STA 3000F — Lecture 5

Recall. $X_1, X_2 \cdots X_n \overset{iid}{\sim} N(\theta^*, I_d)$

$H_0 : \theta^* = 0$

$H_1 : \|\theta^*\|_2 \geq \varepsilon$

$\exists$ a test φ. s.t. $\mathbb{E}_0[\varphi(X)] + \sup_{\theta \in H_1} \mathbb{E}[1-\varphi(X)] \leq \frac{1}{10}$.

as long as $\varepsilon \geq \frac{c \cdot d^{1/4}}{n^{1/2}}$.

Thm (Lower bound) $\exists c_1 > 0$, for $\varepsilon < \frac{c_1 d^{1/4}}{n^{1/2}}$

$$\mathbb{E}_0[\varphi(X)] + \sup_{\theta \in H_1} \mathbb{E}[1-\varphi(X)] \geq \frac{1}{3}$$

for any test φ.

Want to use H_0 vs. H_1

$$\mathbb{E}_0[\varphi(X)] + \mathbb{E}_1[1-\varphi(X)] \geq 1 - d_{TV}(P_0, P_1)$$

First attempt : $H_0 : \theta = 0$

$H_1' : \theta = \theta_1$

$$Y = \frac{\theta_1^T}{\|\theta_1\|_2} \cdot \left(\frac{1}{n}\sum_{i=1}^n X_i\right).$$

Optimal testing radius $\asymp \frac{1}{\sqrt{n}}$.

H_0 : $\theta = 0$

v.s.

H_1' : $\theta \sim \mathrm{Unif}\left(\left\{\frac{\pm\varepsilon}{\sqrt{d}}, \frac{\pm\varepsilon}{\sqrt{d}}, \cdots, \frac{\pm\varepsilon}{\sqrt{d}}\right\}\right)$

2^d vectors in total

$X_1, \cdots X_n \overset{iid}{\sim} \mathcal{N}(\theta, I_d)$

Task : decide $\begin{cases} \theta = 0 & (P_0)\ H_0 \\ \text{Hierarchical process} & (P_1)\ H_1' . \end{cases}$

$$\mathbb{E}_0[\phi(X)] + \sup_{\theta \in H_1} \mathbb{E}_\theta[1-\phi(X)]$$

$$\geq \mathbb{E}_0[\phi(X)] + \underbrace{\frac{1}{2^d} \sum_{\theta \in \left(\frac{\pm\varepsilon}{\sqrt{d}}\right)^{\otimes d}} \mathbb{E}_\theta[1-\phi(X)]}_{= \mathbb{E}_{H_1'}[1-\phi(X)].}$$

$$\geq 1 - d_{TV}(P_0, P_1).$$

$$d_{TV}(P_0, P_1) = \mathbb{E}_{P_0}\left[\left|\frac{dP_1}{dP_0}(x) - 1\right|\right].$$

$$\overset{(C\text{-}S.)}{\leq} \sqrt{\mathbb{E}_{P_0}\left[\left(\frac{dP_1}{dP_0}(x) - 1\right)^2\right]}.$$

$$= \sqrt{\mathbb{E}_{P_0}\left[\frac{dP_1}{dP_0}(X)^2\right] - 1} \quad \leq \frac{1}{3}$$

"χ^2 divergence".

$$L(X_1^n) = \frac{dP_1}{dP_0}(X_1, X_2 \cdots X_n)$$

$$= \frac{\mathbb{E}_{z \sim \mathrm{Unif}(\{\pm 1\}^d)}\left[\exp\left(-\frac{1}{2}\sum_{i=1}^n \left\|X_i - \frac{\varepsilon z}{\sqrt{d}}\right\|_2^2\right)\right]}{\exp\left(-\frac{1}{2}\sum_{i=1}^n \|X_i\|^2\right)},$$

$$= \mathbb{E}_{z \sim \mathrm{Unif}(\{\pm 1\}^d)}\left[\exp\left(\frac{\varepsilon}{\sqrt{d}} \cdot n \bar{X}_n^T z - \frac{\varepsilon^2 n}{2}\right)\right].$$

$$\boxed{(\mathbb{E}(f(z)))^2 = \mathbb{E}\left[f(z) \cdot f(z')\right] \quad \underline{z, z' \overset{iid}{\sim}}}$$

