STA 3000F Lecture 9

Recap $\quad \hat{\theta}_n = \arg\min \{ P_n f_\theta \}$

$$\|\hat{\theta}_n - \theta^*\|^2 \le P f_{\hat{\theta}_n} - P f_{\theta^*} \le \left| (P_n - P)(f_{\hat{\theta}_n} - f_{\theta^*}) \right|$$

(assumption) $\qquad$ (optimality of $\hat{\theta}_n$)

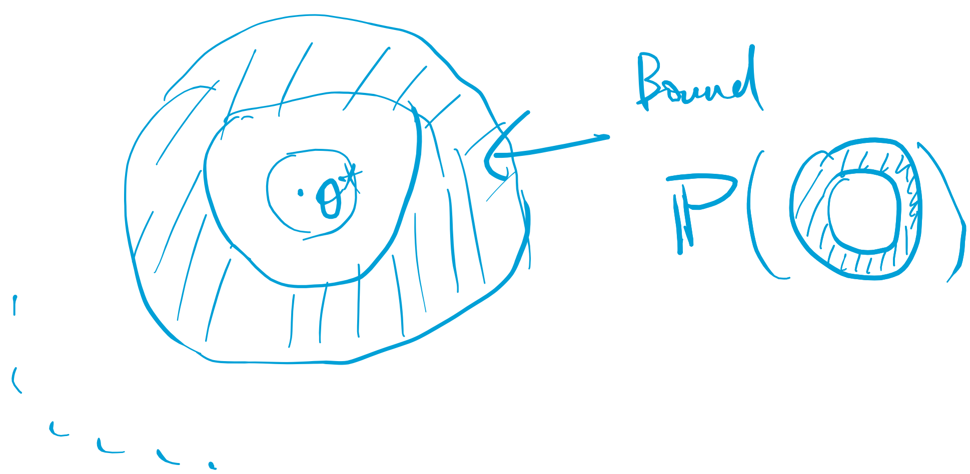
Suppose we have

$$\mathbb{E}\left[\sup_{\substack{\theta \in \Theta \\ \|\theta - \theta^*\| \le r}} \left| (P_n - P)(f_\theta - f_{\theta^*}) \right| \right] \le \phi_n(r).$$

then the rate thm tells us that convergence rate is given by δ_n

where $\delta_n^2 = \phi_n(\delta_n)$.

(Proof idea:



)

Application:

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \le M(x)\cdot\|\theta_1 - \theta_2\|,$$

$F(\theta) - F(\theta^*) \ge |\theta-\theta^*|^2.$

$\theta \in \mathbb{R}^d$. $\quad \mathbb{E}[M(x)^2] < +\infty.$

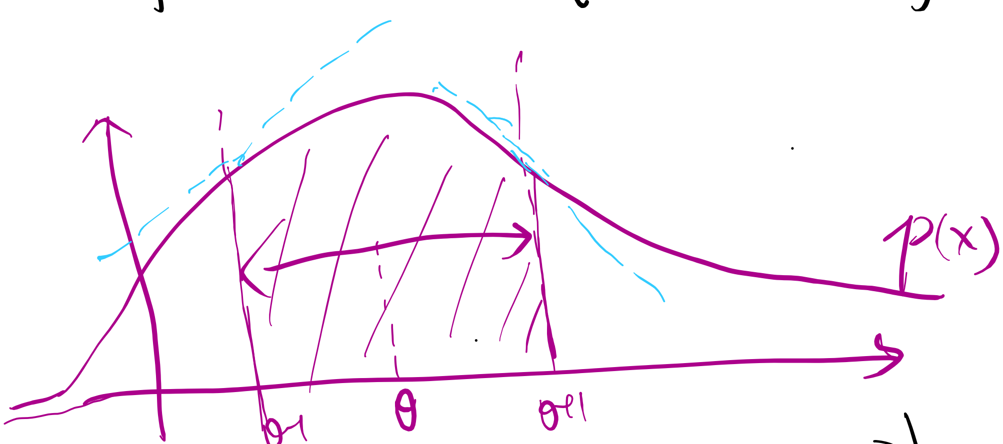
Rate thm implies that

$$\|\hat{\theta}_n - \theta^*\| \le c\cdot\sqrt{\frac{d}{n}}\cdot\|M\|_{L^2(P)} \quad \text{w.h.p.}$$

