01/18. 36709 Lec 1

not cover everything $\begin{cases}\text{pick a paper. connection/report the paper} \\ \quad \text{how to prove. how cool idea} \\ \text{scribing. everyone}\end{cases}$

[Non-asymptotic high-dimensional statistics

1. High-dimensional (Linear) classification
2. Estimating high-dim matrices
3. Non-parametric regression

1. High-dim classification.

Fisher's Linear Discriminant Analysis

$\mu_1, \mu_2 \in \mathbb{R}^d$

$d >> n$

$y=1$. $X_1 \cdots X_{n_1} \sim N(\mu_1, \Sigma)$

$y=2$, $X_{n_1+1} \cdots X_{n_1+n_2} \sim N(\mu_2, \Sigma)$

If know $\mu_1\, \mu_2\, \Sigma$ Assume balanced classes

Bayes-optimal classifier

X, $(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) \leq (x-\mu_2)^T\Sigma^{-1}(x-\mu_2)$ then Label 1. otw. Label 2

Replace μ_1 by $\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$, μ_2 by $\hat{\mu}_2$. Σ by $\hat{\Sigma} = \frac{1}{2}\left[\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\hat{\mu}_1)(X_i-\hat{\mu}_1)^T + \frac{1}{n_2-1}\sum_{i=n_1+1}^{n_1+n_2}(X_i-\hat{\mu}_2)(X_i-\hat{\mu}_2)^T\right]$

$x \to$ pick y, $\max p(Y=y \mid X=x)$

Classifier: $\psi(x) = \langle \mu_1-\mu_2, \Sigma^{-1}(x - \frac{\mu_1+\mu_2}{2})\rangle$, Label 1 if $\psi(x) \geq 0$

$\hat{\psi}(x) \quad \langle \hat{\ } \quad \hat{\ } \quad \hat{\ } \quad - \quad \hat{\ } \quad \hat{\ } \quad \rangle$

$\mathrm{Err}(\hat{\psi}) = \frac{1}{2}\mathbb{P}_{X\sim N(\mu_1,\Sigma)}[\psi(x) < 0] + \frac{1}{2}\mathbb{P}_{X\sim N(\mu_2,\Sigma)}[\psi(x) \geq 0]$

$\mathrm{err} = \mathbb{P}(\hat{\psi}(x) \neq y) = \mathbb{P}(\hat{\psi}(x) \neq 1 \mid y=1)\, p(y=1) + \mathbb{P}(\hat{\psi}(x) \neq 2 \mid y=2)\, p(y=2)$

Story: *classical asymptotics $n \to +\infty$ d fixed

$\to$ consistent estimate $\hat{\mu}_1$, $\hat{\Sigma}$, $\hat{\mu}_2$

$\to \mathrm{Err}(\hat{\psi}) \xrightarrow{P} \mathrm{Err}(\psi)$ $\hookrightarrow$ Bayes-opt

$\mathrm{Err}(\psi) = \Phi\left(-\frac{\gamma}{2}\right)$, $\gamma = \|\mu_1-\mu_2\|$

* high-dim. asymp. Kolmogorov.

↳ $n \to +\infty$, $\left[\frac{n_1}{d}, \frac{n_2}{d}\right] \to \lambda$, (proportional) $\frac{\log d}{n} \to 0$ ($d \asymp \exp\left(\frac{n}{\log n}\right)$) $\frac{d}{n} \to c$, $\frac{d^2}{n} \to c$

Kolmogorov. $Err(\hat{\psi}) \xrightarrow{P} \Phi\left(\frac{-\gamma^2}{2\sqrt{\gamma^2+2\lambda}}\right)$, $\gamma = \|\mu_1 - \mu_2\|_2$ $\lambda \to \frac{n_1}{d}$, $\frac{n_2}{d} \to \lambda$ (λ_1, λ_2)

$\Sigma = I_d$

Exercise. $\hat{\mu}_1 = \mu_1 + Z_1$, $Z_1 \sim N(0, \frac{I_d}{n_1})$, $\hat{\mu}_2 = \mu_2 + Z_2$, $Z_2 \sim N(0, \frac{I_d}{n_2})$

$\hat{\psi}(x) = \langle \mu_1 - \mu_2 + Z_1 - Z_2, x - \frac{\mu_1+\mu_2}{2} - \frac{Z_1+Z_2}{2} \rangle$, prove Kolmogorov. $\Phi\left(\frac{-\gamma^2}{2\sqrt{\gamma^2+2\lambda}}\right)$

2. high-dim. matrices

$X_1, \dots, X_n \sim N(0, \Sigma)$, want to est. Σ. $\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} X_i X_i^T$

$\|\hat{\Sigma} - \Sigma\|_{op} := \lambda_{max}(\hat{\Sigma} - \Sigma) = \sup_{\|v\|_2 = 1} v^T(\hat{\Sigma} - \Sigma)v$ (Largest eigenvalue)

operator norm

$n/d \to \lambda$. or $n/d \to 0$ d fixed $n \to \infty$ (classically $\|\hat{\Sigma} - \Sigma\|_{op} \to 0$) d-fixed. $n \to \infty$

$d/n \to \lambda$. $\hat{\Sigma} - \Sigma \to$ eigenvalues. don't all converge to 0.

Top eigenvector $\hat{v}$ of $\hat{\Sigma}$ won't be "close" v of Σ

Eigenvalues' distribution: Marchenko-Pastur law. random mat

↳ covariance. $f(x) \propto \sqrt{\frac{(r-x)(x-\delta)}{x}}$

$r = (1+\sqrt{\lambda})^2$, $\delta = (1-\sqrt{\lambda})^2$

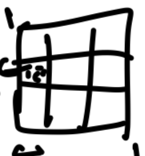
3. Non-parametric regression.

d can be fixed.

est. Lipschitz. funct. $n^{-\frac{2}{2+d}}$, $d = 10$, $n^{-\frac{1}{6}}$, much slower the $1/\sqrt{n}$ err in parameter

Bias-Variance $y_i = f(x_i) + \varepsilon_i$

regressogram



average y-values

# of bins $\left(\frac{1}{\xi}\right)^d$

exponential curse of dim

"Bad things" in high-dim

structure/structural assumptions

* sparsity of vector. * pattern in non-zero - low rank - sparse
* functions. smoothness. additive decompositions

$f(x) = \sum_j f_j(x_j)$

↳ only depends on S variables.

Lec 2 Chapt5

Recap:

1. Constrasting (1) classical asymptotic: $\frac{d}{n} \to 0$

   (2). high-dim asym, $\frac{d}{n} \to \alpha$

   (3) high-dim non asym $\frac{d}{n} \to +\infty$

2. Example.

   (1). vector estimation (classification)

   (2) matrix estimate (PCA/cov. matrices)

   (3) function estimation. (L-Lips)

3. Structure.

   (1) Sparsity

   (2) low-rankness

   (3) Smoothness. sparsity

Motivating example:

$y = \theta^* + \varepsilon$, $\varepsilon \sim N(0, I_d)$. estimate θ^* (goal)

$\theta^* \in K$ (maybe sparse) encode structure → convex set

Maximize likelihood: $\hat{\theta} = \underset{\theta \in K}{\operatorname{argmax}} \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left\{-\frac{(y-\theta)^T(y-\theta)}{2}\right\}$

$= \underset{\theta \in K}{\operatorname{argmin}} \frac{1}{2}\|y-\theta\|^2$

Basic inequality

$\frac{1}{2}\|y-\hat{\theta}\|^2 \le \frac{1}{2}\|y-\theta^*\|$, $\hat{\theta}$: best estimate

$\Delta = \hat{\theta} - \theta^*$, $y = \theta^* + \varepsilon$, $\frac{1}{2}\|\hat{\theta} - \theta^* - \varepsilon\|_2^2 \le \frac{1}{2}\|\varepsilon\|^2 \Rightarrow \underbrace{\frac{1}{2}\|\Delta\|^2}_{\text{bound on this}} \le \langle \Delta, \varepsilon \rangle$ ↳ study this (upper bound) this

→ If Δ was fixed, $\langle \Delta, \varepsilon \rangle \sim N(0, \|\Delta\|^2)$

→ Δ depends on ϵ,

→ One valid upper bd: $\langle \Delta, \varepsilon \rangle \le \sup_{\Delta \in K - \theta^*} \langle \Delta, \varepsilon \rangle = \sup_{\Delta \in K} \langle \Delta, \varepsilon \rangle$ ↳ just an example

Gaussian complexity of K (K)

$\mathcal{G}(K) = \mathbb{E} \sup_{\theta \in K} \langle \theta, \varepsilon \rangle$

Measure of size of K, useful for statistical applications

Examples. $K = B_2(R) = \{\theta : \|\theta\|_2 \le R\}$

$\mathcal{G}(K) \le \mathbb{E}\|\theta\|_2 \|\varepsilon\|_2$ by Cauchy Schwarz

$\le R\,\mathbb{E}\|\varepsilon\|_2 \underset{\text{Jensen}}{\le} R\sqrt{\mathbb{E}\|\varepsilon\|^2} \le R\sqrt{d}$

Metric entropy

K, ρ (metric)

1) $\rho(\theta,\theta') \geq 0$, 0 only if $\theta = \theta'$
2) Triangle ineq: $\rho(\theta,\theta') \leq \rho(\theta,\tilde\theta) + \rho(\tilde\theta,\theta')$
3) $\rho(\theta,\theta') = \rho(\theta',\theta)$

$K \subseteq \mathbb{R}^d$, metric $\rho(\theta,\theta') = \|\theta-\theta'\|$

Hamming metric on $\{0,1\}^d$, $\rho(\theta,\theta') = \frac{1}{d}\sum_{j=1}^{d} \mathbb{1}\{\theta_j \neq \theta_j'\}$

Covering number:

$N(\delta; K, \rho)$: minimum # of radius δ balls needed to "cover" K (最小全包围 ball 数)

$\{\theta^1 \ldots \theta^N\}$, minimal N, covering numbers

$\forall \theta \in K, \exists j, \rho(\theta^j,\theta) \leq \delta$, (任意K中θ都在某ball)

Metric entropy: $\log N(\delta; K, \rho)$. ($\delta\downarrow$, $N\uparrow$)

Example: $K = [-1,1] \subseteq \mathbb{R}$, $\rho(\theta,\theta') = |\theta-\theta'|$

Upper bounds on cover: one valid cover

$\theta^i = -1 + 2(i-1)\delta$, $N = \lfloor \frac{1}{\delta} \rfloor + 1$

$N(\delta; [-1,1], \rho) \leq \lfloor \frac{1}{\delta} \rfloor + 1 \leq \frac{1}{\delta} + 1 \to \log(\frac{1}{\delta})$

$N(\delta; [-1,1]^d, \|\cdot\|_\infty) \leq (\frac{1}{\delta}+1)^d \to d\log(\frac{1}{\delta})$

Packing numbers

δ-packing of K, $\{\theta^1 \ldots \theta^M\} \subseteq K$

$\rho(\theta^i,\theta^j) > \delta$ 两两不相交

$\to M(\delta; K, \rho)$ maximal δ-packing of K

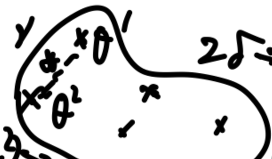
若 $\theta \notin \theta^i, \theta^j$, δ 之外的 θ 也能 packing

add to packing (contradicts maximum)

$\{\theta^1 \ldots \theta^M\}$ δ-packing, claim also a δ-cover

Lemma. $M(2\delta; K, \rho) \leq N(\delta; K, \rho) \leq M(\delta; K, \rho)$

Take any $\nu \in$ covering



$2\delta \to \{\theta^1, \ldots \theta^M\}$

every packing at least has an associated cover

1) $\theta^* = \arg\inf_{\theta \in M} |\theta - \nu|$

$\theta \neq \theta^*$, $\theta \in M$: $\rho(\theta,\nu) \geq \rho(\theta,\theta^*) - \rho(\theta^*,\nu) \geq 2\delta - \delta = \delta$

If $\rho(\theta^1,\theta^2) > 2\delta$, then $\to$ $\rho(\nu,\theta^1) < \delta$, $\rho(\nu,\theta^2) < \delta$

$\rho(\nu,\theta^1) + \rho(\nu,\theta^2) > \delta$ 不可能都在 θ^1, θ^2 ball 里

$N(\delta; K, \delta)$: upper bds, construct cover; Lower bd, construct 2δ packing, $\forall$ any δ

Unit cube

$K = [-1,1] \subseteq \mathbb{R}$, $\rho = |\theta - \theta'|$

$\lfloor \frac{2}{3\delta} \rfloor + 1 \leq N(\delta, K, \rho) \leq \frac{1}{\delta} + 1$

valid 2δ packing

$\cup B(\theta^i, \delta/2) \subseteq K + B(\delta/2)$

Recap

office

F-noon,
T-3-4pm.

1. Estimating a structured vector:

$y=\theta^*+\varepsilon$, $\theta^*\in K$

$\hat\theta = \underset{\theta\in K}{\arg\min} \|y-\theta\|_2^2$

Basic inequality: $\|\hat\theta-\theta^*\|_2^2 \le 2\langle\hat\theta-\theta^*,\varepsilon\rangle \le 2\sup_{\theta\in K-\theta^*}\langle\theta,\varepsilon\rangle$

$\mathbb{E}\|\hat\theta-\theta^*\|_2^2 \le 2\,\mathbb{E}\left(\sup_{\theta\in K-\theta^*}\langle\theta,\varepsilon\rangle\right)$ Gaussian (width) complexity $(K-\theta^*)$

2. Packing & covering:

$N(\delta;K,\rho)$ → min # of δ-balls in metric ρ that we need to cover K

$M(\delta,K,\rho)$ → max # of points that can be placed on K such that $\rho(\theta^i,\theta^j)\ge\delta$

Metric entropy $=\log N(\delta,K,\rho)$

$$M(2\delta,K,\rho)\le N(\delta,K,\rho)\le M(\delta,K,\rho)$$

Metric entropy of B, $\|\cdot\|$ using balls B', $\|\cdot\|'$

→ Examples $\|\cdot\|$ & $\|\cdot\|'$ are $\|\cdot\|_2$

$B\subseteq \bigcup_{i=1}^{N} B'(\theta^i,\delta)$ $\{\theta^1\cdots\theta^N\}$ → covers B

Lower bd covering number

$vol(B)\le N\, vol(\delta B')$, $N\ge \frac{vol(B)}{vol(\delta B')}=\frac{1}{\delta^d}\frac{vol(B)}{vol(B')}$

Upper bd covering number:

$B\to\|\cdot\|$ $\{x:\|x\|\le1\}$ use B' to cover B, (diff norm)

$B'\to\|\cdot\|'$ $\{x:\|x\|'\le1\}$, vol of a set: $vol(C)=\int\int\int\cdots dx$

$N\le M(\delta,K,\|\cdot\|')$ $\{\theta^1\cdots\theta^M\}$

$M\,vol(\frac{\delta}{2}B')\le vol(B+\frac{\delta}{2}B')$ Minkowski sum $C+C'$: $\{x+y: x\in C, y\in C'\}$

$$\left(\frac1\delta\right)^d\frac{vol(B)}{vol(B')}\lesssim N\le \frac{vol(B+\frac\delta2 B')}{vol(\frac\delta2 B')}=\left(\frac2\delta\right)^d\frac{vol(B+\frac\delta2B')}{vol(B')}$$

Examples. $\|\cdot\|$, same norm $(\delta\le1)$

$B+\frac\delta2 B'\subseteq 2B$, $N\le\left(\frac2\delta\right)^d\frac{vol(2B)}{vol(B)}\le\left(\frac4\delta\right)^d$

Cover unit sphere with δ balls $(4/\delta)^d$, cubes of side-length δ to cover unit cubes $\{\|\cdot\|_\infty\le1\}$

Takeaway: $\log N[\delta; \|\cdot\|, \|\cdot\|] \le d \log(\frac{1}{\delta})$

Lipschitz functions on $[0,1]$, $\mathcal{F}_L = \{f(0)=0, |f(x)-f(x')| \le L|x-x'|, \forall (x,x') \in [0,1]\}$

$\|f-g\|_\infty = \sup_{x\in[0,1]} |f(x)-g(x)|$, $\underbrace{\{f^1 \cdots f^N\}}_{\text{cover of } \mathcal{F}_L}$ such that $\underline{f \in \mathcal{F}_L}$, $\exists j\ \|f-f^j\|_\infty \le \delta$

Takeaway $\log N(\delta, \mathcal{F}_L \|\cdot\|_\infty) \asymp (\frac{L}{\delta})$ and d-dims $\asymp (\frac{L}{\delta})^d$

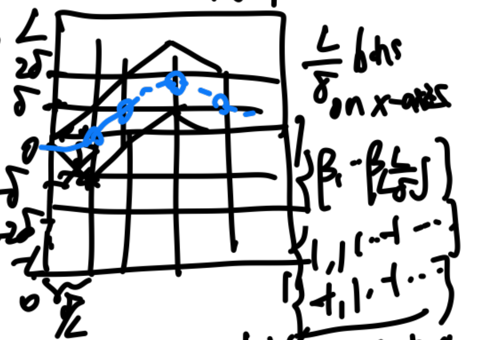
$\hookrightarrow$ Lipschitz >> "unit balls", much much large metric entropy for $\mathcal{F}_L$

Lower bound: suffices to construct 2δ packing

each binary string $\to f^j$

$\|f^j - f^k\|_\infty \ge 2\delta$

$\hookrightarrow 2^{\lfloor L/\delta \rfloor}$ binary strings, $\log N \gtrsim \lfloor \frac{L}{\delta} \rfloor$



Covering argument:

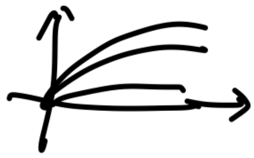
$\to$ at grid points $\{\frac{\delta}{L}, \frac{2\delta}{L}, \dots\}$, $|f^j - f| \le \delta$ $\quad$ $f^j(x) = \sum_{i=1}^{j} \beta_i^j \phi_i$ (L-function)

$\to$ at non-grid points $|f^j(x) - f(x)| \le |f^j(x) - f^j(x_0)| + |f^j(x_0) - f(x_0)| + |f(x_0) - f(x)|$

(x_0: nearest grid pt) $\le 3\delta$

Example. parametric class

$f_\theta(x) = 1 - \exp(-\theta x)$, $\theta \in [0,1]$



indexed by single param $\theta \in \mathbb{R}$

$\log N \le d \log(\frac{1}{\delta})$ $(\theta \in \mathbb{R}^d)$

High-level: to cover $\mathcal{F}_\theta$, just need to "cover" θ. intuition: change θ a little bit f_θ changes a little bit

$\rho(f_\theta, f_{\theta'}) \le \rho(\theta, \theta')$

$\sup_{x\in[0,1]} |f_\theta(x) - f_{\theta'}(x)| \le C(\theta - \theta')$

Smooth function. Holder smooth. S, $\alpha \in (0,1)$, $s=1, \alpha=0 \to$ Lipschitz

$|f^{(s)}(x)| \le L$, $|f^{(s)}(x) - f^{(s)}(x')| \le L|x-x'|^\alpha$ $\quad$ $|f'(x)| < L$ $\quad$ $(f(x) - f(x')) \le L(x-x')$

$\log N \lesssim (\frac{L}{\delta})^{\frac{d}{s+\alpha}}$

Le Cam's equation.

Rate of convergence.

$n\epsilon^2 \asymp \log N(\epsilon, \mathcal{F}, \|\cdot\|)$, ϵ: rate of convergence of $\|\cdot\|$

$\epsilon \downarrow \to \uparrow$, $n\epsilon^2 \asymp (\frac{L}{\epsilon})^d$, $\epsilon \asymp (\frac{1}{n})^{\frac{1}{2+d}}$ typical non-param rate

$n\epsilon^2 \asymp d \log(\frac{1}{\epsilon})$ $\epsilon \asymp \sqrt{\frac{d \log d}{n}}$, Large metric complexity $\to$ small conv rate

Recap.

