be a sub-Gaussian process w.r.t. metric $d(t,s)$. Define $\operatorname{diam}(\mathcal{T}) = \sup_{t,s\in\mathcal{T}} d(t,s)$. Fix $\delta > 0$. Denote $\operatorname{diam}(\mathcal{T}) = \sup_{t,s\in\mathcal{T}} d(t,s)$. Then:
$$\mathbb{E}\left[\sup_{t\in\mathcal{T}} X_t\right] \le 2\mathbb{E}\left[\sup_{\substack{t,s\in\mathcal{T}\\ d(t,s)\le\delta}} d(t,s)\right] + 16\int_{\frac{\delta}{4}}^{\frac{\operatorname{diam}(\mathcal{T})}{2}} \sqrt{\log\mathcal{N}(\mathcal{T},d,\epsilon)}\,\mathrm{d}\epsilon$$
- [Chaining] Let N be a net of $\mathcal{T}$. Define $N_j \subset \mathcal{T}$ at scale $\frac{1}{2^j}\operatorname{diam}(\mathcal{T})$. Let m be the smallest integer s.t. $\frac{1}{2^m} \le \delta$. Then: $\sup_{t,s\in N}\{X_t - X_s\} \le 16\int_{\frac{\delta}{4}}^{\frac{\operatorname{diam}(\mathcal{T})}{2}} \sqrt{2\log\mathcal{N}(\mathcal{T},d,\epsilon)}$
- [Nonparametric Classes Result] Let $\mathcal{F}$ be a nonparametric class with parameter p. Then:
  - If $p < 2$, $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}[f(X)]\right\}\right] \le \frac{C}{\sqrt{n}}$
  - If $p > 2$, $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}[f(X)]\right\}\right] \le cn^{-\frac{1}{p}}$
- [Ledoux-Talagrand] Let $\phi_1,\dots,\phi_n$ be L-Lipschitz functions with $\phi_i:\mathbb{R}\to\mathbb{R}$ and $\phi_i(0)=0$. Then, for any $\mathcal{T}\subset\mathbb{R}^n$:
$$\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\phi_i(t_i)\right|\right] \le 2L\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i t_i\right|\right]$$
  - i.e. Lipschitz functions can be "erased" in place of their Lipschitz constants
- [Contraction] Let $\phi_1,\dots,\phi_n$ be L-Lipschitz functions with $\phi_i:\mathbb{R}\to\mathbb{R}$. Then, for any $\mathcal{T}\subset\mathbb{R}^n$:
$$\mathbb{E}\left[\sup_{t\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^n \epsilon_i\phi_i(t_i)\right] \le L\mathbb{E}\left[\sup_{t\in\mathcal{T}}\frac{1}{n}\sum_{i=1}^n \epsilon_i t_i\right]$$
- [Dudley Bound for Empirical Processes] Let $\mathcal{F}$ be a class of functions s.t. $\forall f\in\mathcal{F}$, $|f(x)|\le b$ a.s. Then, with probability $1-\delta$, $\sup_{f\in\mathcal{F}}\left|\mathbb{E}[f(x)] - \frac{1}{n}\sum_{i=1}^n f(x_i)\right| \le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}[f(x)] - \frac{1}{n}\sum_{i=1}^n f(x_i)\right|\right] + b\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}$
  - *Prove by bounded difference inequality*
- $\mathcal{N}\left(\mathcal{F},\|\cdot\|_{L_q(\mathcal{P})},\epsilon\right) \le \mathcal{N}_{[\,]}\left(\mathcal{F},\|\cdot\|_{L_q(\mathcal{P})},\epsilon\right)$
  - The set of $\left\{f_{l,u}(x) = \frac{u(x)+l(x)}{2}\right\}_{l,u}$ is an ϵ-cover of $\mathcal{F}$.
- [Bracketing Theorem] Let $\mathcal{F}$ be a class of functions s.t. $\|f\|_{L^\infty(\mathcal{P})} \le m$. Let $X_1,\dots,X_n$ be i.i.d. sample of X, then $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}[f(x)]\right|\right] \le \frac{c}{\sqrt{n}}\int_0^m \sqrt{\log\mathcal{N}_{[\,]}\left(\mathcal{F},\|\cdot\|_{L_2(\mathcal{P})},\epsilon\right)}\,\mathrm{d}\epsilon$
  - No Rademacher terms as compared to symmetrisation + Dudley
  - No need to find $\sup_{\mathcal{P}_n}\mathcal{N}\left(\mathcal{F},\|\cdot\|_{L_2(\mathcal{P}_n)},\epsilon\right)$ i.e. over all empirical distributions
  - Rates of convergence may be bad for some function classes $\mathcal{F}$

Applications

- [Excess Risk] Let $R(f) = \mathbb{E}[l(f(X),Y)]$ where $l(\cdot,Y)$ is L-Lipschitz. Then:
$$\mathbb{E}[\hat{\xi}] \le \mathbb{E}\left[\sup_{f\in\mathcal{F}}\{R(f)-R_n(f)\}\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}}\{R_n(f)-R(f)\}\right] \le 4\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i l(f(X_i),Y_i)\right\}\right]$$
$$\le 4L\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right\}\right]$$
  - *Prove by symmetrisation and contraction*

