Basics:

$X \overset{d}{=} Y$ iff $F_X(t) = F_Y(t)$ for all t.

$Cov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y$

Law of total expectation: $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$

variance: $Var(Y) = Var[\mathbb{E}(Y|X)] + \mathbb{E}[Var(Y|X)]$

MGF: $M_X(t) = \mathbb{E}(e^{tX})$

$M_X^{(n)}(t)|_{t=0} = \mathbb{E}(X^n)$

Markov: $\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}[X]}{t}$ $(X \geq 0)$

Chebyshev: $\mathbb{P}(|X - \mathbb{E}(X)| \geq k\sigma) \leq \frac{1}{k^2}$

Chernoff bound: $P((X-\mu) \geq u) \leq \inf_{0 \leq t \leq b} \frac{\mathbb{E}[\exp(tX)]}{\exp(t(\mu+u))}$

Gaussian tail bounds.

$M_X(t) = \mathbb{E}[\exp(tX)] = \exp(t\mu + t^2\sigma^2/2)$

$\mathbb{P}(X-\mu \geq u) \leq \inf_{t \geq 0} \exp(-t(\mu+u))\exp(t\mu + t^2\sigma^2/2)$

$\leq \exp(-u^2/2\sigma^2)$, $t = u/\sigma^2$

sub-Gaussian:

$\mathbb{E}[\exp(t(X-\mu))] \leq \exp(t^2\sigma^2/2)$.

Rademacher R.V. $\mathbb{E}[\exp(tX)] \leq \exp(\frac{t^2}{2})$, 1-subGaussian.

$[a,b]$ bounded, $b-a$-sub gaussian.

$\mathbb{E}_X[\exp(tX)] = \mathbb{E}_X[\exp(t(X - \mathbb{E}(X')))] \leq \mathbb{E}_{X,X'}[\exp(t(X-X'))]$

$= \mathbb{E}_{X,X'}[\cancel{\exp}\, \mathbb{E}_\epsilon[\exp(t\epsilon(X-X'))]] \leq \mathbb{E}_{X,X'}[\exp(t^2(X-X')^2/2)]$

$\leq \exp(t^2(b-a)^2/2)$

$\mathbb{P}(|\frac{1}{n}\sum X_i - \mu| \geq t) \leq 2\exp(-\frac{nt^2}{2(b-a)^2})$ or

$\leq 2\exp(-\frac{2nt^2}{(b-a)^2})$

Bernstein.

$P(|\hat{\mu}-\mu| \geq t) \leq 2\exp(-\frac{nt^2}{2(\sigma^2 + (b-a)t)})$

McDiarmid's

$P(|f(X_1,\dots,X_n) - \mathbb{E}[f(X_1,\dots,X_n)]| \geq t) \leq 2\exp(\frac{-2t^2}{\sum L_k^2})$

$|f(x_1,\dots,x_n) - f(x_1,\dots,x_k',\dots,x_n)| \leq L_k$.

U statistic: $U(X_1,\dots,X_n) = \frac{1}{\binom{n}{2}} \sum_{j<k} g(X_j, X_k)$

Levy's

$|f(X_1,\dots,X_n) - f(Y_1,\dots,Y_n)| \leq L\sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$

$P(|f(X_1,\dots,X_n) - \mathbb{E}[f(X_1,\dots,X_n)]| \geq t) \leq 2\exp(\frac{-t^2}{2L^2})$

χ^2 tail bound

$\mathbb{P}(|\frac{1}{n}\sum^n Z_k^2 - 1| \geq t) \leq 2\exp(-nt^2/8)$ for all $t \in (0,1)$.

Converge in probability: $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$

WLLN: $\lim_{n\to\infty} \mathbb{P}(|\frac{1}{n}\sum X_i - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$

Converge in quadratic mean

$\lim_{n\to\infty} \mathbb{E}(X_n - X)^2 \to 0$

Converge in dist. $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$.

Slutsky's $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c$. $X_n + Y_n \xrightarrow{d} X + c$, $X_nY_n \xrightarrow{d} cX$

$a_n = o(1)$

$a_n = o(b_n)$ $a_n/b_n \to 0$

CLT: $\frac{\sqrt{n}(\hat{\mu}-\mu)}{\sigma} \xrightarrow{d} N(0,1)$

Fact: $\lim_{n\to\infty}(1 + x/n)^n \to \exp(x)$.

