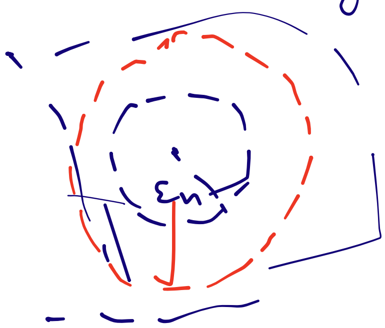
STA3000 Lecture 9

Posterior distribution (given prior π, data $X_1, X_2, \dots, X_n$)

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

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

- Consistency: $\forall \varepsilon > 0$, $\Pi_n(\theta : \|\theta - \theta^*\| > \varepsilon \mid X_1^n) \xrightarrow{P} 0$.
- Contraction rate $(\varepsilon_n)_{n \geq 0}$: $\Pi_n(\theta : \|\theta - \theta^*\| > M_n \varepsilon_n \mid X_1^n) \xrightarrow{P} 0$ (for any seq $M_n \to +\infty$).
- Asymptotic posterior: $d_{TV}(\Pi_n(\cdot \mid X_1^n), ??) \xrightarrow{P} 0$



Thm (L. Schwartz).

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

(ii) $\forall \varepsilon > 0$, $\exists \phi_n$ s.t.

$$\sup_{\|\theta - \theta^*\| \geq \varepsilon} \mathbb{E}_\theta[1 - \phi_n] \to 0$$

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

Then posterior consistency holds true.

Rmk.

- Can be made non-asym, w/ rate of convergence.
- In many cases. N-P + covering + union bound

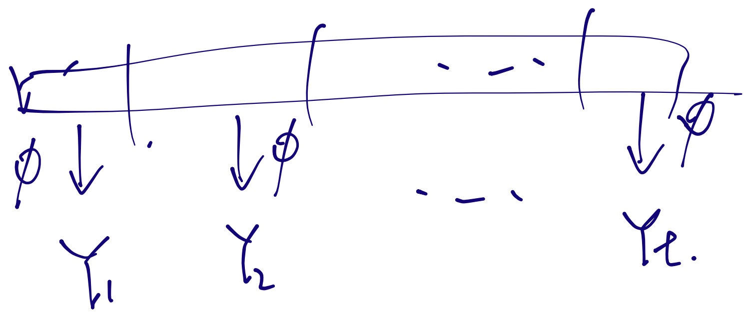
Proof – Step I. Boost the error prob.

n_0 ~~s.t.~~ $\mathbb{P}_{\theta^*}(\phi_{n_0}=1) < \frac{1}{4}$

$\mathbb{P}_{\theta}(\phi_{n_0}=0) < \frac{1}{4}$ for $\|\theta-\theta^*\| > \varepsilon$.

$\forall n$. Let $\ell = \lfloor n/n_0 \rfloor$ divide data into ℓ subgroups.

$Y_1, Y_2, \dots, Y_\ell$ test decisions from each group.



$$\tilde{\phi}_n := \mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^{\ell} Y_i > \frac{1}{2}\right\}.$$

Fact. $\mathbb{E}_{\theta^*}[\tilde{\phi}_n]$, $\sup_{\|\theta-\theta^*\|>\varepsilon} \mathbb{E}[1-\tilde{\phi}_n] \le e^{-cn}$ $\left(c = \frac{c_0}{n_0}\right)$

(for some $c>0$).

$Y_1, \dots, Y_\ell$ iid Bernoulli(p). use Hoeffding bound.

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

"magic"

$$\Pi_n(U^c \mid X_1^n) \le \underbrace{\phi_n}_{\xrightarrow{P} 0 \text{ under } \mathbb{P}_{\theta^*}} + \frac{(1-\phi_n) \int_{U^c} \prod_{i=1}^n p_\theta(X_i)/p_{\theta^*}(X_i)\, \pi(\theta)\, \cancel{d\theta}}{\int_{\Theta} \prod_{i=1}^n p_\theta(X_i)/p_{\theta^*}(X_i)\, \pi(\theta)\, d\theta}$$

Later

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

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

$$\le \sup_{\theta \in U^c} \mathbb{E}_\theta[1-\phi_n].$$

Step III. Subset $\Theta_0 \subseteq \Theta$, $\pi_0(\theta) := \frac{\pi(\theta)}{\pi(\Theta_0)}$.