- [Bounded, Parametric $\mathcal{F}$] Let $\mathcal{F}$ be a bounded, parametric class of functions with $\|f\|_\infty \leq 1$ and dimension p. Then $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\}\right] \leq C\sqrt{\frac{p}{n}}$
- [] Let $\mathcal{F}$ be a class of $\{0,1\}$-valued functions with VC dimension d. Then:
  - Then $\exists C > 0$ absolute constant s.t. $\sup_{\mathcal{P}_n} \mathcal{N}\left(\mathcal{F}, \|\cdot\|_{L_2(\mathcal{P}_n)}, \epsilon\right) \leq \left(\frac{C}{\epsilon}\right)^{4d}$
  - $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\}\right] \leq C\sqrt{\frac{d}{n}}$
  - *Prove by bounding packing number using probabilistic method, then apply Dudley*
- Let $\mathcal{F} = \{f_w : x \mapsto \langle w, x\rangle | w \in b \cdot B_2^d\}$ and $\|x\| \leq r$ a.s. Then:
  - $\mathbb{E}\left[\sup_{w\in b\cdot B_2^d}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i \langle w, x_i\rangle\right\}\right] \leq \frac{br}{\sqrt{n}}$ (direct optimisation)
  - $\mathbb{E}\left[\sup_{w\in b\cdot B_2^d}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i \langle w, x_i\rangle\right\}\right] \leq \frac{dbr}{\sqrt{n}}$ (Dudley integral)

Few Moments Estimators

Definitions

- [Median-of-Means Estimator]
  - Let $X_1, \dots, X_n$ be n i.i.d. observations with mean μ and variance σ^2. Split n points into k non-intersection blocks $B_1, \dots, B_k$ with $|B_j| = \frac{n}{k} = m$. Let $\bar{X}_j = \frac{1}{m}\sum_{i \in B_j} X_i$. Let $\hat{\mu} = \mathrm{Median}(\bar{X}_1, \dots, \bar{X}_k)$.
  - Let f be a function. Then, $\mathrm{MOM}(f) = \mathrm{Median}\left(\frac{1}{m}\sum_{i \in B_1} f(x_i), \dots, \frac{1}{m}\sum_{i \in B_k} f(x_i)\right)$
- [Hypercontractivity]
  - Let X be a random variable. Then X is <u>(p,q)-hypercontractive</u> if $\exists L_{p,q}$ s.t. for $q \geq p$, $\|X\|_q \leq L_{p,q}\|X\|_p$
  - Let $X \in \mathbb{R}^d$ be a random vector. Then X is <u>(p,q)-hypercontractive</u> if $\exists L_{p,q}$ s.t. $\forall v \in S^{d-1}$, $\|\langle X, v\rangle\|_q \leq L_{p,q}\|\langle X, v\rangle\|_p$

Theorems

- [MOM #1] Let $X_1, \dots, X_n$ be i.i.d. copies of a random variable with mean μ and variance σ^2. Let $\hat{\mu}$ be the median-of-means estimator, with $k = 8\log\left(\frac{1}{\delta}\right)$. Then, with probability $1-\delta$, $|\hat{\mu} - \mu| \leq \sigma\sqrt{\frac{32\log\left(\frac{1}{\delta}\right)}{n}}$
  - *First, do intra-block analysis with Chebyshev*
  - *Then, do inter-block analysis with Hoeffding*
- [MOM #2] Let $k = 8\log\left(\frac{1}{\delta}\right)$ and ϵ_i be Rademacher variables, then with probability $1-\delta$:
  - $\sup_{f \in \mathcal{F}}\{|\mathrm{MOM}(f) - \mathbb{E}[f]|\} \leq 64\mathbb{E}\left[\sup_{f \in \mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right\}\right] + 2\sqrt{\frac{128\sup_{f \in \mathcal{F}}\{\mathrm{Var}[f(X)]\}\log\left(\frac{1}{\delta}\right)}{n}}$
  - *Prove by indicator method, bounded difference, symmetrisation, contraction*
- [MOM Variant] Let X be a random variable with only two known moments $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$. Let the mean estimator be $\hat{\mu} \coloneqq \arg\min_{\nu \in \mathbb{R}^d}\left\{\sup_{v \in B_2^d}|\langle \nu, v\rangle - \mathrm{MOM}(\langle X, v\rangle)|\right\}$. Then $\|\hat{\mu} - \mu\|_2 \leq C\left(\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_{\mathrm{op}}\log\left(\frac{2}{\delta}\right)}{n}}\right)$
  - Find vector ν^* that best approximates the MOM estimator, as measured by the worst difference along any projection
- [One-Sided Tail Bound] Let $X_1, \dots, X_n$ be i.i.d. random variables s.t. $X_i \geq 0\ \forall i$, $\mathbb{E}[X_i] = \mu$ and $\mathbb{E}[X_i^2] = \sigma^2$. Then, $\forall t > 0$, $\mathbb{P}\left[\mu - \frac{1}{n}\sum_{i=1}^n X_i \geq t\right] \leq e^{-\frac{t^2 n}{2\sigma^2}}$
  - *Prove by Taylor expansion on* $\mathbb{E}\left[e^{-\lambda X}\right]$
- [Paley Zygmund] Let $Z \geq 0$ be a random variable and $c \in (0,1)$. Then: $\mathbb{P}\left[Z \geq c\mathbb{E}[Z]\right] \geq (1-c)^2\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}$
  - *Prove by* $\mathbb{E}[Z] = \mathbb{E}[Z\mathbb{1}\{Z \geq c\mathbb{E}[Z]\}] + \mathbb{E}[Z\mathbb{1}\{Z < c\mathbb{E}[Z]\}]$, *then Cauchy Schwarz*