Lyapunov CLT:

$\mu_i = \mathbb{E}[X_i]$, $\sigma_i^2 = Var(X_i)$

$s_n^2 = \sum \sigma_i^2$

$\lim_{n\to\infty} \frac{1}{s_n^3}\sum_{i=1}^n \mathbb{E}[|X_i - \mu_i|^3] = 0$

$\frac{1}{s_n}\sum^n [X_i - \mu_i] \xrightarrow{d} N(0,1)$

Multivariate CLT: $\sqrt{n}(\hat{\mu}-\mu) \xrightarrow{d} N(0, \Sigma)$

CLT with $\hat{\sigma}_n$: $\frac{\sqrt{n}(\hat{\mu}-\mu)}{\hat{\sigma}_n} \xrightarrow{d} N(0,1)$

Delta: $\frac{\sqrt{n}(g(\bar{X}_n) - g(\mu))}{\sigma} \xrightarrow{d} N(0, [g'(\mu)]^2)$

Multivariate Delta:

$\sqrt{n}(g(\hat{\mu}_n) - g(\mu)) \xrightarrow{d} N(0, \nabla_\mu(g)^T \Sigma \nabla_\mu(g))$

$\Delta = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_X(x)|$

$\Delta(\mathcal{A}) = \sup_{A\in\mathcal{A}} |P_n(A) - P(A)|$

$\Delta(\mathcal{F}) = \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum^n f(X_i) - \mathbb{E}f|$

Shattering:

$N_\mathcal{A}(z_1,\dots,z_n) \leq 2^n$

n-th shattering coefficient of $\mathcal{A}$

$s(\mathcal{A}, n) = \max_{z_1\dots z_n} N_\mathcal{A}(z_1 \dots z_n)$

Vc theorem: $\Delta(\mathcal{A}) \leq \sqrt{\frac{32}{n}[d\log(n+1) + \log(8/\delta)]}$

$\mathbb{P}(\Delta(\mathcal{A}) \geq t) \leq 8 s(\mathcal{A}, n)\exp(-nt^2/32)$

GC theorem:

$P(\sup_x |\hat{F}_n(x) - F_X(x)| \geq t) \leq 8(n+1)\exp(-nt^2/32)$

Vc dimension: largest d. $s(\mathcal{A}, d) = 2^d$.

Sauer's Lemma: $s(\mathcal{A}, n) \leq (n+1)^d$, for $n > d$.

$P(\Delta(\mathcal{A}) \geq t) \leq 8(n+1)^d \exp(-nt^2/32)$

$\Delta(\mathcal{A}) \leq \sqrt{\frac{32}{n}[d\log(n+1) + \log(8/\delta)]}$

empirical Rademacher complexity

$\mathcal{R}(x_1,\dots,x_n) = \mathbb{E}_\epsilon[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum^n \epsilon_i f(x_i)|]$

$\mathcal{R}(\mathcal{F}) = \mathbb{E}_\epsilon \mathbb{E}_X[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum^n \epsilon_i f(X_i)|]$

$\mathbb{E}[\Delta(\mathcal{F})] \leq 2\mathcal{R}(\mathcal{F})$ Rademacher theorem

Finite class bound: $\mathcal{F} = \{f_1,\dots,f_N\}$. $\|f\|_\infty \leq b$

$\mathcal{R}(\mathcal{F}) \leq 2b\sqrt{\frac{\log(2N)}{n}}$

Sufficient Statistic:

$p(x_1,\dots,x_n | T(x_1,\dots,x_n) = t; \theta)$ does not depend on θ

The factorization theorem.

$T(X_1,\dots,X_n)$ is sufficient for θ iff

$p(x_1,\dots,x_n;\theta) = h(x_1,\dots,x_n) \times g(T(x_1,\dots,x_n);\theta)$

Minimal sufficiency:

$\frac{p(y_1,\dots,y_n;\theta)}{p(x_1,\dots,x_n;\theta)}$ does not depend on θ iff

$T(x_1,\dots,x_n) = T(y_1,\dots,y_n)$

Rao-Blackwell theorem:

$\hat{\theta}$, $\tilde{\theta} = \mathbb{E}[\hat{\theta}|T]$

$R(\tilde{\theta},\theta) \leq R(\hat{\theta},\theta)$

$R(\tilde{\theta},\theta) = \mathbb{E}[(\mathbb{E}(\hat{\theta}|T) - \theta)^2]$

$= \mathbb{E}[(\mathbb{E}(\hat{\theta}-\theta)|T)^2]$

$\leq \mathbb{E}\,\mathbb{E}(\hat{\theta}-\theta)^2|T$

$= R(\hat{\theta},\theta)$.

