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

Lecture 13: Shannon lower bound, Fano's method

Lecturer: Yihong Wu *Scribe: Daewon Seo, Mar 8, 2016 [Ed. Mar 11]*

In the last class, we learned minimax risk bounding technique by data processing inequality of mutual information such that for $\theta - X - \hat{\theta}$,

$$\inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\ell(\theta,\hat{\theta})]\le R_\pi^*} I(\theta;\hat{\theta}) \le I(\theta;\hat{\theta}) \le I(\theta;X) \le \text{capacity} = \sup_{P_\theta} I(\theta;X). \tag{13.1}$$

Because the exact characterization of the LHS is intractable in most cases, we need an appropriate technique that further lower bounds the LHS, which is called the Shannon lower bound. Another technique to get a minimax lower bound, called Fano's method, will be discussed as well.

13.1 Shannon lower bound

13.1.1 Shannon lower bound

Suppose that the loss function is rth power of an arbitrary norm over $\mathbb{R}^p$, i.e., $\ell(\theta,\hat{\theta}) = \|\theta-\hat{\theta}\|^r$, and let $R_\pi^* = D$. Then, the LHS can be written as

$$\begin{aligned}
\inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\ell(\theta,\hat{\theta})]\le D} I(\theta;\hat{\theta}) &= \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\|\theta-\hat{\theta}\|^r\le D} I(\theta;\hat{\theta}) \\
&= \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\|\theta-\hat{\theta}\|^r\le D} h(\theta) - h(\theta|\hat{\theta}) \\
&= \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\|\theta-\hat{\theta}\|^r\le D} h(\theta) - h(\theta-\hat{\theta}|\hat{\theta}) \\
&\ge \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\|\theta-\hat{\theta}\|^r\le D} h(\theta) - h(\theta-\hat{\theta}) \\
&= h(\theta) - \sup_{\mathbb{E}\|W\|^r\le D} h(W) \triangleq \text{SLB}.
\end{aligned}$$

where $W \triangleq \theta - \hat{\theta}$ and the very last quantity is called the Shannon lower bound. To evaluate the supremum term, any convex optimization technique such as Lagrange multiplier can be applied.

A special case of the lower bound for Euclidean norm is given by

$$\text{SLB} = h(\theta) - \sup_{\mathbb{E}\|W\|_2^2\le D} h(W) = h(\theta) - h\left(\mathcal{N}\left(0, \frac{D}{P}I_P\right)\right) = h(\theta) - \frac{P}{2}\log\left(2\pi e\frac{D}{P}\right),$$

where we used the fact that Gaussian maximizes differential entropy when the second moment is bounded.

Theorem 13.1 (Shannon's Lower Bound)**.** *Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^p$ and $r > 0$. Then*

$$\inf_{P_{\hat{\theta}|\theta}:\mathbb{E}\|\theta-\hat{\theta}\|^r\le D} I(\theta;\hat{\theta}) \ge h(\theta) - \log\left\{V_p \cdot \left(\frac{Dre}{p}\right)^{\frac{p}{r}} \cdot \Gamma\left(1+\frac{p}{r}\right)\right\},$$

where V_p is the volume of the unit radius ball, i.e.,

$$V_p \triangleq vol(B_{\|\cdot\|}) = vol(\{x \in \mathbb{R}^p : \|x\| \leq 1\}).$$

The proof will be given in homework.

Note: The Shannon lower bound is asymptotically tight as $D \to 0$.

Example 13.1 (GLM)**.** Consider the p-dimensional n-sample GLM, i.e., $(X_1, \cdots, X_n) \overset{iid}{\sim} \mathcal{N}(\theta, I_p)$ or equivalently $\bar{X} \sim \mathcal{N}\left(\theta, \frac{1}{n} I_p\right)$. Then the minimax risk with respect to $\| \cdot \|^r$ is

$$R^* \gtrsim \frac{1}{(cn)^{r/2}} V_p^{-r/p}.$$

Proof. Take a prior $\theta \sim \pi = \mathcal{N}(0, sI_p)$. Then the inequality chain (13.1) is rewritten as

$$\begin{aligned}\frac{p}{2}\log(1+ns) \geq I(\theta, X) \geq I(\theta; \hat{\theta}) &\geq \inf_{P_{\hat{\theta}|\theta}: \mathbb{E}[\ell(\theta, \hat{\theta})] \leq R_\pi^*} I(\theta; \hat{\theta}) \\ &\geq \text{SLB} = \frac{p}{2}\log(2\pi e s) - \log\left\{V_p \cdot \left(\frac{R_\pi^* r e}{p}\right)^{\frac{p}{r}} \cdot \Gamma\left(1 + \frac{p}{r}\right)\right\}.\end{aligned}$$

