Eddie Aamari
04/10/24

Chapter IV (alternative)

Learning & Minimax theory in a nutshell

I Empirical risk minimization : A simple oracle bound

II Minimax lower bounds : A bayesian method applied to regression

I | Empirical risk minimization

We observe $(X_1, Y_1), \dots, (X_n, Y_n)$ iid with distribution $P_{(X,Y)}$ over $\mathcal{X} \times \mathcal{Y}$.

The goal is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ with small risk

$$\boxed{R(f) := \mathbb{E}_P\big[\ell(Y, f(X))\big]},$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a prescribed loss function

Ex: Classification
$\mathcal{Y} = \{-1, 1\}$
$\ell(y, y') = \mathbb{1}_{y \neq y'}$

Regression
$\mathcal{Y} = \mathbb{R}$
$\ell(y, y') = (y - y')^2$

We choose to search f in a prescribed set $\mathcal{F}$ of (measurable) functions

Ex: $\mathcal{F}$:= linear functions on $\mathcal{X} = \mathbb{R}^d$
:= A class of neural nets
:= sum of N fourier functions

We try an mimick R by the empirical risk

$$\boxed{\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))}$$

The ERM on $\mathcal{F}$ is then

$$\boxed{\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}(f)}$$

The goal is to compare the (expected) performances of $\hat{f}$, compared to the best possible over the class $\mathcal{F}$

$$\boxed{f^* \in \operatorname*{argmin}_{f \in \mathcal{F}} R(f)}$$

We have, almost surely,

$$\underbrace{R(\hat{f})}_{\text{(Random! (performance)}} = \underbrace{R(f^*)}_{\text{(Deterministic) approximation error}} + \underbrace{R(\hat{f}) - R(f^*)}_{\text{Estimation error}}$$

The approximation error depends on $\mathcal{F}$ and $P_{(X,Y)}$. In regression, it has been studied in chapter 2. We focus on the estimation term

$$R(\hat{f}) - R(f^*) = \big(R(\hat{f}) - \hat{R}(\hat{f})\big) + \underbrace{\big(\hat{R}(\hat{f}) - \hat{R}(f^*)\big)}_{\leq 0 \text{ by definition of } \hat{f}} + \big(\hat{R}(f^*) - R(f^*)\big)$$

$\hookrightarrow \leq \delta$ if risk minimization is done approximately

$$\leq 2 \sup_{f \in \mathcal{F}} \big|\hat{R}(f) - R(f)\big|$$

From the central limit theorem we know that for fixed $f \in \mathcal{F}$,

$$\sqrt{n}\big(\hat{R}(f) - R(f)\big) \xrightarrow[n \to \infty]{d} \mathcal{N}\big(0, \mathrm{Var}(\ell(Y, f(X)))\big)$$

so that $|\hat{R}(f) - R(f)| \lesssim \frac{1}{\sqrt{n}}$ whp

Here we need a stronger bound holding whp on all $f \in \mathcal{F}$

$\Longrightarrow$ Depends on the complexity of the class $\mathcal{F}$

For simplicity, assume that $\mathcal{Y} = \mathbb{R}^p$

- $\forall f \in \mathcal{F}$, $\|f\|_\infty \leq M$ (B)
- ℓ is G-lipschitz in the second variable (L)

$$|\ell(y,y') - \ell(y,y'')| \leq G\|y' - y''\|$$

From (L), we get that $\forall f, f' \in \mathcal{F}$,

$$|R(f) - R(f')| = |\mathbb{E}[\ell(Y,f(X)) - \ell(Y,f'(X))]|$$

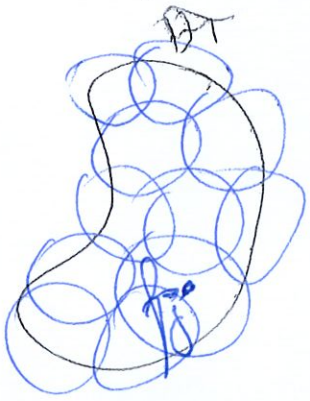
$$\leq G[\mathbb{E}\,|f(X) - f'(X)|]$$

$$\leq G\|f - f'\|_\infty$$

and similarly, $|\hat{R}(f) - \hat{R}(f')| \leq G\|f - f'\|_\infty$

This will allow us to discretize the supremum $\sup_{f \in \mathcal{F}}$ and simplify it to a finite family.

