2.8 General loss functions and random design

Goal: To quantify $L(\hat{\theta}) - L(\theta^*)$ for a general $L(\theta)$.

For an example $z = (x, y)$, $\ell(z, \theta)$ is the loss function and $\theta \in \mathbb{R}^d$.

Denote $\mathcal{Z}$ as the set of all examples. Let $p^* = \Delta(\mathcal{Z})$.

Let $\theta^* \in \mathbb{R}^d$ be the minimal of expected risk:

$$\theta^* := \underset{\theta \in \mathbb{R}^d}{\mathrm{argmax}}\, L(\theta), \quad L(\theta) := E_{z \sim p^*}[\ell(z, \theta)]$$

Let $\hat{\theta} \in \mathbb{R}^d$ be the minimizer of the empirical risk:

$$\hat{\theta} := \underset{\theta \in \mathbb{R}^d}{\mathrm{argmax}}\, \hat{L}(\theta), \quad \hat{L}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(z^{(i)}, \theta)$$

where $z^{(i)}$ i.i.d. drawn from p^*.

Assumptions on $\ell(z, \theta)$:

(i) $\ell(z, \theta)$ is twice differentiable in θ.

(ii) $\nabla \ell(z, \theta) \in \mathbb{R}^d$ means the gradient at θ.

(iii) $\nabla^2 \ell(z, \theta) \in \mathbb{R}^{d \times d}$ means the Hessian at θ.

(iv) $E_{z \sim p^*}[\nabla^2 \ell(z, \theta)] \succ 0$ is positive definite for all θ.

[Definition 3] Well-specified model

(i) $\ell(x, y; \theta) := -\log p_\theta(y|x)$

(ii) $\{p_\theta : \theta \in \mathbb{R}^d\}$ is conditionally well-specified if

$$p^*(x, y) = p^*(x)\, p_{\theta^*}(y|x) \quad \text{for some } \theta^* \in \mathbb{R}^d.$$

(iii) Suppose each θ specifies a $p_\theta(x, y)$. $\{p_\theta : \theta \in \mathbb{R}^d\}$ is jointly well-specified if $p^*(x, y) = p_{\theta^*}(x, y)$ for some $\theta^* \in \mathbb{R}^d$.

[Theorem 3] Bartlett identity.

In the well-specified case (conditionally, thus jointly), the following holds:

following holds:

$$\nabla^2 L(\theta^*) = \mathrm{Cov}[\nabla \ell(z,\theta^*)]$$

Pf: $1 = \int p^*(z)\, dz = \int p^*(x) P_{\theta^*}(y|x)\, dz$

$$\Rightarrow \quad \int p^*(x) e^{-\ell(z,\theta^*)} dz = 1$$

differentiate w.r.t. θ^*:

$$\int p^*(x) e^{-\ell(z,\theta^*)} (-\nabla \ell(z,\theta^*))\, dz = 0$$

which implies $E[\nabla \ell(z,\theta^*)] = 0$. Differentiate again:

$$0 = \int p^*(x) [-e^{-\ell(z,\theta^*)} \nabla^2 \ell(z,\theta^*) + e^{-\ell(z,\theta^*)} \nabla \ell(z,\theta^*) \nabla \ell(z,\theta^*)^T]\, dz$$

$$= -E[\nabla^2 \ell(z,\theta^*)] + E[\nabla \ell(z,\theta^*) \nabla \ell(z,\theta^*)^T]$$

$$= -E[\nabla^2 \ell(z,\theta^*)] + \mathrm{Cov}[\nabla \ell(z,\theta^*)] \quad \text{Since } E[\nabla \ell(z,\theta^*)] = 0.$$

$$\Rightarrow \quad \nabla^2 L(\theta^*) = E[\nabla^2 \ell(z,\theta^*)] = \mathrm{Cov}[\nabla \ell(z,\theta^*)] \qquad \square$$

[Example 2] well-specified random design linear regression.

Model:

(i) $x \sim p^*(x)$ for some arbitrary $p^*(x)$

(ii) $y = \theta^* \cdot x + \varepsilon$ where $\varepsilon \sim N(0,1)$

Loss function: $\ell(x,y;\theta) := \frac{1}{2}(\theta \cdot x - y)^2$

Property 1: $\nabla^2 L(\theta) = \mathrm{Cov}[\nabla \ell(z,\theta^*)]$

(i) $\nabla^2 L(\theta) = \nabla^2 E_{x\sim p^*, \varepsilon \sim N(0,1)} [\frac{1}{2}(\theta \cdot x - \theta^* \cdot x - \varepsilon)^2]$

$$= E[xx^T]$$

(ii) $\mathrm{Cov}[\nabla \ell(z,\theta^*)] = \mathrm{Cov}[-\varepsilon x^T]$

$$= E[\varepsilon x x^T \varepsilon] - E[\varepsilon x] E[\varepsilon x]^T$$

$$= E[xx^T] \quad \text{Since } x, \varepsilon \text{ are independent.}$$

Now we study $\hat{\theta} - \theta^*$:

