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

Lecture 10: Le Cam's Method, Two-point Argument and Assouad's Lemma

Lecturer: Yihong Wu *Scribe: Harsh Gupta, February 27, 2016*

Recap:

Theorem 10.1 (Le Cam's Method)**.** *If* $l(\theta_0, \theta_1) \leq \alpha\{l(\theta_0, \hat{\theta}) + l(\theta_1, \hat{\theta_1})\}, \forall \hat{\theta}$ *then*

$$\Rightarrow R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} E_\theta[l(\theta, \hat{\theta})] \geq \frac{l(\theta_0, \theta_1)}{4\alpha}(1 - TV(P_{\theta_0}, P_{\theta_1})) \tag{10.1}$$

Note:

- For n samples, the total variation increases and hence we get a smaller lower bound.
- For different loss functions we have:

$$l = ||.|| \Rightarrow \alpha = 1$$
$$l = ||.||^q \Rightarrow \alpha = 2^{q-1}$$

- If $l(\theta_0, \hat{\theta}) = ||\theta_0 - \hat{\theta}||_2^2$, using Theorem 10.1, we have:

$$\Rightarrow R^* \geq \frac{||\theta_0 - \theta_1||_2^2}{8}(1 - TV(P_{\theta_0}, P_{\theta_1})) \tag{10.2}$$

  Can we improve the factor of 8 in the above inequality? The answer is YES as we shall see in the next section!

10.1 Reduction of factor from 8 to 4

We view Θ as an inner product space. Therefore, $l(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_2^2 = \langle \theta - \hat{\theta}, \theta - \hat{\theta} \rangle$.

Theorem 10.2 (Reduction of factor in (10.2) from 8 to 4)**.**

Proof. We use minimiax risk $\geq$ Bayes risk:

$$\Rightarrow R^* \geq R_\pi^*$$

Using $\pi = \bar{\lambda}\delta_{\theta_0} + \lambda\delta_{\theta_1}$ as the prior, where $\lambda \in [0, 1], \bar{\lambda} = 1 - \lambda$, we have:

$$\Rightarrow R_\pi^* = \inf_{\hat{\theta}} \bar{\lambda} E_{\theta_0} ||\theta_0 - \hat{\theta}||_2^2 + \bar{\lambda} E_{\theta_1} ||\theta_1 - \hat{\theta}||_2^2 \tag{10.3}$$

$$\Rightarrow R_\pi^* = \int_X \mu(dx)\{\inf_{\hat{\theta}} \bar{\lambda} P_{\theta_0}(x) ||\theta_0 - \hat{\theta}(x)||_2^2 + \lambda P_{\theta_1} ||\theta_1 - \hat{\theta}(x)||_2^2\} \tag{10.4}$$

We first consider the following general problem:

$$\Rightarrow \inf_{\hat{\theta}} \{\bar{\alpha}||\theta_0 - \hat{\theta}||_2^2 + \alpha||\theta_1 - \hat{\theta}||_2^2\}$$
$$\Rightarrow \inf_{\hat{\theta}} \{||\hat{\theta}||_2^2 - 2\hat{\theta}(\bar{\alpha}\theta_0 + \alpha\theta_1) + ||\bar{\alpha}\theta_0 + \alpha\theta_1||_2^2 - ||\bar{\alpha}\theta_0 + \alpha\theta_1||_2^2 + \bar{\alpha}||\theta_0||_2^2 + \alpha||\theta_1||_2^2\}$$
$$\Rightarrow \inf_{\hat{\theta}} \{\alpha\bar{\alpha}||\theta_0 - \theta_1||_2^2 + ||\hat{\theta} - (\bar{\alpha}\theta_0 + \alpha\theta_1)||_2^2\} = \alpha\bar{\alpha}||\theta_0 - \theta_1||_2^2$$

So we basically have the conditional mean as the estimate for the above problem which is intuitively correct. We now normalize (10.4) and use the above result to get:

$$\begin{aligned}\Rightarrow R_\pi^* &= \lambda\hat{\lambda}||\theta_0 - \theta_1||_2^2 \int_X \mu(dx) \frac{P_{\theta_0} P_{\theta_1}}{\bar{\lambda}P_{\theta_0} + \lambda P_{\theta_1}} \\ &= \lambda\hat{\lambda}||\theta_0 - \theta_1||_2^2 E_{\theta_0}\{\frac{P_{\theta_1}}{\bar{\lambda}P_{\theta_0} + \lambda P_{\theta_1}}\}\end{aligned}$$