Given $\delta > 0$, write $f_1, \dots, f_N$ for a minimal δ-covering of $\mathcal{F}$ in L^∞ norm. That is $\begin{cases} N := N(\mathcal{F}, \|\cdot\|_\infty, \delta) \\ \mathcal{F} \subset \bigcup_{j=1}^{N} B_{\|\cdot\|_\infty}(f_j, \delta) \end{cases}$



For all $f \in \mathcal{F}$, there exists f_j with $\|f - f_j\|_\infty \leq \delta$, and hence

$$|R(f) - \hat{R}(f)| \leq |R(f) - R(f_j)| + |R(f_j) - \hat{R}(f_j)| + |\hat{R}(f_j) - \hat{R}(f)|$$

$$\leq 2G\|f - f_j\|_\infty + |R(f_j) - \hat{R}(f_j)|$$

$$\leq 2G\delta + |R(f_j) - \hat{R}(f_j)|$$

Hence $\boxed{\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \leq 2G\delta + \max_{j \leq N_\delta} |R(f_j) - \hat{R}(f_j)|}$

Let now f_j be fixed. We assume that the loss function is bounded

$$\boxed{\forall y, y' \in Y, \ |\ell(y, y')| \leq \ell_\infty}$$

Then $R(f_j) - \hat{R}(f_j) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}[\ell(Y_i, f_j(X_i))] - \ell(Y_i, f_j(X_i)) \right)$

is the sum of iid recentered variables bounded by $\frac{\ell_\infty}{n}$. From Hoeffding's concentration inequality, $\forall t > 0$,

$$P\left(|R(f_j) - \hat{R}(f_j)| > t \right) \leq 2 \exp\left(- \frac{2t^2}{n\left(\frac{2\ell_\infty}{n}\right)^2} \right)$$

$$= 2 \exp\left(- \frac{nt^2}{2\ell_\infty^2} \right)$$

At the end of the day, we apply a union bound to get

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} |R(f)-\hat{R}(f)| > t + 2G\delta\Big) \leq \mathbb{P}\Big(\max_{j\leq N_\delta} |R(f_j)-\hat{R}(f_j)| > t\Big)$$

$$\leq N_\delta \max_{j\leq N_\delta} \mathbb{P}\big(|R(f_j)-\hat{R}(f_j)| > t\big)$$

$$\leq N_\delta \times 2\exp\left(-\frac{nt^2}{2b_\infty}\right)$$

En posant $\xi = 2N_\delta \exp\left(-\frac{nt^2}{2b_\infty}\right)$, on obtient donc qu'avec proba au moins $1-\xi$,

$$\xi = 2N_\delta \exp\left(-\frac{nt^2}{2b_\infty}\right) \iff \frac{nt^2}{2b_\infty} = \log\left(\frac{2N_\delta}{\xi}\right) \iff t = \sqrt{\frac{2b_\infty \log\left(\frac{2N_\delta}{\xi}\right)}{n}}$$

$$\sup_{f\in\mathcal{F}} |R(f)-\hat{R}(f)| \leq 2G\delta + \sqrt{\frac{2b_\infty \log\left(\frac{2N_\delta}{\xi}\right)}{n}}$$

Coming back to the initial bound, with proba $\geq 1-\xi$,

$$\boxed{R(\hat{f}) \leq R(f^*) + \sqrt{\frac{8b_\infty \log\left(\frac{2N_\delta}{\xi}\right)}{n}} + 4G\delta}$$

Oracle bound in probability

Using that $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z\geq t)\,dt$ if $Z\geq 0$, we get a similar expected bound.

II Minimax lower bound technique

Given a particular estimator $\hat{\theta}$ that we designed over n-samples $Z_1, \dots, Z_n \sim P$, on may wonder whether one couldn't get a more precise one, for instance on average. Minimax lower bounds aims at doing this, by showing some intrinsic information-theoretic limitations of specific problems

Framework: . Model $\mathcal{P}$

. Parameter of interest $\theta : \mathcal{P} \longrightarrow (\mathcal{H})$

. Metric $d : (\mathcal{H}) \times (\mathcal{H}) \longrightarrow \mathbb{R}_{\geq 0}$

The minimax risk for estimating $\theta(P)$ over $\mathcal{P}$ is

$$\boxed{R_n(\mathcal{P}) := \inf_{\hat{\theta}_n} \sup_{P \in \mathcal{P}} E_{P^{\otimes n}}\left[d(\theta(P), \hat{\theta}_n) \right]}$$

where $\hat{\theta}_n = \hat{\theta}_n(Z_1, \dots, Z_n)$ ranges among all the estimators over n points $Z_1, \dots, Z_n$