Applications

- [Estimation of Mean of Random Vector] Take $\mathcal{F} = \{f_v : v \in B_2^d\}$ where $f_v(x) = \langle x, v\rangle$ where $x \in \mathbb{R}^d$. Then $\|\hat{\mu} - \mu\|_2 \leq c\left(\sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \sqrt{\frac{\|\Sigma\|_{\mathrm{op}}\log\left(\frac{1}{\delta}\right)}{n}}\right)$
- [Estimation of Higher Moments] Let p be an even integer. Let X be a zero-mean random vector in $\mathbb{R}^d$ s.t. $\forall v \in S^{d-1}$, $\|\langle X, v\rangle\|_{2p} \leq L\|\langle X, v\rangle\|_p$. Then, with probability $1-\delta$, $\forall v \in S^{d-1}$, $|\mathrm{MOM}(\langle X, v\rangle^p)| \leq c(2L)^p\mathbb{E}[\langle X, v\rangle^p]\sqrt{\frac{d + \log\left(\frac{1}{\delta}\right)}{n}}$.

- [Least Eigenvalue of Sample Covariance] Let $X \in \mathbb{R}^d$ be a zero-mean random vector with $\Sigma = \mathbb{E}[XX^T]$. Assume that $\exists c \in (0,1)$, $\beta \in (0,1)$ s.t. $\forall v \in S^d$, $\mathbb{P}\left[|\langle X, v\rangle| > c\sqrt{\mathbb{E}[\langle X, v\rangle^2]}\right] \geq \beta$. Then, $\lambda_{min}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^T\right) \geq \frac{c\beta^2}{2}\lambda_{min}(\Sigma)$ for $n \geq \frac{c'}{\beta^2}\left(d + \log\left(\frac{1}{\delta}\right)\right)$.
    - The assumption $\forall v \in S^d$, $\mathbb{P}\left[|\langle X, v\rangle| > c\sqrt{\mathbb{E}[\langle X, v\rangle^2]}\right] \geq \beta$ means that in any direction v, $|\langle X, v\rangle|$ is not too small relatively as compared to $\sqrt{\mathbb{E}[\langle X, v\rangle^2]}$
    - *Prove using Warren's lemma and VC bound*

Nonparametric Least Squares

Definitions (Fixed Design)

- [Set-Up] Observe $y_i = f^*(x_i) + \epsilon_i$ where $(x_i)_{i=1}^n$ are fixed design vectors and $\epsilon \sim N(0,1)$. Know that $f^* \in \mathcal{F}$. The goal is to bound $\|\hat{f} - f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^n \left(\hat{f}(x_i) - f^*(x_i)\right)^2$.
- [$\|\cdot\|_n$] Define $\|f - y\|_n^2 := \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2 = \frac{1}{n}\|f(X) - Y\|_2^2$
- [Star-Shaped] A class of functions $\mathcal{F}$ is <u>star-shaped</u> around f^* if for any $\alpha \in [0,1]$ and $f \in \mathcal{F}$, $\alpha(f - f^*) \in \mathcal{F} - f^* = \{f - f^* : f \in \mathcal{F}\}$.
  - Convex ⇒ star-shaped

Propositions (Fixed Design)

- [Localisation] Let $\mathcal{F}$ be star-shaped around f^* and $\epsilon \sim N(0, \mathbb{I}_n)$ be the Gaussian noise. Let t^* be the fixed-point solution to $t = \frac{2}{nt} \sup_{f \in \mathcal{F}, \|f - f^*\|_n \le t} \langle \epsilon, f - f^* \rangle$. Then $\forall t \ge t^*$, we have: $\|\hat{f} - f^*\|_n^2 \le \left(2t + \frac{2u}{t}\right)^2$ with probability at least $1 - e^{-\frac{u^2 n}{2t^2}}$.
  - Only care about f close to f^*, rather than the complexity of the whole class $\mathcal{F}$
  - <u>Prove using Gaussian concentration</u>
- Let $\mathcal{F}$ be a nonparametric class with parameter p, then with probability $1 - e^{-\frac{nt^2}{2}}$, $\|\hat{f} - f^*\|_n^2 \le Cn^{-\frac{2}{p+2}}$.
  - *Prove using localisation with $u = t^2$ and Dudley integral.*
- Let $\mathcal{F} = \{f_\beta : x \to \langle x, \beta \rangle\}_\beta$ where $\beta \in \mathbb{R}^p$ be a star-shaped, parametric class. Then, with probability $1 - \delta$, $\|\hat{f} - f^*\|_n^2 \le c\left(\frac{p + \log\left(\frac{1}{\delta}\right)}{n}\right)$.