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

$$\ge \log \pi(\Theta_0) + \log \int_{\Theta_0} \prod_{i=1}^n \frac{p}{p_{\theta^*}}(X_i)\, \pi_0(\theta)\, d\theta.$$

iid sum,

$$\overset{\text{(Jensen)}}{\ge} \log \pi(\Theta_0) + \int_{\Theta_0} \sum_{i=1}^n \left(\log p_\theta(X_i) - \log p_{\theta^*}(X_i)\right) \pi_0(\theta)\, d\theta.$$

$$\frac{1}{n}\int_{\Theta_0} \sum_{i=1}^{n} \left(\log p_\theta(X_i) - \log p_{\theta^*}(X_i)\right) \pi_0(\theta)\,d\theta.$$

$$\xrightarrow{P} \mathbb{E}_{\theta^*}\left[\int_{\Theta_0} \log p_\theta(X)/p_{\theta^*}(X) \quad \pi_0(\theta)\,d\theta\right].$$

$$= -\int_{\Theta_0} D_{KL}\left(p_{\theta^*} \| p_\theta\right) \pi_0(\theta)\,d\theta.$$

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

($\pi(\Theta_0) > 0$ by assumption).

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

$(*)$.

Putting them together

$$\mathbb{P}\left((1-\phi_n)\frac{\int_{U^c} \prod_{i=1}^{n} \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \pi(\theta)\,d\theta}{\int_{\Theta} \prod_{i=1}^{n} \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \pi(\theta)\,d\theta} > \Delta\right).$$

$$\leq \mathbb{P}\left((1-\phi_n)\int_{U^c} \prod_{i=1}^{n} \frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \pi(\theta)\,d\theta > \Delta \cdot \pi(\Theta_0) \cdot e^{-\frac{n\varepsilon}{2}}\right)$$

$$+ \mathbb{P}(*).$$

Markov ineq.

$$\leq \frac{\sup_{\theta \in U^c} \mathbb{E}_\theta[1-\phi_n]}{\Delta \cdot \pi(\Theta_0) \cdot e^{-\frac{n\varepsilon}{2}}}.$$

Asymptotic posterior (Bernstein – von Mises).

Assuming enough regularity.

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\theta^*)^{-1} \nabla \log p_{\theta^*}(X_i)$$

$$\left(\approx \sqrt{n}\left(\hat{\theta}_n - \theta^*\right), \quad \hat{\theta}_n \text{ MLE} \right).$$

$$d_{TV}\left(\mathcal{L}\left(\sqrt{n}(\theta - \theta^*) \,\middle|\, X_1^n\right), \mathcal{N}\left(\Delta_n, I(\theta^*)^{-1}\right)\right) \xrightarrow{P} 0$$

Nonparametric estimation.

- Density estimation $X_1, X_2, \dots, X_n \overset{iid}{\sim} p^* \in \mathcal{P}$.

  estimate p^*

- Nonpar regression. $(X_i, Y_i) \overset{iid}{\sim}$

  $$Y_i = f^*(X_i) + \varepsilon_i \qquad \mathbb{E}[\varepsilon_i \mid X_i] = 0.$$

  $$f^* \in \mathcal{F}.$$

  (fixed design: X_i are deterministic, e.g. $X_i = i/n$)

  (Random design: X_i iid r.v.).

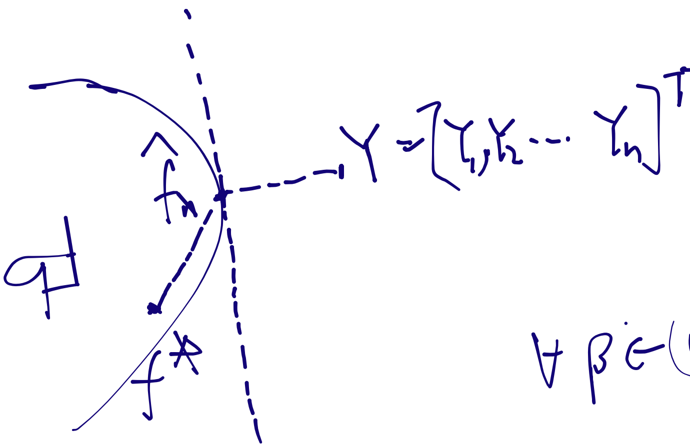
Fixed design iid $\mathcal{N}(0, 1)$ noise.

Natural choice.

(MLE / least sq).

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \right\}.$$

(Rmk: computationally feasible if $\mathcal{F}$ is convex).



$\forall \beta \in (0,1)$.

$$\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{f}_n(X_i)\right)^2$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \beta f^*(X_i) - (1-\beta) \hat{f}_n(X_i)\right)^2.$$

Take $\beta \to 0$

$$(*). \quad \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{f}_n(X_i)\right) \cdot \left(f^*(X_i) - \hat{f}_n(X_i)\right) \leq 0.$$

Define $\hat{\Delta}_n := \hat{f}_n - f^*$.

