3 – Asymptotic Statistics

– Supervised learning setup:

– (Y, X) pairs $\sim p(Y, X)$ joint distribution

$y \in \mathcal{Y}$ is response, $x \in \mathcal{X}$ is inputs (feature)

$\in \mathbb{R}$ or $\{0,1\}$ or $\{\pm 1\}$, $\in \mathbb{R}^d$

– Observe data: $(Y_i, X_i) \overset{iid}{\sim} p(Y, X)$ for $i = 1, 2, \ldots, n$.

Goal: Find a function $f : \mathcal{X} \to \mathcal{Y}$ s.t.

$$Y \approx f(X)$$

using data.

! This is the focus of next six lectures !

– Need to measure the quality of the function f.

* Loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$

– measures distance between Y and $f(X)$

– Ex: $\ell(Y, f(X)) = (Y - f(X))^2$ or $|Y - f(X)|$

* Risk function: $R : \mathcal{F} \to \mathbb{R}_+$ for $\mathcal{F} = \{f\}$.

– measures expected loss of function f.

– $R(f) = \mathbb{E}[\ell(Y, f(X))]$.

$(Y, X) \sim p(Y, X)$

– Depends on ℓ, p, f.

Goal (revised): Find $f \in \mathcal{F}$ s.t. $R(f)$ is small.

* Bias – Variance Decomposition

$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$

– Squared error loss: $\ell(a, b) = (a - b)^2$

$\ell(Y, f(X)) = (Y - f(X))^2$

– Risk: $R(f) = \mathbb{E}[(Y - f(X))^2] = \mathbb{E}\left[\mathbb{E}[(Y - f(X))^2 | X]\right]$

$\pm \mathbb{E}[Y|X]$

$$R(f) = \mathbb{E}\Big[\mathbb{E}\big[\underbrace{(Y - \mathbb{E}[Y|x]}_{\text{I}} + \underbrace{\mathbb{E}[Y|x] - f(x)}_{\text{II}})^2 \,|x\big]\Big]$$

$$= \mathbb{E}\Big[\mathbb{E}\big[\underbrace{(Y-\mathbb{E}[Y|x])^2}_{\text{I}^2} + \underbrace{(\mathbb{E}[Y|x]-f(x))^2}_{\text{II}^2} + \underbrace{2\cdot(Y-\mathbb{E}(y|x))(\mathbb{E}[Y|x]-f(x))}_{2\,\text{I}\cdot\text{II}} \,|x\big]\Big]$$

$$= \mathbb{E}\Big[\mathbb{E}\big[(Y-\mathbb{E}[Y|x])^2|x\big] + \mathbb{E}\big[(\mathbb{E}[Y|x]-f(x))^2|x\big] + 2\,\mathbb{E}\big[(Y-\mathbb{E}(y|x))(\mathbb{E}[Y|x]-f(x))\,|x\big]\Big]$$

($\mathbb{E}[Y|x]$ and $f(x)$ are fncs of x.)

$$= 2\,\big(\mathbb{E}[Y|x] - \mathbb{E}[Y|x]\big)\big(\mathbb{E}[Y|x]-f(x)\big)$$

$$= 0. \qquad (\mathbb{E}[Y|x] - \mathbb{E}[Y|x] = 0)$$

$$= \mathbb{E}\Big[\mathbb{E}\big[(Y-\mathbb{E}[Y|x])^2|x\big] + \mathbb{E}\big[(\mathbb{E}[Y|x]-f(x))^2|x\big]\Big]$$

$$= \mathbb{E}\big[(Y-\mathbb{E}[Y|x])^2\big] + \mathbb{E}\big[(\mathbb{E}[Y|x]-f(x))^2\big]$$

$\mathbb{E}[\mathrm{Var}(Y|x)]$ — irreducible

Bias^2 — minimized at $f(x) = \mathbb{E}[Y|x]$

* Parametric Models

- Need to restrict $\mathcal{F}$

  e.g. 2-layer NNs, linear funcs, etc.

