36-789: Minimax Theory **Spring 2017**

Lecture 2: January 23

Lecturer: Alessandro Rinaldo *Scribes: Ilmun Kim*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

2.1 Recap

Notations

- Let $\mathcal{P}$ be a class of probability distributions on $(\mathcal{X}, \mathcal{B})$. For example, $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{B}$ is a Borel set.
- Let θ be a function $\theta : \mathcal{P} \to \Theta$ where Θ is a parameter space. Here, θ can be a parametrization as $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, but it can be $\theta(P) = \theta(P')$ even if $P \neq P'$.
- P_θ indicates an arbitrary $P \in \mathcal{P}$ such that $\theta(P) = \theta$.
- $d : \Theta \times \Theta \to [0, \infty)$ is a metric on a set Θ.
- $w : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-decreasing function such that $w(x) \neq 0$ for $x \neq 0$ and $w(0) = 0$.
- $X = (X_1, \ldots, X_n)$ are i.i.d. sample from $P \in \mathcal{P}$.
- $\hat{\theta}(X) = \hat{\theta}(X_1, \ldots, X_n)$ is a function $\hat{\theta} : \mathcal{X}^n \to \Theta$ from a sample space to a parameter space.
- A risk of $\hat{\theta}$ at $P \in \mathcal{P}$ is denoted by $\mathbb{E}_P\big[w\big(d(\hat{\theta}(X), \theta(P))\big)\big]$ where $\mathbb{E}_P[\cdot] = \mathbb{E}_{X_1,\ldots,X_n \sim P}[\cdot]$.

A typical example is given by

$$w\big(d(\hat{\theta}, \theta(P))\big) = ||\hat{\theta} - \theta(P)||_2^2$$

where $w(x) = x^2$ and d is the Euclidean norm.

Definition 2.1 (Maximum risk) *The maximum risk for an estimator $\hat{\theta}$ is*

$$r_n(\hat{\theta}) = \sup_{P \in \mathcal{P}} \mathbb{E}_P\big[w\big(d(\hat{\theta}, \theta)\big)\big].$$

For certain estimators $\hat{\theta}$, the maximum risk is upper bounded by $C\psi_n$ where $\psi_n \to 0$ as $n \to \infty$.

Definition 2.2 (Minimax risk) *The mimimax risk R_n^* is the infimum of r_n over all estimators. That is,*

$$R_n^* = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\big[w\big(d(\hat{\theta}, \theta)\big)\big].$$

Our goal is to lower bound the minimax risk. This lower bound depends on $(\mathcal{P}, \Theta, \theta, d, w)$ but not on $\hat{\theta}$. We may allow everything to depend on n such as $\mathcal{P} = \mathcal{P}_n$ and $\Theta = \Theta_n$.

2.2 Reduction scheme

A general reduction scheme is based on the following three steps:

Step 1. Reduction to a bound in probability

For fixed $P \in \mathcal{P}, \hat{\theta}$ and $\delta > 0$, we have

$$\begin{aligned}\mathbb{E}_P\big[w\big(d(\hat{\theta},\theta)\big)\big] &\geq w(\delta)P\left(w\big(d(\hat{\theta},\theta)\big) \geq w(\delta)\right) && \text{by the Markov inequality,}\\ &\geq w(\delta)P\left(d(\hat{\theta},\theta) \geq \delta\right) && \text{since } w \text{ is a non-decreasing function.}\end{aligned}$$

Therefore, if we can establish that

$$\inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} P\left(d(\hat{\theta},\theta) \geq \delta\right)$$

is bounded away from 0, then a minimax lower bound is $w(\delta)$ up to a constant. To be clear, $\delta = \delta_n$ is a function of n such that $\delta_n \to 0$ as $n \to \infty$.