Now, we observe that $\bar{\lambda}P_{\theta_0} + \lambda P_{\theta_1} \leq P_{\theta_0} \vee P_{\theta_1}$. Using this fact, we have:

$$\begin{aligned}R_\pi^* &\geq \lambda\hat{\lambda}||\theta_0 - \theta_1||_2^2 (\int_X \mu(dx)(P_{\theta_0} \vee P_{\theta_1})) \\ &= \frac{1}{4}||\theta_0 - \theta_1||_2^2 (1 - TV(P_{\theta_0}, P_{\theta_1}))\end{aligned}$$

where we used $\lambda = \bar{\lambda} = \frac{1}{2}$. □

10.2 Two-point method

For two-point method, we strip off the uncertainty by choosing only 2 possible values of the parameters. So we have:

$$\begin{aligned}\Rightarrow R_\pi^* &\geq R^*(\{\theta_0, \theta_1\}) \\ &= \sup_\pi R_\pi^*\end{aligned}$$

where the last equality follows from minimax theorem (which holds here since we consider a finite set of parameters). Now, for the optimal Bayes Risk we have:

$$\begin{aligned}\Rightarrow R_\pi^* &= \inf_{\hat{\theta}:\mathcal{X}\to\Theta} \bar{\lambda}E_{\theta_0} l(\theta_0, \hat{\theta}) + \lambda E_{\theta_1} l(\theta_1, \hat{\theta}) \\ &= E_{\theta_0} \inf_{\hat{\theta}:\mathcal{X}\to\Theta} \{\bar{\lambda} l(\theta_0, \hat{\theta}) + \lambda \frac{P_{\theta_1}}{P_{\theta_0}} l(\theta_1, \hat{\theta})\}\end{aligned}$$

Note: We could change the order of expectation and infimum in the above equation as the infimum is over $\hat{\theta}$ which depends only on data.

We now define $\bar{\lambda} l(\theta_0, \hat{\theta}) + \lambda \frac{P_{\theta_1}}{P_{\theta_0}} l(\theta_1, \hat{\theta}) \ = F(\frac{P_{\theta_1}}{P_{\theta_0}})$. Therefore, we have:

$$\Rightarrow R_\pi^* = E_{\theta_0}\{F(\frac{P_{\theta_1}}{P_{\theta_0}})\}$$

Example 10.1 (Quadratic Loss Function)**.** If $l(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_2^2$, then R_π^* = Expected value of a f-divergence between P_{θ_0} and P_{θ_1}.

We can choose a f-divergence which suits our needs.

So for two-point method, we have:

$$\begin{aligned}\Rightarrow R^*(\Theta) &\geq R^*(\{\theta_0, \theta_1\}) \\ &\geq \text{Function of (separation between } \theta_0 \text{ and } \theta_1, \text{ separation between } P_{\theta_0} \text{ and } P_{\theta_1})\end{aligned}$$

Remark 10.1. Since the separation between P_{θ_0} and P_{θ_1} is quantified using f-divergences, we can lower bound the minimax risk in terms of f-divergences other than total variation as well as follows:

- Using Le Cam's method, we can find a bound using total variation and then replace total variation with other f-divergences like χ^2 or hellinger distance.
- We can also use some other f-divergence directly instead of using total variation.

10.3 How good is Le Cam's bound?

In this section, we try to understand how tight Le Cam's bound is. To gain insight, we first consider the following example:

Example 10.2 (p-dimensional, n-sample Gaussian Location Model)**.** For p-dimensional, n-sample GLM, we use $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ as the estimate. So we have $\bar{X} \sim N(\theta, \frac{1}{n}I_p)$. We also know from previous lectures that for $l(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_2^2$, we have $R^* = \frac{p}{n}$. Let us compare this result with the lower bound calculated using Le Cam's method:

$$\begin{aligned}\Rightarrow R^* &\geq \sup_{\theta_0, \theta_1 \in R^p} \frac{1}{4}||\theta_0 - \theta_1||_2^2 (1 - TV(N(\theta_0, \frac{1}{n}I_p), N(\theta_1, \frac{1}{n}I_p))) \\ &= \sup_{\theta \in R^p} \frac{1}{4}||\theta||_2^2 (1 - TV(N(0, \frac{1}{n}I_p), N(\theta, \frac{1}{n}I_p)))\end{aligned}$$

where the last step follows from the fact that we can replace θ_0 by 0 and θ_1 by θ with out any loss of generality. Therefore, converting the above inequality to one involving standard normals, we have:

$$\Rightarrow R^* \geq \sup_{\theta \in R^p} \frac{1}{4n}||\theta||_2^2 (1 - TV(N(0, I_p), N(\theta, I_p)))$$

Clearly, the RHS above is independent of p, which is very poor since the lower bound doesn't scale with the dimension.

Note: Here it is easy to compute the total variation unlike other cases. We simply rotate the vector θ to reduce the problem to that of one-dimensional total variation calculation. We have:

$$\begin{aligned}\Rightarrow TV(N(0, I_p), N(\theta, I_p)) &= TV(N(0, I_p), N(||\theta||e, I_p) \\ &= TV(N(0, 1), N(||\theta||e, 1))\end{aligned}$$

where $||\theta||e$ is the component of θ left after rotating it. Hence, the calculation of total variation for this special case reduces to a one-dimensional problem. So we have:

$$\Rightarrow R^* \geq \sup_{s \geq 0} \frac{1}{4n} s^2 (1 - TV(N(0,1), N(s,1)))$$

How to scale R^* with p? We observe that we have considered a similar model as previous lectures and hence using tensorization of 1-dimensional n-sample GLM, we can conclude R^* should linearly grow in p. Explanation: Since $l(\theta, \hat{\theta}) = \sum_{i=1}^{p} l(\theta_i, \hat{\theta}_i)$, and each dimension of vector θ is estimated using corresponding dimension of the vector $\bar{X}$. Hence, as each dimension has a constant lower bound, the vector should have a lower bound scaling linearly with p as its lower bound is the sum of respective one-dimensional lower bounds. Therefore, we have $pR^*_{\pi_{1-d}} \leq R^*_p \leq pR^*_{1-d}$.