$$\mathbb{E}_0\left[L(X_1^n)^2\right]$$

$$= e^{-\varepsilon^2 n} \cdot \mathbb{E}_0 \, \mathbb{E}_{z, z' \overset{iid}{\sim} \{\pm 1\}^d}\left[\exp\left(\frac{\varepsilon}{\sqrt{d}} \cdot n \bar{X}_n (z + z')\right)\right].$$

$$= e^{-\varepsilon^2 n} \cdot \mathbb{E}_{z, z' \overset{iid}{\sim} \{\pm 1\}^d}\left[\exp\left(\frac{n\varepsilon^2}{2d} \cdot \|z + z'\|_2^2\right)\right].$$

$$= \mathbb{E}_{z, z' \overset{iid}{\sim} \{\pm 1\}^d}\left[\exp\left(\frac{n\varepsilon^2}{d}\langle z, z'\rangle\right)\right].$$

Let $\zeta_i = z_i \cdot z_i'$

$$\mathbb{E}\left[L(x_1^n)^2\right] = \left(\mathbb{E}\left[\exp\left(\frac{n\varepsilon^2}{d}\cdot\zeta_1\right)\right]\right)^d$$

$$\left(\frac{1}{2}\exp\left(\frac{n}{d}\varepsilon^2\right) + \frac{1}{2}\exp\left(-\frac{n\varepsilon^2}{d}\right)\right)^d.$$

$$\leq \exp\left(\frac{1}{2}\left(\frac{n\varepsilon^2}{d}\right)^2\cdot d\right).$$

$$= \exp\left(\frac{n^2\varepsilon^4}{2d}\right). \quad \leq \frac{10}{9}.$$

Requires $\quad \varepsilon \leq \dfrac{c_i d^{1/4}}{n^{1/2}}$

Preliminaries. Convergence of r.v.'s.

(normed vector space, $\|\cdot\|$).

- Convergence in prob. $X_n \xrightarrow{P} X$

  $\forall \varepsilon > 0$, $\mathbb{P}\left(\|X_n - X\| > \varepsilon\right) \to 0$.

- L^p $\quad X_n \xrightarrow{L^p} X$. $\quad \mathbb{E}\left[\|X_n - X\|^p\right] \to 0$.

  (By Markov, $L^p \to p$).

- a.s. $\quad \mathbb{P}\left(\lim_{n\to+\infty} \|X_n - X\| \to 0\right) = 1$.

- We say X_n converges weakly (in distribution) to X $(X_n \xrightarrow{d} X)$

if for any function f that is bounded & continuous, there is $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$.

(cts. means $\forall x$, $\lim_{\|x'-x\|\to 0} f(x') = f(x)$).

(Portmanteau. $X_n \xrightarrow{d} X$ is equivalent to: $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$. for any bdd, Lip(1) f),

(Lip(1) $|f(x') - f(x)| \le \|x'-x\|$).

(in 1-D: when limiting cdf is cts, weak convergence $\Leftrightarrow$ cdf convergence).

Thm. (cts mapping thm) g cts function.

$X_n \xrightarrow{*} X$ then $g(X_n) \to g(X)$

for $* \in \{ P, a.s., d \}$

Thm (Slutsky).

(i) If c is deterministic then $X_n \xrightarrow{d} c \Leftrightarrow X_n \xrightarrow{P} c$.

(ii) If $X_n \xrightarrow{d} X$, $\|X_n - Y_n\| \xrightarrow{P} 0$

then $Y_n \xrightarrow{d} X$.

Proof: (i) Take $f(x) = \min\{\|x-c\|, 1\}$. $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(c)] = 0$.

(ii). $f \in$ Bounded, Lip(1).

$$\left| \mathbb{E}[f(Y_n)] - \mathbb{E}[f(X_n)] \right| \le \mathbb{E}\left[\min(\|X_n - Y_n\|, 2)\right] \to 0 \quad \text{(by DCT)}.$$

Consequences. $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c$.

then $X_n + Y_n \xrightarrow{d} X + c$, $X_n Y_n \xrightarrow{d} c \cdot X$.

$o_p(\cdot)$ and $O_p(\cdot)$ notations. (for R_n positive, deterministic)

- $X_n = o_p(R_n)$ — $X_n / R_n \xrightarrow{P} 0$
- $X_n = O_p(R_n)$ if $Y_n = X_n / R_n$ is uniformly tight.

i.e., $\forall \varepsilon > 0$, $\exists M > 0$, s.t. $\left(\sup_{n \ge 0} \mathbb{P}(\|Y_n\| \ge M) \le \varepsilon \right)$.