- Assume $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ where Θ is the parameter set.

  e.g. $f_\theta(x) = \langle x, \theta \rangle$

$$\ell(y, f_\theta(x)) = (y - \langle x, \theta\rangle)^2 \triangleq \ell((y,x), \theta)$$

(y, x: data; θ: parameter; (y,x): data; θ: parameter)

- Recall: Want $\theta_* = \underset{\theta \in \Theta}{\operatorname{argmin}} R(\theta) = \mathbb{E}\left[\ell((y,x),\theta)\right]$

  - we can't do this, since we don't know P.
  - Can do: Observe data $(y_i, x_i) \overset{iid}{\sim} P \quad i = 1 \dots n$

    Estimate $R(\theta)$ w/ $\hat{R}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell((y_i,x_i),\theta)$

    Find $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \hat{R}(\theta)$

    Hope $\hat{\theta} \approx \theta_*$

$\hat{R}(\hat{\theta})$: training error

$R(\hat{\theta})$: test error

Def (Excess risk): $R(\hat{\theta}) - R(\theta_*)$

Goal (revised+): Quantify how small excess risk is.

* MLE

- Assume $y|x \sim P_{\theta_*}(y|x)$ where $\theta_* \in \Theta$

  ↑ unknown true parameter.

  but we know the distributional form, that is $P_\theta(y|x)$

- Loss: $\ell((y,x),\theta) = -\log P_\theta(y|x)$

  Ex: $y|x \sim N(\langle \theta_*, x\rangle, 1)$ $\quad P_\theta(y|x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y - \langle x,\theta\rangle)^2\right\}$

  $\ell((y,x),\theta) = \frac{1}{2}(y - \langle x,\theta\rangle)^2 + \text{constant}.$

- MLE problem: - Observe data $(x_i, y_i) \sim p(x,y)$
$= p_{\theta_*}(y|x) \cdot p(x)$

- Want θ_*

or $\theta_{**} = \arg\min_\theta R(\theta)$

- Can do $\hat{\theta} = \arg\min_\theta \hat{R}(\theta)$

$= \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} -\log p_\theta(y_i|x_i)$

- θ_* or θ_{**} ?

* $\nabla R(\theta_*) = -\mathbb{E}\left[\nabla \log p_{\theta_*}(y|x)\right] = 0$ (hw1)

* $\nabla^2 R(\theta_*) = -\mathbb{E}\left[\nabla^2 \log p_{\theta_*}(y|x)\right] = I_{\theta_*} \geq 0$

- Can assume $I_{\theta_*} > 0 \Rightarrow \theta_*$ is a min of $R(\theta)$.

- Identifiability condition:

$\theta \neq \theta'$, $p_\theta \neq p_{\theta'} \Rightarrow \theta_*$ is global minimizer $R(\theta)$.

(exercise)

- Asymptotics of MLE

Def (Conv. in prob): $X_n \xrightarrow{p} X$ if $\forall \epsilon > 0$,

$\mathbb{P}(|X_n - X| > \epsilon) \to 0$.

Def (Conv. in dist.): Let F_n and F be the CDFs of X_n and X, respectively.

$X_n \xrightarrow{d} X$ if $\lim_n F_n(x) = F(x)$ for $\forall x$ s.t F is cntns.

Slutsky's thm: Let $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} a$, $Z_n \xrightarrow{P} b$,
then $X_n Y_n + Z_n \xrightarrow{d} aX + b$.

Def (Consistency): $\{\hat{\theta} \equiv \hat{\theta}_n\}_{n \in \mathbb{N}}$ is consistent if $\hat{\theta} \xrightarrow{P} \theta_*$.

Theorem (Asymp of MLE): Assume $\{\hat{\theta}\}_n$ is consistent, and $I_{\theta_*} \succ 0$ and $\|\nabla^3 \log p_\theta\|_{op} < B$, $\forall \theta$.
Then,

1. $\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} \mathcal{N}(0, I_{\theta_*}^{-1})$
2. $n(R(\hat{\theta}) - R(\theta_*)) \xrightarrow{d} \frac{1}{2}\chi^2_d$.

Remarks: 1- $I_{\theta_*} \uparrow \Rightarrow$ less variance $\Rightarrow$ more info on θ_*.

2- Asymptotic efficiency

3- $\mathbb{E}\chi^2_d = d \Rightarrow R(\hat{\theta}) - R(\theta_*) \approx O(d/n)$

(d: dimension, n: num of samples)

generalization error (excess risk)

Tensor notation:

$A \in \mathbb{R}^{n \times n}$, $u \in \mathbb{R}^n$

$Au \in \mathbb{R}^n \longrightarrow A[u]$

$u^T A u \in \mathbb{R} \longrightarrow A[u,u]$

$A \in \mathbb{R}^{n \times n \times n}$, $u \in \mathbb{R}^n$

$A[u] \in \mathbb{R}^{n \times n}$

$A[u,u] \in \mathbb{R}^n$

$A[u,u,u] \in \mathbb{R}$.

