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

Lecture 1: January 18

Lecturer: Alessandro Rinaldo *Scribes: Jaehyeok Shin*

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

1.1 Minimax lower bounds

For statistical problems such as estimation, testing, confidence set, model selection, etc, we start with a given procedure, and establish

- Consistency
- Rates of convergence (as a function of n and other parameters)

The last one upper bound the minimax risk which we will define later. The *minimax theory* is about a quantifying how hard a problem is by producing rates that lower bound the convergence rates of any procedure.

Example Normal means $Y = \theta^* + \epsilon \ \in \mathbb{R}^d, \quad \epsilon \sim N(0, \sigma^2 I_d)$
Observation : $Y_1, \dots, Y_n \overset{i.i.d.}{\sim} N(\theta^*, \sigma^2 I_d)$
Let $\hat{\theta}(Y_1, \dots, Y_n) \in \mathbb{R}^d$. We are interested in the expected squared loss $\mathbb{E}\left[\|\hat{\theta} - \theta^*\|^2\right]$
For the penalized regression : $\hat{\theta} \in \arg\min_{\theta \in \Theta} \frac{1}{2n} \sum_{i=1}^n \|Y_i - \theta\|^2$, we get

$$\mathbb{E}\left[\|\hat{\theta} - \theta^*\|^2\right] \lesssim \begin{cases} \frac{\sigma^2 d}{n} & \Theta = \mathbb{R}^d \\ \frac{\sigma^2 \log d}{n} & \Theta = B_1 := \{\theta \in R^d, \|\theta\|_1 \leq 1\} \\ \frac{\sigma^2 k}{n} \log \frac{ed}{n} & \Theta = B_0(k) := \{\theta \in R^d, \|\theta\|_0 \leq k\} \end{cases}$$

Example Covariance estimation $X_1, \dots, X_n \overset{i.i.d.}{\sim} P \quad (e.g., P = N(0, \Sigma_{d\times d}))$

$$\begin{aligned} \Sigma_{d\times d} \in C(\alpha, M_0, M_1) = \{&\Sigma_{d\times d} = (\sigma_{ij})_{i,j=1,\dots,d} \text{ Symmetric \& P.S.D} \\ &\text{such that} \max_j \sum_{|i-j|>k} |\sigma_{ij}| \leq M_0 k^{-\alpha}, \forall k \ \lambda_{\max}(\Sigma) \leq M_1\} \end{aligned}$$

Bickel, Levina (2008) estimated Σ with $\hat{\Sigma}_{BL} = \hat{\Sigma}_{BL}(X_1, \dots, X_n)$, and got

$$\sup_{\Sigma \in C(\alpha, M_0, M_1)} \mathbb{E}\|\Sigma - \hat{\Sigma}_{BL}\|_{op} \lessapprox \left(\frac{\log d}{n}\right)^{\frac{\alpha}{\alpha+1}}$$

Cai, Zhang, and Zhou (2010) obtained a different estimator $\hat{\Sigma}_{CZZ}$ with

$$\sup_{\Sigma \in C(\alpha, M_0, M_1)} \mathbb{E}\|\Sigma - \hat{\Sigma}_{CZZ}\|_{op} \lessapprox \min\left\{\left(\frac{1}{n}\right)^{\frac{2\alpha}{2\alpha+1}} + \frac{\log d}{n}, \frac{p}{n}\right\}$$

Better rates, in fact, they are optimal.

$$\inf_{\hat{\Sigma}} \sup_{\Sigma \in C(\alpha, M_0, M_1)} \mathbb{E}\|\Sigma - \hat{\Sigma}\|_{op} \text{ is the same order as } \hat{\Sigma}_{CZZ}$$

1.1.1 Set-up

Let $\mathcal{P}$ collection of probability distributions on $(\mathcal{X}, \mathcal{A})$
Let $\theta : \mathcal{P} \longrightarrow \Theta$ functional, and $\theta(P)$ parameter.