Step 1: perform a Taylor expansion of ∇L around θ^*:

$$\nabla\hat{L}(\hat{\theta}) = \nabla\hat{L}(\theta^*) + \nabla^2\hat{L}(\theta^*)(\hat{\theta}-\theta^*) + O_p(\|\hat{\theta}-\theta^*\|_2^2)$$

Using the fact $\nabla\hat{L}(\hat{\theta})=0$ since $\hat{\theta}$ is optimal:

$$\hat{\theta}-\theta^* = -\nabla^2\hat{L}(\theta^*)^{-1}\left(\nabla\hat{L}(\theta^*) + O_p(\|\hat{\theta}-\theta^*\|_2^2)\right) \quad (\mathrm{I})$$

As $n\to\infty$, by the weak law of large numbers:

$$\nabla^2\hat{L}(\theta^*) \xrightarrow{P} \nabla^2 L(\theta^*)$$

Since $\nabla^2 L(\theta^*) > 0$, $\nabla^2 L(\cdot)^{-1}$ is smooth around θ^*:

$$\nabla^2\hat{L}(\theta^*)^{-1} \xrightarrow{P} \nabla^2 L(\theta^*)^{-1} \qquad (\mathrm{CMT})$$

By central limit theorem:

$$\sqrt{n}\,\nabla\hat{L}(\theta^*) \xrightarrow{d} N(0, \mathrm{Cov}[\nabla\ell(z,\theta^*)])$$

Suppose $\hat{\theta}-\theta = O_p(f(n))$, then by (I), $f(n)$ decays at a rate of $O(\frac{1}{\sqrt{n}})$ or $f^2(n)$, which implies $f^2(n)=\frac{1}{\sqrt{n}}$. Thus:

$$\sqrt{n}\cdot O_p(\|\hat{\theta}-\theta^*\|_2^2) \xrightarrow{P} 0.$$

By Slutsky's theorem, with (I):

$$\sqrt{n}\cdot(\hat{\theta}-\theta^*) \xrightarrow{d} N(0, \nabla^2 L(\theta^*)^{-1}\mathrm{Cov}[\nabla\ell(z,\theta^*)]\nabla^2 L(\theta^*)^{-1}) \cdots (\mathrm{II})$$

Due to Property 1:

$$\sqrt{n}(\hat{\theta}-\theta^*) \xrightarrow{d} N(0, E[xx^T]^{-1})$$

Step 2: analysis of excess risk: By Taylor expansion

$$L(\hat{\theta}) = L(\theta^*) + \nabla L(\theta^*)^T(\hat{\theta}-\theta^*) + \frac{1}{2}\|\hat{\theta}-\theta^*\|^2_{\nabla^2 L(\theta^*)} + O_p(\|\hat{\theta}-\theta^*\|_2^3)$$

Since θ^* is the optimal, $\nabla L(\theta^*)=0$. Multiply by n for both sides:

$$n(L(\hat{\theta})-L(\theta^*)) = \frac{1}{2}\sqrt{n}(\hat{\theta}-\theta^*)^T\nabla^2 L(\theta^*)\sqrt{n}(\hat{\theta}-\theta^*) + O_p(n\|\hat{\theta}-\theta^*\|_2^3) \cdots (\mathrm{III})$$

Define $x_n = \sqrt{n}(\nabla^2 L(\theta^*))^{\frac{1}{2}}(\hat{\theta}-\theta^*)$,

by (II), we have $x_n \sim N(0, \nabla^2 L(\theta^*)^{-\frac{1}{2}}\mathrm{Cov}[\nabla\ell(z,\theta^*)]\nabla^2 L(\theta^*)^{-\frac{1}{2}})$.

Let $\Sigma = \nabla^2 L(\theta^*)^{-\frac{1}{2}}\mathrm{Cov}[\nabla\ell(z,\theta^*)]\nabla^2 L(\theta^*)^{-\frac{1}{2}}$, then

$$x_n x_n^T \xrightarrow{d} W(\Sigma, 1),$$ where W is the Wishart distribution.

Furthermore, $x_n^T x_n = tr(x_n x_n^T) \xrightarrow{d} tr(W(\Sigma, 1))$.

Therefore, by (III), we have:

$$n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2} tr(W(\Sigma, 1))$$

In the well-specified models, by Theorem 3:

$$Cov[\nabla \ell(z, \theta^*)] = \nabla^2 L(\theta^*),$$

then we have $\Sigma = I_{d \times d}$, resulting in

$$n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2} tr(W(I_{d\times d}, 1))$$

RHS shows a distribution of the sum of d chi-squared r.v.,
whose distribution is the same as $\frac{1}{2}\sum_{j=1}^{d} v_j^2$, where $v_j \sim N(0,1)$.

Then $E[n(L(\hat{\theta}) - L(\theta^*))] \to \frac{d}{2}$

$$Var[n(L(\hat{\theta}) - L(\theta^*))] \to d$$

which implies

$$L(\hat{\theta}) - L(\theta^*) \sim d/2n.$$

□