1. Metric entropy: $\log N(\delta; K, C)$ (详阅)

2. For $\mathbb{B}=\{\theta : \|\theta\| \le 1\}$ & $\mathbb{B}'=\{\theta, \|\theta\|' \le 1\}$

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

Roughly, $N(\delta; \mathbb{B}, \|\cdot\|') = \left(\frac{1}{\delta}\right)^d \frac{vol(\mathbb{B})}{vol(\mathbb{B}')}$

$\rightarrow N(\delta; \mathbb{B}_2(1), \|\cdot\|_2) \asymp \left(\frac{1}{\delta}\right)^d$ 

$N(\delta, \mathbb{B}_\infty(1), \|\cdot\|_\infty) \asymp \left(\frac{1}{\delta}\right)^d \rightarrow$ 

3. $\mathcal{F}_L = \{f(0)=0, |f(x)-f(x')| \le L|x-x'|, x\in[0,1]\}$

$\log N(\delta; \mathcal{F}_L, \|\cdot\|_\infty) \asymp \frac{L}{\delta}$

4. "Parametric functions" $\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \log\left(\frac{1}{\delta}\right)$

5. Le Cam: $n\varepsilon^2 \asymp \log N(\varepsilon, \mathcal{F}, \|\cdot\|)$ $\frac{d\log(\frac{1}{\delta})}{}$ $\hookrightarrow$ d-dim

Ellipses: $f(x)=\sum_{i=1}^{\infty}\theta_i\varphi_i(x)$ $x\in[0,1]$, $\int_0^1 \varphi_i^2(x)dx=1$, $\int_0^1 \varphi_i(x)\varphi_j(x)dx=0$ $(i\ne j)$

$\int_0^1 f^2(x)dx<\infty$ Basic functions

Example. Fourier basis  orthonormal function 

orthogonal series estimators, θ.

smooth functions: "most of the high-order coefficients are nearly zero"

coefficient decay

Lipschitz functions: Fourier basis, $\sum_{j=1}^{\infty} j^{2\alpha}\theta_j^2 \le 1$ (Holder functions) α-times diff α derivative

$\alpha=100$, $j^{200}\theta_j^2\le 1$. $\theta_j \le \frac{1}{j^{2\alpha}}$

$\alpha\uparrow$, θ_j decay faster

L_2 functions $\sum_{j=1}^{\infty}\theta_j^2<\infty$.

Lipschitz $\sum_{j=1}^{\infty} j^2\theta_j^2 \le L$ (related to L)

$y_i=\theta_i+\varepsilon_i$, $\varepsilon_i\sim N(0,1)$

$\{\sum_{j=1}^{\infty}\theta_j^2 j^{2\alpha}\le 1\}$. Orthogonal series estimators 

axis of ellipse, $j^{-\alpha}$, $1^{-\alpha}, 2^{-\alpha}, 3^{-\alpha}\ldots$ high-dim collapse ...

Naive est: y $Ey=0$, $Var=\infty$. (bias $\to$ unbias. Var $\uparrow$)

Truncated series estimator: pick T

Bias $\sum_{j=T+1}^{\infty}\theta_j^2 \le T^{-2\alpha}\sum_{j=T+1}^{\infty}\theta_j^2 j^{2\alpha}$, Variance: $\frac{T}{n}=\sum_{j=1}^{T}E(y_j-\theta_j)^2=\sum_{j=1}^{T}Var(\varepsilon_j)=\frac{T}{n}$

$(j^{2\alpha}\ge T^{2\alpha})$

$\le T^{-2\alpha}$

$MSE \leq T^{-2\alpha} + \frac{T}{n} \asymp n^{-\frac{2\alpha}{2\alpha+1}}$, $T \asymp n^{\frac{1}{2\alpha+1}}$

$y_i = f(x_i) + \varepsilon_i$, $f(x_i) = \sum_{j=1}^{\infty} \theta_j \varphi_j(x_i)$

$z_1 = \frac{1}{n}\sum_{i=1}^{n} y_i \varphi_1(x_i)$, $\mathbb{E} z_1 = \theta_1$ $\quad \int_0^1 f(x)\varphi_j(x)dx = \theta_j$

$z_j = \frac{1}{n}\sum_{i=1}^{n} y_i \varphi_j(x_i)$, $\mathbb{E} z_j = \theta_j$

Metric entropy of $\{\sum_{j=1}^{\infty} \theta_j^2 j^{2\alpha} \leq 1\} \Rightarrow K$, $\sum_{j=1}^{\infty} \theta_j^2 \leq 1 \to$ infinite metric entropy

$\log N(\delta, K, \|\cdot\|_2) \asymp (\frac{1}{\delta})^{1/\alpha}$

- truncate ellipse at some T can "ignore" all the remaining axis...
- cover finite dimensional ellipse

Metric entropy & sub-Gaussian processes.

1. X. sub-Gaussian if $\mathbb{E}\exp(tX) \leq \exp(\frac{t^2\sigma^2}{2})$
2. $\theta \in K, \to X_\theta$, X_θ mean 0

Sub-Gaussian process w.r.t. ρ

$\mathbb{E}\exp[t(X_\theta - X_{\theta'})] \leq \exp\left(\frac{t^2\rho^2(\theta,\theta')}{2}\right)$, $\forall t \in \mathbb{R}, \theta, \theta' \in K$

$\to$ canonical GP

Bound. $\mathbb{E}\sup_{\theta \in K} \langle \theta, \varepsilon \rangle$, $\varepsilon \sim N(0, I_d)$. $\quad \langle \theta, \varepsilon \rangle = X_\theta$

X_θ: sub-GP w.r.t. $\rho(\theta,\theta') = |\theta - \theta'|_2$

$X_\theta - X_{\theta'} = \langle \theta - \theta', \varepsilon \rangle = N(0, |\theta - \theta'|_2^2)$

X_θ. $\mathbb{E}\sup_\theta X_\theta$

1-step discretization.

$N(\delta, K, \rho)$

If N Gaussians

$\mathbb{E}\sup X_\theta \asymp \sqrt{\log N}$

$\sqrt{\log N(\delta, K, \rho)}$ + Naive δ-dep bound of net points

$$\mathbb{E}\sup_{\theta} X_\theta \leq 2\,\mathbb{E}\left[\sup_{\substack{\gamma,\gamma' \in K \\ \rho(\gamma,\gamma') < \delta}} (X_\gamma - X_{\gamma'})\right] + 4\sqrt{D^2 \log N(\delta, K, \rho)}$$

$D = \sup_{\theta,\theta'} \rho(\theta,\theta')$

$$\mathbb{E}\sup_\theta X_\theta = \mathbb{E}\sup_\theta (X_\theta - X_{\theta_0}) \quad (\mathbb{E}X_{\theta_0} = 0)$$

$$\leq \mathbb{E}\sup_{\theta,\theta'} (X_\theta - X_{\theta'})$$

θ. netpoint θ^i such that. $\rho(\theta, \theta^i) \leq \delta$

$$X_\theta - X_{\theta_0} = X_\theta - X_{\theta^i} + X_{\theta^i} - X_{\theta_0}$$

$$\leq \sup_{\substack{\theta,\theta' \in K \\ \rho(\theta,\theta') \leq \delta}} (X_\theta - X_{\theta'}) + \max_{i=1\cdots N} |X_{\theta^i} - X_{\theta_0}|$$

$$X_\theta - X_{\theta'} \leq 2 \sup_{\substack{\theta,\theta' \in K \\ \rho(\theta,\theta') < \delta}} [X_\theta - X_{\theta'}] + 2\max_{j=1\cdots N} |X_{\theta^j} - X_{\theta_0}|$$

$4\sqrt{D^2 \log N(\delta, K, \rho)}$

$$\mathbb{E}\sup_{\theta,\theta'}(X_\theta - X_{\theta'}) \leq 2\,\mathbb{E}\ \_\_\_ + 2\,\mathbb{E}\max_{j=1\cdots N}|X_{\theta^j} - X_{\theta_0}|$$

$$\max_{i=1\cdots N} |X_{\theta^i} - X_{\theta^1}|$$

$$E\exp t\underbrace{(X_{\theta^i} - X_{\theta^1})} \le \exp\left(\frac{t^2 \rho^2(\theta^i, \theta^1)}{2}\right) \le \exp\left(\frac{t^2 D^2}{2}\right)$$

$X_1 \ldots X_N$, σ-sub-Gaussian..

$$E\max|X_i| \le 2\sigma\sqrt{\log N}$$

Recap.

1. Gaussian complexity of K

$$G(K)=\mathbb{E}\sup_{\theta\in K}\theta^T\varepsilon,\ \varepsilon\sim N(0,I_d)$$

$\|\theta-\theta'\|_2$ L_2 norm

Sub-Gaussian process with L_2

2. Sub-Gaussian process $\{X_\theta,\theta\in K\}$ mean 0

$$\mathbb{E}\exp(t(X_\theta-X_{\theta'}))\le\exp\left(\frac{t^2\rho^2(\theta,\theta')}{2}\right),\ \text{e.g. } \underline{\underline{X_\theta=\theta^T\varepsilon}},\ \mathbb{E}\exp(t(X_\theta-X_{\theta'}))\le\exp\left(\frac{t^2\|\theta-\theta'\|_2^2}{2}\right)$$

1b Function spaces,

$f=\sum_{i=1}^{\infty}\theta_i\varphi_i$, $\sum\theta_i^2<\infty$

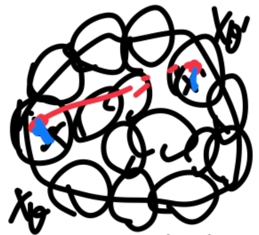
Smoothness $\sum_{i=1}^{\infty}i^{2\alpha}\theta_i^2\le1\to\log N(\delta,K,\|\cdot\|_2)\asymp\left(\frac{1}{\delta}\right)^{\frac{1}{\alpha}}$

$y_i=\theta_i+\varepsilon_i$, $\varepsilon_i\sim N(0,\frac{1}{n})$, then $\mathbb{E}\|\hat\theta-\theta\|^2\sim\frac{T}{n}+T^{-2\alpha}$ → bias, variance; $n^{-\frac{2\alpha}{2\alpha+1}}$

truncated series

3. One-step. $\mathbb{E}\sup_\theta X_\theta\le\mathbb{E}\sup_{\theta,\theta'}(X_\theta-X_{\theta'})\le\mathbb{E}\sup_{\theta,\theta'}\rho(\theta,\theta')\le\delta$

$$X_\theta-X_{\theta'}+\sqrt{D^2\log N(\delta,K,\rho)}$$



1. center distance
2. X to center.

Expected operator norm of a random matrix

n. W. $W_{ij}\sim N(0,1)$, $\langle W,uv^T\rangle=\sum_{ij}W_{ij}u_iv_j$, $u\in\mathbb{R}^n$ $v\in\mathbb{R}^d$

$$\mathbb{E}\|W\|_{op}=\mathbb{E}\sup_{\|u\|_2\le1,\ \|v\|_2\le1}u^TWv=\mathbb{E}\sup_{\|u\|\le1,\ \|v\|_2\le1}\langle W,uv^T\rangle$$

$\to=\mathbb{E}\sup_{M\in\mathcal{M}}\langle W,M\rangle=G(\mathcal{M})$

sub-Gaussian process with metric $\rho(M,M')=\|M-M'\|_F$

$\mathcal{M}=\{M\in\mathbb{R}^{n\times d},\ \mathrm{rank}(M)=1,\ \|M\|_F\le1\}$

$D=\sup_{M,M'}\|M-M'\|_F\le2$ diameter of $\mathcal{M}$

first term

nuclear norm $\sum_i\sigma_i(M)\le\sqrt{s}\sqrt{\sum_i\sigma_i^2(M)}$, rank ≤2

$$\mathbb{E}\|W\|_{op}\le\mathbb{E}\sup_{M,M':\ \|M-M'\|_F\le\delta}\langle W,M-M'\rangle+\sqrt{\log N(\delta,\mathcal{M},\|\cdot\|_F)}\le\mathbb{E}\|W\|_{op}\|M-M'\|_{nuclear}\le\mathbb{E}\|W\|_{op}\sqrt{\mathrm{rank}(M-M')}\|M-M'\|_F$$

$\le\sqrt{2}\,\delta\,\mathbb{E}\|W\|_{op}$

second term

$\mathbb{E}\sup\{X_1\cdots X_N\}\le\sigma\sqrt{\log N}$

$U=\{u_1\cdots u_M\}\subset\mathbb{R}^n$ $\forall u,\ \|u\|_2=1$, $\|u'-u\|\le\delta/2$

$V=\{v_1\cdots v_N\}\subset\mathbb{R}^d$

$\{u_iv_j^T\cdots\}$ $i\in\{1\cdots M\},\ j\in\{1\cdots N\}$, covering $\mathcal{M}$ using U,V.

$\mathcal{U}^\delta$ covers $\mathcal{U}$, $\mathcal{V}^\delta$ covers $\mathcal{V}$

$$\|UV^T - U^\delta V^{T\delta}\|_F = \|UV^T - UV^{\delta} + UV^{\delta T} - U^\delta V^{\delta T}\|_F$$
$$\leq \|U\|_2 \|V - V^\delta\|_2 + \|U - U^\delta\|_2 \|V^\delta\|_2 \leq \delta$$

$$M \times N : \left(\frac{2}{\delta}\right)^n \times \left(\frac{2}{\delta}\right)^d$$

pick $\delta = \frac{1}{2\sqrt{2}}$

$$\mathbb{E}\|W\|_{op} \leq \sqrt{2}\,\delta\, \mathbb{E}\|W\|_{op} + C\sqrt{(n+d)\log\left(\frac{1}{\delta}\right)} \Longrightarrow \mathbb{E}\|W\|_{op} \lesssim \sqrt{n+d}$$

intuition. u.v. $u^T W v \sim \mathcal{N}(0,1)$

union bound $\left(\frac{1}{\delta}\right)^{n+d}$

$$\mathbb{E}\|W\|_{op} \leq \mathbb{E}\|W\|_F \leq \sqrt{n \times d}$$

$\underset{L_\infty}{\downarrow}$ $\quad \approx \sqrt{n \times d}$ $\quad \underset{L_2}{\downarrow}$ $\quad L_1$ norm

$W(\sigma_1 \ldots \sigma_d)$.

$\|W\|_{op} = \max \sigma_i$

$\|W\|_{nuc} = \sum \sigma_i$

$\|W\|_F = \sqrt{\sum \sigma^2}$

Non-parametric regression

$y_i = f^*(x_i) + \varepsilon_i \quad \varepsilon_i \sim \mathcal{N}(0,1)$

$f \in \mathcal{F}_L = \{ f(0) = 0,\ |f(x) - f(y)| \leq L|x-y| \}$

Non-para. L.S.

$$\hat{f} = \underset{f \in \mathcal{F}_L}{\arg\min} \|y - f\|_2^2$$

Representer thm.

finite-dim opt. prob.

Basic ineq:
$$\frac{1}{n}\sum_{i=1}^n (\hat{f}(x_i) - y_i)^2 \leq \frac{1}{n}\sum_{i=1}^n (f^*(x_i) - y_i)^2$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))^2\right] \leq \mathbb{E}\frac{2}{n}\sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))\,\varepsilon_i \rightarrow \text{Gaussian complexity}$$

{in-sample error / fix design error / $\|\hat{f} - f\|^2_{L_2(P_n)}$} $\quad \int (\hat{f} - f)^2 \phi(x)\,dx = \|\hat{f} - f\|^2_{L_2(P)}$

Now: bound Gaussian complexity

$$\mathbb{E}\frac{1}{\sqrt{n}}\sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))\varepsilon_i \leq \mathbb{E}\sup_{f \in \mathcal{F}_L} \frac{1}{\sqrt{n}}\sum_{i=1}^n (f(x_i) - f^*(x_i))\varepsilon_i$$
$$\leq \mathbb{E}\sup_{f, f' \in \mathcal{F}^n} \frac{1}{\sqrt{n}}\langle f - f', \varepsilon\rangle \qquad \mathcal{F}^n = \{ (f(x_1), \ldots, f(x_n)),\ f \in \mathcal{F}_L \}$$

Natural metric: $\rho^2(f, f') = \frac{1}{n}\sum_{i=1}^n (f(x_i) - f'(x_i))^2 = \|f - f'\|^2_{L_2(P_n)}$

$$\leq \mathbb{E}\sup_{\substack{f, f' \\ \|f - f'\|_{L_2(P_n)} \leq \delta}} \langle \frac{f - f'}{\sqrt{n}}, \varepsilon \rangle + \sqrt{D \cdot \log N(\ \ )}.$$

$\|f' - f\|_{L_2(P_n)} \leq \|f' - f\|_\infty$

$D = \sup_{f, f' \in \mathcal{F}_L} \|f - f'\|_{L_2(P_n)} \leq L$ (can drop it)

$$\leq \sqrt{n} \cdot \sqrt{n}\,\delta + \sqrt{\frac{1}{\delta}}$$

$\lesssim n^{2/3}$ $\qquad \delta^{3/2} \asymp n^{-1/2}$

$\delta \asymp n^{-1/3}$

Recap

1. Sub-Gaussian process $\{X_\theta : \theta \in K\}$

$$\mathbb{E}\exp(t(X_\theta - X_{\theta'})) \le \exp\left(\frac{t^2 \rho(\theta,\theta')}{2}\right)$$

2. One-step discretization. mean $X_\theta = 0$

$$\mathbb{E}\sup_{\theta\in K} X_\theta \le \mathbb{E}\sup_{\substack{\theta,\theta' \\ \rho(\theta,\theta')\le\delta}} X_\theta - X_{\theta'} + \sqrt{D^2 \log N(\delta, K, \rho)}$$

Applications

1. $\mathbb{E}\|W\|_{op} \le \sqrt{n} + \sqrt{d}$, $G(\{M : \mathrm{rank}(M)=1,\ \|M\|_F = 1\})$

Nonparam: local smoothing $\to$ 3570S

2. Lipschitz fn estimation

n-par LS: $\hat f = \arg\min_{f\in F} \|y - f\|_2^2$, $\frac{1}{n}\sum_{i=1}^n (\hat f(x_i) - f^*(x_i))^2 \le \frac{2}{n}\sum_{i=1}^n (\hat f(x_i) - f^*(x_i))\varepsilon_i$

$$\mathbb{E}\frac{1}{n}\sum_{i=1}^n (\hat f(x_i) - f^*(x_i))^2 \le \frac{1}{\sqrt n}\mathbb{E}\sup_{f\in F_L} \frac{1}{\sqrt n}\langle \hat f - f^*, \varepsilon\rangle, \quad \rho(f,f') = \|f-f'\|_{L_2(P_n)} = \sqrt{\frac{1}{n}\sum_{i=1}^n (f(x_i) - f'(x_i))^2}$$

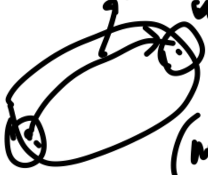
$\le \|f - f'\|_\infty$

$$\le \frac{1}{\sqrt n}\mathbb{E}\sup_{\substack{f,f'\in F_L \\ \|f-f'\|_{L_2(P_n)}\le\delta}} \left\langle \frac{f-f'}{\sqrt n}, \varepsilon\right\rangle + \ (\text{Lathy}) \text{ check...?} \quad \longleftrightarrow \ \log N_\infty : \log N \approx \sqrt{\frac{L}{\delta}}$$

$\le \mathbb{E}\|\varepsilon\|\,\|f-f'\|_{L_2(P_n)}$

$\le \sqrt{n}\times\delta$

$$\le \frac{1}{\sqrt n}(\sqrt{n}\times\delta) + \sqrt{\frac{L}{\delta}} \lesssim n^{-\frac{1}{3}}$$



scale up width D

union bound (not optimal)

now increasing the resolution of covering

δ, 2δ, 4δ

$|X_\theta - X_1| + |X_1 - X_2| + |X_2 - X_3|$

$\le \boxed{|X_\theta - X_1|} + |X_1 - X_2| + |X_2 - X_3|$

distance. $\sqrt{\delta^2\log(\delta)} + \sqrt{(2\delta)^2\log(2\delta)} \ldots$

Dudley's bound

$$\mathbb{E}\sup_\theta X_\theta \lesssim \mathbb{E}\sup_{\theta,\theta':\rho(\theta,\theta')\le\delta} X_\theta - X_{\theta'} + \int_\delta^D \sqrt{\log N(u, K, \rho)}\,du \le \int_0^D \sqrt{\log N(u, K, \rho)}\,du$$

entropy integral

$$\mathbb{E}\frac{1}{n}\sum_{i=1}^n (\hat f - f^*(x))^2 \le \frac{1}{\sqrt n}\int_0^D \frac{1}{\sqrt u}\,du \lesssim \frac{1}{\sqrt n},$$

better than $n^{-1/3}$, but if ~~d dim~~

lipschitz smoothing $n^{-\frac{2}{2+d}}$

$\int_0^D \left(\frac{L}{\delta}\right)^d d\delta$. $d > 2$. diverge

$D = \sup_{f,f'} \|f - f'\|_{L_2(P_n)} \le 2L$

Density estimation $X_1 \cdots X_n \sim p$ on $[0,1]$ $\to$ make assumptions on p: $\to$ non-para. method. (local smoothing).