Rmk: dimension dependence can be suboptimal (from $\|M\|_{L^2(P)}$)

"non-regular" M-estimator. $(x, \theta \in \mathbb{R})$.

$$f_\theta(x) = -\mathbb{1}\{x \in [\theta-1, \theta+1]\}$$



$$P f_\theta = -\mathbb{P}(x \in [\theta-1, \theta+1])$$

$$P_n f_\theta = -\frac{\#\text{ data } \in [\theta-1, \theta+1]}{n}.$$

$$\frac{d^2}{d\theta^2} Pf_\theta \Big|_{\theta=\theta^*} = p'(\theta^*-1) - p'(\theta^*+1) > 0.$$

(assume)

(assume p is cts differentiable, so $\frac{d^2}{d\theta^2} Pf_\theta$ is cts in θ).

So $\exists\, \delta_0, c_0 > 0$, s.t.

$$Pf_\theta - Pf_{\theta^*} \geq c_0 |\theta - \theta^*|^2$$

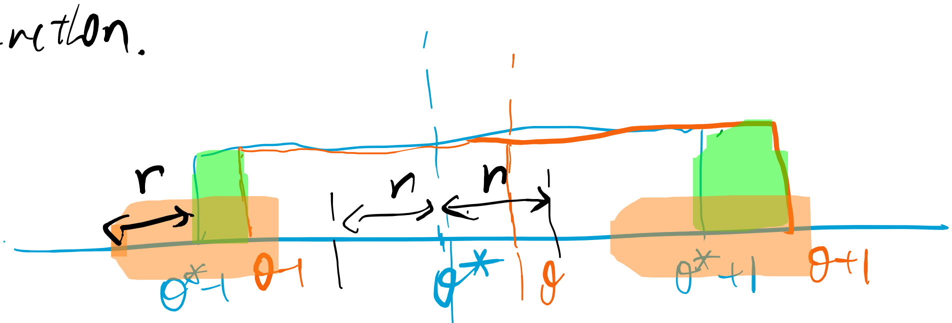
$$\forall \theta \in (\theta^* - \delta_0, \theta^* + \delta_0).$$

By consistency, we can focus on this interval when n is large.

It remains to study

$$\mathbb{E}\left[\sup_{|\theta-\theta^*| \leq r} \left| (P_n - P)(f_\theta - f_{\theta^*}) \right| \right].$$

· Envelope function.



$$G(x) := \mathbb{1}\left\{ x \in [\theta^* + 1 - r, \theta^* + 1 + r] \text{ or } x \in [\theta^* - 1 - r, \theta^* - 1 + r] \right\}$$

$$\mathbb{E}\left[\sup |\cdots|\right]$$

$$\leq C \cdot \sqrt{\frac{\mathbb{E}[G(X)^2]}{n}} \int_0^1 \sqrt{\log \sup_Q N(\delta \cdot \|G\|_{L^2(Q)}; \mathcal{F}_r, \|\cdot\|_{L^2(Q)})}\, d\delta$$

where $\mathcal{F}_r = \{ f_\theta - f_{\theta^*} : |\theta^* - \theta| \leq r \}$.