Thm (Prohorov). (i). $X_n \xrightarrow{d} X$ then $\{X_n : n \in \cancel{\mathbb{N}}\}$ is uniformly tight.

(ii) $\{X_n\}_{n \ge 0}$ uniformly tight, $\exists$ subseq $X_{n_j} \xrightarrow{d} X$ for some X.

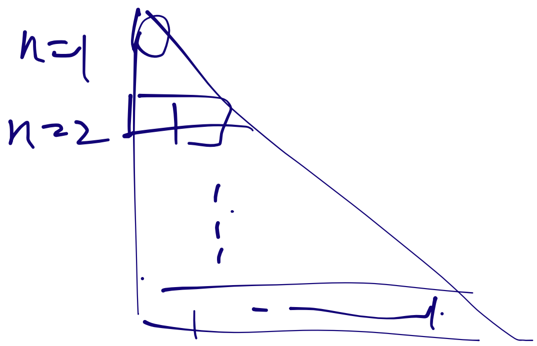
(For this class, uniform tightness restricts to finite-dim)

· LLN. $\mathbb{E}[|X|] < +\infty$. $\frac{1}{n}\sum_{1}^{n} X_i \xrightarrow[a.s.]{P} \mathbb{E}[X]$

· CLT. If $\cancel{\mathrm{var}(X)} < +\infty$. $X_1, \cdots \cancel{X_n \text{ iid}}$

$$\frac{1}{\sqrt{n}} \sum_{1}^{n} \left(X_i - \mathbb{E}[X]\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{var}(X)).$$

Lindeberg - Feller condition for triangle arrays.



$Y_{n1}, Y_{n2}, \cdots Y_{nK_n}$ indp. zero-mean.

$$\sum_{i=1}^{K_n} \mathrm{cov}(Y_{nK_i}) \xrightarrow{n\to+\infty} \Sigma.$$

and $\forall \varepsilon > 0$,

$$\sum_{i=1}^{K_n} \mathbb{E}\left[\|Y_{ni}\|_2^2 \, \mathbb{1}\{\|Y_{ni}\|_2 > \varepsilon\}\right] \longrightarrow 0.$$

then

$$\sum_{i=1}^{K_n} Y_{ni} \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Thm (Delta method). $(r_n \to +\infty)$.

If $r_n (T_n - \theta) \xrightarrow{d} T$.

then for $\phi : \mathbb{R}^k \to \mathbb{R}^m$, differentiable at θ.

we have $r_n \cdot \left(\phi(T_n) - \phi(\theta)\right) \xrightarrow{d} \nabla\phi(\theta) \cdot T$.

Proof — $\phi(t) = \phi(\theta) + \nabla\phi(\theta)^T (t - \theta) + R(t)$.

$$\left(\frac{R(t)}{\|t-\theta\|_2} \to 0\right).$$

$$r_n(T_n-\theta) \xrightarrow{d} T. \qquad \|T_n-\theta\|_2 = O_p(1/r_n).$$

$$\|T_n-\theta\|_2 = o_p(1).$$

$$\Longrightarrow \quad R(t) = o_p(\|T_n-\theta\|_2).$$

So we have

$$r_n(\phi(T_n)-\phi(\theta)) = \underbrace{r_n\cdot \nabla\phi(\theta)^T(T_n-\theta)}_{\xrightarrow{d} \nabla\phi(\theta)^T T.} + \underbrace{r_n\cdot R(T_n)}_{\xrightarrow{?\,p} 0}$$

$$r_n\cdot R(T_n) = r_n\cdot o_p(\|T_n-\theta\|_2) = o_p(r_n\cdot\|T_n-\theta\|_2)$$

$$= o_p(O_p(1)) = o_p(1).$$

e.g. $X_1, X_2, \dots, X_n \overset{iid}{\sim} \mathbb{P}$. $\quad \mathbb{E}[X]=\theta$, $\mathrm{cov}(X) = \Sigma$.

$$\sqrt{n}(\bar{X}_n-\theta) \xrightarrow{d} \mathcal{N}(0,\Sigma). \qquad (\theta \neq 0).$$

$$\phi(h) = \frac{1}{2}\|h\|_2^2.$$

$$\sqrt{n}\cdot\left(\frac{1}{2}\|\bar{X}_n\|_2^2 - \frac{1}{2}\|\theta\|_2^2\right) \xrightarrow{d} \mathcal{N}(0, \theta^T\Sigma\theta).$$

Thm (second-order Delta method).