Exponential Families.

$p(x;\theta) = \exp[\sum^s \eta_i(\theta)T_i(x) - A(\theta)]\,h(x)$

$= \exp[\sum^s \theta_i T_i(x) - A(\theta)]\,h(x)$.

$A(\theta) = \log\int_x \exp[\sum^s \theta_i T_i(x)]\,h(x)\,dx$

natural parameter space: θ, $A(\theta) < \infty$

preserved exponential family:

$p(x_1,\dots,x_n;\theta) = \prod^n h(x_i)\exp[\sum \theta_i \sum_j T_i(x_j) - nA(\theta)]$

$\frac{\partial A(\theta)}{\partial\theta_i} = \mathbb{E}[T_i(X)]$ $\frac{\partial^2 A(\theta)}{\partial\theta_i\partial\theta_j} = Cov(T_i(X), T_j(X))$

A is convex.

$LL(\theta, x_1,\dots,x_n) \propto [\sum_i^s \theta_i \sum_j^n T_i(x_j) - nA(\theta)]$

concave. $-n\frac{\partial^2 A(\theta)}{\partial\theta_i\partial\theta_j}$

minimal representation: no redundant ss

$\sum^s \alpha_i T_i(x) = $ const. no set of $\alpha \in \mathbb{R}^s$.

Non-minimal - over complete. not identifiable

full rank. θ_i is s-dimensional.

Curved: if θ_is are related. $\theta_2 = \theta_1^2$

for full rank family, ss turn out to be MSS.

$p^* = \arg\max_p H(p) = -\int_{x\in\mathcal{X}} p(x)\log(p(x))\,dx$

subject to $\hat{\mu}_i = \mathbb{E}_p[T_i(X)]$ for $i \in \{1,\dots,s\}$

$p^*(x) = \exp[\sum^s \theta_i T_i(x) - A(\theta)]\,h(x)$

MLE estimator and MoM coincide:

$\frac{\partial LL(\theta; x_1\dots x_n)}{\partial\theta} = 0 \Rightarrow \mathbb{E}[T_i(X)] = \frac{1}{n}\sum_j^n T_i(x_j)$

MLE equivariant. $\eta = g(\theta)$, then $\hat{\eta} = g(\hat{\theta})$

$L^*(\eta) = L(\theta)$ if invertible.

$L^*(\eta) = \sup_{\theta: g(\theta)=\eta} L(\theta)$.

Bayes estimator: $p(\theta|x_1,\dots,x_n) \propto L(\theta)p(\theta)$

$E_\theta(\hat{\theta}-\theta)^2 = (E_\theta(\hat{\theta}) - \theta)^2 + Var_\theta(\hat{\theta})$

$S(\theta) = \nabla_\theta LL(\theta)$, $I(\theta) = \mathbb{E}[S(\theta)S(\theta)^T]$

$\mathbb{E}_{p(x_1,\dots,x_n;\theta)}[S(\theta)] = \sum^n \int \nabla_\theta \log p(x_i;\theta)\,p(x_1,\dots,x_n;\theta)\,dx_1 dx_2 \dots dx_n$

$= \sum^n \int \nabla_\theta \log p(x_i;\theta)\,p(x_i;\theta)\,dx_i$

$= 0$

$I(\theta) = Var(S(\theta)) = -\mathbb{E}[\nabla_\theta^2 \log p(x;\theta)]$

Cramér-Rao Bound.

$Var(\hat{\theta}) \geq \frac{1}{nI_1(\theta)}$

KL: $L(a,\theta) = \mathbb{E}_{X\sim p(x;\theta)}\log(\frac{p(X;\theta)}{p(X;a)})$

$R(\theta, \hat{\theta}(X)) = \mathbb{E}_\theta L(\hat{\theta}(X), \theta) = E_\theta(\hat{\theta}-\theta)^2$

$B_\pi(\hat{\theta}) = \int R(\theta,\hat{\theta})\pi(\theta)\,d\theta$. Bayes risk.