$\to$ make assumptions on metric weaker 7 of

$\hat p$, $\hat F(x) = \int_{-\infty}^x \hat p(x)\,dx$, $\rho(\hat p, p) = \sup_x |\hat F(x) - F(x)|$

$= \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{x_i \le x\}$

Wasstein-1 : norm 1

$$W_1(p,q) = \sup_{|f(x)-f(y)| \le |x-y|} |E_p f - E_q f| \quad , \quad TV(p,q) = \sup_{\|f\|_\infty \le 1} |E_p f - E_q f|$$

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = x\}$$

$$E\, W_1(\hat{p}, p^*) = \sup_{\|f\|_L \le 1,\ f(0)=0} \left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - E_{p^*} f\right| = \sup_{f \in \mathcal{F}_L} X_f, \quad X_f = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - E_{p^*} f$$

$$X_f - X_g = \frac{1}{n}\sum_{i=1}^{n} \left(f(X_i) - g(X_i)\right) - E_{p^*}(f-g) \rightarrow$$ Average of bounded RVs. MGF

$X_f - X_g$ changes by at most $\frac{2\|f-g\|_\infty}{n}$

Azuma's bound

$$E \exp\left(t(X_f - X_g)\right) \le \exp\left(\frac{c t^2 \|f-g\|_\infty^2}{n}\right)$$

Hoeffding

$(b-a)^2 = (X_f - X_g)^2$

∴ X_f is a sub-Gaussian process with $\rho(f,g) = \frac{\|f-g\|_\infty}{\sqrt{n}}$

Apply Dudley bound + Lipschitz

$$E\, W_1(\hat{p}, p^*) \le \frac{1}{\sqrt{n}}$$

: !!! $\frac{1}{n}\sum_{i=1}^{n} f(X_i) - Ef = O\left(\frac{1}{\sqrt{n}}\right)$ for all f 1-dim

d-dim : rate $n^{-\frac{1}{2d}}$

VC classes

$$S(\mathcal{C}) = \sup_{C \in \mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in C\} \to P(C)$$

$\mathcal{C}$ has VC-dim d

$$E \le \sqrt{\frac{d \log n}{n}}$$

?

Covering number for VC collections of sets

$$N(\delta, \mathcal{C}, \|\cdot\|_{L_2(P_n)}) \le d\, 2^d \left(\frac{1}{\delta}\right)^d$$

sub-Gaussian

$\le \sqrt{\frac{d}{n}}$

Recap.

1. Dudley's chaining bound

$$\mathbb{E}\sup_{\theta\in K} X_\theta \le \mathbb{E}\sup_{\substack{\theta,\theta'\in K\\ \rho(\theta,\theta')\le\delta}} (X_\theta - X_{\theta'}) + \int_\delta^D \sqrt{\log N(u,K,\rho)}\,du$$

Entropy Integral

2. Applications.

1. Lipschitz regression $\mathbb{E}\frac{1}{n}\sum_{i=1}^n (\hat f(x_i) - f(x_i))^2 \lesssim \frac{1}{\sqrt{n}},\ n^{-\frac{2}{3}}\ \ n^{-\frac{1}{3}}$

tight $\{$ 2. W_1-distance between P_n & P: $\mathbb{E}W_1(P_n,P) \lesssim \frac{1}{\sqrt{n}}$ in 1-D

3. VC-classes $\mathbb{E}\sup_{C\in\mathcal{C}} |P_n(C) - P(C)| \lesssim \sqrt{\frac{d}{n}}$, $\left(\text{vs } \sqrt{\frac{d\log n}{n}}\right)$

1. Gaussian comparison ineq. N $\{X_1 \cdots X_n\}$

$\mathbb{E}\sup_\theta X_\theta$ Gaussian process $\to \{Y_1 \cdots Y_n\}$, $\mathbb{E}\sup_i X_i$ v/s $\mathbb{E}\sup_i Y_i$

Sudakov-Fernique. Intrinsic metric $\rho_X^2(i,j) = \mathbb{E}(X_i - X_j)^2$

$\rho_X(i,j) \le \rho_Y(i,j)\ \forall i,j \Rightarrow \mathbb{E}\sup_i X_i \le \mathbb{E}\sup_i Y_i$

Gaussian contraction ineq.

$\mathbb{E}\sup_{\theta\in K} \langle\theta,\varepsilon\rangle$ $\quad \phi(\theta) = (\phi_1(\theta_1) \cdots \phi_d(\theta_d))$ $\hookrightarrow \phi(K)$

v/s. $\mathbb{E}\sup_{\substack{\alpha\in\phi(K)\\ \theta\in K}} \langle\phi(\theta),\varepsilon\rangle$

$\int (\hat f(x) - f(x))^2 p(x)\,dx$

$\frac{1}{n}\sum (\hat f(x_i) - f(x_i))^2$

$G(\mathcal{F}_1^n) = \mathbb{E}\sup_{f\in\mathcal{F}_1^n} \langle f,\varepsilon\rangle$, $H(\mathcal{F}_2^{n,2}) = \mathbb{E}\sup_{f\in\mathcal{F}_2^n} \langle f^2,\varepsilon\rangle$ $\hookrightarrow \left|\frac{1}{n}\sum_{i=1}^n f^2(x_i) - \mathbb{E}f^2\right|$ small for all fs.

If ϕ_i — 1-Lipschitz & $\phi_i(0) = 0$ then $\mathbb{E}\sup_\theta\langle\phi(\theta),\varepsilon\rangle \lesssim \mathbb{E}\sup_\theta\langle\theta,\varepsilon\rangle$ $\to$ proof.

e.g. $\|f\|_\infty \le b$, $\phi_i(x) = \begin{cases} \frac{x^2}{2b} & |x|\le b \\ \frac{b}{2} & o/w \end{cases}$. $H(\mathcal{F}_2^{n,2}) \le 2b \cdot G(\mathcal{F}_1^n)$?

$\mathbb{E}(\theta_i - \theta_j)^2 \ge \mathbb{E}(\phi(\theta_i) - \phi(\theta_j))^2$

1-Lipschitz. 1-Lipschitz ✓

Sudakov minoration.

$\mathbb{E}\sup_{\theta\in K} X_\theta \ge \sup_\delta \frac{\delta}{2}\sqrt{\log M(\frac{\delta}{2}, K, \rho_X)}$, $\rho_X(\theta,\theta') = \mathbb{E}(X_\theta - X_{\theta'})^2$

Proof: $Y_\theta \sim N(0, \delta^2/2)$

$\mathbb{E}(X_{\theta^i} - X_{\theta^j})^2 > \delta^2$

$\mathbb{E}[Y_{\theta^i} - Y_{\theta^j}]^2 = \delta^2$

$$\mathbb{E}\sup_{\theta\in K} X_\theta \geq \mathbb{E}\sup_{i\in\{1\cdots m\}} X_{\theta^i} \geq \mathbb{E}\sup_{i\in\{1\cdots m\}} Y_{\theta^i} \geq \delta\sqrt{\log M}$$

Metric entropy of an l_1-ball 

$$\omega(K) = \mathbb{E}\sup_{\|\theta\|_1\leq 1}\langle\theta, \varepsilon\rangle = \mathbb{E}\|\theta\|_1\|\varepsilon\|_\infty \leq 2\sqrt{\log d}$$

$$\log N(\delta, K, \|\cdot\|_2) \lesssim \underbrace{\frac{\log d}{\delta^2}}_{L_1\text{-ball}, \delta\leq 1}$$

$$\log N(\delta, \underbrace{\|\cdot\|_2}_{L_2\text{-ball}}, \|\cdot\|_2) \lesssim d\log\left(\frac{1}{\delta}\right)$$

Matrix estimation.
$\begin{cases} 1.\ \text{covariance matrix est. } X_1\cdots X_n\sim P,\ \Sigma=\mathbb{E}(X-\mu)(X-\mu)^T,\ \hat{\Sigma}=\frac{1}{n}\sum_{i=1}^n (X_i-\hat{\mu})(X_i-\hat{\mu})^T \\ 2.\ \text{Matrix completion} \to M, S.\ M_{S}+\text{noise} \\ 3.\ \text{Matrix + Noise} \end{cases}$

$y_i = \operatorname{tr}(M^T X_i) + \varepsilon_i$. $\beta^{*T}X_i + \varepsilon_i$

Matrix completion $X_i = \mathbb{1}_{jk}$

$$\downarrow$$

$$Y = M^* + W$$

want to est M^*, observe Y

Stochastic Block Models

$M^*_{ij} = \boxed{\begin{matrix} p & q \\ q & p\end{matrix}}$ $(p>q)$ $\quad \underbrace{Y_{ij}}_{\text{observe.}} = \begin{cases} 1, & \text{with prob. } M^*_{ij} \\ 0, & \text{with prob. } 1-M^*_{ij}\end{cases}$

$E(Y_{ij}) = M^*_{ij}$, $Y = M^* + \underbrace{W}_{\text{mean } 0}$

Reminder - Low rank matrix est. / sparse-vect est. 7.05

hard thresholding $\leftarrow$ w.p. $\geq 1-\delta$ $\|\hat{\theta}-\theta^*\|_2^2 \lesssim \sum_{i=1}^d \min\left\{\theta_i^{*2}, \frac{\sigma^2\log(d/\delta)}{n}\right\}$

$\to \theta^*$ is s sparse, $\frac{s\log\frac{d}{\delta}}{n}$

θ^* is l_1 sparse $\leq R$ $\quad \sqrt{\frac{R\log(d/\delta)}{n}}$

Matrix analog (symmetric) $M^* = \sum_{i=1}^n \alpha_i v_i v_i^T$, $Y = \sum_{i=1}^n \beta_i u_i u_i^T$, $\hat{M} = \sum_{i=1}^n \beta_i \mathbb{1}\{|\beta_i|\geq t\} u_i u_i^T$

$Y = M^* + W$. If $\|W\|_{op}\leq t/2$, $\|\hat{M}-M^*\|_F^2 \lesssim \sum_{i=1}^n \min\{\underbrace{\alpha_i^2}_{\text{bias}}, \underbrace{t^2}_{\text{Var}}\}$

Recap.

1. Sudakov-Fernique. $\{X_1 \dots X_N\}$, $\{Y_1 \dots Y_N\}$ Gaussian process.

If $\mathbb{E}(X_i - X_j)^2 \le \mathbb{E}(Y_i - Y_j)^2 \ \forall i, j$.

Then. $\mathbb{E}\sup_i X_i \le \mathbb{E}\sup_j Y_j$

1a) Gaussian contraction.

$\mathbb{E}\sup_\theta \langle \varphi(\theta), \varepsilon \rangle \le \mathbb{E}\sup_\theta \langle \theta, \varepsilon \rangle$, If $\varphi(\theta)$ is a contraction

e.g. $\varphi(\theta) = (\varphi_1(\theta) \dots \varphi_n(\theta))$

φ_i is 1-Lipschitz

1b) Sudakov minoration. (packing)

$\mathbb{E}\sup_{\theta \in K} X_\theta \ge \sup_\delta \frac{\delta}{2}\sqrt{\log M(\delta; K, \rho_X)}$

where $\rho_X(\theta, \theta') = \mathbb{E}(X_\theta - X_{\theta'})^2$

2. Levy's ineq.

If f is L-Lipschitz. $\mathbb{P}(|f(z) - \mathbb{E}f(z)| \ge t) \le 2\exp\left(-\frac{t^2}{2L^2}\right)$

z is std. normal. $\left|\sup_\theta X_\theta - \mathbb{E}\sup_\theta X_\theta\right|$

Matrix. Est.

$Y = \underset{\text{Low rank}}{M^*} + W \rightarrow$ indep. sub-Gaussian mean 0 entries.

e.g.

$Y = \sum_{i=1}^n \beta_i v_i v_i^T$, $\hat{M} = \sum_{i=1}^n \beta_i \mathbb{1}\{|\beta_i| \ge 2t\} v_i v_i^T$, $M^* = \sum_{i=1}^n \lambda_i u_i u_i^T$

Suppose $\|W\|_{op} \le t$, then $\|\hat{M} - M^*\|_F^2 \lesssim \sum_{i=1}^n \min\{t^2, \lambda_i^2\}$ $\Rightarrow$ prove it!

$\beta_1 \le \beta_2 \le \dots \le \beta_n$, $\lambda_1 \le \lambda_2 \le \dots \le \lambda_n$.

$S = \{i, |\beta_i| \ge 2t\}$. Weyl's Law. (matrix perturbation) $\Rightarrow |\lambda_i - \beta_i| \le \|W\|_{op}$

eigenvalue shift

$\|\hat{M} - M^*\|_F \le \|\hat{M} - M_S^*\|_F + \|M_S^* - M^*\|_F$, $M_S^* = \sum_{i \in S} \lambda_i u_i u_i^T$

$\|M_S^* - M^*\|_F \le \sqrt{\left(\sum_{i \in S^c} \lambda_i u_i v_i^T\right)^2} \le \sqrt{\sum_{i \in S^c} \lambda_i^2}$

$\|\hat{M} - M_S^*\|_F \le \sqrt{2|S|}\,\|\hat{M} - M_S^*\|_{op}$?

$\le \sqrt{2|S|}\left(\underset{\le 2t}{\|\hat{M} - Y\|_{op}} + \underset{\le t}{\|Y - M^*\|_{op}} + \|M^* - M_S^*\|_{op}\right)$, $\|M^* - M_S^*\|_{op} = \max_{i \in S^c} |\lambda_i| \le 3t$

$\sqrt{a} + \sqrt{b} \le 2\sqrt{a+b}$ $\lesssim t\sqrt{|S|}$

$\Rightarrow \|\hat{M} - M^*\|_F \lesssim \sqrt{t^2|S| + \sum_{i \in S^c} \lambda_i^2} \le \sqrt{\sum_{i=1}^n \min\{t^2, \lambda_i^2\}}$ ∎ $\begin{cases} \text{on } S, |\lambda_i| \ge t & \text{by Weyl's} \\ \text{on } S^c, |\lambda_i| \le 3t \end{cases}$

Suppose: $\mathbb{E}\|W\|_{op} \lesssim \sqrt{n}$, $t \asymp \sqrt{n}$

$\|\hat{M} - M^*\|_F^2 \lesssim \sum_{i=1}^n \min\{n, \lambda_i^2\}$ Suppose M^* has rank r, $\|\hat{M} - M^*\|_F^2 \lesssim nr$

$\text{nuc} \sum_{i=1}^n |\lambda_i| \le R$ $(n-r)$ 0 eigenvalue

Oracle Ineq. form

$\|\hat{M}-M^*\|_F^2 \le \inf_\theta \{ \underbrace{\|M^*-\theta\|_F^2}_{\text{bias}} + \underbrace{n\,\mathrm{rank}(\theta)}_{\text{variance}} \}$, M^* not necessarily low-rank, θ is low rank

prove. $\|\hat{M}-M^*\|_F^2 \lesssim \|M^*-M^*_r\|_F^2 + r\times n$

$\le \sum_{i=1}^n \min\{\lambda_i^2, n\} \le \sum_{i\in S^c} \lambda_i^2 + \sum_{i\in S} n$

Universal SVT

Strongly stochastically transitive matrices (SST) Nihar. Martin. Siva

$\{1\dots n\}$ $P(i\succ j)$. parametric model. $\{\theta_1\dots\theta_n\}$. Bradley-Terry-Luce $\propto \exp(\theta_i-\theta_j)$

non-parametric. Π^*

If i is better j.

$P(i\succ k) \ge P(j\succ k), \forall k$.

Best — increase — worst

decrease:

$$\begin{array}{c} \text{Best} \\ \\ \text{worst} \end{array} \begin{pmatrix} 1/2 & x & \\ 1-x & \ddots & \\ & & 1/2 \end{pmatrix}$$

$Y = \underbrace{M^*}_{\text{not necess. low rank}} + \text{Bernoulli noise}$

Sourav Chatterjee

$\|M^*-M^*_{\sqrt{n}}\|_F^2 \lesssim n\sqrt{n} \Rightarrow \|\hat{M}-M^*\|_F^2 \lesssim n\sqrt{n}$

$\sqrt{n}$ 组 $\sqrt{n}$

$\|M^*-\tilde{M}\|_F^2 \le \sum_{i=1}^n \sum_{j=1}^{\sqrt{n}} \underbrace{(M^*_{ij}-\tilde{M}_{ij})^2}_{\text{positive}}$

$\le \sum_{i=1}^n \sum_{j=1}^{\sqrt{n}} (M^*_{ij}-\tilde{M}_{ij})$

$\le n\sqrt{n}$

$\tilde{M} = \begin{pmatrix} r_1 \\ r_1 \\ r_2 \\ r_2 \\ \vdots \\ r_{\sqrt{n}} \end{pmatrix} \sqrt{n}$

$r_1 r_2$ 这些 rows

W. with prob $\ge 1-\delta$, $\|W\|_{op} \sqrt{n} + \sqrt{\log(\frac{1}{\delta})}$

Symmetric Gaussian. $\|W\|_{op} = \sup_{|u|=1} u^T W u$ ★ $N_{\frac{1}{2}}$ net of Sphere. S^n — 2^n $\left(\frac{1}{\delta}\right)^d$

$\sup_{\|u\|_2=1} u^T\theta \le 2 \sup_{u\in N} u^T\theta$ ⇐ 套路, u^* max over S^n. $\tilde{u}$ closest net point

$\tilde{u}^T\theta = u^{*T}\theta + \frac{(\tilde{u}-u^*)^T\theta}{\|\tilde{u}-u^*\|_2}\|\tilde{u}-u^*\|_2$

$\ge -|u^{*T}\theta|\frac{1}{2}$

$\tilde{u}^T\theta \ge \frac{u^{*T}\theta}{2}$

$u^{*T}\theta \le 2\tilde{u}^T\theta \le 2\sup_N \tilde{u}^T\theta$

Sub Gauss

$P(\sup_{u\in N} u^T W u \ge t) \le 2^n \exp(t^2)$

Recap.

