ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 21: Functional estimation & testing

Lecturer: Yihong Wu *Scribe: Ashok Vardhan, Apr 14, 2016*

In this chapter, we will be interested in analyzing the sample complexity and minimax rates for the functional estimation problem introduced earlier. We will also consider the hypothesis testing paradigm so that we can utilize important tools such as LeCam's method for proving bounds and analyzing the minimax rate.

Formally, the setting is as follows: Assume that θ is an unknown parameter in the parameter space Θ and θ generates the data X according to the distribution P_θ. For a fixed real valued functional T on θ, i.e, $T : \Theta \to \mathbb{R}$, we wish to estimate T based on the observations through the estimator $\hat{T}(X)$.

In the estimation paradigm, the setting can be pictorially represented as follows:

$$\begin{array}{ccc} \theta & \xrightarrow{P_\theta} & X \\ \downarrow & & \downarrow \\ T = T(\theta) & & \hat{T}(X) \end{array}$$

In the hypothesis testing paradigm, we are interested in determining the class of parameters that gave rise to the observations. Formally, given $t_0, t_1 \in \mathbb{R}$, the problem is formulated as:

$$H_0 : T \leq t_0,$$
$$H_1 : T \geq t_1.$$

Equivalently, we can also think of the above hypothesis testing as a composite hypothesis testing of θ as follows:

$$H_0 : \theta \in \Theta_0 = \{\theta : T(\theta) \leq t_0\},$$
$$H_1 : \theta \in \Theta_1 = \{\theta : T(\theta) \geq t_1\}.$$

Example 21.1. Consider the Gaussian location model $X \sim \mathcal{N}(\theta, I_p), \theta \in \mathbb{R}^p$. Let $T : \mathbb{R}^p \to \mathbb{R}$ be given by $T(\theta) = \|\theta\|$. A possible test would be determining if $\|\theta\|$ is too small or too large given some thresholds. Specifically,

$$H_0 : \|\theta\| \leq 1,$$
$$H_1 : \|\theta\| \geq 3.$$

21.1 Lower bounds on minimax risk for functional estimation

Since T takes only real values where as Θ can be arbitrary high dimensional space, such as Euclidean space $\mathbb{R}^p$, T can be thought of as a low dimensional representation of the parameter space Θ. Thus

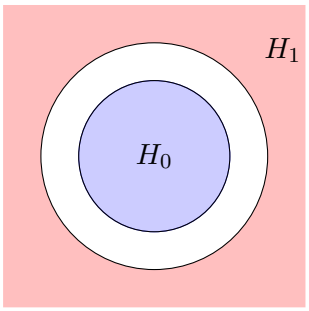

Figure 21.1: $\|\theta\| \le 1$ vs. $\|\theta\| \ge 3$.

it suggests if we can employ techniques such as LeCam's two-point argument to prove lower bounds on the minimax risk estimation of $T(\theta)$.

To this end, recall the LeCam's two-point method discussed in Lecture 9. The key idea in the two-point argument is the fact that if we can estimate a parameter, we can also test it efficiently. We reduced the task of estimation to that of the binary hypothesis testing, i.e, for fixed $\theta_0, \theta_1 \in \Theta$,

$$\begin{aligned} H_0 &: \theta = \theta_0, \\ H_1 &: \theta = \theta_1 \end{aligned}$$

to derive lower bounds on the minimax risk R^* (under quadratic loss). In particular, we showed that

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[(\theta - \hat{\theta})^2 \right] \gtrsim \sup_{\theta_0 \neq \theta_1} \|\theta_0 - \theta_1\|^2 \left(1 - d_{\mathrm{TV}}\left(P_{\theta_0}, P_{\theta_1}\right)\right).$$

In a similar vein, we can consider the following binary hypothesis testing for T (for some $t_0, t_1 \in \mathbb{R}$)

$$\begin{aligned} H_0 &: T \le t_0, \\ H_1 &: T \ge t_1 \end{aligned} \tag{21.1}$$

to obtain a lower bound on the minimax risk for estimation of T. Specifically, if $\pi_0 \in \mathcal{M}(\Theta_0), \pi_1 \in \mathcal{M}(\Theta_1)$ are any two priors on Θ_0 and Θ_1 respectively, we obtain

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[(T(\theta) - \hat{T}(X))^2 \right] \gtrsim (t_0 - t_1)^2 \left(1 - d_{\mathrm{TV}}\left(P_{\pi_0}, P_{\pi_1}\right)\right). \tag{21.2}$$

Thus our task reduces to finding two priors π_0, π_1 so that the lower bound in (21.2) would be maximized, or roughly speaking, we want to pick two priors that would ensure maximum confusion in testing of the two hypothesis.