Example

- $\theta(P) = \mathbb{E}_P[X], \ X \sim P$
- If P has a density f, $\theta(P) = f(x_0)$, or $\theta(P) = \int \left(f^{'}(x)\right)^2 dx$

In particular, we may have that $\theta(P) = \theta(Q)$ for $P, Q \in \mathcal{P}$

Simple case
θ parametrize $\mathcal{P}$, in which case $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, and $\theta(P) \neq \theta(Q)$ iff $P \neq Q$.
If $\Theta \subset \mathbb{R}^d$, $\mathcal{P}$ is a parametric family.

Examples

1. $\mathcal{P} = \{N(\theta, I) : \theta \in \Re^d\}$
2. $\Theta = \{\text{Set of smooth functions on} [0,1]^d\}$,
   $\mathcal{P}$ consists of probabilities P for $(Y, X) \in \Re \times \Re^d$ such that
   $$Y = f(X) + \epsilon, \ f \in \Theta, \ \epsilon \sim (0, \sigma^2) \perp\!\!\!\perp X$$
   $\Longrightarrow$ Non-parametric function estimation problem.
3. $\mathcal{P}$ of $(Y, X) \in \Re \times \Re^d, \quad Y = X^T\theta + \epsilon, \quad \theta \in \Theta \subset \Re^d$

Let $\underset{\sim}{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} P \in \mathcal{P}$. We will estimate $\theta(P)$ using $\underset{\sim}{X}$.
You may allow $\mathcal{P}$ to change with n, i.e., for each n, we will have $\mathcal{P}_n, \Theta_n$ depending on n

Let $d : \Theta \times \Theta \longrightarrow [0, \infty)$ be a metric

1. $d \geq 0$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

4. $d(x, y) = 0 \Leftrightarrow x = y$

More generally, we will consider $w\left(d(\theta, \theta^{'})\right)$ where

$$w : [0, \infty) \longrightarrow [0, \infty) \text{ non-decreasing}, \quad w(0) = 0, \quad w \not\equiv 0$$

Example $d(\theta, \theta^{'}) = \|\theta - \theta^{'}\|$

- $w(x) = x^2$: square error
- $w(x) = \mathbf{1}\{x > c\}, \quad c > 0$

Assuming $\underset{\sim}{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} P \in \mathcal{P}$. We are concerned with $\theta(P) \in \Theta$.
For a given procedure $\hat{\theta} : \mathcal{X}^n \longrightarrow \Theta$, its (point-wise) risk at P is

$$\mathbb{E}_{\underset{\sim}{X} \sim P^n}\left[w\left(d(\hat{\theta}(\underset{\sim}{X}), \theta(P)\right)\right]$$

Example

- $\mathcal{P} = \left\{N(\theta, I), \theta \in \Re^d\right\}, \quad d(x, y) = \|x - y\|_2, \quad w(x) = x^2$
  Risk : $\mathbb{E}\|\hat{\theta} - \theta\|_2^2$
- $\mathcal{P} = \{P_0, P_1\}, \quad \theta(\underset{\sim}{X}) = 1 \text{ or } 0$
  Risk : $l_i P_i\left(\hat{X} \neq i\right), \quad i = 0, 1, \quad l_i > 0$

To measure how well $\hat{\theta}$ does, let's take sup over all $P \in \mathcal{P}_n$

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

Upper bound calculations entail finding a constant $C = C(\mathcal{P}_n)$ and a sequence $\psi_n \longrightarrow 0$ as $n \longrightarrow \infty$ such that

$$r_n(\hat{\theta}, \mathcal{P}_n) \leq C\psi_n$$

The minimax risk is

$$R_n(\mathcal{P}_n) = \inf_{\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)} \sup_{P \in \mathcal{P}_n} \mathbb{E}_P\left[w\left(d(\hat{\theta}, \theta(P))\right)\right]$$

If we can show that

$$R_n(\mathcal{P}_n) \geq C^{'}\psi_n^{'}, \quad C^{'} = C^{'}(\mathcal{P}_n), \quad \psi_n^{'} \longrightarrow 0$$

Then, $C^{'}\psi_n^{'}$ is a lower bound on minimax risk, $\forall n$
If $\hat{\theta}$ is such that $\frac{\psi_n}{\psi_n^{'}} = \Theta(1)$, then $\hat{\theta}$ is minimax rate optimal.