Definitions (Random Design)

- [Set Up] Let $(x, y) \sim P_{X,Y}$ some unknown distribution and $\mathcal{F}$ be a convex class of functions. Observe $(x_i, y_i)_{i=1}^n$ i.i.d. samples. $|y| \le m$, $|f(x)| \le m \ \forall f \in \mathcal{F}$. $\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2$. Let $R(f) = \mathbb{E}_{x,y \sim P_{X,Y}}\left[\left(y - f(x)\right)^2\right]$. Want to analyse $R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f)$.
- [Notation]
  - $P_n((y - f)^2) = \frac{1}{n}\sum_{i=1}^n \left(y_i - f(x_i)\right)^2$
  - $P((y - f)^2) = \mathbb{E}_{(x,y) \sim P}\left[\left(y - f(x)\right)^2\right]$

Propositions (Random Design)

- Let $\mathcal{F}$ be convex. Then $\forall f \in \mathcal{F}$, $R_n(f) - R_n(\hat{f}) \ge \frac{1}{n}\sum_{i=1}^n \left(f(x_i) - \hat{f}(x_i)\right)^2$ where $R_n(f) = \frac{1}{n}\sum_{i=1}^n \left(y_i - f(x_i)\right)^2$.
- [Process with Quadratic Penalty] Let $\mathcal{G}$ be a class of functions with covering number $\mathcal{N}(\mathcal{F}, L_2(P_n), \gamma)$ for some $\gamma > 0$ and the function $0 \in \mathcal{N}$. Then, for any $\alpha \ge 0$, $\mathbb{E}_\epsilon\left[\sup_{g \in \mathcal{G}}\left\{\frac{1}{n}\sum_{i=1}^n (\epsilon_i g(x_i) - c' g(x_i)^2)\right\}\right] \le C\left(\alpha + \frac{1}{\sqrt{n}}\int_\alpha^\gamma \sqrt{\log \mathcal{N}(\mathcal{G}, L_2(P_n), \epsilon)}\, d\epsilon + \frac{1}{c'}\frac{\log \mathcal{N}(\mathcal{G}, L_2(P_n), \gamma)}{n}\right)$ where C is an absolute constant and c' is a tuneable parameter.
- Let $\gamma > \alpha \ge 0$, then: $\mathbb{E}[R(\hat{f})] - \inf_{f \in \mathcal{F}} R(f) \le C\mathbb{E}\left[m\left(\alpha + \frac{1}{\sqrt{n}}\int_\alpha^\gamma \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \epsilon)}\, d\epsilon + \frac{m}{n}\log \mathcal{N}(\mathcal{F}, L_2(P_n), \gamma)\right)\right]$ where C is an absolute constant.

- $\mathbb{E}\left[R(\hat{f})\right] - \inf_{f \in \mathcal{F}} R(f) \leq C \sup_{P_n} \left\{ m \left(\alpha + \frac{1}{\sqrt{n}} \int_{\alpha}^{\gamma} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \epsilon)}\, \mathrm{d}\epsilon + \frac{m}{n} \log \mathcal{N}(\mathcal{F}, L_2(P_n), \gamma) \right) \right\}$
- When $\mathcal{N}(\mathcal{F}, L_2(P_n), \epsilon) \sim \epsilon^{-p}$ for $\epsilon \in (0,2)$ then $\mathbb{E}\left[R(\hat{f})\right] - \inf_{f \in \mathcal{F}} R(f) \leq C n^{-\frac{2}{p+2}}$. (optimise with respect to γ; balance the powers of n)

Online Learning

Definitions

- [Regret] Let $\mathcal{F}$ be a class of functions and $L(f(x), y)$ be a loss function. Let $\left(\hat{f}_n\right)_n$ be a sequence of predictors with $\hat{f}_n \in \mathcal{F}$ s.t. $\hat{f}_n$ is trained on $(x_i, y_i)_{i=1}^{n-1}$. Then, regret is $\sum_{i=1}^{n} L\left(\hat{f}_i(x_i), y_i\right) - \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} L(f(x_i), y_i)$
    - Intuition: how much you regret is compared with the best fixed function.
- [Exponential Weights Algorithm] Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a family of distributions parametrised by $\Theta \subset \mathbb{R}^d$. Let $\pi(\theta)$ be a prior distribution over Θ. Let $\eta > 0$ be a fixed learning rate.
    - $\hat{\rho}_n \coloneqq \frac{e^{-\eta \sum_{i=1}^{n} L(f_\theta(x_i), y_i)} \pi(\theta)}{\mathbb{E}_{\theta \sim \pi}\left[e^{-\eta \sum_{i=1}^{n} L(f_\theta(x_i), y_i)}\right]}$
    - $\hat{\rho}_0 = \pi(\theta)$
    - Denominator is just a normalising constant
- [Mixture] $f(x) = \int_\alpha f_\alpha(x) g(\alpha) \mathrm{d}\alpha$
- [Mix Loss] $-\frac{1}{\eta} \sum_{i=1}^{n} \log\left(\mathbb{E}_{\theta \sim \hat{\rho}_{i-1}}\left[e^{-\eta L(f_\theta(x_i), y_i)}\right]\right)$
    - Mixing together all the loss from the n steps
- [Potential] $-\frac{1}{\eta} \log\left(\mathbb{E}_{\theta \sim \pi}\left[e^{-\eta \sum_{i=1}^{n} L(f_\theta(x_i), y_i)}\right]\right)$
- [KL Divergence] Let f be the true density and $\hat{f}$ be the predicted density. Then the excess risk is: $KL(f \| \hat{f}) \coloneqq \int \log\left(\frac{f(x)}{\hat{f}(x)}\right) f(x) \mathrm{d}x = \mathbb{E}_{X \sim f}\left[-\log \hat{f}(X) + \log f(X)\right]$
- [Covering] Let $\mathcal{F}$ be a collection of densities. Then:
    - $\mathcal{N}(\mathcal{F}, KL, \epsilon) = \min\{n \in \mathbb{N} : \exists q_1, \dots, q_n \text{ s.t. } \forall f \in \mathcal{F}, \exists i \in [n] \text{ s.t. } KL(f \| q_i) \le \epsilon^2\}$
    - Note the ϵ^2 instead of ϵ.
    - i.e. if f is the true density, then using q_i will incur a KL loss of at most ϵ^2
- [Exponentially Concave] A loss L is <u>exponentially concave</u> with $\eta > 0$ if $\mathbb{E}_{\theta \sim \hat{\rho}_{i-1}}\left[e^{-\eta L(f_\theta(x_i), y_i)}\right] \le e^{-\eta L\left(\mathbb{E}_{\theta \sim \hat{\rho}_{i-1}}[f_\theta(x_i)], y_i\right)}$ holds $\forall i$.
    - Intuitively, can bring $\mathbb{E}$ into the exponent and into the loss function; the predictor is of a nicer form.
    - If L is exponentially concave, then can upper bound total loss as if the predictors are the expectations of the mixture of predictors at each step.
    - $-\log$ is exponentially concave with $\eta = 1$
    - $L(f_\theta(x_i), y_i) = (f_\theta(x_i) - y_i)^2$ with $|f_\theta(x_i)|, |y_i| \le m$ is exponentially concave with $\eta = \frac{1}{8m^2}$
- [Clip] Define $\mathrm{clip}_m(x) = \begin{cases} \min(m, x), & x \ge 0 \\ \max(-m, x), & x < 0 \end{cases}$.