Now we study a closely related concept of sample complexity for the analysis of the same.

21.2 Estimation of $\|\theta\|$ in GLM

Our aim is to prove that for the p-dimensional GLM where the data $X \sim \mathcal{N}(\theta, \frac{1}{n} I_p), \theta \in \mathbb{R}^p$, the minimax risk R^* for the estimation of $T(\theta) = \|\theta\|$ obeys $R^* \asymp \frac{\sqrt{p}}{n}$.

First we give a preview of this result and other estimation tasks in terms of a closely related concept: sample complexity. The proofs of these results is similar to that of those concerning average and minimax risk. Recall from Lecture 3, where we defined the sample complexity to be the minimum number of samples required to achieve a prescribed estimation error, either in expectation or in probability with high confidence.

Estimation tasks	Sample complexity
$T(\theta) = \theta$	$n^* \asymp p$
$T(\theta) = \theta_1$	$n^* \asymp 1$
$T(\theta) = \theta_{\max}$	$n^* \asymp \log p$
$T(\theta) = \|\theta\|_2$	$n^* \asymp \sqrt{p}$

One important observation is the fact that to estimate $\|\theta\|_2$, one can employ a plug-in estimator where we first estimate θ and then compute $\|\theta\|_2$. However, this naive procedure requires as many samples as that are required to estimate θ. Instead, we can perform much better by using only $\sqrt{p}$ samples to estimate $\|\theta\|_2$.

Instead of the setting in (21.1), where both the hypotheses are composite, we consider a simplified testing scenario where only one of the hypotheses is composite and hence more tractable.

$$H_0 : \theta = 0,$$
$$H_1 : \|\theta\|_2 \geq \rho$$

Pictorially,

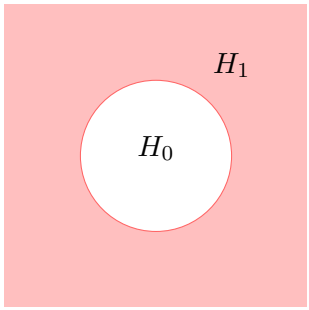

Figure 21.2: $\theta = 0$ vs. $\|\theta\| \geq \rho$.

We can further simplify this to the observation of one sample case making use of the fact that, to incur a minimum probability of error (say 0.01), $\max \rho$ for n-sample GLM $= \frac{\max \rho \text{ for 1 sample GLM}}{\sqrt{n}}$. Thus our model reduces to

$$X \sim \mathcal{N}(\theta, I_p), \theta \in \mathbb{R}^p.$$

21.2.1 Draw backs of two-point argument

A naive application of LeCam's two-point argument for the estimation of $\|\theta\|$ through the binary hypothesis testing of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ would yield

$$\begin{aligned} R^* &\gtrsim (\|\theta_0\| - \|\theta_1\|)^2 \left(1 - d_{\mathrm{TV}}\left(\mathcal{N}(\theta_0, I_p), \mathcal{N}(\theta_1, I_p)\right)\right) \\ &= 2\left(\|\theta_0\| - \|\theta_1\|\right)^2 Q\left(\frac{\|\theta_0 - \theta_1\|}{2}\right). \end{aligned}$$

Since $\|\theta_0\| - \|\theta_1\| \le \|\theta_0 - \theta_1\|$ by triangle inequality and $\sup \|\theta_0 - \theta_1\|_2 Q\left(\frac{\|\theta_0 - \theta_1\|_2}{2}\right) \asymp 1$, we see that this approach does not yield any useful lower bound. Thus we lose the much needed dependence of the minimax risk R^* on the dimension of our data, p.

Forgoing the two-point approach wherein we assume uniform distribution on two fixed parameters θ_0, θ_1, we want to choose a prior π supported on $\{\theta : \|\theta\|_2 \ge t\}$ such that the total variation distance $d_{\mathrm{TV}}(P_0, P_\pi)$ is bounded away from 1. In other words, we want to choose a prior π so that P_π closely resembles P_0 in the sense that the probability of error for testing is bounded away from 0. Note that P_π denotes the distribution on the data X which is given by $P_\pi = \pi * \mathcal{N}(0, I_p)$ whereas $P_0 = \mathcal{N}(0, I_p)$.