- $\mathbb{E}[G(X)^2] = \mathbb{P}(X \in [\theta^* - 1 - r, \theta^* - 1 + r] \text{ or } X \in [\theta^* + 1 - r, \theta^* + 1 + r])$

  $\leq 4 p_{max} r.$

- $VC(\text{interval}) \leq 2 \Rightarrow VC(\mathcal{F}_r) \leq 8.$

  $$\log \sup_Q N(\delta \|G\|_{L^2(Q)}; \mathcal{F}_r, \|\cdot\|_{L^2(Q)}) \leq C' \cdot \log(1/\delta).$$

Substituting back, we have

$$\mathbb{E}\left[\sup_{|\theta - \theta^*| \leq r} |(P_n - P)(f_\theta - f_{\theta^*})|\right] \leq C \sqrt{\frac{p_{max} \cdot r}{n}}.$$

Fixed-pt eq:

$$\left(p'(\theta^*-1) - p'(\theta^*+1)\right) \cdot r^2 = c \cdot \sqrt{\frac{p_{max}\, r}{n}}.$$

By rate thm, we have

$$\left|\hat{\theta}_n - \theta^*\right| \leq c \cdot \left\{\frac{p_{max}}{n \cdot \left(p'(\theta^*-1) - p'(\theta^*+1)\right)^2}\right\}^{1/3}.$$

Indeed, we can show

$$n^{1/3}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{d} \text{something}.$$

"something" is not Gaussian,

it is a functional of Gaussian process

$$\underset{t \in \mathbb{R}}{\operatorname{argmax}} \left\{ G(t) - ct^2 \right\}$$

where G is a GP.

Basic idea behind this:

$$\sup_{f \in \mathcal{F}} \left|(\mathbb{P}_n - P) f\right| \xrightarrow{P} 0 \qquad (\text{ULLN})$$

we also have convergence rates.

How about UCLT?

$$\left(\sqrt{n}\,(\mathbb{P}_n - P) f_\theta\right)_{\theta \in \Theta} \xrightarrow{d} \text{something?}$$

We know that for $\theta_1, \theta_2, \cdots \theta_K \in \Theta$

$$\left(\sqrt{n}\,(\mathbb{P}_n - P) f_{\theta_k}\right)_{k=1}^{K} \xrightarrow{d} \text{multivariate normal.}$$

So we can define GP:

- mean $\quad \mathbb{E}[G(\theta)] = 0 \quad \forall \theta \in \Theta$
- cov $\quad \mathbb{E}\left[G(\theta_1) \cdot G(\theta_2)\right] = \mathrm{cov}\left(f_{\theta_1}(X), f_{\theta_2}(X)\right)$
  $(\forall \theta_1, \theta_2 \in \Theta)$.

Additional technical ingredient:
"stochastic equicontinuity".

For a seq of processes $(X_n(t))_{t \in T}$
$n = 1, 2, \cdots$

$$\lim_{n\to+\infty} \mathbb{E}\left[\sup_{\substack{|s-t|\le\eta \\ s,t\in T}} |X_n(s)-X_n(t)|\right] \longrightarrow 0$$

(as $\eta \to 0$).

(can be verified using empirical process tools we discussed).

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \{ P_n f_\theta \}$$

$$= \arg\min_{\theta\in\Theta} \{ \underbrace{P f_\theta}_{\text{Locally Looks like quadratic}} + \underbrace{(P_n - P) f_\theta}_{\xrightarrow{d} \text{GP (after rescale)}} \}$$

We get results of the form

$$n^\alpha(\hat{\theta}_n - \theta^*) \xrightarrow{d} \arg\min_h \{ \tfrac{1}{2} h^T A h + GP(h) \}$$

- Regular case, GP is degenerate (finite rank)

$$h \longmapsto g^T h \quad \text{where} \quad g \sim N(0, \Sigma).$$

(need weaker assumption than the classical Taylor expansion method)

- Irregular case, GP can be Brownian motion / bridge / ....

Bayesian methods.