$B_\pi(\hat{\theta}) = \inf_{\tilde{\theta}} B_\pi(\tilde{\theta})$ Bayes estimator

$\sup_\theta R(\theta,\hat{\theta}) = \inf_{\tilde{\theta}}\sup_\theta R(\theta,\tilde{\theta})$ MinMax

$B_\pi(\hat{\theta}) = \int R(\theta,\hat{\theta})\pi(\theta)\,d\theta =$

$\iint L(\theta,\hat{\theta}(x^n))\,p(x^n|\theta)\,dx\,d\theta = \iint L(\theta,\hat{\theta}(x^n))\,p(x^n;\theta)\,dx\,d\theta = \iint L(\theta,\hat{\theta}(x^n))\,\pi(\theta|x^n)\,d\theta\, m(x^n)\,dx^n$

if $\hat{\theta}$ is Bayes estimator, $R(\theta,\hat{\theta})$ is constant

$R(\theta,\hat{\theta}) \leq B_\pi(\hat{\theta})$ for all θ. then $\hat{\theta}$ is minimax estimator

least favorable prior

MLE consistency:

$\to \inf_{\hat{\theta}: |\theta - \hat{\theta}| \geq \epsilon} KL(p(x;\theta)\|p(x;\hat{\theta})) > 0$.

$\to \sup_{\hat{\theta}} |R_n(\hat{\theta},\theta) - R(\hat{\theta},\theta)| \xrightarrow{P} 0$.

$KL(p(x;\theta)\|p(x;\hat{\theta})) = R(\hat{\theta},\theta) = R(\hat{\theta},\theta) - R_n(\hat{\theta},\theta) + \underbrace{R_n(\hat{\theta},\theta)}_{\leq 0} \leq R(\hat{\theta},\theta) - R_n(\hat{\theta},\theta) \xrightarrow{P} 0$

$= \frac{1}{n}\sum \log\frac{p(X_i;\theta)}{p(X_i;\hat{\theta})} \leq 0$.

MLE Asymptotic: $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0, 1/I_1(\theta))$

$0 = LL'(\hat{\theta}) = LL'(\theta) + (\hat{\theta}-\theta)LL''(\hat{\theta})$

$\sqrt{n}(\hat{\theta}-\theta) = \frac{LL'(\theta)/\sqrt{n}}{-\frac{LL''(\hat{\theta})}{n}}$ $\xrightarrow{d} N(0, I(\theta))$, $\xrightarrow{P} I(\theta)$.

Type I error: reject H_0 when H_0 true

power function: $\beta(\theta) = \mathbb{P}_\theta(X_1,\dots,X_n \in R)$

size α. level α $\sup_{\theta\in\Theta_0}\beta(\theta) \leq \alpha$

C_α, uniformly most powerful (N-P)

$T_n = \frac{L(\theta_1)}{L(\theta_0)} > k$, $P_{\theta_0}(X^n \in R) = P_{\theta_0}(T_n > k) = \alpha$

(Wald) $T_n = \frac{\sqrt{n}(\hat{\theta}-\theta_0)}{\hat{\sigma}_0} \xrightarrow{d} N(0,1)$

(LRT) $\lambda(x_1,\dots,x_n) = \frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\in\Theta} L(\theta)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$

$T_n = -2\log\lambda(X^n) \xrightarrow{d} \chi_1^2$

$-2\log\lambda(x^n) = -LL''(\hat{\theta})(\theta_0 - \hat{\theta})^2 = \frac{-\frac{1}{n}LL''(\hat{\theta})}{I_1(\theta)}(\sqrt{nI_1(\theta)}(\hat{\theta}-\theta_0))^2$

(Universal inference)

$\hat{\theta}_0 = \arg\max_{\theta\in\Theta_0} L_0(\theta; x_1,\dots,x_n)$ $H_0: \theta\in\Theta_0$

$\hat{\theta}_1 = \arg\max_{\theta\in\Theta_1} L_1(\theta; x_{n+1},\dots,x_{2n})$ $H_1: \theta\in\Theta_1$

$U_n = \frac{L_0(\hat{\theta}_0; x_1,\dots,x_n)}{L_0(\hat{\theta}_1; x_1,\dots,x_n)} \leq \alpha$.

$P(U_n \leq \alpha) \leq \alpha E_\theta[\frac{L_0(\hat{\theta}_1)}{L_0(\hat{\theta}_0)}] \leq \alpha E_\theta[\frac{L_0(\hat{\theta}_1)}{L_0(\theta)}]$

$\leq E[E[\frac{L_0(\hat{\theta}_1)}{L_0(\theta_0)}|D]] \leq 1$

posterior risk.