1. Matrix estimation, signal noise model

$Y = M^* + W$ → independent mean 0, sub-gaussian

2. Single value thresholding.

$\hat{M} = \sum_{i=1}^{n} \beta_i \mathbb{1}\{|\beta_i| \ge 2t\} u_i u_i^T$, where $Y = \sum_{i=1}^{n} \beta_i u_i u_i^T$

3. Guarantee If $\|W\|_{op} \le t$, $\|\hat{M} - M^*\|_F \lesssim \sum_{i=1}^{n} \min\{\lambda_i^2, t^2\}$

where $M^* = \sum_{i=1}^{n} \lambda_i v_i v_i^T \lesssim \arg\min_{\theta} \{ \underbrace{\|M^* - \theta\|_F^2}_{\text{approx. error}} + \underbrace{n\,\mathrm{rank}(\theta)}_{\text{est. error}} \}$

Oracle inequality.

$\Rightarrow$ w.p. $\ge 1-\delta$

$\|W\|_{op} \lesssim \sqrt{n + \log(\frac{1}{\delta})}$

4. Example. SST. matrice increasing/decreasing, $n\sqrt{n}$ close to $\sqrt{n}$-rank matrix

Gaussian covariance matrix

$X_1 \cdots X_n \sim N(0, \Sigma) \to \underset{n \times d}{X}$, $\hat{\Sigma} = \frac{X^T X}{n} = R^T R$, $R = \frac{X}{\sqrt{n}}$, how close Σ & $\hat{\Sigma}$

Thm. $\mathbb{P}\left(\sigma_{max}\left(\frac{X}{\sqrt{n}}\right) \ge (1+\delta)\|\sqrt{\Sigma}\|_{op} + \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}\right) \le \exp\left(-\frac{n\delta^2}{2}\right)$ → proof

$\mathbb{P}\left(\sigma_{min}\left(\frac{X}{\sqrt{n}}\right) \le (1-\delta)\lambda_{min}(\sqrt{\Sigma}) - \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}\right) \le \exp\left(-\frac{n\delta^2}{2}\right)$

$X_1 \cdots X_n \sim N(\mu, \Sigma)$, $\mathbb{E}\|\hat{\mu}_n - \mu\|_2^2 = \mathbb{E}\|N(0, \frac{\Sigma}{n})\|_2^2 = \frac{\mathrm{tr}(\Sigma)}{n}$, $\frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|_{op}}$ eff. rank of Σ

$\sum_{i=1}^{d} y_i^2$

Σ is identity. with prob $\ge 1 - 2\exp\left(-\frac{n\delta^2}{2}\right)$

$\left.\begin{array}{l} \lambda_{max}\left(\frac{X^T X}{n}\right) \le \Big(1 + \underbrace{\delta + \sqrt{\tfrac{\mathrm{tr}(\Sigma)}{n}}}_{\varepsilon}\Big)^2 \le 1 + \varepsilon^2 + 2\varepsilon \\ \lambda_{min}\left(\frac{X^T X}{n}\right) \ge 1 + \varepsilon^2 - 2\varepsilon \end{array}\right\} \to \left\|\frac{X^T X}{n} - I\right\|_{op} \le \varepsilon^2 + 2\varepsilon$

$\underset{n \times d}{X} = W\sqrt{\Sigma}$, $W_i \sim N(0, I_d)$, $X_i \sim N(0, \Sigma)$

$\|\hat{\Sigma} - \Sigma\|_{op} = \left\|\Sigma^{1/2}\left(\frac{W^T W}{n} - I\right)\Sigma^{1/2}\right\|_{op} \underset{\text{Cauchy.}}{\le} \|\Sigma\|_{op} \left\|\frac{W^T W}{n} - I\right\|_{op}$

$\frac{\|\hat{\Sigma} - \Sigma\|_{op}}{\|\Sigma\|_{op}} \le \varepsilon^2 + 2\varepsilon$

w.p. $\ge 1 - 2\exp\left(-\frac{n\delta^2}{2}\right)$, $\left\|\frac{W^T W}{n} - I\right\|_{op} \le \sqrt{\frac{d}{n}} + \frac{d}{n} + \delta$

$(X = W\sqrt{\Sigma})$

$$\mathbb{P}\left(\sigma_{max}\left(\frac{X}{\sqrt{n}}\right) \geq \mathbb{E}\,\sigma_{max}\left(\frac{X}{\sqrt{n}}\right) + t\|\sqrt{\Sigma}\|_{op}\right) \leq \exp\left(\frac{-nt^2}{2}\right)$$

$$\left(\left|\sigma_{max}\left(\frac{X}{\sqrt{n}}\right) - \sigma_{max}\left(\frac{Y}{\sqrt{n}}\right)\right| \leq \frac{\|X-Y\|_{op}}{\sqrt{n}} \leq \frac{\|X-Y\|_F}{\sqrt{n}},\ \frac{1}{\sqrt{n}}\text{-Lipschitz}\right)$$

↳ Weyl's Inequ.

$\sqrt{\Sigma}\, v = \tilde{v}$.

$$\mathbb{E}\,\sigma_{max}\left(\frac{X}{\sqrt{n}}\right) = \mathbb{E}\,\sigma_{max}\left(\frac{W\sqrt{\Sigma}}{\sqrt{n}}\right) = \mathbb{E}\sup_{\|u\|,\|v\|_2\leq 1} \frac{u^T W\sqrt{\Sigma}\, v}{\sqrt{n}} = \mathbb{E}\sup_{\substack{\|\Sigma^{-1/2}\tilde{v}\|_2\leq 1\\ \|u\|\leq 1}} u^T W \tilde{v}$$

$$Z_{u,v} = u^T W v,\quad \mathbb{E}(Z_{u,v} - Z_{\tilde{u},\tilde{v}})^2 = \mathbb{E}(u^T W v - \tilde{u}^T W \tilde{v})^2 = \sum_{ij}(u_i v_j - \tilde{u}_i \tilde{v}_j)^2$$

$$= \|u v^T - \tilde{u}\tilde{v}^T\|_F^2$$

$$= \|(u-\tilde{u})\tilde{v}^T + u(v-\tilde{v})^T\|_F^2$$

$$\leq \|\tilde{v}\|_2^2\|u-\tilde{u}\|_2^2 + \|u\|_2^2\|v-\tilde{v}\|_2^2$$

$$\leq \|v-\tilde{v}\|_2^2 + \|\Sigma\|_{op}\|u-\tilde{u}\|_2^2$$

as $\|u\|\leq 1$ as $\|\Sigma^{-1/2}\tilde{v}\|\leq 1$

cross term.

$$\mathrm{tr}\left(\tilde{v}(u-\tilde{u})^T u(v-\tilde{v})^T\right)$$

$$= (\tilde{v}^T v - \|\tilde{v}\|_2^2)(\|u\|_2^2 - \tilde{u}^T u) \leq 0$$

$(\|\tilde{v}\|\leq\|v\|) \downarrow \leq 0$ $\quad (\|\tilde{u}\|\|u\|\leq\|u\|_2^2)\ \downarrow \geq 0$

$\|\tilde{v}\|\leq\|v\|$

$\|\tilde{u}\|\leq\|u\|$

$$Y_{u,v} = \|\Sigma\|_{op}\langle u, g\rangle + \langle v, h\rangle$$

$g\sim N(0,I)$ $\quad h\sim N(0,I)$

$$\mathbb{E}(Y_{u,v} - Y_{\tilde{u},\tilde{v}})^2 = \|v-\tilde{v}\|_2^2 + \|\Sigma\|_{op}\|u-\tilde{u}\|_2^2$$

$$\mathbb{E}\sup_{u,v} Z_{u,v} \leq \mathbb{E}\sup_{u,v} Y_{u,v} \leq \|\sqrt{\Sigma}\|_{op}\,\mathbb{E}\sup_{\|u\|\leq 1}\langle u, g\rangle + \mathbb{E}\sup_{\|\Sigma^{-1/2}v\|\leq 1}\langle v, h\rangle$$

$$= \sup_{\|a\|_2\leq 1}\langle \Sigma^{1/2}a, h\rangle = \sup_{\|u\|\leq 1} u^T\Sigma^{1/2}h \leq \|u\|\sqrt{h^T\Sigma h}$$

$$= \|\Sigma\|_{op}\sqrt{n} + \mathbb{E}\sqrt{h^T\Sigma h}\ \left(\leq\sqrt{\mathrm{tr}(\Sigma)}\right)$$

$$\mathbb{E}\,\sigma_{max}\left(\frac{W\sqrt{\Sigma}}{\sqrt{n}}\right) \leq \|\sqrt{\Sigma}\|_{op} + \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}}$$

Qqwww

Recap.

1. Gaussian cov est. $X_1 \cdots X_n \sim N(0, \Sigma)$

$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ with probability $1 - 2\exp\left(\frac{-n\delta^2}{2}\right)$

$$\frac{\|\hat{\Sigma} - \Sigma\|_{op}}{\|\Sigma\|_{op}} \le 2\sqrt{\frac{d}{n}} + 2\delta + \left(\delta + \sqrt{\frac{d}{n}}\right)^2$$

Alternatively, w.p. $\ge 1-\delta$ if $n \ge d$, $\frac{\|\hat{\Sigma} - \Sigma\|_{op}}{\|\Sigma\|_{op}} \lesssim \sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}$ $\left(+\frac{d}{n} + \cdots\right)$

2. (Not covered)

Sub-Gaussian case. suppose each X_i is sub-Gaussian

$$\mathbb{E}\exp(\lambda \langle X, v \rangle) \le \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$$

$\|v\|_2 = 1$

$\frac{\|\hat{\Sigma} - \Sigma\|_2}{\sigma^2} \lesssim \sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}} + \frac{d}{n}$ $(n \ge d) \to$ 2-step discretization

$Q_1 \cdots Q_n$ symmetric, mean 0, matrices, study $\underbrace{\|\frac{1}{n}\sum_{i=1}^{n} Q_i\|_{op}}_{=Y}$

special. $Q_i = X_i X_i^T - \Sigma$, $\|\hat{\Sigma} - \Sigma\|_{op}$

[MV] chernoff method.

$P(Y \ge t) = \mathbb{P}(\exp(sY) \ge \exp(st)) \le \inf_{s>0} \exp(-st)\,\mathbb{E}\exp(sY)$

[BG] $\log(M)$ $\exp(M)$, $M = U^T S U$, $\exp M = U^T \begin{pmatrix} \exp s_1 & & \\ & \exp s_2 & \\ & & \ddots \end{pmatrix} U$

S: eigenvalues $s_1, s_2, \ldots$

$\bar{Q} = \frac{1}{n}\sum_{i=1}^{n} Q_i$, $\mathbb{E}(s\|\bar{Q}\|_{op}) = \mathbb{E}\|\exp(s\bar{Q})\|_{op} \le \mathbb{E}[\mathrm{tr}(\exp(s\bar{Q}))] = \mathrm{tr}[\mathbb{E}(\exp(s\bar{Q}))]$

Sub-Gaussian, $\exists V$, $\mathbb{E}\exp(\lambda Q) \preceq \exp\left(\frac{\lambda^2 V}{2}\right)$, $\forall \lambda$, $\left(\mathbb{E}\exp(\lambda X) \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right)\right)$

Q: Def: $\exists$ square symmetric

similar to vector X, σ

B symmetric, $Q = \varepsilon B$, $\varepsilon = \{1, -1\}$. Q is B^2 sub Gaussian

$\to \{+b, -b\}$ equal probability b-sub-Gaussian.

$\mathbb{E}\exp(\lambda\varepsilon B) = \frac{1}{2}[\exp(\lambda B) + \exp(-\lambda B)]$, $\exp(\lambda B) = \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} B^j$

$\mathbb{E}\exp(\lambda\varepsilon B) = \sum_{j=0}^{\infty} \frac{(\lambda^2 B^2)^j}{(2j)!} \le \sum_{j=0}^{\infty} \left(\frac{\lambda^2 B^2}{2}\right)^j \Big/ j! \le \exp\left(\frac{\lambda^2 B^2}{2}\right) \Rightarrow \underbrace{\varepsilon B}_{Q}$ is B^2 sub Gaussian

$(2j)! \ge 2^j j!$

$\mathrm{tr}(\mathbb{E}\exp(s\bar{Q}))$

$\bar{Q} = \frac{1}{n}\sum_{i=1}^{n} Q_i$. Q_i is V_i-sub-Gaussian.

For vector: $\mathbb{E}\exp(t\bar{X}) = \prod_{i=1}^{n} \mathbb{E}\exp\left(\frac{tX_i}{n}\right)$

But not true for matrix

$\mathrm{tr}\,\mathbb{E}\exp(A+B+C) \ne \mathrm{tr}\,\mathbb{E}\exp A\,\mathbb{E}\exp B\,\mathbb{E}\exp C$

Lieb's theorem. $\forall$ symmetric H

$f(A) = tr\left(\exp(H + \log A)\right)$, f is a concave fn of A

推论 concave

$A \leq B$, $\exp(A) \preceq \exp(B)$ ✗

$\log(A) \preceq \log(B)$ ✓

A is PD

$$tr\,\mathbb{E}\left(\exp(s\bar{Q})\right)$$
$$= \mathbb{E}\,tr\left(\exp\left(s\frac{\sum_{i=1}^{n-1} Q_i}{n} + \log\exp\left(\frac{sQ_n}{n}\right)\right)\right)$$
$$\underset{\text{Jensen}}{\leq} \mathbb{E}\,tr\left(\exp\left(s\frac{\sum_{i=1}^{n-1} Q_i}{n} + \log\mathbb{E}\exp\left(\frac{sQ_n}{n}\right)\right)\right)$$
$$\underset{(n\text{次})}{\leq} tr\exp\left(\sum_{i=1}^{n}\log\mathbb{E}\exp\left(\frac{sQ_i}{n}\right)\right)$$

Matrix Hoeffding $Q_1 \dots Q_n$ symmetric mean 0, sub Gaussian Q_i with V_i

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n} Q_i\right\| \geq t\right) \leq 2\exp\left(\frac{-nt^2}{2\sigma^2}\right), \quad \sigma^2 = \left\|\frac{1}{n}\sum_{i=1}^{n} V_i\right\|_{op}$$

Chernoff.

$$\mathbb{P}\left(\lambda_{max}\left(\frac{1}{n}\sum_{i=1}^{n} Q_i\right) \geq t\right) \leq \inf_{s>0} \exp(-st)\, tr\left(\mathbb{E}\exp(s\bar{Q})\right) \leq tr\exp\sum\log\mathbb{E}\exp\frac{s}{n}Q_i$$

sub-Gaussian

$$\mathbb{E}\exp(tQ_i) \preceq \exp\frac{t^2 V_i}{2}$$

$$\underbrace{\sum_{i=1}^{n}\log\mathbb{E}\exp\left(\frac{sQ_i}{n}\right) \preceq \sum_{i=1}^{n}\log\exp\left(\frac{s^2 V_i}{2n^2}\right)}_{\text{sub-Gaussian}} = \frac{s^2\sum_{i=1}^{n} V_i}{2n^2}$$

$$tr\exp\left(\frac{s^2\sum_{i=1}^{n} V_i}{2n^2}\right) \leq d\exp\left(\frac{s^2\sigma^2}{2n}\right), \quad \sigma^2 = \left\|\frac{1}{n}\sum_{i=1}^{n} V_i\right\|_{op}$$

$$\mathbb{P}(\cdot \geq t) \leq \inf_{s>0}\exp(-st)\, d\exp\left(\frac{s^2\sigma^2}{2n}\right) \leq d\exp\left(\frac{-nt^2}{2\sigma^2}\right)$$

$d\begin{pmatrix} 0 & & 0 \\ & \ddots & \\ 0 & & \end{pmatrix} \varepsilon_1$ d of the.

$d\begin{pmatrix} 0 & & 0 \\ & 1 \; 0 & \\ 0 & & \ddots \end{pmatrix} \varepsilon_2$. $\bar{Q} = \frac{1}{d}\sum_{i=1}^{d} \varepsilon_i e_i e_i^T$, $\|\bar{Q}\|_{op} = \frac{1}{d}$

$V_i = e_i e_i^T$, $\sigma^2 = \|\frac{1}{d} Id\|_{op} = \frac{1}{d}$

$\mathbb{P}(\|\bar{Q}\|_{op} \geq t) \leq 2d\exp\left(\frac{-d^2 t^2}{2}\right)$, w p $2/3$ $\|\bar{Q}\|_{op} \lesssim \frac{\sqrt{\log d}}{d}$

if ε_i is $g_i \sim \mathcal{N}(0,1)$, $\bar{Q} = \frac{1}{d}\sum_{i=1}^{d} g_i e_i e_i^T \Rightarrow \|\bar{Q}\|_{op} \lesssim \frac{\sqrt{\log d}}{d}$, tight upper bound

Matrix Bernstein. $Q_1 \dots Q_n$. $\|Q_i\| \leq b$ $Var(Q_i) = \mathbb{E}Q^2 - (\mathbb{E}Q)^2$

$\sigma^2 = \|\frac{1}{n}\sum_{i=1}^{n} Var(Q_i)\|_{op}$

$$\left\|\frac{1}{n}\sum_{i=1}^{n} Q_i\right\|_{op} \geq t \leq 2d\exp\left(\frac{-nt^2}{2(bt+\sigma^2)}\right), \quad \sigma^2 = \left\|\frac{1}{n}\sum_{i=1}^{n} V_i\right\|_{op}$$

σ^2小时 Bernstein 比 Hoeffding tight

Recap

Covariance $\|\hat{\Sigma} - \Sigma\|_{op} \lesssim \sqrt{\frac{d}{n}}$

x_i is 1-sub-Gaussian

σ-subGau $\lesssim \sigma^2 \sqrt{\frac{d}{n}}$

1. Matrix Hoeffding $Q_1 \cdots Q_n$. symmetric

$$P\left(\left\|\frac{1}{n}\sum_i Q_i\right\|_{op} \ge t\right) \le 2d \exp\left(\frac{-nt^2}{2\sigma^2}\right),\quad \sigma^2 = \left\|\frac{1}{n}\sum_{i=1}^n V_i\right\|_{op}$$

2. Matrix Bernstein. suppose $\|Q_i\|_{op} \le b$.

Q_i is V_i sub-Gaussian

$\mathbb{E}\exp(tQ_i) \preceq \exp\left(\frac{t^2 V_i}{2}\right)$

$$P\left(\left\|\frac{1}{n}\sum_i Q_i\right\|_{op} \ge t\right) \le 2d \exp\left(\frac{-nt^2}{2(bt+\sigma^2)}\right)$$

markov $\le \frac{\mathbb{E}\|\frac{1}{n}\sum Q_i\|_{op}}{t}$

$$\sigma^2 = \left\|\frac{1}{n}\sum_{i=1}^n Var(Q_i)\right\|_{op}$$

$x_1 \dots x_n \sim P$. mean 0 cov Σ. $Q_i = x_i x_i^T - \Sigma$, $\|x_i\|_2 \le \sqrt{b}$

Then $P(\|\hat{\Sigma} - \Sigma\|_{op} \ge t) \le 2d\exp\left(\frac{-nt^2}{4(\|\Sigma\|_{op} + bt)}\right) \le \delta$, $\|\hat{\Sigma} - \Sigma\|_{op} \lesssim \sqrt{\frac{b\|\Sigma\|_{op}\log(d/\delta)}{n}}$

proof: To use Bernstein: $\|Q_i\|_{op} \le \|x_i x_i^T\|_{op} + \|\Sigma\|_{op} \le \|x_i\|^2 + \|\Sigma\|_{op} \le b + \|\Sigma\|_{op}$

$Var(Q_i) = Var(x_i x_i^T) = \mathbb{E}(x_i x_i^T)^2 - \Sigma^2 \preceq \mathbb{E}\, x_i x_i^T \|x_i\|^2 \preceq b\Sigma$

$\|\Sigma\|_{op} \le \sup_{\|v\|_2 = 1} \mathbb{E}\langle v, x\rangle^2 \le b$ $\qquad \sigma^2 = b\|\Sigma\|_{op}$

- $x_i \to \sqrt{d}$-sphere, $\|x_i\|_2 = \sqrt{d}$, $\mathbb{E}(x_i x_i^T) = I_d$, $\|\Sigma\|_{op} = 1$
- C-subGaussian. $\|\hat{\Sigma} - \Sigma\|_{op} \lesssim \sqrt{\frac{d}{n}} + O\left(\frac{d}{n}\right)$ ↳ tighter unionbound; $\|\hat{\Sigma} - \Sigma\|_{op} \lesssim \sqrt{\frac{d\log d}{n}} + O\left(\frac{d\log d}{n}\right)$ → lower order
- $x_i \to \sqrt{d}\, e_j$, $j \sim unif[d]$ ↳ sparse, only 1 coordinate; $\mathbb{E}(x_i x_i^T) = I_d$, d-sub-Gaussian, $\frac{d^{3/2}}{\sqrt{n}}$ → union bound is worse than Bernstein,

Estimating Structured Cov Matrix (Sparse Matrix).