Ex: $\mathcal{P} = \{ \mathcal{D}(X,Y) \mid$. $X \sim \mathrm{Unif}[0,1]^d$
. $Y = f^*(X) + \varepsilon$
. $X \perp\!\!\!\perp \varepsilon$
. $\mathrm{Var}(\varepsilon) \leq \sigma^2$, $E[\varepsilon] = 0$
. $f \in \mathcal{C}^\beta \}$

. $\theta(P) = f^*$

. $d(f, f') = \| f - f' \|_{L^2([0,1]^d)}$

Studying $\mathcal{R}_n(\mathcal{P})$ is twofold

Upper bounds : Exhibit as particular $\hat{\theta}_n$ which has good estimation rate uniformly over $\mathcal{P}$
$\mathcal{R}_n(\mathcal{P}) \leq \dots$

Lower bounds : Shows that one cannot do better, at least up to constants
$\mathcal{R}_n(\mathcal{P}) \geq \dots$

For lower bounds, the main idea is to replace the $\sup_{P \in \mathcal{P}}$ by an average $\int_{\mathcal{P}} \dots \pi(dP)$.

Indeed, if π is a probability distribution over $\mathcal{P}$, then for all $\hat{\theta}_n$,

(π : Prior distribution)

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{P^n}\left[d(\theta(P), \hat{\theta}_n)\right] \geq \int_{\mathcal{P}} \mathbb{E}_{P^n}\left[d(\theta(P), \hat{\theta}_n)\right] \pi(dP)$$

The right-hand term can usually be studied relatively easily, since
- WE choose π
- The best possible estimator $\hat{\theta}_\pi$ (bayes estimator) can be explicit

To cover the main ideas, we will present the simplest case where

$$\boxed{\pi = \frac{1}{2}(\delta_{P_0} + \delta_{P_1})}$$

← Two "hypotheses"

Thm: (Le Cam's lemma)

For all $P_0, P_1 \in \mathcal{P}$,

$$R_n(\mathcal{P}) \geq \frac{1}{2} d(\theta(P_0), \theta(P_1)) \int P_0^{\otimes n} \wedge P_1^{\otimes n} \, d\nu^{\otimes n}$$

where $P_0, P_1 \ll \nu$

$$P_j = \frac{dP_j}{d\nu}$$

Proof:

$$R_n(\mathcal{P}) \geq \inf_{\hat{\theta}_n} \frac{1}{2} \left(\mathbb{E}_{P_0^n}\left[d(\theta(P_0), \hat{\theta}) \right] + \mathbb{E}_{P_1^{\otimes n}}\left[d(\theta(P_1), \hat{\theta}_n) \right] \right)$$

$$= \inf_{\hat{\theta}_n} \frac{1}{2} \int_{Z^n} \left(d(\theta(P_0), \hat{\theta}(y_1, \dots, y_n)) P_0^{\otimes n}(\vec{y}) + d(\theta(P_1), \hat{\theta}_n) P_1^{\otimes n}(\vec{y}) \right) \nu^{\otimes n}(d\vec{y})$$

$$\geq \frac{1}{2} d(\theta(P_0), \theta(P_1)) \int P_0^{\otimes n} \wedge P_1^{\otimes n} \, d\nu^{\otimes n}$$

□

Rk: Actually, $\int P_0 \wedge P_1 = 1 - \frac{1}{2} \int |P_0 - P_1|$ → Total Variation

We can further bound the right-hand term

Lemma: $$\int P_0^{\otimes n} \wedge P_1^{\otimes n} \geq \frac{1}{2} \exp\left(-n KL(P_0, P_1)\right),$$

where $$KL(P_0, P_1) = \begin{cases} \int \log\left(\frac{P_0}{P_1}\right) P_0 & \text{if } P_0 \ll P_1, \\ \infty & \text{otherwise} \end{cases}$$

Proof: Jensen, See lemma 2.6 in Tsybakov [Introduction to Nonparametric estimation]

Rk: For instance, gaussian satisfy $$KL(N(0,\sigma^2), N(m,\sigma^2)) = \frac{m^2}{2\sigma^2}$$

Application to regression:

Say that $\mathcal{X} = [-1,1]^d$, and $d(f,f') = |f(0) - f'(0)|$

$\mathcal{Y} = \mathbb{R}$

$$Y = f^*(X) + \varepsilon \quad \text{with } \varepsilon \sim N(0,\sigma^2), \ X \perp\!\!\!\perp \varepsilon, \ X \sim \text{Unif}([-1,1]^d)$$