$\|f\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} f(X_i)^2$. $\quad \mathcal{F}^* = \{f - f^*, f \in \mathcal{F}\}$.

(*) can be re-written as

$$\|\hat{\Delta}_n\|_n^2 \leq \frac{1}{n}\sum_{i=1}^n \varepsilon_i \hat{\Delta}_n(x_i).$$

$$\left(\leq \sup_{\substack{\|h\|_n \leq \|\hat{\Delta}_n\|_n \\ h \in \mathcal{F}^*}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(x_i) \right).$$

Define $\quad \mathcal{G}_n(r) := \mathbb{E}\left[\sup_{\substack{\|h\|_n \leq r \\ h \in \mathcal{F}^*}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(x_i)\right].$

where $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \overset{iid}{\sim} \mathcal{N}(0,\nu)$.

Thm:— Suppose $\mathcal{G}_n(r) \leq \phi_n(r)$

$\boxed{\text{s.t. } \phi_n(cr) \leq c^\alpha \phi_n(r) \text{ for some } \alpha < 2.}$

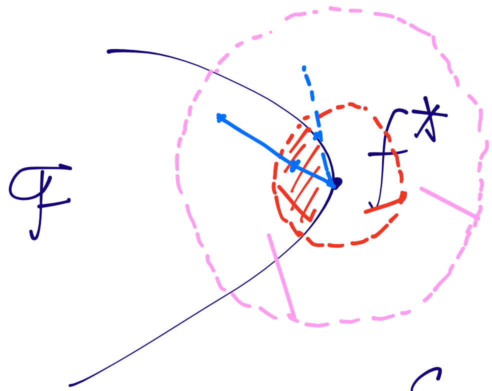
If δ_n solves $\delta_n^2 = \phi_n(\delta_n)$

then $\forall \varepsilon > 0$, $\exists\, C_\varepsilon > 0$, s.t.

$$\|\hat{f}_n - f^*\|_n \lesssim_\varepsilon \delta_n \quad \text{w.p. } 1-\varepsilon.$$

- Proof: See Lecture 7. Annulus decomposition.

- $\boxed{\dots\dots}$ condition is automatically satisfied for convex $\mathcal{F}$ with $\alpha = 1$.



$f \in \mathcal{F}^* \cap \mathbb{B}(c \cdot r)$

(convexity) $(c > 1)$.

$f/c \in \mathcal{F}^* \cap \mathbb{B}(r)$.

So $\quad \mathcal{G}(r) \geq \dfrac{\mathcal{G}(cr)}{c}$

How to bound Gaussian complexity.

Thm.

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(X_i)\right|\right] \leq \frac{c}{\sqrt{n}} \int_0^{\mathrm{diam}_n(\mathcal{H})} \sqrt{\log N(\delta; \mathcal{H}, \|\cdot\|_n)}\, d\delta.$$

(Proof: see Lecture 6).

Thm for any $\delta_0 \in (0, \mathrm{diam}_n(\mathcal{H}))$,

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(X_i)\right|\right] \leq c\delta_0 + \frac{c}{\sqrt{n}} \int_{\delta_0}^{\mathrm{diam}_n(\mathcal{H})} \sqrt{\log N(\delta; \mathcal{H}, \|\cdot\|_n)}\, d\delta.$$

When applied to local Gaussian complexity $\mathcal{G}_n(r)$

$\mathrm{diam}_n(\mathcal{H}) \leq 2r$.

Concrete examples:

- $\mathcal{F} = \{ f : [0,1] \to [0,N],\ |f(x) - f(y)| \leq |x-y|,\ \forall x, y \in [0,N] \}$.

Want :— $\delta_n^2 = \frac{1}{\sqrt{n}} \int_0^{\delta_n} \sqrt{\log N(\delta;\ \mathcal{F}^* \cap B(\delta_n), \|\cdot\|_n)}\, d\delta$.

$$N(\delta;\ \mathcal{F}^* \cap \mathbb{B}_n(u), \|\cdot\|_n)$$

$$\leq N(\delta;\ \mathcal{F}^*, \|\cdot\|_n).$$

$$\boxed{\mathcal{F}' = \{ f : [0,N] \to [-1,1],\ |f(x) - f(y)| \leq 2|x-y|,\ x, y \in [0,N] \}}$$

$$\leq N(\delta;\ \mathcal{F}', \|\cdot\|_n).$$

$$\begin{pmatrix}\text{Rudelson} \\ \text{-Vershynin}\end{pmatrix} \leq \exp\left(c_1 \operatorname{fat}_{c_2\delta}(\mathcal{F}')\right).$$

Fat-shattering dim of $\mathcal{F}'$.