Propositions

- [Online To Batch] Let $L(\cdot, y)$ be convex and $(x_i, y_i)_{i=1}^{n} \sim P_{X,Y}$ be i.i.d. samples. Let $\left(\hat{f}_i\right)_{i=1}^{n}$ be sequential estimators satisfying $\sum_{i=1}^{n} L\left(\hat{f}_i(x_i), y_i\right) - \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} L(f(x_i), y_i) \le R^{(n)}$ a.s., where $R^{(n)}$ is a constant. Then:
    - $\mathbb{E}_{(x_i, y_i)_{i=1}^{n} \sim P_{X,Y}}\left[\mathbb{E}_{(X,Y) \sim P_{X,Y}}\left[L\left(\frac{1}{n} \sum_{i=1}^{n} \hat{f}_i(X), Y\right)\right] - \inf_{f \in \mathcal{F}} \mathbb{E}[L(f(X), Y)]\right] \le \frac{R^{(n)}}{n}$
    - Using the batch estimator $\frac{1}{n} \sum_{i=1}^{n} \hat{f}_i(X)$ leads to the convergence of regret to 0
    - Can be thought of as averaging the trajectory density
    - $\mathbb{E}_{(x_i, y_i)_{i=1}^{n} \sim P_{X,Y}}\left[R\left(\frac{1}{n} \sum_{i=1}^{n} \hat{f}_i\right) - \inf_{f \in \mathcal{F}} R(f)\right] \le \frac{R^{(n)}}{n}$
        - Any bound on regret is a bound on excess risk
- [Unfolding Lemma] Equivalence between potential and mix-loss; deterministic result.

- $-\frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\pi}\left[e^{-\eta\sum_{i=1}^n L(f_\theta(x_i),y_i)}\right]\right) = -\frac{1}{\eta}\sum_{i=1}^n \log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}\left[e^{-\eta L(f_\theta(x_i),y_i)}\right]\right)$
    - Property of the exponential weights algorithm
- Let $\left(\hat{f}_n\right)_n$ be a sequence such that the predictors satisfy $L\left(\hat{f}_i(x_i), y_i\right) \le -\frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}\left[e^{-\eta L(f_\theta(x_i),y_i)}\right]\right)$, then $\sum_{i=1}^n L\left(\hat{f}_i(x_i), y_i\right) \le -\frac{1}{\eta}\log\left(\mathbb{E}_{\theta\sim\pi}\left[e^{-\eta\sum_{i=1}^n L(f_\theta(x_i),y_i)}\right]\right)$
    - Total loss is upper bounded by the total loss