$\phi: \mathbb{R}^d \to \mathbb{R}$ twice differentiable at θ.

$r_n(T_n-\theta) \xrightarrow{d} T$. and $\nabla\phi(\theta)=0$.

then $$r_n^2\cdot(\phi(T_n)-\phi(\theta)) \xrightarrow{d} \frac{1}{2}T^T\nabla^2\phi(\theta)\cdot T.$$

(Proof: second-order Taylor, generalizable to higher order)

e.g. (ct'd). If $\theta = 0$. $\phi(\theta) = \frac{1}{2}\|\theta\|_2^2$, $\nabla^2 \phi(\theta) = I_d$.

$$n \cdot \left(\frac{1}{2} \|\bar{X}_n\|_2^2 \right) \xrightarrow{d} \frac{1}{2} g^T g.$$

where $g \sim \mathcal{N}(0, \Sigma)$.

In particular, $\Sigma = I_d$. $n \cdot \|\bar{X}_n\|_2^2 \xrightarrow{d} \chi^2(d)$.

Def. (M-estimator). $\hat{\theta}_n := \underset{\theta \in \Theta}{\arg\min} \underbrace{\frac{1}{n} \sum_{i=1}^n f(\theta; X_i)}_{F_n(\theta)}$.

(e.g. MLE where $f = -\log p_\theta(X_i)$)

$F(\theta) = \mathbb{E}[f(\theta; X)]$. $\theta^* = \arg\min \{ F(\theta) \}$.

$$F(\hat{\theta}_n) - F(\theta^*)$$

$$= \boxed{F(\hat{\theta}_n) - F_n(\hat{\theta}_n)} + \underbrace{F_n(\hat{\theta}_n) - F_n(\theta^*)}_{\leq 0} + \underbrace{F_n(\theta^*) - F(\theta^*)}_{\frac{1}{n}\sum_{i=1}^n \left(f(\theta^*; X_i) - \mathbb{E}[f(\theta^*; X_i)] \right)}$$

$$\frac{1}{n} \sum_{i=1}^n \left(-f(\hat{\theta}_n; X_i) + F(\hat{\theta}_n) \right).$$

$$\left|F(\hat{\theta}_n) - F_n(\hat{\theta}_n)\right| \le \sup_{\theta \in \Theta} \left| \frac{1}{n}\sum_{1}^{n} f(\theta; X_i) - F(\theta)\right|.$$

Def/ Let $\mathcal{F}$ be a collection of functions.

we say ULLN is satisfied if

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{1}^{n} f(X_i) - \mathbb{E} f(X)\right| \xrightarrow{P} 0.$$

Cor (argmax convergence)

If $\forall \varepsilon > 0$, $\displaystyle\inf_{\|\theta - \theta^*\|_2 \ge \varepsilon} F(\theta) > F(\theta^*)$ $(+\delta)$.

then ULLN implies

$$\hat{\theta}_n \xrightarrow{P} \theta^*.$$

Proof: $\mathbb{P}(\|\hat{\theta}_n - \theta^*\|_2 \ge \varepsilon) \le \mathbb{P}\left(F(\hat{\theta}_n) - F(\theta^*) \ge \delta\right) \xrightarrow{P} 0.$

Warmup. $|\Theta| < +\infty$.

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |F_n(\theta) - F(\theta)| \ge \varepsilon\right).$$

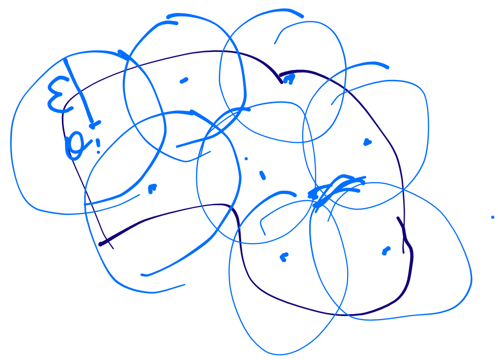
$$\le \sum_{\theta \in \Theta} \mathbb{P}\left(|F_n(\theta) - F(\theta)| \ge \varepsilon\right)$$

$$\to 0.$$

Key idea: discretization.