Step 2. Reduction to a finite number of hypotheses

Choose $(M+1)$ points $\{\theta_0, \theta_1, \ldots, \theta_M\}$ in Θ and $(M+1)$ probability distributions $\{P_{\theta_0}, P_{\theta_1}, \ldots, P_{\theta_M}\}$ in $\mathcal{P}$ such that $\theta(P_{\theta_i}) = \theta_i$ where M can be a function of n. Now, we need a lower bound on

$$\inf_{\hat{\theta}} \max_{\theta\in\{\theta_0,\ldots,\theta_M\}} P_\theta\left(d(\hat{\theta},\theta) \geq \delta\right).$$

Each θ_i is a hypothesis and our next goal is to study the testing problem of recovering the correct hypothesis. We consider a multiple hypothesis test

$$\phi(X) : \mathcal{X}^n \to \{0, 1, \ldots, M\}$$

where $\phi(X) = i$ means that we think $X \sim P^n_{\theta_i}$. Given any estimator $\hat{\theta}$, define the minimum distance test

$$\phi^*(X) = \underset{i\in\{0,1,\ldots,M\}}{\operatorname{argmin}}\; d(\hat{\theta}(X), \theta_i).$$

Step 3. Choice of 2δ-separated hypotheses

If we consider $d(\theta_i, \theta_j) \geq 2\delta,\ \forall i \neq j$, then, for any $\hat{\theta}$ and $i = 0, \ldots, M$,

$$P_{\theta_i}\left(d(\hat{\theta},\theta_i) \geq \delta\right) \geq P_{\theta_i}\left(\phi^*(X) \neq i\right) \geq \inf_{\phi} P_{\theta_i}(\phi \neq i),$$

where the triangle inequality is used to obtain the result.

Summary of the reduction scheme

If we can choose (M+1) hypotheses $P_{\theta_0}, \dots, P_{\theta_M}$ that are 2δ-separated, then

$$\begin{aligned} & \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} P\left(d(\hat{\theta}, \theta(P)) \geq \delta\right) \\ \geq\ & \inf_{\hat{\theta}} \max_{\theta \in \{\theta_0, \dots, \theta_M\}} P_\theta\left(d(\hat{\theta}, \theta) \geq \delta\right) \\ \geq\ & \inf_{\phi} \max_{i \in \{0, \dots, M\}} P_{\theta_i}\left(\phi(X) \neq i\right) \\ =\ & P_{e,M,\delta} \end{aligned}$$

where ϕ is a test function mapping $\mathcal{X}$ into $\{0, \dots, M\}$. Thus, the final lower bound on minimax risk is $w(\delta)P_{e,M,\delta}$. If $P_{e,M,\delta} > c$, then $w(\delta)c$ is a minimax lower bound.

Remarks

1. If $\delta_n \to 0$ as $n \to \infty$ and $P_{e,M_n,\delta_n} \geq c > 0$ for all large n, then $w(\delta_n)$ is a lower bound on minimax rate.
2. This bound needs not to be tight. It is just a lower bound. To show that it is optimal, we need to find one of $\hat{\theta}$ with the matching upper bound.
3. This is an art: you need to pick (M+1) 2δ-separated hypotheses that are far apart in the distance d, but whose corresponding probability distributions are very close.

2.3 Distance between probability distributions

Let P, Q be probability distributions on $(\mathcal{X}, \mathcal{B})$ with a common dominance measure μ (e.g. $\mu = P + Q$) and their Radon–Nikodym derivative ($dP/d\mu = p$ and $dQ/d\mu = q$).