Recall from Lecture 5, the following chain of inequalities for KL divergence, χ^2 distance and total variation obtained using the concept of joint range. We have

$$\chi^2(P\|Q) \ge \log\left(\chi^2(P\|Q) + 1\right) \ge D(P\|Q) \ge d_{\mathrm{TV}}(P, Q) \log \frac{1 + d_{\mathrm{TV}}(P, Q)}{1 - d_{\mathrm{TV}}(P, Q)}$$

for any two distributions P, Q. This relation suggests that a sufficient condition to ensure $d_{\mathrm{TV}}(P_0, P_\pi)$ to be bounded away from 1, or equivalently $1 - d_{\mathrm{TV}}(P_0, P_\pi) \gtrsim 0$, is to make $\chi^2(P_0, P_\pi) \lesssim 1$. In this regard, we need the following lemma which gives an alternative characterization of χ^2-distance.

Lemma 21.1 (Ingster-Suslina method)**.** *Let Θ be a parameter space and for each $\theta \in \Theta$, let P_θ be a family of probability distributions on a measure space $\mathcal{X}$ and let Q also be a distribution on $\mathcal{X}$. Then*

$$\chi^2(P_\pi\|Q) = \mathbb{E}[G(\theta, \tilde{\theta})] - 1,$$

where $\theta, \tilde{\theta} \overset{i.i.d.}{\sim} \pi$, $G(\theta, \tilde{\theta}) = \int \frac{P_\theta P_{\tilde{\theta}}}{Q}$ and $P_\pi = \int P_\theta \pi(d\theta) = \int P_{\tilde{\theta}} \pi(d\tilde{\theta})$.

Proof. For any two distributions P and Q, we have

$$\chi^2(P\|Q) = \int \frac{(P - Q)^2}{Q} = \mathrm{Var}_Q\left(\frac{P}{Q}\right) = \mathbb{E}_Q\left(\frac{P}{Q}\right)^2 - 1 = \int \frac{P^2}{Q} - 1.$$

Thus, $\chi^2(P_\pi\|Q) = \int \frac{P_\pi^2}{Q} - 1$ and

$$\begin{aligned} \int \frac{P_\pi^2}{Q} = \int \frac{P_\pi P_\pi}{Q} &= \int \frac{\int P_\theta(x)\pi(d\theta) \int P_{\tilde{\theta}}(x)\pi(d\tilde{\theta})}{Q(x)} \mu(dx) \\ &\overset{\text{Fubini}}{=} \int\int \pi(d\theta)\pi(d\tilde{\theta}) \int \frac{P_\theta P_{\tilde{\theta}}}{Q} \mu(dx) \\ &= \mathbb{E}[G(\theta, \tilde{\theta})], \end{aligned}$$

□

In the case of GLM, $\mathbb{E}[G(\theta, \tilde{\theta})]$ can be computed explicitly and hence we obtain the following corollary.

Corollary 21.1. *If $P_\theta = \mathcal{N}(\theta, I_p)$ and $Q = \mathcal{N}(0, I_p)$, then*

$$\mathbb{E}[G(\theta, \tilde{\theta})] = \mathbb{E}[\exp\langle \theta, \tilde{\theta} \rangle].$$

Proof. Since $P_\theta = \mathcal{N}(\theta, I_p)$, we have

$$\begin{aligned}
G(\theta, \tilde{\theta}) &= \int \frac{\frac{1}{\sqrt{(2\pi)^p}} \exp\left(\frac{-\|x-\theta\|^2}{2}\right) \frac{1}{\sqrt{(2\pi)^p}} \exp\left(\frac{-\|x-\tilde{\theta}\|^2}{2}\right)}{\frac{1}{\sqrt{(2\pi)^p}} \exp\left(\frac{-\|x\|^2}{2}\right)} \\
&= \int \frac{1}{\sqrt{(2\pi)^p}} \exp\left(\frac{-1}{2}\left(\|x-\theta\|^2 + \|x-\tilde{\theta}\|^2 - \|x\|^2\right)\right) \\
&= \int \frac{1}{\sqrt{(2\pi)^p}} \exp\left(\frac{-1}{2}\left(\|x\|^2 - 2\langle x, \theta + \tilde{\theta}\rangle + \|\theta + \tilde{\theta}\|^2 - 2\langle \theta, \tilde{\theta}\rangle\right)\right) \\
&= \exp(\langle \theta, \tilde{\theta} \rangle) \int \frac{1}{\sqrt{(2\pi)^p}} \exp\left(\frac{-\|x-\theta-\tilde{\theta}\|^2}{2}\right) \\
&= \exp(\langle \theta, \tilde{\theta} \rangle).
\end{aligned}$$

□

References