- "~~covering number~~" for a set K, under metric ρ, $\varepsilon > 0$.

$$N(K; \rho, \varepsilon) = \min\{N : \exists \{\theta_i\}_{i=1}^N \subseteq K, \text{ s.t. } K \subseteq \bigcup_{i=1}^N \mathbb{B}_\rho(\theta_i, \varepsilon)\}$$



- "packing number":

$\{\theta_i\}_{i=1}^M \subseteq K$ is ε-packing if $\rho(\theta_i, \theta_j) \geq \varepsilon$ $(\forall\, i \neq j)$.

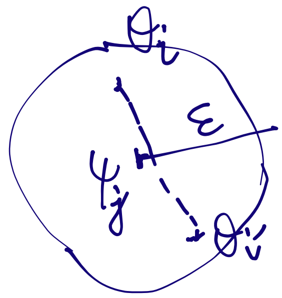
$$M(K; \rho, \varepsilon) := \max\{M : \exists\ \varepsilon\text{-packing } \{\theta_i\}_{i=1}^M \text{ of } K \text{ under } \rho\}.$$

Thm (duality).

$$M(K; \rho, 2\varepsilon) \overset{(i)}{\leq} N(K; \rho, \varepsilon) \overset{(ii)}{\leq} M(K; \rho, \varepsilon).$$

Proof: (i). $\{\theta_j\}_{j=1}^M$ is 2ε-packing

$\{\psi_j\}_{j=1}^N$ is ε-covering.

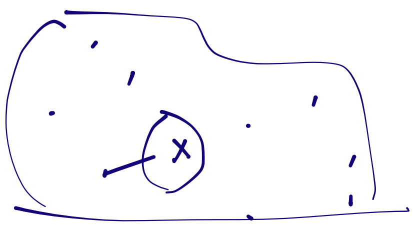


For each ball $\mathbb{B}_\rho(\psi_j; \varepsilon)$ contains at most one of θ's.

$\forall\ \theta_i,\ \exists\ \psi_j$ s.t. $\theta_i \in \mathbb{B}_\rho(\psi_j, \varepsilon)$

$M \le N$.

(ii). $\{\theta_j\}_{j=1}^{M}$ be a maximal ε-packing.



$\forall \theta \in K,\ \exists j \in [M]$ s.t. $\rho(\theta; \theta_j) \le \varepsilon$.

So $\{\theta_j\}_{j=1}^{M}$ is also a covering.

$N(K; \rho, \varepsilon) \le M(K; \rho, \varepsilon)$.

Prop/ (Volumn ratio argument).

$\mathbb{B} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le 1\}$

$N(\mathbb{B}^d; \|\cdot\|_2, \varepsilon)$?

$\rho = \|\cdot\|_2$,

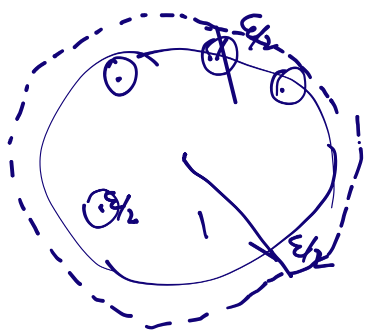
— $\mathbb{B} \subseteq \bigcup_{i=1}^{N} \mathbb{B}(\theta_i; \varepsilon)$.

$$\mathrm{Vol}(\mathbb{B}) \leq \sum_{i=1}^{N} \mathrm{Vol}(\mathbb{B}(\theta_i; \varepsilon)).$$

$$C_d \cdot 1^d \leq N \cdot C_d \cdot \varepsilon^d$$

$$N \geq \left(\frac{1}{\varepsilon}\right)^d.$$

— $N \leq M(K; \rho, \varepsilon)$.



$$\mathrm{Vol}(\mathbb{B}(0, 1+\varepsilon/2))$$

$$\geq \sum_{i=1}^{M} \mathrm{Vol}(\mathbb{B}(\theta_i; \varepsilon/2)).$$

$$\left(1+\frac{\varepsilon}{2}\right)^d \geq M \cdot \left(\frac{\varepsilon}{2}\right)^d.$$

$$\left(\frac{1}{\varepsilon}\right)^d \leq N \leq \left(1+\frac{2}{\varepsilon}\right)^d.$$