Given prior π, $X_1, X_2, \dots X_n \overset{iid}{\sim} \mathbb{P}_{\theta^*}$

$$\Pi(\theta \mid X_1, X_2 \dots X_n) = \frac{\pi(\theta) \cdot \prod_{i=1}^{n} \mathbb{P}_\theta(X_i)}{\int \pi(\theta') \cdot \prod_{i=1}^{n} \mathbb{P}_{\theta'}(X_i)\, d\theta'}$$

- Consistency $\forall \varepsilon > 0$,

$$\Pi(\theta : |\theta - \theta^*| > \varepsilon \mid X_1^n) \xrightarrow{\mathbb{P}} 0.$$

- Contraction rate, w.h.p.

$$\Pi(\theta : |\theta - \theta^*| \geq \varepsilon_n \mid X_1^n) \leq \delta$$

for ε_n decaying at certain rate

- Asymptotic posterior

$$d_{TV}(\Pi(\cdot \mid X_1^n), \text{something}) \xrightarrow{\mathbb{P}} 0$$

Thm (Schwartz).

Suppose (i) $\forall \varepsilon > 0$, $\pi(\theta : D_{KL}(P_{\theta^*} \| P_\theta) \leq \varepsilon) > 0$.

(ii) $\forall \varepsilon > 0$, $\exists$ sequence of tests $\{\phi_n\}_{n \geq 1}$

s.t. $\sup\limits_{\|\theta - \theta^*\| \geq \varepsilon} \mathbb{E}_\theta[1 - \phi_n(X_1^n)] \to 0$

$$\mathbb{E}_{\theta^*}[\phi_n(X_1^n)] \to 0.$$

Then we have posterior consistency

Rmk:

- $(\phi_n)_{n \geq 1}$ is only a theory device.
  alg does not need to know the test.
- With more careful arg. this idea can be turned into sharp convergence rates.

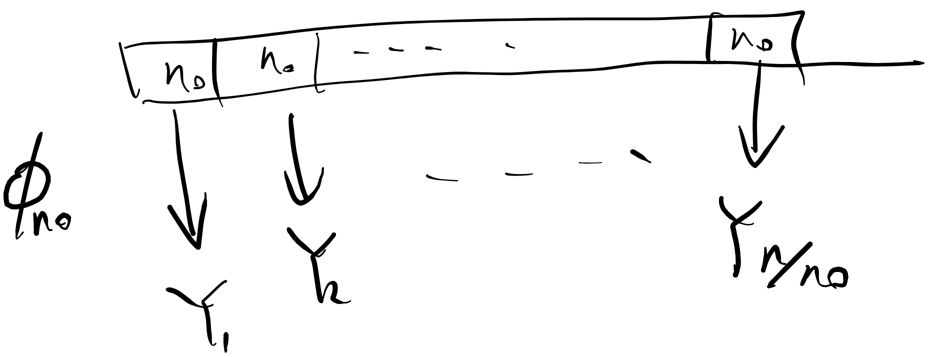
Step I. Bound the error prob.

By assumption, $\exists\ n_0 > 0$, s.t.

$$\mathbb{P}_{\theta^*}(\text{type } I \text{ err for } \phi_{n_0}) \leq \frac{1}{4}.$$

$$\sup_{|\theta-\theta^*|>\varepsilon} \mathbb{P}_\theta\left(\text{type II err for } \phi_{n_0}\right) \le \frac{1}{4}.$$

For large n, divide data into blocks
n/n_0 blocks.



Y_i are iid r.v. (binary-valued).

$$\tilde{\phi}_n(X_1^n) := \mathbb{1}\left\{\frac{n_0}{n}\sum_{1}^{n/n_0} Y_i > \frac{1}{2}\right\}.$$

(majority vote).

By Hoeffding bound, $$\mathbb{E}_{\theta^*}\left[\tilde{\phi}_n(X_1^n)\right] \le \exp\left(-\frac{cn}{n_0}\right)$$

$$\sup_{|\theta-\theta^*|>\varepsilon} \mathbb{E}_\theta\left[1-\tilde{\phi}_n(X_1^n)\right] \le \exp\left(-\frac{cn}{n_0}\right).$$

(n_0 is a const, $n\to+\infty$, exponential decay rate)