Density Estimation

- [Donsker-Varadhan Variational Formula] $\log\left(\mathbb{E}_{\theta\sim\pi}\left[e^{h(\theta)}\right]\right) = \sup_\rho\left\{\mathbb{E}_{\theta\sim\rho}[h(\theta)] - \mathrm{KL}(\rho\|\pi)\right\}$
- Let l be a bounded loss i.e. $l(f_\theta(x), y) \le m\ \forall\theta, x, y$. Then: $\sum_{i=1}^n \mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}[L(f_\theta(x_i), y_i)] \le \frac{n\eta m^2}{8} + \inf_\gamma\left\{\mathbb{E}_{\theta\sim\gamma}\left[\sum_{i=1}^n L(f_\theta(x_i), y_i)\right] + \frac{\mathrm{KL}(\gamma\|\pi)}{\eta}\right\}$
    - Works for any $\eta > 0$; any distribution γ sets an upper bound
    - π is the prior distribution; typically pick the uniform distribution.
- $\sum_{i=1}^n -\log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}[f_\theta(z_i)]\right) \le \inf_\rho\left\{\mathbb{E}_{\theta\sim\rho}\left[-\sum_{i=1}^n \log\left(f_\theta(z_i)\right)\right] + \mathrm{KL}(\rho\|\pi)\right\}$
    - In particular, any density ρ' yields an upper bound.
- [Progressive Mixture] Let $\mathcal{F} = \{f_1, \ldots, f_M\}$ be a class of densities. Assume that $z_1, \ldots, z_n$ are sampled from $f^* \in \mathcal{F}$. Then, $\exists \hat{f}$ based on $z_1, \ldots, z_n$ s.t. $\mathbb{E}_{z_1,\ldots,z_n}\left[KL\left(f^*\|\hat{f}\right)\right] \le \frac{\log(M)}{n}$.
- [Yang-Barron] Let $\mathcal{F}$ be a collection of densities with $\{q_1, \ldots, q_{|\mathcal{N}_\epsilon|}\}$ be a cover. Let f be the progressive mixture on $q_1, \ldots, q_{|\mathcal{N}_\epsilon|}$. Then: $\mathbb{E}_{z_1,\ldots,z_n}\left[\mathrm{KL}\left(f^*\|\hat{f}\right)\right] \le \inf_{\epsilon>0}\left\{\epsilon^2 + \frac{\log\left(\mathcal{N}(\mathcal{F}, KL, \epsilon)\right)}{n}\right\}$
- [Lemma] Let $Q(\theta) = \theta^T A\theta + b^T\theta + c$ and A positive semi-definite. Then $\int_{\mathbb{R}^d} e^{-Q(\theta)} \mathrm{d}\theta = \frac{\pi^{\frac{d}{2}}}{\sqrt{\det A}} e^{-\inf_{\theta\in\mathbb{R}^d} Q(\theta)}$.

Workflow

- Check exp-concavity: $x \mapsto e^{-\eta(x-y)^2}$ is concave for some $\eta > 0\ \forall x, y$ in range considered
- Write out the equivalence of mix losses = potential
- Bound potential by Donsker-Varadhan or reduce to pure Gaussian integrals
- When the loss is convex, perform online-to-batch to bound excess risk

Examples

- Let $\Theta = \{1, \ldots, M\}$ be finite with $|\Theta| = M$. Let $f^* = \arg\min_{f_j: j\in\Theta}\left\{-\sum_{i=1}^n \log\left(f_j(z_i)\right)\right\}$. With $\pi \sim \mathrm{Uniform}(\Theta)$ and $\rho = \delta_{f^*}$. Then: $-\sum_{i=1}^n \log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}\left[f_j(z_i)\right]\right) - \left(-\sum_{i=1}^n \log\left(f^*(z_i)\right)\right) \le \log M$
- Let $\mathcal{F} = \{f_+ \equiv 1, f_- \equiv -1\}$. Then $\sum_{i=1}^n \mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}[\mathbb{1}\{f_\theta(x_i) \ne y_i\}] \le \min_{f\in\mathcal{F}}\{\sum_{i=1}^n \mathbb{1}\{f_\theta(x_i) \ne y_i\}\} + \sqrt{\frac{n\log 2}{2}}$.
    - *Use Donsker-Varadhan with bounded loss*
- Let $\mathcal{F} = \{f_1, \ldots, f_K\}$ where $f_i \equiv i$. Then $\sum_{i=1}^n \mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}[\mathbb{1}\{f_\theta(x_i) \ne y_i\}] \le \min_{f\in\mathcal{F}}\{\sum_{i=1}^n \mathbb{1}\{f_\theta(x_i) \ne y_i\}\} + \sqrt{\frac{n\log K}{2}}$
    - *Use Donsker-Varadhan with bounded loss*
- [Logistic Regression] Consider the logistic regression with $-\log$ loss i.e. the loss at each step is $-\log\hat{p} = -\log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}[\sigma(y_i\langle x_i, \theta\rangle)]\right)$. Let θ^* be the MLE solution with $\|\theta^*\|_2 \le b$ and $\|x_i\| \le r$. Then:
    - $-\sum_{i=1}^n \log\left(\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}[\sigma(y_i\langle x_i, \theta\rangle)]\right) \le -\sum_{i=1}^n \log\left(\sigma(y_i\langle x_i, \theta\rangle)\right) + d + \frac{d}{2}\log\left(1 + \frac{ab^2r^2}{8d^2}\right)$
    - MLE θ^* helps in simplifying Taylor expansion of total loss about θ^*