$= \iint r(\hat{\theta}|x^n)\,m(x^n)\,dx^n$.

p-value: $p = \inf\{\alpha : \{x_1, \dots, x_n\} \in R_\alpha\}$

$P = 1 - \Phi(T(x_1, \dots, x_n)) = \sup_{\theta \in \Theta_0} P_\theta(T_n(X_1, \dots, X_n) \geq T_n(x_1, \dots, x_n))$

$\mathbb{P}_0(\text{p-value} \le u) = \mathbb{P}_0(\Phi(-T_n) \le u) = \mathbb{P}_0(-T_n \le \Phi^{-1}(u)) = \Phi(\Phi^{-1}(u)) = u$

χ^2 test: $T = \sum^k \frac{(z_i - np_{0i})^2 - np_{0i}}{np_{0i}} \to \chi^2_{k-1}$

Two sample testing: $H_0: P = Q$, $H_1: P \neq Q$ $X_1, \dots, X_{n_1} \sim P$, $Y_1, \dots, Y_{n_2} \sim Q$

$\hat{C}_i = \frac{z_i + z_i'}{n_1 + n_2}$

$T_n = \sum^k \left(\frac{(z_i - n_1 \hat{C}_i)^2}{n_1 \hat{C}_i} + \frac{(z_i' - n_2 \hat{C}_i)^2}{n_2 \hat{C}_i} \right) \to \chi^2_{k-1}$

Permutation: $N = m + n$. $N!$ permutation of $\{X_1, \dots, X_m, Y_1, \dots, Y_n\}$

p-value $= \frac{1}{N!} \sum^{N!} \mathbb{1}(T_i > T_{obs})$

$\phi_{perm}(z_{obs}) = \mathbb{1}\left[\frac{1}{N!} \sum^{N!} \mathbb{1}(T_i \geq T_{obs}) < \alpha\right]$

$P_{H_0}(\phi_{perm}(z_{obs}) = 1) \le \alpha$. controls type I error

Multiple testing: FWER $= P$(falsely reject any null).

Sidak: reject if p-value $\le 1 - (1-\alpha)^{1/d}$.

p-value independent, FWER $= 1 - ((1-\alpha)^{1/d})^d = \alpha$

Bonferroni: reject if p-value $\le \frac{\alpha}{d}$

FWER $= P(\bigcup^d \text{reject } H_{0i}) \le \sum^d p(\text{reject } H_{0i}) \le \sum^d \frac{\alpha}{d} = \alpha$

Holm's, order, $i^* = \min\{i : P_{(i)} > \frac{\alpha}{d-i+1}\}$ reject all $H_{(i)}$ $i < i^*$

False Discovery Rate: FDP $= \begin{cases} \frac{V}{R}, & V \text{: false rejection.} \\ 0, & \end{cases}$

FDR $= \mathbb{E}[\text{FDP}]$ FWER $= \mathbb{P}(V \geq 1)$

BH: order, $t_i = \frac{i\alpha}{d}$, $i_{max} = \arg\max_i \{i : P_{(i)} < t_i\}$

reject all nulls upto and include H_i

at the cut-off $\frac{i_{max}\alpha}{d}$, $\frac{d_0 i_{max}\alpha}{d}$ null rejected.

Bootstrap: $X_{1B}^*, \dots, X_{nB}^* \sim P_n$. $\theta_{n,B}^* = g(X_{1B}^*, \dots, X_{nB}^*)$

$\hat{s}^2 = \frac{1}{B} \sum^B (\theta_{n,B}^* - \bar{\theta})^2$, $\bar{\theta} = \frac{1}{B} \sum^B \theta_{n,i}^*$

$F_n(t) = \mathbb{P}(\sqrt{n}(\hat{\mu}_n - \mu) \le t)$.

$\hat{F}_n(t) = \mathbb{P}(\sqrt{n}(\hat{\mu}_n^* - \mu) \le t \mid X_1, \dots, X_n)$.

theorem: $\sup_t |\hat{F}_n(t) - F_n(t)| = O_P(\frac{1}{\sqrt{n}})$, $\mathbb{E}|X_i|^3 < \infty$.

Berry-Esseen: $\sup_z |\mathbb{P}(Z_n < z) - \Phi(z)| \le \frac{33}{4} \frac{\mu_3}{\sigma^3 \sqrt{n}}$.