To improve upon the lower bound obtained using Le Cam's method, we consider more than two points to obtain the minimax bound. In next section, we shall discuss Assouad's Lemma which consider a hypercube instead of a line.

10.4 Assouad's Lemma

Lemma 10.1 (Assouad's Lemma)**.** *If each coordinate consists of binary testing, i.e.* $\theta \in \{0,1\}^p \subset \Theta = R^p$ *and* $l(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_1$*, then:*

$$\Rightarrow R^* \geq \frac{p}{4}(1 - \max_{d(\theta,\theta')=1} TV(P_\theta, P_{\theta'}))$$

Proof. Since minimiax risk is greater than Bayes risk, we have $\Rightarrow R^* \geq R^*_\pi$. Also we consider a uniform prior over $\{0,1\}^p$. We also define $\tilde{\theta}_i$ as follows:

$$\Rightarrow \tilde{\theta}_i = \begin{cases} 0, & \hat{\theta}_i < \frac{1}{2} \\ 1, & \text{otherwise} \end{cases}$$

Therefore, $\forall \hat{\theta} : X \to R^p$, we have:

$$\begin{aligned} \Rightarrow E||\theta - \hat{\theta}||_1 &= \sum_{i=1}^{p} E|\theta_i - \hat{\theta}_i| \\ &\geq \frac{1}{2} \sum_{i=1}^{p} E|\theta_i - \tilde{\theta}_i| \\ &= \frac{1}{2} \sum_{i=1}^{p} P(\theta_i \neq \tilde{\theta}_i) \\ &\geq \frac{1}{2} \sum_{i=1}^{p} \inf_{\hat{\theta}_i = \hat{\theta}_i(X)} P(\theta_i \neq \hat{\theta}_i) \end{aligned}$$

Since, $\theta_i \in \{0,1\}$, we have:

$$\Rightarrow E||\theta - \hat{\theta}||_1 \geq \frac{1}{4} \sum_{i=1}^{p} (1 - TV(P_{X|\theta_i=0}, P_{X|\theta_i=1})) \tag{10.5}$$

We now try to upper bound the total variation expression in the above inequality. From Bayes rule, we get:

$$\Rightarrow TV(P_{X|\theta_i=0}, P_{X|\theta_i=1}) = TV(\frac{1}{2^{p-1}} \sum_{\theta:\theta_i=1} P_\theta, \frac{1}{2^{p-1}} \sum_{\theta:\theta_i=0} P_\theta)$$

Using convexity of total variation, we have:

$$\begin{aligned} \Rightarrow\Rightarrow TV(P_{X|\theta_i=0}, P_{X|\theta_i=1}) &\leq \frac{1}{2^{p-1}} \sum_{\theta_{\setminus i} \in \{0,1\}^{p-1}} TV(P_{\{\theta_{\setminus i},1\}}, P_{\{\theta_{\setminus i},0\}}) \\ &\leq \max_{d(\theta,\theta')=1} TV(P_\theta, P_{\theta'}) \end{aligned}$$

Using the above result in (10.5) and using the fact that $l(\theta, \hat{\theta} = \sum_{i=1}^{p} l(\theta_i, \hat{\theta}_i)$, we get:

$$\Rightarrow R^* \geq \frac{l(0,1)p}{4}(1 - \max_{d(\theta,\theta')=1} TV(P_\theta, P_{\theta'}))$$

For l_1 loss function, $l(0,1) = 1$, hence we obtain the result. □

Example 10.3 (p-dimensional, n-sample Gaussian Location Model(GLM))**.** We consider $l(\theta, \hat{\theta}) = \sum_{i=1}^{p} (\theta_i - \hat{\theta}_i)^2, \theta \in \{0, \epsilon\}^p$. Using Assoud's Lemma, we get:

$$\begin{aligned} \Rightarrow R^* &\geq \frac{\epsilon^2 p}{4}\{1 - \max_{\theta,\theta' \in \{0,\epsilon\}^p, d(\theta,\theta')=1} TV(N(\theta, \frac{1}{n}I_p), N(\epsilon, \frac{1}{n}I_p))\} \\ &= \frac{\epsilon^2 p}{4}\{1 - TV(N(0, \frac{1}{n}I_p), N(\epsilon, \frac{1}{n}I_p))\} \end{aligned}$$

Using $\epsilon = \frac{1}{\sqrt{n}}$ and scaling by $\frac{1}{n}$, we get:

$$\Rightarrow R^* \geq \frac{kp}{n}$$

where $k = 1 - TV(N(0,1), N(1,1))$ is a constant (~ 0.7).

In the next lecture we will talk more about Assouad's Lemma which considers a hypercube of parameters. We will also introduce Fano's Lemma which uses a pyramid of parameters instead of a hypercube.