Σ is sparse: $A_{ij} = \begin{cases} 1 & \text{if } \Sigma_{ij} \ne 0 \\ 0 & \text{o/w} \end{cases}$ $\|A\|_{op}$ is small, $\|A\|_F$ is small.

use $\hat{\Sigma}$ threshold in entries

with prob $\ge 1-\delta$, $\|\hat{\Sigma} - \Sigma\|_\infty \le \alpha$, $\hat{\Sigma}_{jk} = \frac{1}{n}\sum_{i=1}^n x_{ij} x_{ik}$, x_{ij} is σ-Sub Gaussian

$\underbrace{x_i \cdot x_i^T}$ → (σ^4, α^2)-sub-exponential $\hat{\Sigma}$ Σ?

$$\left(P(|\hat{\Sigma}_{jk} - \Sigma_{jk}| \ge t) \lesssim \exp\frac{-nt^2}{2\sigma^4}\right)$$

$\Rightarrow \|\hat{\Sigma} - \Sigma\|_\infty \le \sigma^2\sqrt{\frac{\log(d/\delta)}{n}}$, pick t to be this with prob $\ge 1-\delta$

$\tilde{\Sigma} = T_{2t}(\hat{\Sigma})$ condition on $\|\hat{\Sigma} - \Sigma\|_\infty \le t$, $t = C\sigma^2\boxed{\sqrt{\frac{\log d}{n}}}$

$$\begin{cases} \text{zero coordinate } \Sigma_{ij} = 0, \\ \text{non-zero coordinate. } |\tilde{\Sigma}_{ij} - \Sigma_{ij}| \le |\tilde{\Sigma}_{ij} - \hat{\Sigma}_{ij}| + |\hat{\Sigma}_{ij} - \Sigma_{ij}| \le 3t \end{cases}$$

$|\tilde{\Sigma} - \Sigma| \le 3tA$
↳ elementwise

$\underline{\|\tilde{\Sigma} - \Sigma\|_{op}} \le 3t\|A\|_{op}$

用 $\|\cdot\|_\infty$ noise t 去 threshold Σ ($\hat{\Sigma}$),

构造 $\tilde{\Sigma}$ 去 estimate Σ
↳ sparse and structured.

不会像 $\hat{\Sigma}$

$\boxed{\sqrt{\frac{\log d}{n}}}$ scale up

$\hat{\Sigma}$ 不 sparse

Recap.

1. Applications of matrix Bernstein to covariance est.

(a). $x_i \sim \text{unif}(\sqrt{d}\text{-sphere})$, union bound better

(b). $x_i \sim \pm\sqrt{d}\, e_j$, $j \sim \text{unif}[d]$. union bound worse

$\|\hat{\Sigma} - I\|_2 \lesssim \sqrt{\frac{d \log d}{n}}$

2. Estimating sparse covariance matrices

$X_{ij} \to \sigma$-sub Gaussian

$\|\hat{\Sigma} - \Sigma\|_\infty \lesssim \sigma^2 \sqrt{\log d / n}$. $\|T(\hat{\Sigma}) - \Sigma\|_2 \lesssim \sigma^2 \sqrt{\frac{\log d}{n}} \|A\|_{op}$.

$A_{ij} = 1$ if $\Sigma_{ij} \neq 0$
$= 0$ w/o

Sparse linear models compressed sensing

$y = X\theta^* + w$. θ^* is s-sparse $\|\theta\|_q \leq R_q$. $q \in [0,1]$

sensing vector $\leftarrow x_1 \cdots x_n$. $n \ll d$ $X: n \times d$

$y_1 = \langle x_1, \theta^* \rangle \cdots y_n = \langle x_n, \theta^* \rangle$

$\theta^* = \Psi^T \beta^*$ $\hookrightarrow$ image
$\hookrightarrow$ wavelet representation, sparse

$\nrightarrow$ now under-constrained linear system. $y = X\theta^*$, ∞-many solutions

$\min \|\theta\|_0$ $y = X\theta$, θ^* is unique

(BP) Basis pursuit $\min\limits_{\|\theta\|_1} y = X\theta$. min-norm linear interpolation

$\theta^* \to S \subseteq \{1, \dots d\}$. $\theta \in \text{null}(X)$, $(X\theta = 0)$ $\theta^* + \theta$ is also a solution

$C(S) = \{\|\theta_{S^c}\|_1 \leq \|\theta_S\|_1\}$. Restricted null space. RN(S) $C(S) \cap \text{null}(X) = \{0\}$

Thm. 1. Basis pursuit works for any θ^* with support S, ($y = X\theta^*$ unique solution (BP/min ℓ_1))

$1 \Leftrightarrow 2$ 2. Restricted null space RN(S) holds for S

Proof: $2 \Rightarrow 1$. $\hat{\theta}$ is a soln to BP. $X\hat{\theta} = y$. $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$, $\Delta = \hat{\theta} - \theta^*$

$\|\theta^*\|_1 \geq \|\Delta + \theta^*\|_1 = \|(\Delta + \theta^*)_S\|_1 + \|\Delta_{S^c}\|_1 \geq \|\theta^*\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1$

$\Rightarrow \|\Delta_{S^c}\| \leq \|\Delta_S\|$, $\Delta \in C(S)$. $\because$ RN $\Rightarrow \Delta = 0$

$\neg 2 \Rightarrow \neg 1$

If RN(S) fails. θ^* with supp S & $\hat{\theta}$ which solves BP

$\theta \in \text{null}(X) \cap C(S)$, $y = X\theta_S$. solves BP (y, X) $\|\theta_{S^c}\|_1 \leq \|\theta_S\|_1$

$= X(-\theta_{S^c})$ $\xrightarrow[X\theta_S = -X\theta_{S^c}]{X\theta = 0}$ $-\theta_{S^c}$ is another better solution

X so that RN(S) holds for $|S| \le s$, BP will exactly reconstruct any θ^*, $\|\theta^*\|_0 \le s$

1. Pairwise incoherence δ, $\left\|\frac{X^TX}{n} - I\right\|_\infty \le \delta_{pw}$

If $\delta_{pw} < \frac{1}{2s}$ then RN(S) holds for for any $|S| \le s$

(if $X_{ij} \sim N(0,1)$), $\|\ \ \|_\infty \lesssim \sqrt{\frac{\log d}{n}} \lesssim \frac{1}{2s}$. ($n \gtrsim s^2 \log d$) with as above).

Proof: $\theta \in \text{null}(X)$, $\left\|\frac{X\theta_S}{\sqrt{n}}\right\|_2^2 = \theta_S^T \frac{X^TX}{n} \theta_S \ge -\theta_S^T\left(I - \frac{X^TX}{n}\right)\theta_S + \|\theta_S\|_2^2$

$\ge \|\theta_S\|_2^2 - \left\|I - \frac{X^TX}{n}\right\|_\infty \|\theta_S\|_1^2$

$\ge \|\theta_S\|_2^2 (1 - \delta_{pw} s)$

$u^T\Sigma v \le u^T |\Sigma_{bd}| |v|_1$

max Σ_{ij} (abs)

$\left\|\frac{X\theta_S}{\sqrt{n}}\right\|_2^2 = \left\langle \frac{X\theta_S}{\sqrt{n}}, \frac{-X\theta_{S^c}}{\sqrt{n}} \right\rangle = -\theta_S^T \frac{X^TX}{n} \theta_{S^c}$

$= -\theta_S^T\left(\frac{X^TX}{n} - I\right)\theta_{S^c} + \theta_S^T\theta_{S^c}$ $\to 0$

$\le \|\theta_S\|_1 \|\theta_{S^c}\|_1 \delta_{pw}$

Combin upper band & low band

$\|\theta_S\|_2^2 \le \frac{\sqrt{s}\|\theta_S\|_2 \delta \|\theta_{S^c}\|_1}{1 - s\delta}$

$\|\theta_S\|_1 \le \sqrt{s}\|\theta_S\|_2 \le \frac{\delta s}{1-\delta s}\|\theta_{S^c}\|_1$, $\|\theta_S\|_1 < \|\theta_{S^c}\|_1$

$\delta \le \frac{1}{2s}$ $\quad \to \theta \notin C(S)$

Restricted Isometry Prop (RIP)

$\left\|\frac{X_S^TX_S}{n} - I\right\|_{op} \le \delta_s$, for all $|S| \le s$

$\to \delta_{2s} \le 1/3$. then RN holds for all $|S| \le s$

e.g. $X_{ij} \in N(0,1)$. $\delta_s \lesssim \sqrt{\frac{s + \log d/s}{n}} \lesssim \sqrt{\frac{s \log d}{n}}$, $n \gtrsim s \log d$

Recap: $y = X\theta^*$, $X \in \mathbb{R}^{n\times d}$, $d \gg n$

θ^* is sparse. (motivation)

1. Basis pursuit $\min \|\theta\|_1$ s.t. $y = X\theta$ — find θ^* uniquely

2. Restricted null space $C(S) = \{\|\Delta_{S^c}\|_1 \le \|\Delta_S\|_1\}$ supports

   $RN(S)$ holds if $C(S) \cap \mathrm{null}(X) = \{0\}$

3. ~~RN~~ RN & BP. (1) RN holds for some S

   (2). BP uniquely recovers any θ^* with support S

   (1) $\Leftrightarrow$ 2

4. Certify RN. (a). Pairwise incoherence $\left\|\frac{X^TX}{n} - I\right\|_\infty \le \delta_{PW}$

   If ($n \gtrsim s^2 \log d$) $\delta_{PW} \le \frac{1}{3s}$ then $RN(S)$ holds for any $|S| \le s$

   (b). Restricted isometry $RIP(s)$ holds if

   $$\left\|\frac{X_S^TX_S}{n} - I\right\|_{op} \le \delta_{RIP,s} \quad \forall |S| \le s$$

   If $\delta_{RIP,2s} \le 1/3$ then $RN(S)$ holds for any $|S| \le s$

   $\rightarrow n \gtrsim s\log d$

$\theta \in \mathrm{null}(X)$. want to show $\theta \notin C(S)$, $\|\theta_S\|_1 < \|\theta_{S^c}\|_1$, $\forall S$

$\theta = \theta_0 + \theta_1 + \cdots \theta_n$

largest entry — θ_0; second largest entry — θ_1

$\Updownarrow$

$\|\theta_0\|_1 < \|\theta_0^c\|_1$

$$\left\|\frac{X\theta_0}{\sqrt{n}}\right\|_2^2 = \theta_0^T\frac{X^TX}{n}\theta_0 = \|\theta_0\|_2^2 + \theta_0^T\left(\frac{X^TX}{n} - I\right)\theta_0 \ge \|\theta_0\|_2^2 - \delta\|\theta_0\|_2^2 \ge (1-\delta)\|\theta_0\|_2^2$$

$$\left\|\frac{X\theta_0}{\sqrt{n}}\right\|_2^2 = -\theta_0\frac{X^TX}{n}\theta_0^c, \quad (X\theta_0 = -X\theta_0^c)$$

$$= -\theta_0\left(\frac{X^TX}{n} - I\right)\sum_{j\ge 1}\theta_j \le \delta\|\theta_0\|_2\sum_{j\ge 1}\|\theta_j\|_2$$

$$\le \frac{\delta}{\sqrt{s}}\|\theta_0\|_2\left\{\sum_{j\ge 0}\|\theta_0\|_1\right\}$$

$\|\theta_j\|_2 \le \sqrt{s}\|\theta_j\|_\infty \le \sqrt{s}\,\frac{\|\theta_{j-1}\|_1}{s}$

$$\Rightarrow \|\theta_0\|_1 \le \sqrt{s}\|\theta_0\|_2 \le \frac{\delta}{1-\delta}\left\{\sum_{j\ge 0}\|\theta_j\|_1\right\} = \frac{\delta}{1-\delta}\left[\|\theta_0\|_1 + \|\theta_0^c\|_1\right]$$

Now: $y = X\theta^* + w$, w mean 0 sub-Gaussian ($X: n\times d$. $d \gg n$)

1. Relaxed BP: $\min \|\theta\|_1$ s.t. $\frac{1}{2n}\|y - X\theta\|_2^2 \le b^2$

2. Lagrangian LASSO $\min_\theta \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1$

3. $\min_\theta \frac{1}{2n}\|y - X\theta\|_2^2$, $\|\theta\|_1 \le R$. constrained LASSO

1. Estimation error $\|\hat\theta-\theta^*\|_2^2$
2. Prediction error. in-sample $\frac{1}{n}\|X\hat\theta-X\theta^*\|_2^2$
3. Support recovery: $\hat S \approx supp(\theta^*)$
4. Inference: pick coordinate j, C_α $\quad P(\theta_j^*\in C_\alpha)\geq 1-\alpha$, $C_\alpha(y,X)$

Restricted eigenvalue cond. $\alpha\geq 1$, $C_\alpha(S)=\{\|\Delta_{S^c}\|_1\leq \alpha\|\Delta_S\|_1\}$ Cone.

RE(α,κ) $\quad \|\frac{X\Delta}{\sqrt{n}}\|_2^2\geq \kappa\|\Delta\|_2^2$, $\forall \Delta\in C_\alpha(S)$: Fisher Infor. matrix is curvatured well

stronger than RN(S)

1. Estimation error $\|\hat\theta-\theta^*\|_2^2$. Constrained LASSO, $R=\|\theta^*\|_1$

$\min \frac{1}{2n}\|y-X\theta\|_2^2$, $\|\theta\|_1\leq\|\theta^*\|_1$

$\Delta=\hat\theta-\theta^*$

$\|\theta^*\|_1\geq\|\Delta+\theta^*\|_1=\|\Delta_S+\theta_S^*\|_1+\|\Delta_{S^c}\|_1\geq\|\theta^*\|_1-\|\Delta_S\|_1+\|\Delta_{S^c}\|_1$

$\Rightarrow \|\Delta_{S^c}\|_1\leq\|\Delta_S\|_1$

BI: $\frac{1}{2n}\|y-X\hat\theta\|_2^2\leq\frac{1}{2n}\|y-X\theta^*\|_2^2$, $\frac{1}{2n}\|X\Delta\|_2^2\leq\frac{1}{n}\langle X^Tw,\Delta\rangle$

$\frac{\kappa\|\Delta\|_2^2}{2}\leq$

$\|\Delta\|_2^2\leq 2\|\frac{X^Tw}{n}\|_\infty\|\Delta\|_1$

$\leq 4\|\frac{X^Tw}{n}\|_\infty \cancel{\frac{\|\Delta_S\|_1}{\kappa}}\sqrt{s}\|\Delta\|_2 \quad \Rightarrow \|\Delta\|_2\leq 4\|\frac{X^Tw}{n}\|_\infty\frac{\sqrt{s}}{\kappa}$

Formal. $R=\|\theta^*\|_1$, $X\to$ RE$(1,\kappa)$

then $\|\Delta\|_2^2\leq 4\|\frac{X^Tw}{n}\|_\infty\frac{\sqrt{s}}{\kappa}$, $\|\Delta_{S^c}\|_1\leq\|\Delta_S\|_1$

X: column normalized $\frac{\|X_j\|_2}{\sqrt{n}}\leq C$, $\frac{X^Tw}{n}\sim\begin{cases}N(0,\sigma^2/n)\\ \vdots\end{cases}$, $\Rightarrow \|\frac{X^Tw}{n}\|_\infty\lesssim\sqrt{\frac{s\log d}{n}}$

Low-dim. $\frac{\|X\Delta\|_2^2}{2n}\leq\frac{1}{n}\langle X^T\varepsilon,\Delta\rangle$

$\lambda_{min}(\frac{X^TX}{n})\|\Delta\|_2^2\leq$ $\quad \leq\|\Delta\|_2\cdot\underbrace{\|\frac{X^T\varepsilon}{n}\|_2}_{\sqrt{d/n}}$

$\Rightarrow \|\Delta\|_2\lesssim\sqrt{\frac{d}{n}}$

Recap. (i.i.d).