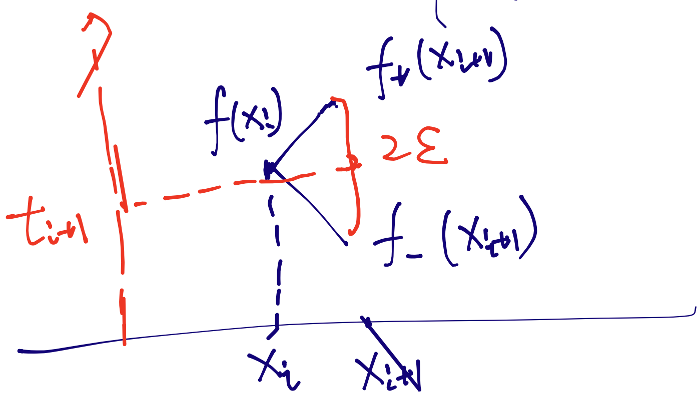
Suppose: $(x_1, t_1), (x_2, t_2), \dots, (x_m, t_m)$

shattered at scale ε

(WLOG $0 \leq \cancel{x_1} < x_2 < \dots < x_m \leq 1$).

$\forall$ binary seq $z \in \{-1,1\}^m$, $\exists f_z \in \mathcal{F}'$

s.t. $f_z(x_i) \begin{cases} \geq t_i + \varepsilon & z_i = 1 \\ \leq t_i - \varepsilon & z_i = -1 . \end{cases}$



By Lip. $|f_+(x_{i+1}) - f(x_i)| \leq 2|x_{i+1} - x_i|$.

$|f_-(x_{i+1}) - f(x_i)| \leq 2|x_{i+1} - x_i|$.

$$|f_+(x_{i+1}) - f_-(x_{i+1})| \geq 2\varepsilon.$$

$$x_{i+1} - x_i \geq \frac{\varepsilon}{2}.$$

So $m < \lceil \frac{2}{\varepsilon} \rceil$.

So $\mathrm{fat}_\varepsilon(\mathcal{F}') \lesssim \frac{1}{\varepsilon}$

$$\int_0^r \sqrt{\log N(\cdots)}\, d\delta$$

$$\leq \int_0^r 1/\sqrt{\delta}\, d\delta \lesssim \sqrt{r}.$$

$$r^2 = \sqrt{\frac{r}{n}} \implies r \asymp n^{-1/3}.$$

In general, β-th order Hölder class.

· $\beta > 0$. $\quad \beta = k + \gamma \quad \begin{matrix} k \in \mathbb{N} \\ \gamma \in [0,1) \end{matrix}$

$$\Sigma(\beta, L) = \left\{ f : [0,1]^d \to [0,1], \ \begin{matrix} \forall \text{ multi-index } \alpha \\ \text{s.t. } |\alpha| = k \\ |\partial^\alpha f(x) - \partial^\alpha f(y)| \leq L \|x-y\|^\gamma \\ \forall x, y \in [0,1]^d \end{matrix} \right\}$$

· If β is integer, β-th order cts diff (w/ uniform bound on derivatives).

$$\boxed{\text{Thm} \quad N(\varepsilon; \Sigma(\beta, 1), \|\cdot\|_{L^2(Q)}) \leq \exp\left(\left(\frac{1}{\varepsilon}\right)^{d/\beta}\right).}$$

Proof in 1-D: $\qquad \delta = \varepsilon^{1/\beta}$

$\vdash \delta \mid \delta \mid \cdots \delta \dashv$ $\qquad \delta$ grid on $[0,1]$

$x_j = j\delta.$

$$A : f \mapsto \left(\left\lfloor \frac{\partial^k f(x_j)}{\delta^{\beta-k}} \right\rfloor \right) \quad \begin{matrix} 0 \leq k \leq \lfloor \beta \rfloor \\ j \in \{0, 1, 2, \ldots, \lceil 1/\delta \rceil\}. \end{matrix}$$

Claim: If $\mathcal{A}f = \mathcal{A}g$

then $\|f - g\|_\infty \leq \varepsilon$.

Cardinality of $\mathcal{A}\mathcal{F}$.

- Product of range of each coordinate.

$$\log N \leq \left(\frac{1}{\varepsilon}\right)^{1/\beta} \underline{\log(1/\varepsilon)}$$

- Clever bound.

  # values a new coordinate can take by knowing the previous ones.