↳ tensor

Ex: $f: \mathbb{R}^d \longrightarrow \mathbb{R}$.

$\nabla f \in \mathbb{R}^d$

$\nabla^2 f \in \mathbb{R}^{d \times d}$

$\nabla^3 f \in \mathbb{R}^{d \times d \times d}$.

* proof: $\nabla \hat{R}(\hat{\theta}) = 0 = \nabla R(\theta_*)$

By Taylor's thm,

1.

$$0 = \nabla \hat{R}(\hat{\theta}) = \nabla \hat{R}(\theta_*) + \nabla^2 \hat{R}(\theta_*)(\hat{\theta}-\theta_*) + \frac{1}{2}\nabla^3 \hat{R}(\bar{\theta})[\hat{\theta}-\theta_*, \hat{\theta}-\theta_*]$$

$\bar{\theta} \in (\theta_*, \hat{\theta})$

$$\Rightarrow -\nabla \hat{R}(\theta_*) = \left\{ \nabla^2 \hat{R}(\theta_*) + \frac{1}{2}\nabla^3 \hat{R}(\bar{\theta})[\hat{\theta}-\theta_*] \right\}(\hat{\theta}-\theta_*)$$

$-\nabla \hat{R}(\theta_*) \in \mathbb{R}^d$; $\nabla^2 \hat{R}(\theta_*) \in \mathbb{R}^{d\times d}$; $\nabla^3 \hat{R}(\bar{\theta}) \in \mathbb{R}^{d\times d\times d}$; $[\hat{\theta}-\theta_*] \in \mathbb{R}^d$; $\nabla^3 \hat{R}(\bar{\theta})[\hat{\theta}-\theta_*] \in \mathbb{R}^{d\times d}$; $(\hat{\theta}-\theta_*) \in \mathbb{R}^d$

$$\Rightarrow \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} \nabla \log p_{\theta_*}(y_i|x_i) = \left\{ \frac{1}{n}\sum_{i=1}^{n} -\nabla^2 \log p_{\theta_*}(y_i|x_i) + \frac{1}{2}\nabla^3 R(\bar{\theta})[\hat{\theta}-\theta_*] \right\} \sqrt{n}(\hat{\theta}-\theta_*)$$

- $\frac{1}{n}\sum_{i=1}^{n} \nabla \log p_{\theta_*}(y_i|x_i) \xrightarrow{d}$ by CLT $\; \mathcal{N}(0, \mathrm{Cov}(\nabla \log p_{\theta_*}(y|x)))$, and $\mathrm{Cov}(\nabla \log p_{\theta_*}(y|x)) = I_{\theta_*}$
- $\frac{1}{n}\sum_{i=1}^{n} -\nabla^2 \log p_{\theta_*}(y_i|x_i) \xrightarrow{p}$ (LLN) $\; -\mathbb{E}[\nabla^2 \log p_{\theta_*}(y|x)] = I_{\theta_*}$
- $\frac{1}{2}\nabla^3 R(\bar{\theta}) \le B$
- $[\hat{\theta}-\theta_*] \xrightarrow{p} 0$
- $\sqrt{n}(\hat{\theta}-\theta_*)$ ← want

by Slutsky's thm

$$\Rightarrow \mathcal{N}(0, I_{\theta_*}) = I_{\theta_*} \cdot \sqrt{n}(\hat{\theta}-\theta_*)$$

$$\Rightarrow \sqrt{n}(\hat{\theta}-\theta_*) \xrightarrow{d} \mathcal{N}(0, I_{\theta_*}^{-1}).$$

> $X \sim \mathcal{N}(\mu, \Sigma)$
> $AX \sim \mathcal{N}(A\mu, A\Sigma A^T)$

2. For excess risk:

$$R(\hat\theta) = \underbrace{R(\theta_*)}_{=0} + \langle \nabla R(\theta_*), \hat\theta - \theta_* \rangle + \frac{1}{2}\langle \hat\theta - \theta_*, \nabla^2 R(\theta_*)(\hat\theta - \theta_*)\rangle$$

$$+ \frac{1}{6}\nabla^3 R(\bar\theta)[\hat\theta - \theta_*, \hat\theta - \theta_*, \hat\theta - \theta_*].$$

$$n\left(R(\hat\theta) - R(\theta_*)\right) = \frac{1}{2}\left\langle \sqrt{n}(\hat\theta - \theta_*), \left\{ \underbrace{\nabla^2 R(\theta_*)}_{I_{\theta_*}} + \frac{1}{3}\nabla^3 R(\bar\theta)[\hat\theta - \theta_*] \right\} \sqrt{n}(\hat\theta - \theta_*) \right\rangle$$