1. BP $\min \|\theta\|_1$, $Y = X\theta$. If $\delta_{RIP,2s} \le 1/3$ then RN holds $\forall |S| \le s$

2. LASSO. $y = X\theta^* + w$ ($X \in \mathbb{R}^{n \times d}, d \gg n$) → Lagrangian $\hat\theta = \arg\min_\theta \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1$

Constrained $\hat\theta = \arg\min_{\|\theta\|_1 \le R} \frac{1}{2n}\|y - X\theta\|_2^2$

Relaxed BP. $\hat\theta = \arg\min \|\theta\|_1$, $\frac{1}{2n}\|y - X\theta\|_2^2 \le b$

3. Restricted eigenvalue (α, κ) $\frac{X^TX}{n}$ $\alpha \ge 1$, Δ: eigen vector, κ: eigenvalue.

$\kappa\|\Delta\|_2^2 \le \frac{\|X\Delta\|_2^2}{n}$, $\forall \{\Delta, \|\Delta_{S^c}\|_1 \le \alpha\|\Delta_S\|_1, \forall |S| \le s\}$

4. Result for constant LASSO. If $X \sim RE(1,\kappa)$. $R = \|\theta^*\|_1$

then $\underset{\|\hat\theta - \theta^*\|_2}{\|\Delta\|_2} \lesssim \|\frac{X^Tw}{n}\|_\infty \frac{\sqrt{s}}{\kappa}$

↓ $\sqrt{\frac{\log d}{n}}$

Lagrangian LASSO, $\lambda \ge 2\|\frac{X^Tw}{n}\|_\infty$. $X \sim RE(3,\kappa)$, then $\|\Delta\|_2 \lesssim \frac{\lambda\sqrt{s}}{\kappa}$

$\frac{1}{2n}\|y - X\hat\theta\|_2^2 \le \frac{1}{2n}\|y - X\theta^*\|_2^2 + \lambda\{\|\theta^*\|_1 - \|\hat\theta\|_1\}$, $\Delta = \hat\theta - \theta^*$

$\frac{\kappa\|\Delta\|_2^2}{2} \le \frac{1}{2n}\|X\Delta\|_2^2 \le \frac{1}{n}\langle X^Tw, \Delta\rangle + \lambda\{\|\theta^*\|_1 - (\|\theta^*\|_1 - \|\Delta_S\|_1) - \|\Delta_{S^c}\|_1\}$

$\le \|\frac{X^Tw}{n}\|_\infty\|\Delta\|_1 + \lambda(\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1)$

$\le \frac{\lambda}{2}\{\underbrace{3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1}_{\ge 0}\}$

Cone condition $\boxed{\|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1}$, ⇒ $\frac{\kappa\|\Delta\|_2^2}{2} \le \frac{3\lambda}{2}\|\Delta_S\|_1 \le 3\lambda\sqrt{s}\|\Delta\|_2$

↳ 这时 λ 大, noise 很大 起 thresholding. eigen vector → sparse

RBP. $b^2 \ge \frac{\|w\|_2^2}{2n}$ RE(1,κ), then $\|\Delta\|_2 \lesssim \|\frac{X^Tw}{n}\|_\infty \frac{\sqrt{s}}{\kappa} + \frac{1}{\sqrt{\kappa}}\sqrt{b^2 - \frac{\|w\|_2^2}{2n}}$

$\min \|\theta\|_1$, $\frac{1}{2n}\|y - X\theta\|_2^2 \le b^2$ $\quad \|\hat\theta\|_1 \le \|\theta^*\|_1 \Leftrightarrow \|\Delta_{S^c}\|_1 \le \|\Delta_S\|_1$

↳ θ^* is feasible

$\frac{1}{2n}\|y - X\hat\theta\|_2^2 \le \frac{1}{2n}\|y - X\theta^*\|_2^2 + b^2 - \frac{\|w\|_2^2}{2n}$

$\frac{1}{2n}\|X\Delta\|_2^2 \le \|\frac{X^Tw}{n}\|_\infty\|\Delta\|_1 + b^2 - \frac{\|w\|_2^2}{2n}$

$(\|\Delta\|_1 \le 2\|\Delta_S\|_1 \le 2\sqrt{s}\|\Delta\|_2)$ ↓

$$\frac{K\|\Delta\|_2^2}{2} \le 2\sqrt{s}\left\|\frac{X^T w}{n}\right\|_\infty \|\Delta\|_2 + b^2 - \frac{\|w\|^2}{2n}$$

Prediction error : $\frac{\|X\hat{\theta} - X\theta^*\|_2^2}{n}$

: slow-rate → very few assumptions

: fast rate → θ^* s-sparse + RE

$$\frac{1}{2n}\|Y - X\hat{\theta}\|_2^2 \le \frac{1}{2n}\|Y - X\theta^*\|_2^2 + \lambda\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\}$$

$$\frac{1}{n}\|X\Delta\|_2^2 \le \left\|\frac{X^T w}{n}\right\|_\infty \|\Delta\|_1 \times 2 + 2\lambda\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\}$$

(pick $\lambda \ge 2\left\|\frac{X^T w}{n}\right\|_\infty$)

$$\le \lambda\{\|\theta^*\|_1 + \|\hat{\theta}\|_1 + 2\|\theta^*\|_1 - 2\|\hat{\theta}\|_1\} \quad (\|\hat{\theta}\|_1 \le 3\|\theta^*\|_1)$$

$$\le 3\lambda\|\theta^*\|_1$$

$$\Rightarrow \frac{1}{n}\|X\Delta\|_2^2 \lesssim \lambda\|\theta^*\|_1 \lesssim \|\theta^*\|_1\sqrt{\frac{\log d}{n}}$$

X - RE , θ^* - s . sparse : $\|\Delta\|_2^2 \le \frac{\|X\Delta\|_2^2}{nK}$

$$\frac{1}{n}\|X\Delta\|_2^2 \lesssim \lambda\sqrt{s}\|\Delta\|_2 \lesssim \lambda\sqrt{\frac{s}{K}}\left(\|X\Delta\|_2/\sqrt{n}\right)$$

$$\le \frac{\lambda^2 s}{K} \lesssim \frac{s\log d}{n \times K} \to \text{fast rate}$$

$Y = X\theta^* + w$, θ^*_j , $\hat{\theta} = (X^TX)^{-1}X^TY = \theta^* + (X^TX)^{-1}X^Tw$ $\quad w \sim N(0, \sigma^2)$

$$\sim N(0, (X^TX)^{-1}(X^TX)(X^TX)^{-1}\sigma^2)$$

$$\theta^*_j \pm \sigma^2 z_{\alpha/2}\sqrt{\frac{\hat{\Sigma}^{-1}_{jj}}{n}}$$

LASSO : not work , sparse , biased . related to λ.

Recap.

1. LASSO theory. ~ θ^* is s-sparse.

- Estimation error ~ X satisfies $RE(k,3)$
~ column normalization.

$$\|\hat{\theta}-\theta^*\|_2 \lesssim \sqrt{\frac{s\log p}{nk}}$$

2. Prediction error: - slow rate - X satisfies col norm

$$\frac{\|X\hat{\theta}-X\theta^*\|_2^2}{n} \lesssim \sqrt{\frac{\log p}{n}}\|\theta^*\|_1$$

- Fast rate X + RE + θ^* - s-sparse

$$\frac{\|X\hat{\theta}-X\theta^*\|_2^2}{n} \lesssim s\frac{\log p}{n}$$

Debiasing the LASSO, $y=X\theta^*+w$. CI for θ_j^*. $w\sim N(0,\sigma_\varepsilon^2)$

$\hat{\theta}$: LASSO est

debiased: $\tilde{\theta}=\hat{\theta}+\frac{MX^T(y-X\hat{\theta})}{n}$, $M\approx\left(\frac{X^TX}{n}\right)^{-1}$

$\therefore \tilde{\theta}=\hat{\theta}-\frac{MX^TX}{n}\hat{\theta}+\frac{MX^TX}{n}\theta^*+\frac{MX^Tw}{n}$

$\approx \theta^*$ + noise (put $M\approx\left(\frac{X^TX}{n}\right)^{-1}$ into it)

$\tilde{\theta}=\theta^*+\left(I-\frac{MX^TX}{n}\right)(\hat{\theta}-\theta^*)+\frac{MX^Tw}{n}$

$\tilde{\theta}\overset{d}{=}N\left(\theta^*,\frac{MX^TXM^T\sigma_\varepsilon^2}{n^2}\right)$ + error

err $\leq \left\|I-\frac{MX^TX}{n}\right\|_\infty\|\hat{\theta}-\theta^*\|_1 \lesssim \sqrt{\frac{\log p}{n}}\sqrt{s}\underbrace{\sqrt{\frac{s\log p}{n}}}_{\sqrt{s}\|\cdot\|_2\geq\|\cdot\|_1}$

$\sim\frac{s\log p}{n}$

under RE + col norm + θ^*: s-sparse

$X\sim N(0,I)$ then $M=I$

if $\frac{s\log p}{n}\ll\frac{1}{\sqrt{n}}$, then CI is correct

Support recovery. $y=X\theta^*+w$. $\text{supp}(\theta^*)$

$\text{supp}(\hat{\theta})=\text{supp}(\theta^*)$ under condi.

$\|\hat{\theta}-\theta^*\|_2\lesssim\sqrt{\frac{s\log p}{n}} \rightarrow \|\hat{\theta}-\theta^*\|_\infty\lesssim\sqrt{\frac{s\log p}{n}}$ will recover any $|\theta_j^*|\gtrsim\sqrt{\frac{s\log p}{n}}$

βmin conds. $\left\{\theta_j^* \text{ min zero} > \sqrt{\frac{s\log p}{n}}\right\} \Rightarrow \hat{\theta}_j$ will be non-zero

+ threshold $\hat{\theta}$ at $\sqrt{\frac{s\log p}{n}}$

(Imagine $X=I$, $HT(y)\sqrt{\frac{\log p}{n}}$
will recover supp $|\theta^*\text{min}|>\sqrt{\frac{\log p}{n}}$)

Mutual Incoherence

$$\max_{j\in S^c} \|(X_S^T X_S)^{-1} X_S^T X_j\|_1 < 1 \quad \text{, predict } X_j \text{ from } X_S$$

$\|\beta_j\|_1 < 1$

If β_{min} + Mutual Incoherence ✓ ⇒ support recovery ✓

KKT of LASSO: $\frac{1}{2n}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1$

$$-\frac{1}{n}X^T(y - X\hat\theta) + \lambda\,\mathrm{sign}(\hat\theta) = 0 \qquad \mathrm{sign}(\theta) = \begin{cases} 1 & \theta>0 \\ [-1,1] & \theta=0 \\ -1 & \theta<0 \end{cases}$$

$\hat\theta$ has support S

$(\hat\theta, \mathrm{sign}(\hat\theta))$ which solves KKT

$\hat\theta$: primal sol, $\mathrm{sign}(\hat\theta)$: dual sol

Minimax lower bounds.

$\theta(P)$, $\hat\theta$, $\mathbb{E}\,\rho(\hat\theta, \theta(P))$

ρ: semi-metric

minimax rate

$$\mathfrak{M}(\mathcal{P}, \rho) = \inf_{\hat\theta} \sup_{P\in\mathcal{P}} \mathbb{E}\,\rho(\hat\theta, \theta(P))$$

From estimation to testing, $\{\theta^1 \cdots \theta^M\}$, $\rho(\theta^i, \theta^j) \geq 2\delta$

$P_{\theta^1} \cdots P_{\theta^M}$

$J \sim \mathrm{unif}[M]$, $Z \sim P_{\theta^J}$ ⇒ find which θ^M? (testing)

testing error $\mathbb{Q}(\psi(Z) \neq J)$

$\psi(Z) \to [M]$

$\mathbb{Q}(Z \mid J=j) = P_{\theta^j}$, $\mathbb{Q}(J=j) = \frac{1}{M}$

$\delta\downarrow$, $\Phi(\delta)\downarrow$, $\inf_\psi \to 1$

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi\circ\rho) \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}(\psi(Z) \neq J)$$

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\, \Phi\circ\rho(\hat\theta, \theta(P)) \geq \Phi(\delta)\, \mathbb{P}(\Phi\circ\rho(\hat\theta, \theta(P)) \geq \Phi(\delta))$$

↳ Markov ineq.

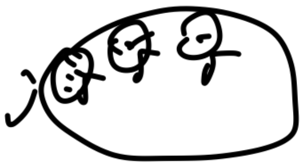
$$\geq \Phi(\delta) \sup_{P\in\mathcal{P}} \mathbb{P}(\rho(\hat\theta, \theta(P)) \geq \delta) \geq \Phi(\delta) \frac{1}{M}\sum_{j=1}^{M} P_{\theta^j}(\rho(\hat\theta, \theta^j) \geq \delta)$$

$\hat\theta$, $\hat\psi(Z) = \arg\min_{j\in[M]} \rho(\hat\theta, \theta^j)$

If $\rho(\hat{\theta}, \theta^j) \le \delta, \Rightarrow \psi(Z) = j$

$\therefore \geq \Phi(\delta) \cdot Q(\psi(Z) \neq j)$

$$\inf_{\hat{\theta}} \sup_{P} R \geq \Phi(\delta) \inf_{\psi} Q(\psi(Z) \neq j)$$



Recap.

r-LASSO. theory.

est. error: $\|\hat\theta - \theta^*\|_2 \lesssim \sqrt{\frac{s\log p}{n}}$ under RE + θ^* s-sparse

pred. error: $\frac{\|X\hat\theta - X\theta^*\|_2^2}{n} \lesssim \frac{s\log p}{n} \lesssim \sqrt{\frac{\log p}{n}}\|\theta^*\|_1$

under col. norm..

Variable selection / supp recovery.

$\max_j \|(X_S^T X_S)^{-1} X_S^T X_j\|_1 < 1$

β_{min} $|\theta^*_{min}| \gtrsim \sqrt{\frac{\log p}{n}}$ then $P(\mathrm{supp}(\hat\theta) = \mathrm{supp}(\theta^*)) \to 1$

Debias.

if $\|I - M\frac{X^TX}{n}\|_\infty \|\hat\theta - \theta^*\|_1 = o(1/\sqrt{n})$

then. $\tilde\theta = \hat\theta + \frac{MX^T(y - X\hat\theta)}{n}$ is asymptotically mean θ^* Gaussian

Min Max.

$\mathbb{E}\,\Phi\circ\rho(\hat\theta, \theta(P))$

$= R(\hat\theta, \theta(P))$

$\mathcal{M}(\mathcal{P}, \Phi\circ\rho) = \inf_{\hat\theta}\sup_{P\in\mathcal{P}} R(\hat\theta, \theta(P))$

$\{\theta^1, \cdots \theta^M\} \to \rho(\theta^i, \theta^j) \geq 2\delta$ 这个 δ

$\mathcal{M} \geq \Phi(\delta) \underbrace{\inf_\psi Q(\psi(Z) \neq J)}$ → testing error: lower bound of est. error

Divergence measures.

1. $TV(P, Q) = \sup_A |P(A) - Q(A)| = \frac{1}{2}\int |p(x) - q(x)|\,dx$
2. $KL(P\|Q) = \mathbb{E}_{x\sim p}\log\frac{p(x)}{q(x)}$
3. $H^2(P, Q) = \frac{1}{2}\int(\sqrt{p(x)} - \sqrt{q(x)})^2\,dx$

Properties: $H^2 \lesssim TV \lesssim H \lesssim \sqrt{KL}$

$x_1 \cdots x_n \sim p$

P_n Q_n

Tensorization. $KL(P^n, Q^n) = n\,KL(P, Q)$.

$H^2(P^n, Q^n) \leq n H^2(P, Q)$

$H^2(P, Q) = 1 - \underbrace{\int\sqrt{pq}}$ Hell affinity

$H^2(P^n, Q^n) = 1 - \left(\int\sqrt{pq}\right)^n = 1 - (1 - H^2)^n \leq nH^2$

$\to \int_{x_1}\cdots\int_{x_n}\sqrt{p(x_1)q(x_1)\times\cdots p(x_n)q(x_n)}\,d\cdots$

Mutual information.

$I(X; Y) = KL(p(X, Y), p(X)p(Y))$

Le Cam's $M \geq \Phi(\delta) \inf_{\psi} Q(\psi(Z) \neq J)$

$$Q = \frac{1}{2} P_{\theta_1} + \frac{1}{2} P_{\theta_2} \geq \frac{\Phi(\delta)}{2}\left(1 - TV(P_{\theta_1}, P_{\theta_2})\right)$$

$X_1 \cdots X_n \sim N(\theta, \sigma^2)$, $R(\hat{\theta}, \theta^*) \geq \frac{\sigma^2}{n}$

$\theta_1 \cdots \theta_1 + 2\delta$ $\quad R(\hat{\theta}, \theta^*) = \Phi(|\hat{\theta} - \theta^*|)$ → θ_1与$\theta_1+2\delta$差别小.

$$M \geq \delta\left(1 - TV(N(\theta_1, \sigma^2), N(\theta_1 + 2\delta, \sigma^2))\right)$$

$TV(\leq \sqrt{\frac{n KL(N(\theta, \sigma^2), N(\theta+2\delta, \sigma^2))}{2}}$, $\quad TV \leq \sqrt{\frac{n}{2} \frac{4\delta^2}{\sigma^2}}$

⇒ LeCam: $M \geq \delta\left(1 - \sqrt{\frac{2n\delta^2}{\sigma^2}}\right)$, $\delta = \sigma/\sqrt{n}$, $M \gtrsim \sigma/\sqrt{n}$

$X_1 \cdots X_n \sim U[\theta, \theta+1]$

$\hat{\theta}_{MLE} = \min_{i=1}^{n} X_i$, $|\hat{\theta}_{MLE} - \theta| \lesssim 1/n$

$\theta_1: U[\theta, \theta+1]$, $\theta_{1+2\delta}: U[\theta+2\delta, \theta+2\delta+1]$

$$H^2(P_{\theta_1}, P_{\theta_1+2\delta}) = \frac{1}{2}(4\delta) = 2\delta$$

$$H^2(P^n, Q^n) \leq 2n\delta$$

$M \geq \frac{\delta}{2}\left(1 - \sqrt{2n\delta}\right)$, $\delta \simeq 1/n$, $M \gtrsim \frac{1}{n}$

Lipschitz fn at a point $X_1 \cdots Y_1 \cdots X_n, Y_n$, $X \sim U(0,1)$

$y_i = f(x_i) + \underset{N(0,1)}{\varepsilon_i}$, $\quad |f(x) - f(y)| \leq L|x - y|$

$\theta = f(0.5)$

$P_1: f_1 = 0 \quad P_2: f_2 =$ Lh 0.5

$$KL(P_1(x,y), P_2(x,y)) = E_{P_1} \log \frac{P_1(y|x)}{P_2(y|x)} = \int_x p(x) KL(P_1(y|x), P_2(y|x))$$

$$= \int_x p(x) (f_1(x) - f_2(x))^2$$

Recap.

1. ~~Est to testing~~

$\mathfrak{M}(\Phi\circ\rho,\mathcal{P}) = \inf_{\hat\theta}\sup_{P\in\mathcal{P}} \mathbb{E}\,\Phi\circ\rho(\hat\theta,\theta(P))$

$\to \mathfrak{M} \geq \Phi(\delta)\inf_{\psi} Q(\psi(Z)\neq J)$, $\theta_1\ldots\theta_m$ are 2δ separated.

$J\sim \mathrm{Unif}[M]$, $Z|J\sim P_{\theta^J}$

3. Le Cam $M=2$,

$\mathfrak{M}\geq \frac{\Phi(\delta)}{2}\left(1-TV(P_{\theta^1},P_{\theta^2})\right)$

4. Ineq. $H^2\lesssim TV\lesssim H\lesssim\sqrt{KL}\lesssim\sqrt{\chi^2}$

$KL(P^n,Q^n)=n\,KL(P,Q)$, $H^2(P^n,Q^n)\leq n H^2(P,Q)$

5. Examples. Univariate Normal Location $\gtrsim \frac{\sigma}{\sqrt{n}}$

Uniform location $\gtrsim \frac{1}{n}$. Lipschitz fn. est. at a point $\gtrsim \frac{1}{n^{1/3}}$

Fano's inequality: $\inf_{\psi} Q(\psi(Z)\neq J)\geq 1-\frac{I(Z;J)+\log 2}{\log M}$ $\quad J\sim[M]$

Intuitively $I(Z;J)\lesssim \log M$, then $\mathfrak{M}\gtrsim\Phi(\delta)$

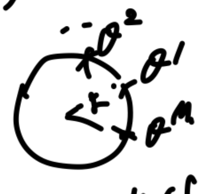
$I(Z;J)=KL(Q(Z,J)\,\|\,Q(Z)Q(J))$

$\left(\bar Q=\frac{1}{M}\sum_{j=1}^{M} P_{\theta^j}\right) = \mathbb{E}_{J\sim\mathrm{Unif}[M],\ Z|J\sim P_{\theta^J}}\left(\log\frac{Q(J)Q(Z|J)}{\bar Q\cdot Q(J)}\right) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{E}_{Z|J}\log\frac{Q(Z|J)}{\bar Q} = \frac{1}{M}\sum_{j=1}^{M} KL(P_{\theta^j}\,\|\,\bar Q)$

convexity of KL $KL(\lambda p_1+(1-\lambda)p_2\,\|\,\lambda q_1+(1-\lambda)q_2)\leq\lambda KL(p_1\|p_2)+(1-\lambda)KL(q_1\|q_2)$

$I(Z;J)\leq\frac{1}{M^2}\sum_{j=1}^{M}\sum_{k=1}^{M} KL(P_{\theta^j}\,\|\,P_{\theta^k})$ 因为 $\left(\bar Q=\frac{1}{M}\sum_{k=1}^{M} P_{\theta^k}\right)$

$\leq \sup_{\theta^j,\theta^k} KL(P_{\theta^j}\,\|\,P_{\theta^k})$



$X_1,\ldots,X_n\sim N(\theta,\sigma^2 I_d)$ local packings

$KL(P^n_{\theta^j},P^n_{\theta^k})=\frac{n\|\theta^j-\theta^k\|_2^2}{2\sigma^2}\leq\frac{2nr^2}{\sigma^2}$

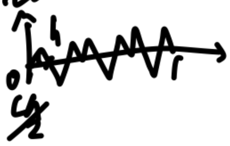
$KL(\ )=\frac{(\theta-\hat\theta)^T(\theta-\hat\theta)}{2\sigma^2}$

$\log M\gtrsim\log\left(\frac{r}{\delta}\right)^d\gtrsim d\log\left(\frac{r}{\delta}\right)\gtrsim d.$

$\Rightarrow$ Let $\frac{2nr^2}{\sigma^2}\lesssim d$, $r\lesssim\sigma\sqrt{\frac{d}{n}}$ $\overset{\text{Fano's}}{\Rightarrow}$ $\mathfrak{M}\gtrsim\frac{\sigma^2 d}{n}$

Lipshitz fn. esti. with l_2^2 loss

$X \sim \mathrm{Unif}(0,1)$ $\quad y = f(X) + \varepsilon$



$\Phi(x) = \begin{cases} x, & x \in (0, \frac{1}{2}) \\ 1-x, & x \in (\frac{1}{2}, 1) \end{cases}$

$\varphi_j = Lh\,\Phi\left(\frac{x-j}{h}\right)$

$f_\omega(x) = \sum_{j=0,h,2h,\dots}^{1} \omega_j \varphi_j(x)$, $\omega_j \in \{\pm 1\}$

$KL(P_{\omega_1}, P_{\omega_2}) = \frac{n\int_0^1 (f_{\omega_1}(x) - f_{\omega_2}(x))^2 dx}{2\sigma^2} \lesssim \frac{nL^2h^3}{\sigma^2} \times d_H(\omega_1, \omega_2)$

hamming distance. number of disagree entries between ω_1 & ω_2

$\rho(f_{\omega_1}, f_{\omega_2}) \gtrsim L^2h^3 d_H(\omega_1, \omega_2)$

Varshamov-Gilbert bound.