Step II. $U = \{\theta : \|\theta - \theta^*\| \le \varepsilon\}$

$$\pi(U^c \mid X_1^n) = \phi_n + (1-\phi_n) \frac{\int_{U^c} \prod_{i=1}^n p_\theta(X_i) / p_{\theta^*}(X_i)\, \pi(\theta)\, d\theta}{\int_{\Theta} \prod_{i=1}^n p_\theta(X_i) / p_{\theta^*}(X_i)\, \pi(\theta)\, d\theta}.$$

$$\mathbb{E}_{\theta^*}\left[(1-\phi_n) \int_{U^c} \prod_{i=1}^n \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \pi(\theta)\, d\theta\right]$$

$$= \int_{U^c} \mathbb{E}_\theta\left[1 - \phi_n(X_1^n)\right] \pi(\theta)\, d\theta.$$

$$\le \sup_{|\theta - \theta^*| > \varepsilon} \mathbb{E}_\theta\left[1 - \phi_n(X_1^n)\right].$$

Step III. For any subset $\Theta_0 \subseteq \Theta$

need to lower bound

$$\log \int_{\Theta} \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i) \cdot \pi(\theta)\, d\theta$$

$$\left(\ge \int_{\Theta} \left(\sum_{i=1}^n \log \frac{p_\theta}{p_{\theta^*}}(X_i) \right) \cdot \pi(\theta)\, d\theta \right)$$

$$\sum_{i=1}^{n} \log \frac{p_\theta}{p_{\theta^*}}(X_i) \approx n \cdot \mathbb{E}_{\theta^*}\left[\log \frac{p_\theta}{p_{\theta^*}}(X_i)\right]$$

$$= -n \cdot D_{KL}(P_{\theta^*} \| P_\theta).$$

$$\geq \log\left(\int_{\Theta_0} \prod_{i=1}^{n} \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \pi(\theta)\, d\theta\right).$$

$$= \log \pi(\Theta_0) + \log\left(\int_{\Theta_0} \prod_{i=1}^{n} \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \cdot \frac{\pi(\theta)}{\pi(\Theta_0)}\, d\theta\right).$$

$$\geq \log \pi(\Theta_0) + \int_{\Theta_0} \sum_{i=1}^{n} \log \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \cdot \frac{\pi(\theta)}{\pi(\Theta_0)}\, d\theta$$

iid sum.

$$\frac{1}{n} \sum_{i=1}^{n} \int_{\Theta_0} \log \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \cdot \frac{\pi(\theta)}{\pi(\Theta_0)}\, d\theta \xrightarrow{P} -D_{KL}(P_{\theta^*} \| P_\theta)$$

$$\geq -\varepsilon.$$

(take $\Theta_0 = \{\theta : D_{KL}(P_{\theta^*} \| P_\theta) \leq \varepsilon\}$)

$$\mathbb{P}\left(\int_{\Theta_0} \sum_{i=1}^{n} \log \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \cdot \frac{\pi(\theta)}{\pi(\Theta_0)}\, d\theta \leq -2n\varepsilon\right) \to 0.$$

$(n \to +\infty)$

So, w.p. $\to 1$,

denominator $\geq \exp(-2n\varepsilon) \cdot \pi(\Theta_0)$

Putting them together, w.h.p.

$$\frac{\text{numerator}}{\text{denominator}} \leq \frac{\exp\left(-c\frac{n}{n_0}\right)}{\pi(\Theta_0)\cdot \exp(-2n\varepsilon)} \longrightarrow 0.$$

(choose ε small enough)

Existence of tests.

- Easy to construct 1 vs 1 test (θ^* vs θ).
- Use union bound to bound failure prob of 1 vs. all test. (along with discretization).

Bernstein-von-Mises thm.

Assuming sufficient regularity (same regularity as needed by MLE asymptotic normality).

$$d_{TV}\left(\pi(\cdot \mid X_1^n),\ \mathcal{N}\left(\hat{\theta}_n,\ n^{-1} I(\theta^*)^{-1}\right)\right) \xrightarrow{P} 0.$$

($\hat{\theta}_n$ is the MLE).

Proof idea: locally expand the posterior density.