2.3.1 Total Variation Distance

Definition 2.3 (Total Variation Distance) *The total variation distance between P and Q is defined as follows:*

$$d_{TV}(P, Q) = ||P - Q||_{TV} = \sup_{B \in \mathcal{B}} |P(B) - Q(B)|.$$

Properties of the total variation distance:

- It is a distance.
- $d_{TV}(P, Q) = 0$ if and only if $P = Q$.
- $d_{TV}(P, Q) = 1$ if and only if P and Q are singular. ($\exists B \in \mathcal{B}, P(B) = 1$ and $Q(B) = 0$)

Lemma 2.4 (Scheffé lemma)

$$\begin{aligned} d_{TV}(P,Q) &= \frac{1}{2}\int_{\mathcal{B}} |p(x)-q(x)|d\mu(x) \\ &= 1 - \underbrace{\int_{\mathcal{B}} \min\{p(x),q(x)\}d\mu(x)}_{affinity} \\ &= 1 - \int \min\{dP, dQ\}. \end{aligned}$$

Proof: *Let $A = \{x \in \mathcal{X} : q(x) \geq p(x)\}$. Then, we can get*

$$\int_{\mathcal{X}} |p(x)-q(x)|d\mu(x) = 2\int_A q(x)-p(x)d\mu(x).$$

Thus,

$$d_{TV}(P,Q) \geq Q(A) - P(A) = \frac{1}{2}\int |p(x)-q(x)|d\mu(x).$$

To show the opposite, we have that for $\forall B \in \mathcal{B}$,

$$\begin{aligned} \Big| \int_B (q-p)d\mu \Big| &= \Big| \int_{B\cap A} (q-p)d\mu + \int_{B\cap A^c} (q-p)\mu \Big| \\ &\leq \max\Big\{ \int_A (q-p)d\mu, \int_{A^c} (p-q)d\mu \Big\} \\ &\leq \frac{1}{2}\int |p-q|d\mu. \end{aligned}$$

■

Remark The supremum is achieved at the set $A = \{x : q(x) \geq p(x)\}$.

2.3.2 Connection with hypothesis testing

Suppose we want to test $H_0 : X \sim P$ vs. $H_a : X \sim Q$. A test function is given as $\phi(X) \in \{0,1\}$, where

$$\begin{cases} \phi(X) = 1 & \text{reject } H_0 \\ \phi(X) = 0 & \text{accept } H_0. \end{cases}$$

For each test ϕ, the type I error and the type II error are provided by

$$\begin{aligned} \text{Type I error} &= \mathbb{E}_P[\phi(X)] \\ \text{Type II error} &= \mathbb{E}_Q[1-\phi(X)]. \end{aligned}$$

Note that, according to the Neyman-Pearson lemma, the optimal test is

$$\phi^*(x) = I(q(x) \geq p(x)) = I(x \in A).$$

This test achieves the infimum as

$$\inf_{\phi} (\text{Type I error} + \text{Type II error}) = 1 - d_{TV}(P,Q) = \int \min\{dP, dQ\}.$$

More facts:

- $\inf_{0\leq f\leq 1} \mathbb{E}_P\left[f(x)\right] + \mathbb{E}_Q\left[1-f(x)\right] = 1 - d_{TV}(P,Q)$
- $\inf_{f,g\geq 0, f+g\geq 1} \mathbb{E}_P\left[f(x)\right] + \mathbb{E}_Q\left[1-f(x)\right] \geq 1 - d_{TV}(P,Q)$

Problem: It does not tensorize well, i.e. $d_{TV}(P^n, Q^n)$ is not trivially related to $d_{TV}(P,Q)$.

2.3.3 Hellinger Distance

Definition 2.5 (Hellinger distance) *The Hellinger distance between P and Q is defined as follows:*

$$H(P,Q) = \sqrt{\int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\mu(x)}$$

Properties of the Hellinger distance:

- It is a distance.
- $0 \leq H^2(P,Q) \leq 2$ where the upper bound holds when P, Q are singular.
- $H^2(P,Q) = 2\left(1 - \underbrace{\int_{\mathcal{X}} \sqrt{p(x)}\sqrt{q(x)} d\mu(x)}_{\text{Hellinger affinity}}\right)$
- If P and Q are product measures, $P = \otimes_{i=1}^n P_i$, $Q = \otimes_{i=1}^n$, then
$$H^2(P,Q) = 2\left(1 - \prod_{i=1}^n \left(1 - \frac{H^2(P_i, Q_i)}{2}\right)\right).$$