$\exists\, \Omega \subseteq \{-1,1\}^d$, $\omega_1, \omega_2 \in \Omega$ $\quad d_H(\omega_1, \omega_2) \geq \frac{d}{8}$, $|\Omega| \geq 2^{\frac{d}{8}}$

$d = \frac{1}{h}$

$KL \lesssim \frac{nL^2h^3}{2\sigma^2} \times \frac{1}{h}$, $\delta^2 \gtrsim L^2h^3 \times \frac{1}{h} \gtrsim L^2h^2$

$\log M \gtrsim \frac{1}{h}$

$\Rightarrow \frac{nL^2h^2}{\sigma^2} \lesssim \frac{1}{h}$, $h \leq n^{-1/3}$. $\delta^2 \gtrsim n^{-2/3}$

Recap

1. Fano's ineq

$$\inf_{\psi} Q\,\psi(Z) \neq J) \geq 1 - \frac{I(Z;J) + \log 2}{\log M}$$

$\Rightarrow$: $P_{\theta^1} \dots P_{\theta^M}$ such that $\rho(\theta^i, \theta^j) \geq 2\delta$, $I(Z;J) \leq \log M$

then $M \geq \Phi(\delta)$

Bounding $I(Z;J)$ mutual information

$$I(Z;J) = \frac{1}{M}\sum_{j=1}^{M} KL\left(P_{\theta^j} \,\Big\|\, \frac{1}{M}\sum_{j=1}^{M} P_{\theta^j}\right)$$

$$\leq \frac{1}{M^2}\sum_j \sum_k KL(P_{\theta^j} \| P_{\theta^k})$$

$$\leq \sup_{j,k} KL(P_{\theta^j} \| P_{\theta^k})$$ local packings

使 $\rho(\theta^j, \theta^k)$ 最小化 $KL(P_{\theta^j}, P_{\theta^k})$ 最大

求 lower bound of minimax.

Varshamo Gilbert:

$\Omega \subseteq \{-1, 1\}^d$, $|\Omega| \geq 2^{d/8}$

& $d_{HW}(w, w') \geq \frac{d}{8}$ 设[illegible]距离

$d(w, w') \in \Omega$

Eg: $X_1 \cdots X_n \sim N(\theta^*, \sigma^2 I)$

Support recovery $\rightarrow$ Est. $\|\theta\|_0 \leq S$ $\rightarrow$ Est. $\|\theta\|_1 \leq R$

1) hard thresholding, HT_λ, $\lambda = \sigma\sqrt{\frac{\log d}{n}}$

$\hat{\theta} = HT_\lambda(\bar{X})$, $supp(\hat{\theta}) = supp(\theta^*)$ if $|\theta^*|_{min} \gtrsim \sigma\sqrt{\frac{\log d}{n}}$

$\mathbb{E}\|\hat{\theta} - \theta^*\|_2^2 \lesssim \frac{s \log d}{n}$

$\lesssim \sigma R\sqrt{\frac{\log d}{n}}$

$\mathbb{1}\{supp(\hat{\theta}) \neq supp(\theta)\}$, $M = \inf_{\hat{\theta}} \sup_{\theta^*} \mathbb{P}(supp(\hat{\theta}) \neq supp(\theta^*))$, minimax risk

$\theta^1 = \begin{pmatrix} \theta_{min} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ $\theta^2 = \begin{pmatrix} 0 \\ \theta_{min} \\ 0 \\ 0 \\ 0 \end{pmatrix}$ $\dots$ $\delta = 1$. $\rho(\theta^i, \theta^j) \geq 1$

$KL(P^n_{\theta^1} \| P^n_{\theta^2}) = \frac{2n\,\theta_{min}^2}{2\sigma^2} \Rightarrow$ if $\frac{n\,\theta_{min}^2}{\sigma^2} \lesssim \log d$, $M \gtrsim C$

$(KL < \log$ of packing, by Fano's ineq. lower bound minimax$)$

2) $\alpha, w \in \{-1, 0, +1\}^d$, $\theta^i = \alpha w^i$ $(H) \to \sum_{t=1}^{s} \binom{d}{t} \to \left(\frac{d}{s}\right)^s$

Sparse Varshamov-Gilbert. $\Omega \subseteq (H)$, $|\Omega| \geq c \sum_{t=1}^{s} \binom{d}{t}$

$d_H(w, w') \geq s/8$, $w, w' \in \Omega$

Sauer's Lemma

$\rho(\theta^i, \theta^j) = \|\theta^i - \theta^j\| \geq \sqrt{\frac{s}{8}}\alpha = \delta$

$\Phi(\delta) = \delta^2$

$\|\theta^i - \theta^j\|_2^2 \geq \frac{s}{8} \times \alpha^2$, $KL(P_{\theta^i} \| P_{\theta^j}) \leq \frac{n s \alpha^2}{2\sigma^2}$

Let KL $\leq$ log of packing.

$\frac{n s \alpha^2}{\sigma^2} \lesssim s \log \frac{d}{s} \Rightarrow \alpha \leq \sqrt{\frac{\log(d/s)}{n}}$, $\Rightarrow M \geq \frac{\sigma^2 s \log(d/s)}{n}$.

3) $\theta^i = \frac{R}{K} w^i$, $\|\theta^i - \theta^j\|_2^2 \geq \frac{K}{8}\left(\frac{R}{K}\right)^2 = \delta^2$, $KL(P_{\theta^i} \| P_{\theta^j}) \leq \frac{nK\left(\frac{R}{K}\right)^2}{\sigma^2}$

$w \in \{1, 0\}$

$\Rightarrow KL \leq \log(d) \Rightarrow K \gtrsim \sqrt{\frac{nR^2}{\sigma^2 \log d}}$

$M \gtrsim \delta^2 = \frac{R^2}{K} = \sigma R \sqrt{\frac{\log d}{n}}$

Yang-Barron

$I(Z, J) \leq \sup_{i,j} KL(P_{\theta^i} \| P_{\theta^j})$

$I(Z, J) \leq \log M$

~~H~~ $H(Z|J)$

YB: $I(Z, J) \leq \inf_{\varepsilon > 0} \{\varepsilon^2 + \log N_{\sqrt{KL}}(\varepsilon)\}$ $N_{\sqrt{KL}} = \{P_{\gamma^1}, \ldots P_{\gamma^N}\}^P$. for any $P \in \mathcal{P}$ $KL(P \| P_{\theta^j}) \leq \varepsilon^2$ for some j

↳ $\leq$ 跟 packing 无关，总有 upper bound

Roadmap.

1. Find ε, s.t. $\varepsilon^2 \geq \log N_{\sqrt{KL}}(\varepsilon)$
   smallest. $I \leq 2\varepsilon^2$
2. $\log M(\delta) >> 4\varepsilon^2$. find δ → global packing (全部)
3. Apply Fano. $M \geq \Phi(\delta)$

local packing (before)
↓
会与 KL 耦合

Non-parametric regression. $f \in \mathcal{F}$. $\log N(\varepsilon, \mathcal{F}, \|\cdot\|_2) \asymp \left(\frac{1}{\varepsilon}\right)^{d/\alpha}$

$y = f(x) + \varepsilon$ $X \sim \text{Uniform}[0, 1]$. $KL(P_{f^1}^n \| P_{f^2}^n) = \frac{n\|f^1 - f^2\|_2^2}{2\sigma^2}$, $\sqrt{KL} \leq \varepsilon$, $\|f^1 - f^2\|_2 \leq \frac{\sigma\varepsilon}{\sqrt{n}}$

$\Rightarrow$ KL covering number
$\log N_{\sqrt{KL}}(\varepsilon) = \left(\frac{\sqrt{n}}{\sigma\varepsilon}\right)^{d/\alpha}$

$\varepsilon^2 \geq \left(\frac{\sqrt{n}}{\sigma\varepsilon}\right)^{d/\alpha}$

1. $I \leq n^{\frac{d}{2\alpha+d}}$, $\left(\frac{1}{\delta}\right)^{d/\alpha} \geq n^{\frac{d}{2\alpha+d}}$. $\delta \leq n^{\frac{-\alpha}{2\alpha+d}}$ $\Rightarrow M \geq n^{\frac{-2\alpha}{2\alpha+d}}$

YB Proof. $\{\theta^1, \ldots \theta^M\}$. $\{P_{\gamma^1}, \ldots P_{\gamma^N}\}$. $I(Z, J) = \frac{1}{M}\sum_{j=1}^{M} KL(P_{\theta^j} \| \bar{Q})$.

Pick any Q, then $I(Z, J) \leq \frac{1}{M}\sum_{j=1}^{M} KL(P_{\theta^j} \| Q)$

$= \frac{1}{M}\sum_{j=1}^{M} \mathbb{E}_{P_{\theta^j}} \log \frac{P_{\theta^j}}{Q} = \frac{1}{M}\sum_{j=1}^{M} \left(\mathbb{E}_{P_{\theta^j}} \log \frac{P_{\theta^j}}{\bar{Q}} + \mathbb{E} \log \frac{\bar{Q}}{Q}\right)$

$= I(Z, J) + KL(\bar{Q} \| Q)$

$$Q = \frac{1}{N}\sum_{j=1}^{N} P_{\nu j}, \quad I \le \frac{1}{M}\sum_{j=1}^{M} KL\left(P_{\theta i} \,\Big\|\, \frac{1}{N}\sum_{k=1}^{N} P_{\nu k}\right)$$

$$\le \mathbb{E}_{P_{\theta j}} \log \frac{P_{\theta j}}{\frac{1}{N} P_{\nu j}} = \underbrace{\mathbb{E}_{P_{\theta j}} \log \frac{P_{\theta j}}{P_{\nu j}}}_{KL(\ )} + \log N$$

$$\le \varepsilon^2 + \log N_{KL}(\varepsilon)$$

Recap.

1. Yang-Barron. $I(Z;J) \le \inf_{\varepsilon > 0} \{ \varepsilon^2 + \log N_{\sqrt{KL}}(\varepsilon) \}$

   与如何选 J, z 无关: $\theta^1 \cdots \theta^M$

   $J \sim \text{Unif}(M)$

   $Z \sim P_{\theta_J}$

2. Strategy. Global packing.

   a. find smallest ε that $\varepsilon^2 \ge \log N_{\sqrt{KL}}(\varepsilon)$

   b. Find largest δ such that. $\log M(\delta) \ge 4\varepsilon^2$

   c. Conclude $M \ge \Phi(\delta)$

$n\varepsilon^2 \asymp \log N(\varepsilon)$

Non-parametric max Likelihood.

$\mathcal{P}, \ p^* \in \mathcal{P}, \ X_1 \cdots X_n \sim p^*$

$$\hat{p} = \underset{p \in \mathcal{P}}{\operatorname{argmax}} \, L_n(p) = \prod_{i=1}^{n} p(X_i)$$

Hellinger. $h^2(\hat{p}, p^*)$?

Informal. find $n\varepsilon^{*2} \asymp \log N(\mathcal{P}, \varepsilon^*, h)$

$$h^2(\hat{p}, p^*) \lesssim \varepsilon^{*2}$$

$L_n p \to$ likelihood of p

Finite collection $\{p_1, \cdots p_N\}$

Lemma. Wong-Shen. If $h^2(p, p^*) \ge \delta^2$, $\mathbb{P}\left(\frac{L_n p}{L_n p^*} \ge \exp\left(\frac{-n\delta^2}{4}\right) \right) \le \exp\left(\frac{-n\delta^2}{8}\right)$

Proof: $\mathbb{P}\left(\sqrt{\frac{L_n p}{L_n p^*}} \ge \exp\left(\frac{-n\delta^2}{8}\right) \right) \le \underline{\mathbb{E}\left[\sqrt{\frac{L_n p}{L_n p^*}}\right]} \exp\left(\frac{n\delta^2}{8}\right)$

$$\int \prod_{i=1}^{n} \sqrt{\frac{p}{p^*}} \, p^*(x_i) = \left(\int \sqrt{p(x) p^*(x)} \, dx \right)^n = (1 - H^2)^n$$

$$\le \exp(-nH^2)$$

$\underline{H = h^2(p, p^*)} \ge \delta^2$

$\{p_1 \cdots p_N\}$ $\quad \mathbb{P}\left(\frac{L_n(\hat{p})}{L_n(p^*)} \geq 1,\ h^2(\hat{p}, p^*) \geq \delta^2\right)$

$\leq \mathbb{P}\left(\sup_{p \in \mathcal{P},\, h^2(p,p^*) \geq \delta^2} \frac{L_n(p)}{L_n(p^*)} \geq 1\right)$

$\overset{\text{Wong-Shen}}{\leq} N \exp\left(\frac{-n\delta^2}{8}\right)$

$\therefore$ If $\frac{n\delta^2}{16} \geq \log N$, then $\mathbb{P}(h^2(\hat{p}, p^*) \geq \delta^2) \leq \exp\left(\frac{-n\delta^2}{16}\right)$

Sieve MLE

$\rightarrow$ Sieve: Hellinger cover, i.e. for any $p \in \mathcal{P}$, $\tilde{p} \in N$, s.t. $h^2(p, \tilde{p}) \leq \varepsilon^2$

Assumption. 1. $\exists\, \tilde{p} \in N$, $\sup_x \frac{p^*(x)}{\tilde{p}(x)} \leq \frac{11}{8}$ $\quad$ 2. $\chi^2(p^*, \tilde{p}) \leq \frac{1}{16}\varepsilon^2$

Proof $\mathbb{P}\left(\frac{L_n(\hat{p})}{L_n(u)} > 1,\ h^2(\hat{p}, p^*) \geq 2\delta^2\right) = \mathbb{P}\left(\frac{L_n \hat{p}}{L_n p^*} \cdot \frac{L_n p^*}{L_n(u)} > 1 \cdots\right)$

$u \rightarrow$ within $c\varepsilon^2$ in H^2 and χ^2

$u \gtrsim p^*$

$\leq \mathbb{P}\left(\frac{L_n \hat{p}}{L_n p^*} \geq \exp\left(\frac{-n\delta^2}{8}\right), h^2(\hat{p}, p^*) \geq \delta^2\right)$

$+ \mathbb{P}\left(\frac{L_n p^*}{L_n u} > \exp\left(\frac{n\delta^2}{8}\right)\right)$

$\Rightarrow \mathbb{P}(h^2(\hat{p}, p^*) \geq \delta^2) \leq 2N(\delta) \exp\left(\frac{-n\delta^2}{8}\right)$ $\quad \leq |N| \exp\left(\frac{-n\delta^2}{8}\right) + \exp\left(\frac{-n\delta^2}{8}\right)\left[\mathbb{E}_{p^*} \frac{p^*}{u}\right]^n$

If $n\delta^2 \geq \log N(\delta)$

then $\mathbb{P}(h^2(\hat{p}, p^*) \geq \delta^2) \leq \exp\left(\frac{-n\delta^2}{16}\right)$

$1 + \chi^2(p^*, u)$ $\leftarrow$ $\chi^2(p^*, u) = \int \frac{(p^* - u)^2}{u} = \int \frac{p^{*2}}{u} - 1$

$\leq \exp\left(\frac{-n\delta^2}{8}\right) \exp(n \chi^2(p^*, u))$

$\leq \exp(-n\delta^2/16)$

(a). $H^2(p^*, u) \leq \varepsilon^2$, $\ln \frac{p^*}{u} \leq C \rightarrow \chi^2(p^*, u) \leq C\varepsilon^2$

why?

$\int \frac{(p^* - u)^2}{u} \leq \int \frac{(\sqrt{p^*} - \sqrt{u})^2 (\sqrt{p^*} + \sqrt{u})^2}{u} \leq 2\int \left(\frac{p^*}{u} + 1\right)(\sqrt{p^*} - \sqrt{u})^2 \leq 2(C+1)\varepsilon^2$

Total variation density est.

$X_1 \cdots X_n \sim p^*$ $\quad H_0: p = p_1,\ H_1: p = p_2$, $A = \{p_1 > p_2\}$, $\hat{p} = \arg\min_{p \in \{p_1, p_2\}} |P_n(A) - p(A)|$

$X_1 \cdots X_n \sim p_1$, $|P_n(A) - P_1(A)| = \left|p_1(A) - \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{X_i \in A\}\right|$

$P_1\left(A - P_n(A) \geq \sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta$

$P_1(A) - P_2(A) = TV(P_1, P_2) \ge$

$\Delta(P) = |P(A) - P_n(A)| = |P(A) - P^*(A) + P^*(A) - P_n(A)|$

$\ge TV(P, P^*) - \underbrace{|P^*(A) - P_n(A)|}_{\sqrt{\frac{\log(1/\delta)}{n}}}$ Hoeffding.

If $TV(P, P^*) \ge \max\left\{\varepsilon, \sqrt{\frac{\log(1/\delta)}{n}}\right\}$

$$P(\Delta(P) \le \Delta(P^*)) \le \underbrace{P\left(\Delta P^* \ge \sqrt{\frac{\log 1/\delta}{n}}\right)}_{\le \delta} + \underbrace{P\left(\Delta P \le \sqrt{\frac{\log 1/\delta}{n}}\right)}_{\le \exp(-n\varepsilon^2)}$$

TV-version. Wong & Shen.

Recap.