Assume also that $f \in \mathcal{C}^\beta(L)$: $|f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)| \leq L\|x-y\|^{\beta - \lfloor\beta\rfloor}$

↳ $P_\beta := \{\mathcal{D}((X,Y)), \ldots\}$

Apply le Cam's lemma to $P_0 = P_{f_0}$, $P_1 = P_{f_1}$:

$$\mathcal{R}_n(\mathcal{P}_\beta) \geq \frac{|f_0(0) - f_1(0)|}{2} \exp\left(-n KL(P_{f_0}, P_{f_1})\right)$$

When $\begin{cases} \varepsilon \sim N(0, \sigma^2), \\ X \sim \mathrm{Unif}([-1,1]^d) \end{cases}$, we can show that $\begin{cases} Y_0 = f_0(X) + \varepsilon \\ Y_1 = f_1(X) + \varepsilon \end{cases}$ satisfy

$$KL(P_{f_0}, P_{f_1}) \lesssim \int_{[-1,1]^d} KL\left(P_{Y_0|x}, P_{Y_1|x}\right) dx$$

$$= \int_{[-1,1]^d} KL\left(N(f_0(x), \sigma^2), N(f_1(x), \sigma^2)\right) dx$$

$$= \frac{\|f_0 - f_1\|_{L^2}^2}{2\sigma^2}$$

Hence, $\mathcal{R}_n(\mathcal{P}_\beta) \geq |f_0(0) - f_1(0)| \exp\left(-n \|f_0 - f_1\|^2_{L^2([-1,1]^d)}\right)$

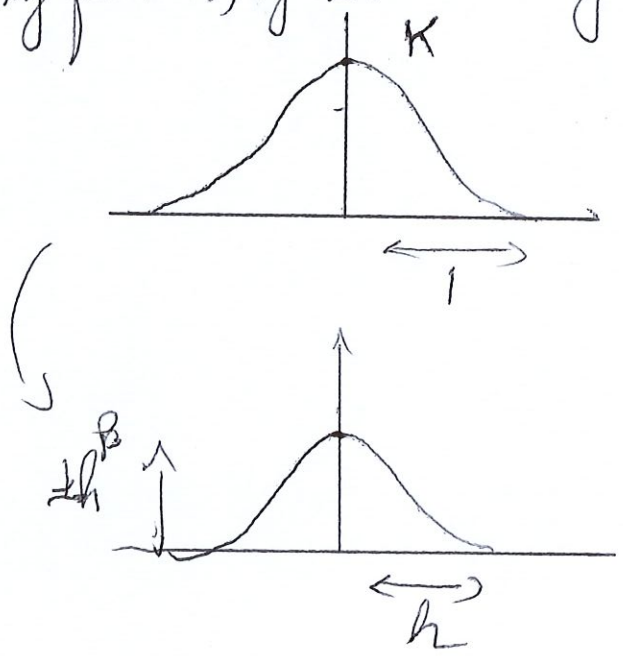
↳ Need to exhibit $f_0, f_1 \in \mathcal{C}^\beta(L)$ such that

- $|f_0(0) - f_1(0)|$ is large
- $\|f_0 - f_1\|^2_{L^2([-1,1]^d)}$ is small.

For this, we let $h>0$ to be chosen later, and

$$\begin{cases} \cdot\ f_0 \equiv 0 \\ \cdot\ f_1(x) = L h^\beta K\left(\frac{x}{h}\right) \end{cases}$$

where $K \in \mathcal{C}^\beta(1)$ is any function, by should be thought of as a localized bump



We have: • $f_0, f_1 \in \mathcal{C}^\beta(L)$ $\forall h>0$

• $|f_0(0) - f_1(0)| = L h^\beta K(0)$

• $\|f_0 - f_1\|_{L^2}^2 = \int_{B(0,h)} L^2 h^{2\beta} K^2\left(\frac{x}{h}\right) dx = L^2 h^{2\beta+d} \|K\|_{L^2}^2$

$$\forall h>0 \Rightarrow R_n(\mathcal{F}_\beta) \gtrsim L h^\beta \exp\left(-n \frac{L^2 h^{2\beta+d}}{\sigma^2}\right)$$

Choosing $\frac{n L^2 h^{2\beta+d}}{\sigma^2} = 1$ (i.e $h = \left(\frac{\sigma^2}{L^2 n}\right)^{\frac{1}{2\beta+d}}$), we get

$$\boxed{R_n(\mathcal{F}_\beta) \gtrsim \left(\frac{\sigma^2}{L^{d/\beta} n}\right)^{\frac{\beta}{2\beta+d}}}$$