$X_1, \dots, X_n$. iid μ, σ^2. $\mu_3 < \infty$. $\bar{X}_n = \frac{1}{n} \sum X_i$,

Φ, cdf of $N(0,1)$. $Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$

Causal inference. $\tau = \mathbb{E}[Y(1) - Y(0)]$

$Y_i^{obs} = Y_i(1) W_i + Y_i(0)(1 - W_i)$

$\hat{\tau} = \frac{1}{m} \sum_{i:W_i=1} Y_i^{obs} - \frac{1}{n-m} \sum_{i:W_i=0} Y_i^{obs}$

$W \perp\!\!\!\perp (Y(0), Y(1))$

$\mathbb{E}(\hat{\tau}) = \frac{1}{m} E(W_i) Y_i(1) - \frac{1}{n-m} E(W_i) Y_i(0)$

$= \tau$

Confounding X

$\tau = \mathbb{E}_X[\mathbb{E}[Y(1) - Y(0) \mid X]]$

$\tau = \mathbb{E}_X[\mathbb{E}[Y^{obs} \mid X, W=1] - \mathbb{E}[Y^{obs} \mid X, W=0]]$

$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$

inverse propensity score.

$\pi(x) = p(W=1 \mid X)$.

$\hat{\tau} = \frac{1}{n} \sum^n \left[\frac{Y_i^{obs} W_i}{\pi(X_i)} - \frac{Y_i^{obs}(1 - W_i)}{1 - \pi(X_i)} \right]$

$\tau_1 = \mathbb{E}[Y^{obs} \mid X, W=1] = \mathbb{E}_W\left[\mathbb{E}\left[\frac{Y^{obs} W}{\pi(X)} \mid X = x\right]\right]$

Gaussian Sequence Model

$y_i = \theta_i + \epsilon_i$ $\epsilon_i \sim N(0, \sigma^2/n)$

$\hat{\theta} = y$, $R(\hat{\theta}, \theta) = \mathbb{E}[\sum \epsilon_i^2] = \frac{d\sigma^2}{n}$

HT: $\hat{\theta}_i = y_i \mathbb{1}(|y_i| \geq t)$.

ST: $\hat{\theta}_i = \text{sign}(y_i) \max\{|y_i| - t, 0\}$

HT. estimator: $\arg\min_\theta \frac{1}{2} \|y - \theta\|_2^2 + \frac{t^2}{2} \sum^d \mathbb{1}(\theta_i \neq 0)$

ST. estimator: $\arg\min_\theta \frac{1}{2} \|y - \theta\|_2^2 + t \sum^d |\theta_i|$

Maximum of Gaussian. $\max^d |\epsilon_i| \le \sigma \sqrt{2 \log(2d/\delta)}$

$R(\hat{\theta}, \theta) \lesssim \sum^d \min\{\theta_i^2, \frac{\sigma^2 \log(d)}{n}\} \to$ HT

S-sparse: $R(\hat{\theta}, \theta) \lesssim \frac{\sigma^2 s \log(d)}{n}$

l_1 sparsity: $\sum^d |\theta_i| \le R$.

$R(\hat{\theta}, \theta) \lesssim 2R\sigma \sqrt{\frac{\log(d)}{n}}$

Binary classification

$\hat{R}_n(f) = \frac{1}{n} \sum^n \mathbb{1}(f(x_i) \neq y_i)$

$\hat{f} = \arg\min_{f \in F} \hat{R}_n(f)$

f^*: best in F

$\Delta = \mathbb{P}(\hat{f}(X) \neq y) - \mathbb{P}(f^*(X) \neq y)$

$= \underline{\mathbb{P}(\hat{f}(X) \neq y) - \hat{R}_n(\hat{f})} + \underbrace{\hat{R}_n(\hat{f}) - \hat{R}_n(f^*)}_{T_2 \le 0} + \underbrace{\hat{R}_n(f^*) - \mathbb{P}(f^*(X) \neq y)}_{\text{small Hoeffding}}$

$\sup_{f \in F} [\mathbb{P}(f(X) \neq y) - \hat{R}_n(f)] \le \Theta$ $1 - \delta/2$.

then $\Delta \le \Theta + \sqrt{\frac{2 \log 2/\delta}{n}}$

$\xrightarrow{P} \Rightarrow \xrightarrow{d}$

$F_{X_n}(x) = \mathbb{P}(X_n < x) \le \mathbb{P}(X \le x + \epsilon) + \mathbb{P}(|X_n - X| \geq \epsilon)$

$F_X(x - \epsilon) = \mathbb{P}(X \le x - \epsilon)$

$F_X(x - \epsilon) - \mathbb{P}(|X_n - X| \geq \epsilon) \le F_{X_n}(x) \le F_X(x + \epsilon) + \mathbb{P}(|X_n - X| \geq \epsilon)$