1. Non-parametric MLE

$X_1 \dots X_n \sim p^* \in \mathcal{P}$, $\hat{p} = \arg\max_{p \in \mathcal{P}} \prod_{i=1}^n p(X_i)$

2. Wong-Shen. If $h^2(p, p^*) \geq \varepsilon$

$$\mathbb{P}\left(\prod_{i=1}^n \frac{p(X_i)}{p^*(X_i)} \geq \exp\left(-\frac{n\varepsilon^2}{6}\right)\right) \leq \exp\left(-\frac{n\varepsilon^2}{8}\right)$$

3. If $n\varepsilon^2 \gtrsim \log N(\varepsilon, \mathcal{P}, h) + \chi^2$ approx enough

then $\mathbb{P}\left(h^2(\hat{p}, p^*) \geq \varepsilon^2\right) \lesssim \exp(-cn\varepsilon^2)$

where $\hat{p} = \arg\max_{N(\varepsilon, \mathcal{P}, h)} \prod_{i=1}^n p(X_i)$

$X_1 \cdots X_n \sim p^*$ $\{p_1, p_2\} \Rightarrow A = \{p_1 > p_2\}$

$\Delta p = |p_n(A) - p(A)|$, $\hat{p} = \arg\min_{p \in \{p_1, p_2\}} \Delta(p)$

$\{p_1 \cdots p_N\} \to$ Yatracos's method. $A_{ij} = \{p_i > p_j\} \to \mathcal{A}$

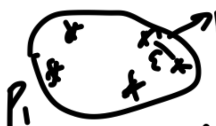
$\Delta(p) = \sup_{A \in \mathcal{A}} |p_n(A) - p(A)|$

$\Rightarrow \hat{p} = \arg\min_{p \in \{p_1 \cdots p_N\}} \Delta(p)$

$$\begin{aligned}
\mathbb{P}(TV(\hat{p}, p^*) \geq \varepsilon) &= \mathbb{P}(\Delta(\hat{p}) \leq \Delta(p^*), TV(\hat{p}, p^*) \geq \varepsilon) \\
&\leq \underline{\mathbb{P}(\Delta(\hat{p}) \leq \varepsilon/2, TV(\hat{p}, p^*) \geq \varepsilon)} + \underbrace{\mathbb{P}(\Delta(p^*) \geq \varepsilon/2)}_{\leq 2\exp\left(-\frac{2n\varepsilon^2}{4}\right) \cdot \binom{N}{2}} \\
\text{union bound} \leftarrow &\leq \sum_{i=1}^N \mathbb{P}(\Delta(p_i) \leq \varepsilon/2, TV(p_i, p^*) \geq \varepsilon) \qquad \text{Hoeffding} \quad \text{Union bound} \\
&\leq \sum_{i=1}^N \mathbb{P}\left(|p_n(A_{i*}) - p_i(A_{i*})| \leq \frac{\varepsilon}{2}, TV(p_i, p^*) \geq \varepsilon\right) \\
&\leq \sum_{i=1}^N \mathbb{P}\Big(|p_n(A_{i*}) - p^*(A_{i*})| \geq \underbrace{|p_i(A_{i*}) - p^*(A_{i*})|}_{TV(p_i, p^*) \geq \varepsilon} - \varepsilon/2\Big) \\
&\leq \sum_{i=1}^N \mathbb{P}(|p_n(A_{i*}) - p^*(A_{i*})| \geq \varepsilon/2) + \binom{N}{2}\exp(-n\varepsilon^2) \\
&\leq \underset{\text{Hoeffding}}{N\exp(-n\varepsilon^2)} + \binom{N}{2}\exp(-n\varepsilon^2)
\end{aligned}$$

If $n\varepsilon^2 \gtrsim \log N$

then $\mathbb{P}(TV(\hat{p}, p^*) \geq \varepsilon) \leq \exp(-n\varepsilon^2)$, $\varepsilon \approx \sqrt{\frac{\log N}{n}}$

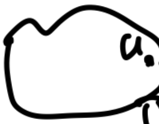
$\to \mathcal{P}$  p^* $n\varepsilon^2 \gtrsim \log N(\mathcal{P}, \varepsilon)$

then $\mathbb{P}(TV(\hat{p}, p^*) \geq \varepsilon) \leq \exp(-n\varepsilon^2)$

$TV(p^*, u) \leq \varepsilon$, $\mathbb{P}(\Delta\hat{p} \leq \Delta(u), TV(\hat{p}, p^*) \geq 10\varepsilon) \geq \mathbb{P}(TV(\hat{p}, p^*) \geq 10\varepsilon)$

$\leq \mathbb{P}(\Delta(\hat{p}) \leq 5\varepsilon, TV(\hat{p}, p^*) \geq 10\varepsilon) + \mathbb{P}(\Delta(u) \geq 5\varepsilon)$

$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - u(A)| \geq 5\varepsilon\right) \leq \mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - p^*(A)| \geq 4\varepsilon\right) \leq \binom{N}{2} \exp(-cn\varepsilon^2)$

robust estimation. $\mathcal{P}$  $u_0 \to p^*$ $X_1 \dots X_n \sim p^*$

$TV(u, p^*) \leq \rho$

$n\varepsilon^2 \gtrsim \log N(\varepsilon)$

$\mathbb{P}\left(TV(\hat{p}, p^*) \leq 3 \inf_{u \in \mathcal{P}} TV(u, p^*) + \varepsilon\right) \leq C\exp(-n\varepsilon^2)$

Huber's model. $p^* = (1-\rho)p + \rho Q$

$TV(p^*, p) \leq \varepsilon$

$X_1 \dots X_n \sim N(\mu, \sigma^2 I_d)$. $\mathbb{E}\|\hat{\mu} - \mu\|^2 \leq \frac{\sigma^2 d}{n}$

$\mu \in \mathbb{R}^d$

Nonparametric L-S $y_i = f(x_i) + \varepsilon_i$, $\varepsilon_i \sim N(0,1)$

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\mathbb{E}\|\hat{f} - f^*\|_n^2 = \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - f^*(x_i))^2$$

Local Gaussian width. $\mathcal{G}(\delta, \mathcal{F}) = \mathbb{E}\left[\sup_{\substack{g \in \mathcal{F} \\ \|g\|_n \leq \delta}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(x_i)\right]$ $\to$ Local.

$\mathcal{F}^* = \{f - f^* : f \in \mathcal{F}\}$

Critical Eq.. $\mathcal{G}(\delta, \mathcal{F}^*) \leq \frac{\delta^2}{2}$

then if $\mathcal{F}^*$ is star-shaped. for any $t \geq \delta^*$

$$\mathbb{P}\left(\|\hat{f} - f^*\|_n^2 \geq 16t\delta^*\right) \leq \exp\left(\frac{-nt\delta^*}{2}\right)$$

$\mathbb{E}\|\hat{f} - f^*\|_n^2 \lesssim \{\delta^{*2} + \frac{1}{n}\}$ $\to$ e.g. $n^{-1/3}$ one-step dis $n^{-1/3}$-localization

$n^{-1/2}$ chaining.

star-shape:

if $f \in \mathcal{F}$ then. $\alpha f + (1-\alpha) f^* \in \mathcal{F}$. $\alpha = (0,1)$ $\to$ f^* is the origin.

If $\mathcal{F}^*$ is not starshaped, $\mathcal{G}(\delta, \mathrm{star}(\mathcal{F}^*))$

Recap.

1. Estimation of TV. $X_1 \cdots X_n \sim p^* \in \mathcal{P}$

construct $\hat{p}$ such that $TV(\hat{p}, p^*)$ is small.

2. Yatracos method.

$\rightarrow$ find smallest ε. $n\varepsilon^2 \geq \log N_{TV}(\mathcal{P}, \varepsilon)$

$\rightarrow$ Construct $p_1 \cdots p_N$ which is an ε-cover of $\mathcal{P}$

$\rightarrow \mathcal{A} = A_{ij} = \{p_i > p_j\}$ $i, j \in \{1 \cdots N\}$

$$\Delta(p) = \sup_{A \in \mathcal{A}} |p_n(A) - p(A)|$$

$$\hat{p} = \arg\inf_{p \in \{p_1 \cdots p_N\}} \Delta(p)$$

3. Guarantee $TV(\hat{p}, p^*) < \sqrt{\cdots}$ robust..

$y_i = f(x_i) + w_i$, $w_i \sim N(0,1)$. $x_i \sim P_x$. $f^* \in \mathcal{F}$

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \rightarrow \mathbb{E} \frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))^2$$

Local Gaussian width.

$$\mathcal{G}(\delta, \mathcal{F}) = \mathbb{E}\Big[\sup_{\substack{g \in \mathcal{F} \\ \|g\|_n \leq \delta}} \frac{1}{n} \sum_{i=1}^n w_i g(x_i)\Big] \qquad \|g\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^n g^2(x_i)}$$

Critical radius. find δ^* $\mathcal{G}(\delta, \mathcal{F}) \leq \frac{\delta^2}{2}$

$\mathcal{F}$ is star-shaped around f^* $\mathcal{F}^* = \{f - f^*, f \in \mathcal{F}\}$. δ^*. $\mathcal{G}(\delta, \mathcal{F}^*) \leq \frac{\delta^2}{2}$

then $t \geq \delta^*$, $\mathbb{P}(\|\hat{f} - f^*\|_n^2 \geq 16 t \delta^*) \leq \exp(-\frac{n t \delta^*}{2})$

$\mathbb{E}\|\hat{f} - f^*\|_n \leq \delta^{*2} + \frac{1}{n}$

Lemma. If $\mathcal{F}$ is star-shaped. then. $\frac{\mathcal{G}(\delta, \mathcal{F})}{\delta}$ is $\downarrow$ fn of δ

Proof: $\mathcal{G}(\delta, \mathcal{F}) = \mathbb{E} \sup_{\substack{h \in \mathcal{F} \\ \|h\|_n \leq \delta}} \frac{1}{n} \sum_{i=1}^n w_i h(x_i) = \mathbb{E} \sup_{\substack{\tilde{h} \in \mathcal{F} \\ \|\tilde{h}\|_n \leq t}} \frac{1}{n} \sum_{i=1}^n w_i \tilde{h}(x_i) \times \frac{\delta}{t}$

$\leq \mathcal{G}(t, F) \times \frac{\delta}{t} \Rightarrow \frac{G(\delta, F)}{\delta} \leq \frac{\mathcal{G}(t, F)}{t}$. decreasing ..

$\frac{1}{n}\sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))^2 \leq \frac{2}{n}\sum_{i=1}^n w_i (\hat{f}(x_i) - f^*(x_i))$

denote. $\hat{\Delta} = \hat{f} - f^*$.

$\frac{\|\hat\Delta\|_n^2}{2} \leq \frac{1}{n}\sum_{i=1}^n w_i \hat\Delta(x_i) \leq \sup_{\|g\|_n \leq \|\hat\Delta\|_n} \frac{1}{n}\sum_{i=1}^n w_i g(x_i)$

$\mathbb{E}\|\hat\Delta\|_n^2 = \delta^2$, $\frac{\delta^2}{2} \leq G(F^*, \delta)$

$\delta \to G(\delta, F^*) \leq \frac{\delta^2}{2}$, $u \geq \delta$, $A(u) = \{g \in F.\ \|g\|_n \geq u, \frac{1}{n}\sum_{i=1}^n w_i g(x_i) \geq 2\|g\|_n \times u\}$

$\mathbb{P}(A(u)) \leq \exp(-\frac{nu^2}{2})$

$t \geq \delta$, $u = \sqrt{t\delta} \geq \delta$,

$\|\hat\Delta\|_n \leq \sqrt{t\delta} \to$ done ✓ $\|\hat\Delta\|_n \geq \sqrt{t\delta}$ cond. on $A(u)^c$ $\nearrow \|\hat\Delta\|_n \leq 4\sqrt{t\delta}$

$\frac{\|\hat\Delta\|_n^2}{2} \leq \frac{1}{n}\sum_{i=1}^n w_i \hat\Delta(x_i)$. $\frac{\|\hat\Delta\|_n^2}{2} \leq 2\|\hat\Delta\|_n \times u \leq 4u \Rightarrow$ on $A(u)^c$ $\|\hat\Delta\|_n^2 \leq 16 t\delta$

$\mathbb{P}(\|\hat\Delta\|_n^2 \geq 16 t\delta) \leq \exp(-\frac{nt\delta}{2})$

$\tilde{g} = \frac{g}{\|g\|_n} \times u$, $\frac{1}{n}\sum_{i=1}^n w_i \tilde{g}(x_i) \geq 2u^2$

$Z(u) = \sup_{\|\tilde g\|_n \leq u} \frac{1}{n}\sum_{i=1}^n w_i \tilde{g}(x_i)$, $\mathbb{P}(A(u)) \leq \mathbb{P}(Z(u) \geq 2u^2)$

$\mathbb{E}(Z(u)) \leq \frac{u^2}{2}$

$\sup_{\|g\|_n \leq u} \frac{1}{n}\sum_{i=1}^n w_i g(x_i) - \sup_{\|g\|_n \leq u} \frac{1}{n}\sum_{i=1}^n w_i' g(x_i)$

$(\mathbb{P}\ A(u) \leq \exp(-\frac{nu^2}{2})) \leq \frac{1}{n}\sum_{i=1}^n (w_i - w_i') g(x_i) \leq \frac{\|w - w'\|_2 \times u}{\sqrt{n}}$

$Z(u) \xrightarrow{\text{wrt.}} W$. L-Lipschitz.

L-Lip: $\mathbb{P}(Z(u) - \mathbb{E}(Z(u)) \geq t) \leq \exp(\frac{-t^2}{2L^2})$

$\mathbb{P}(A(u)) \leq \mathbb{P}(Z(u) \geq 2u^2) \leq \mathbb{P}(Z(u) - \mathbb{E}Z(u) \geq \frac{3}{2}u^2) \leq \exp(\frac{-n \cdot 9u^4}{2 \times 4u^2}) \leq \exp(\frac{-nu^2}{2})$

$\mathcal{G}(\delta, F) \leq \frac{16}{\sqrt{n}}\int_{\delta^2/4}^{\delta} \sqrt{\log N(u, F_\delta)}\, du + \frac{\delta^2}{4}$ $\quad \frac{16}{\sqrt{n}}\int_{\delta^2}^{\delta} \sqrt{\log N(u)}\, du \leq \delta^2/4$

eg. L-Lipschitz fns.

$\log N(F, u) \leq \log N_\infty(F, u) \lesssim \frac{L}{u}$, $\mathbb{E}\|\hat f - f^*\|_n^2 \lesssim \frac{1}{\sqrt{n}}\int_0^L \sqrt{\frac{L}{u}}\, du \lesssim \frac{1}{\sqrt{n}} f(L)$

$\frac{1}{\sqrt{n}} \times \sqrt{\delta} \lesssim \frac{\delta^2}{4}$ $\quad \delta^{3/2} \gtrsim \frac{1}{\sqrt{n}}$,

$\delta \gtrsim \frac{1}{n^{1/3}}$

$\mathbb{E}\|\hat f - f^*\|_n^2 \lesssim \frac{1}{n^{2/3}}$

Another example: $\sqrt{d\log(k_0)}$

Recap

1. Nonparametric LS. $\hat f = \arg\min_{f\in\mathcal F} \frac{1}{n}(y_i - f(x_i))^2$

2. Local Gaussian width

$$G(\delta, \mathcal F) = \mathbb E \sup_{\substack{\|g\|\le\delta \\ g\in\mathcal F}} \frac{1}{n}\sum_{i=1}^n g(x_i) w_i$$

3. Critical radius (δ^*)

Any positive soln to $G(\delta, \mathcal F^*) \le \frac{\delta^2}{2}$

4. Main result: $\mathbb E\|\hat f - f^*\|_n^2 \lesssim \delta^{*2} + \frac{1}{n}$

5. Can upper bound $G(\delta, \mathcal F^*) \le \frac{16}{\sqrt n}\underbrace{\int_{\delta^2/4}^{\delta} \sqrt{\log N(u, \mathcal F^*, \|\cdot\|_n)}\,du}_{I} + \frac{\delta^2}{4}$

Sufficient to find δ, $I \le \delta^2/4$

Estimating functional function

$X_1 \cdots X_n \sim f$. Entropy $= \int f\log f$. Expected density $= \int f^2$

Given $Y_1 \cdots Y_n \sim g$. Goodness-of-fit $\int (f - f_0)^2$

↳ KL: $\int f \log\frac{f}{g}$, H^2: $\int(\sqrt f - \sqrt g)^2$

Causal Inf: (X, A, Y) $X \sim p_X$, $P(A=1|X)$ = prop score $= \pi(X)$, $Y | A, X \sim N(\mu_0, 1)$, $N(\mu_1, 1)$

ATE $= \int (\mu_1(x) - \mu_0(x))\, p_X\, dx$

Plug-in estimator: $\hat f$, $T(f) = \int f^2$, $\hat T = \int \hat f^2$

$R = \mathbb E(\hat T - T)^2$ $\quad |\hat T - T| = |\int f^2 - \hat f^2| = |\int (\hat f - f)(\hat f + f)| \le (\|\hat f\|_\infty + \|f\|_\infty)\int|\hat f - f|$

$|\hat f - f| \lesssim n^{-\frac{\beta}{2\beta + d}}$ assume $X_1 \cdots X_n \sim f$, $[0,1]^d$, Hölder-β

Not obvious → parametric rate not achievable

$$\int f^2 = \int \hat f^2 + \int 2\hat f(f - \hat f) + \int(\hat f - f)^2 = -\int \hat f^2 + 2\int f\hat f + \int(\hat f - f)^2$$

$$\hat T = \int f^2 = -\int \hat f^2 + \frac{2}{n}\sum_{i=n+1}^{2n} \hat f(X_i)$$ ($2n$ $X_1 \cdots X_{2n}$, n total, $\hat f$ is fixed)

$$|\mathbb E\hat T - T| = |-\int \hat f^2 + 2\int f\hat f - \int f^2| = \int(\hat f - f)^2$$

$R_{plugin} \lesssim n^{-\frac{2\beta}{2\beta+d}}$, $R_{first\text{-}order} \lesssim \frac{1}{n} + n^{-\frac{4\beta}{2\beta+d}}$ if $\beta > d/2$ then $1/n$; $\beta < d/2$ then nonparametric rate

$R \lesssim \frac{1}{n} + n^{-\frac{8\beta}{4\beta+d}}$

$\beta > d/4$: $\frac{1}{n}$ (parametric) $\beta < d/4$: non-parametric

Parametric. $X_1 \dots X_n \sim N(\theta^*, I_d)$ $\quad \bar{X}, \bar{Y} \sim N(\theta^*, \frac{I_d}{n})$

$T(\theta^*) = \|\theta^*\|_2^2 \quad \hat{T} = \frac{1}{\binom{n}{2}} \sum_{i<j} X_i^T X_j = \bar{X}^T \bar{Y}$ $\quad \bar{X} = \theta^* + z, \; z \sim N(0, \frac{I_d}{n})$

$E\hat{T} - T = 0 \quad Var(\hat{T}) = \sum_{j=1}^{d} Var(\bar{X}_j \cdot \bar{Y}_j) = \sum_{j=1}^{d} \left(E(\bar{X}_j \cdot \bar{Y}_j)^2 - \theta_j^{*4} \right) = \sum_{j=1}^{d} \left(E(\theta_j^* + z_j)(\theta_j^* + \tilde{z}_j) - \theta_j^{*4} \right)$

$\leq \sum_{j=1}^{d} \left(\frac{1}{n^2} + \frac{\theta_j^{*2}}{n} \right) \lesssim \frac{\|\theta^*\|^2}{n} + \frac{d}{n^2}$

$R_{\hat{T}} \lesssim \frac{1}{n} + \frac{d}{n^2}$. If $d < n$, parametric 1D rate

$E\|\hat{\theta} - \theta^*\|_2^2 \lesssim d/n$

with decay: $\theta^*: \sum_{j=1}^{\infty} j^{\frac{2\beta}{d}} \theta_j^{*2} \leq C$

Truncate J: Bias: $\sum_{j=J+1}^{\infty} \theta_j^{*2}$, Var $\lesssim \frac{1}{n} + \frac{J}{n^2}$

Now how good is the estimator?

$\hat{T} = \sum_{j=1}^{J} \bar{X}_j \bar{Y}_j \quad R \lesssim \left(\sum_{j=J+1}^{\infty} \theta_j^{*2} \right)^2 + \frac{C}{n} + \frac{J}{n^2}$

$\lesssim J^{-\frac{4\beta}{d}} \left(\sum_{j=J+1} j^{\frac{2\beta}{d}} \theta_j^{*} \right)^2 + \frac{C}{n} + \frac{J}{n^2} \qquad J \asymp n^{\frac{2d}{4\beta+d}}$

$f = \sum_{j=1}^{\infty} \theta_j \varphi_j, \; \int f^2 = \sum_{j=1}^{\infty} \theta_j^2, \; \theta_j = \int f \varphi_j, \; \hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \varphi_j(X_i)$

$R_{\text{truncated series}} \leq R_{\text{minimax}}$