$\sqrt{n}(\hat\theta - \theta_*) \xrightarrow{d} N(0, I_{\theta_*}^{-1})$ ← same r.v. → $\sqrt{n}(\hat\theta - \theta_*) \xrightarrow{d} N(0, I_{\theta_*}^{-1})$

$\nabla^3 R(\bar\theta)[\hat\theta - \theta_*] \le B$, $\xrightarrow{?} 0$

by Slutsky's thm:

$$= \frac{1}{2}\left\langle N(0, I_{\theta_*}^{-1}), I_{\theta_*} \cdot N(0, I_{\theta_*}^{-1}) \right\rangle$$

$I_{\theta_*} = I_{\theta_*}^{1/2} I_{\theta_*}^{1/2}$

$$= \frac{1}{2}\left\langle I_{\theta_*}^{1/2} N(0, I_{\theta_*}^{-1}), I_{\theta_*}^{1/2} N(0, I_{\theta_*}^{-1}) \right\rangle$$

$$= \frac{1}{2}\left\langle N(0, I), N(0, I) \right\rangle$$

same r.v.

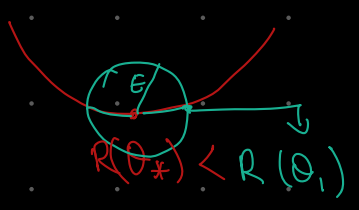
$$\sim \frac{1}{2}\chi_d^2.$$

If $\xi_i \overset{iid}{\sim} N(0,1)$, $\sum_{i=1}^{d} \xi_i^2 \sim \chi_d^2$

∎

– Consistency of MLE

Theorem: Assume 1- uniform convergence $\sup_{\theta \in \Theta} |\hat R(\theta) - R(\theta)| \xrightarrow{P} 0$



2 - identifiability $\inf_{\theta : \|\theta - \theta_*\| \ge \epsilon > 0} R(\theta) > R(\theta_*)$

3- $\mathcal{U}$ is compact, non-empty and ℓ is cnts.

Then, $\hat{\theta} = \underset{\theta \in \mathcal{U}}{\operatorname{argmin}}\, \hat{R}(\theta)$ is consistent for θ_*.

proof: - Item 3 $\Rightarrow \hat{\theta} \in \mathcal{U}$ and $\theta_* \in \mathcal{U}$

- $\hat{\theta}$ minimizes $\hat{R}(\theta)$ in $\mathcal{U}$

$$\Rightarrow \hat{R}(\hat{\theta}) \leq \hat{R}(\theta_*) - R(\theta_*) + R(\theta_*)$$

$$\leq \sup_{\theta \in \mathcal{U}} \left|\hat{R}(\theta) - R(\theta)\right| + R(\theta_*) \xrightarrow{P} R(\theta_*) \quad \text{(by item 1)}$$

$$\sup_{\theta \in \mathcal{U}} \left|\hat{R}(\theta) - R(\theta)\right| \xrightarrow{P} 0$$

$$\Rightarrow \hat{R}(\hat{\theta}) \approx R(\theta_*) \quad \text{as } n \to \infty.$$

$$(*) \quad 0 \leq R(\hat{\theta}) - R(\theta_*) \leq R(\hat{\theta}) - \hat{R}(\hat{\theta}) \quad (\text{when } n \to \infty)$$

$$\leq \sup_{\theta \in \mathcal{U}} \left|\hat{R}(\theta) - R(\theta)\right| \xrightarrow{P} 0 \quad \text{(by item 1)}$$

- Then, compare the events

$$\forall \varepsilon, \exists \delta: \quad \left\{ \|\hat{\theta} - \theta_*\| \geq \epsilon \right\} \overset{\text{by item 2}}{\subseteq} \left\{ R(\hat{\theta}) - R(\theta_*) > \delta \right\}$$

but $\mathbb{P}(\quad) \xrightarrow{P} 0$

by $*$ above. ∎