- [Squared Loss] Let (x_i, y_i) be i.i.d. samples with $|y|, |f| \leq m$, then: $\mathbb{E}_{x,y\sim P_{X,Y}}\left[R\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta\sim\hat{\rho}_{i-1}}[f_\theta(x)]\right)\right] \leq \inf_{f\in\mathcal{F}} R(f) + \frac{8m^2\log(M)}{n}$
- [Vork-Azoury-Warmuth] Let $(x_i, y_i)_{i=1}^n$ be a deterministic sequence with $\|x_i\|_2 \leq r, |y_i| \leq m$ and $x_i \in \mathbb{R}^d$. Let $\theta^* = \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{n}(y_i - \langle\theta, x_i\rangle)^2$. Assume $\|\theta^*\|_2 \leq b$. Let $\hat{\theta}_{i-1} = \arg\min_{\theta\in\mathbb{R}^d} \sum_{j=1}^{i-1}\left(y_j - \langle\theta, x_j\rangle\right)^2 + \lambda\|\theta\|_2^2$. Then, there is a choice of λ s.t. $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \leq \sum_{i=1}^{n}(y_i - \langle\theta^*, x_i\rangle)^2 + m^2\left(d + 4d\log\left(1 + \frac{nr^2b^2}{d^2m^2}\right)\right)$

Statistical Models

Classification (Lecture 5)

- Let $\mathcal{X}$ be a set and $\mathcal{F}$ be a finite family of classifiers i.e. $\mathcal{F} = \{f: \mathcal{X} \to \{0,1\}\}$. Let $M = |\mathcal{F}|$. Assume that the true classifier $f^* \in \mathcal{F}$. Observe $\{(X_i, f(X_i))\}_{i=1}^n$. Define $R(f) = \mathbb{P}[f(X) \neq f^*(X)]$ and $R_n(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{f(X_i) \neq f^*(X_i)\}$. Then, with probability $1-\delta$, $R(\hat{f}) \leq C\frac{\log(M)+\log\left(\frac{1}{\delta}\right)}{n}$
  - *Prove by Bernstein (Bernoulli) or union bound*
  - *Exploit finite $|\mathcal{F}|$ and $R_n(f) - R(f) = R_n(f) - \mathbb{E}[R_n(f)]$*

Kernel Density Estimation (Lecture 6)

- Observe $X_1, \dots, X_n$ i.i.d. from density f over $\mathbb{R}$. Let $K: \mathbb{R} \to \mathbb{R}$ be a kernel function i.e. $K(x) \geq 0$ and $\int_{-\infty}^{\infty} K(x)\mathrm{d}x = 1$. Let $\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$ be the kernel estimator. The loss is $L(\hat{f}, f) = \int_{-\infty}^{\infty}|\hat{f}_n(x) - f(x)|\mathrm{d}x$. Then $\mathbb{P}\left[L(\hat{f}, f) - \mathbb{E}_{X_1,\dots,X_n}[L(\hat{f}, f)] \geq t\right] \leq 2e^{-\frac{t^2 n}{2}}$.
  - *Prove by bounded difference inequality*

Fixed Design Linear Regression (Lecture 8)

- [Oracle] Let $K \subset \mathbb{R}^d$ be a convex set. Let $y_i = \langle x_i, \beta^*\rangle + \epsilon_i$ be the true model, where ϵ_i is zero-mean, independent and σ-sub-Gaussian. Let $\tilde{\beta} = \arg\inf_{\beta \in K}\|X\beta - X\beta^*\|^2$.
$$\mathbb{E}\left[\frac{1}{n}\left\|X\hat{\beta} - X\beta^*\right\|_2^2\right] \leq \frac{1}{n}\left\|X\tilde{\beta} - X\beta^*\right\|_2^2 + \frac{4\sigma^2 d}{n}$$
- [Lecture 9] $\mathbb{E}\left[\frac{1}{n}\left\|X\hat{\beta} - X\beta^*\right\|_2^2\right] \leq \inf_{\beta \in K}\left\{\frac{1}{n}\left\|X\tilde{\beta} - X\beta^*\right\|_2^2\right\} + \frac{2\sqrt{2\log(2d)}\max\|X_i\|_2}{n}$
- [HW1 P8] Let $x_1, \dots, x_n \in \mathbb{R}^d$ be fixed design vectors and $Y_1, \dots, Y_n$ independent, sub-Gaussian with parameter σ. Then $\xi(\hat{\beta}) \leq \frac{\sigma^2}{n}\left(\sqrt{d} + \sqrt{2\log\left(\frac{1}{\delta}\right)}\right)^2$.
  - Exploit the sub-Gaussian vector HY

Sparse Linear Regression

- Let $K = \{x: \|x\|_0 \leq s\}$ with $s \ll d$ and $\beta^* \in K$.
  - With probability $1-\delta$, $\frac{1}{n}\left\|X\hat{\beta} - X\beta^*\right\|^2 \leq \frac{C\sigma^2\left(s\log\left(\frac{ed}{2s}\right) + \log\left(\frac{1}{\delta}\right)\right)}{n}$
- [Binomial Bound] $\sum_{j=0}^{2s}\binom{d}{j} \leq \left(\frac{ed}{2s}\right)^{2s}$

Kolmogorov Smirnov Statistic (Lecture 14)

- Let $\mathcal{F}$ be a uniformly bounded family of functions s.t. $|f| \leq M$. Then, with probability $1-\delta$, we have: $\sup_{f\in\mathcal{F}}\left|\mathbb{E}[f(X)] - \frac{1}{n}\sum_{i=1}^n f(X_i)\right| \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}[f(X)] - \frac{1}{n}\sum_{i=1}^n f(X_i)\right|\right] + M\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n}}$
  - *Prove by symmetrisation*
- With probability $1-\delta$, $\sup_{t\in\mathbb{R}}|F_n(t) - F(t)| \leq 2\sqrt{\frac{2\,log(2n+2)}{n}} + M\sqrt{\frac{2\,log\left(\frac{1}{\delta}\right)}{n}}$
  - *Prove by bounded difference inequality*

Classification (Lecture 16)

- Let $\mathcal{X}$ be a set and $\mathcal{F}$ be a family of classifiers i.e. $\mathcal{F} = \{f: \mathcal{X} \to \{0,1\}\}$. Let $P_{X,Y}$ be an unknown distribution and $\{(X_i, Y_i)\}_{i=1}^n$ be the train set. Then, w.p. $1-\delta$: $R(\hat{f}) - R(f^*) \leq C\left(\sqrt{\frac{d\log\left(\frac{en}{d}\right)}{n}} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}}\right)$
  - *Exploit Sauer-Shelah lemma*

Random Design Regression for Non-Parametric Model

- Let $\mathcal{F}$ be a convex class of functions. Let $Y = f^*(X) + \xi$ be the true model and $\max(|\xi|, Y) \leq m$ and $|f(X)| \leq m$ a.s. Let $\log \mathcal{N}(\mathcal{F}, L_2(P_n), \epsilon) \leq C\epsilon^{-p}$. Then:
  - If $p > 2$, $\mathbb{E}[R(\hat{f})] - R(f^*)$ converges on the order of $n^{-\frac{1}{p}}$

Problem Solving

Common Ideas

- Dimensional analysis; balancing powers
  - With an infimum or supremum bound, sometimes substituting in nice values work.
- Brute-force; whack; algebra; direct optimisation (norm, linear class)
  - Taylor expansion (***)
- Jensen; Cauchy-Schwarz (inner-product); Hölder $\mathbb{E}[|XY|] \leq \mathbb{E}[X^p]^{\frac{1}{p}}\mathbb{E}[Y^q]^{\frac{1}{q}}$ for $\frac{1}{p}+\frac{1}{q}=1$
- Bounded difference inequality; contraction, Lipschitz-ness; Symmetrisation
- Subtle but important ideas:
  - $\sup\{a+b\} \leq \sup a + \sup b$
  - $\sup\{a-b\} \leq \sup a + \sup b$
  - $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ (when square roots start to become annoying)
- Directions:
  - Sub-Gaussians → use Hoeffding
  - Sub-exponential → try Bernstein → Self-bounding functions technique
  - Anything Gaussian → use Gaussian concentration
  - Function class has a covering number → Dudley integral
  - Bounded difference property → bounded difference inequality
  - Prove Lipschitz property → Find a function $f(\xi)$ where $\xi \sim N(0,\sigma^2\mathbb{I}_d)$. ($f$ can be some complicated function involving $\sup$) Prove $|f(\xi)-f(\nu)| \leq L\|\xi-\nu\|_2$ and you can already apply Gaussian concentration.
- Sometimes, just bound one part of the term in an expression e.g.
  - $\mathbb{E}[X^2\mathbb{1}\{A\}] = \mathbb{E}[X^{1+\epsilon}\mathbb{1}\{A\}X^{1-\epsilon}] \leq R\mathbb{E}[X^{1-\epsilon}]$
  - $\mathbb{E}[(XX^T-\Sigma)^2] \preccurlyeq \mathbb{E}[XX^TXX^T] \preccurlyeq \mathbb{E}[X(r^2)X^T] = r^2\Sigma$
- Think in high probability form $\mathbb{P}[X \geq t]$
  - Union bound
  - Exploit sub-Gaussianity (anything else sub-Gaussian)
- Think in moment form $\mathbb{E}[X^p]$
- [Workflow] Choose loss function, check exponential concavity, predict with exponential weights for each round
- Exotic:
  - ϵ-Net + Union Bound
  - Donsker-Varadhan

Algebraic Gymnastics

- [Stirling Approximation] $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$
- [Gamma] $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$
  - $\Gamma(n) = (n-1)!$
  - $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
  - $\Gamma(x) \leq x^x$
    - Prove by Stirling's
- $\frac{1}{2}\left(e^\lambda + e^{-\lambda}\right) \leq e^{\frac{\lambda^2}{2}}$ for $\lambda \in \mathbb{R}$
  - Rademacher random variables are sub-Gaussian
- [Miscellaneous]
  - $1+x \leq e^x$
    - When x is small and you want to facilitate multiplication
  - $1-x \leq e^{-x}$
  - $\frac{1}{1-x} \leq e^{2x}$ for $x \in \left[0,\frac{1}{2}\right]$

- $\frac{x}{2+x} \leq \log(1+x)$ for $x \geq 0$
- $-x + \frac{x^2}{2} \leq (1-x)\log(1-x)$ for $x \in (0,1)$
- $e^x \leq 1 + x + \frac{\frac{x^2}{2}}{1-\frac{|x|}{3}}$ if $|x| < 3$
  - Used to prove matrix Bernstein
- $e^x \leq x + e^{x^2}$
  - Proving $\mathbb{E}[X] = 0$ of sub-Gaussian equivalency
- $2\lambda x \leq \lambda^2 + x^2$
  - Proving $\mathbb{E}[X] = 0$ of sub-Gaussian equivalency
- $|x|^p \leq p^p(e^x + e^{-x})$

- [Jensen]
  - $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$
  - $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$