Then, rearranging terms, taking limit $s \to \infty$, and using the Stirling's formula we get

$$R_\pi^* \gtrsim \frac{1}{(cn)^{r/2}} V_p^{-r/p} \Rightarrow R^* \gtrsim \frac{1}{(cn)^{r/2}} V_p^{-r/p}. \tag{13.2}$$

□

Note that for $r = 2$,

$$R^* \gtrsim \frac{1}{n} V_p^{-2/p},$$

while the exact bound (see Sec. 3.2) is $R^* = \frac{\mathbb{E}\|Z\|^2}{n} = \frac{p}{n}$. In the next example, we will see volumes for ℓ_q norm.

Example 13.2 (ℓ_q-norm)**.** Consider ℓ_q-norm, i.e., for $1 \leq q \leq \infty$

$$\|x\|_q = \left(\sum_{i=1}^p |x_i|^q\right)^{1/q}.$$

See the volume for several q's.

- ($q = 2$) (Cont'd from the previous) Note that $R^* = \frac{p}{n}$ for the quadratic loss $\| \cdot \|_2^2$. The n-dimensional volume of a unit Euclidean ball B_2 is given by

$$V_p(B_2)^{1/p} = \frac{\pi^{1/2}}{\left(\Gamma\left(1 + \frac{p}{2}\right)\right)^{1/p}} \asymp \frac{1}{\sqrt{p}},$$

which follows from the Stirling's approximation,

$$\left(\Gamma\left(1+\frac{p}{2}\right)\right)^{1/p} \asymp \left(\left(\frac{p}{2e}\right)^{p/2}\left(\frac{p}{2}\right)^{1/2}\right)^{1/p} \asymp \left(\frac{p}{2e}\right)^{1/2}\left(\frac{p}{2}\right)^{1/2p} \asymp \sqrt{p}.$$

Plugging in (13.2) with $r = 2$,

$$R^* \gtrsim \frac{1}{n} V_p^{-1/2} = \frac{p}{n}.$$

Hence in this case the SLB is tight.

- ($1 \leq q < \infty$) Consider ℓ_q norm, where $1 \leq q < \infty$, the volume of a unit ℓ_q ball is given by

  $$V_p(B_q) = \frac{\left[2\Gamma\left(1+\frac{1}{q}\right)\right]^p}{\Gamma\left(1+\frac{p}{q}\right)}.$$

  So using (13.2) and the Stirling's formula, the minimax bound for a loss function $\|\cdot\|_q^2$ is given by

  $$R^* \gtrsim \frac{p^{2/q}}{n}.$$

  Another way to get the same bound is that

  $$R^* \gtrsim \frac{1}{n}\mathbb{E}\|Z\|_q^2 \asymp \frac{p^{2/q}}{n}.$$

  Here the property that if $Z \sim \mathcal{N}(0, I_p)$, $\|Z\|_q^q = \Theta_P(p)$ is used.

- ($q = \infty$) Recall a unit hypercube in $\mathbb{R}^p$. Then, $V_p(B_\infty) = 2^p$, hence, $R^* \gtrsim \frac{1}{n}$ by the SLB. On the other hand, we know the exact risk,

  $$R^* = \frac{1}{n}\mathbb{E}\|Z\|_\infty^2 \asymp \frac{\log p}{n}.$$

  So in this case the SLB is not tight. Here, the equality follows from the fact that if $Z \sim \mathcal{N}(0, I_p)$, $\|Z\|_\infty = \Theta_P(\sqrt{\log p})$.

Note: In the case that we have restriction on θ such that $\theta \in \Theta \subset \mathbb{R}^p$, where Θ is a convex set with non-empty interior, the only thing to be changed is the SLB part. Upper bound by capacity remains unchanged. As an example of uniform prior over some $\Theta \subset \mathbb{R}^p$,

$$\text{capacity} \geq \text{SLB} = h(\theta) - \log[\cdots R_\pi^* \cdots] = \log vol(\Theta) - \log[\cdots R_\pi^* \cdots].$$

We get the bound of minimax risk connecting this SLB with capacity formula.

Also note that the exact characterization of $R^*(\Theta)$ is open even for a convex set Θ.

13.1.2 Gaussian width of a convex body K

Suppose $Z \sim N(0, I_p)$ and a set K is convex and symmetric. Define the *Gaussian width* of K

$$w(K) \triangleq \mathbb{E}\left[\sup_{x \in K} \langle x, Z\rangle\right].$$

Lemma 13.1 (Urysohn)**.**

$$\mathrm{vol}(K)^{1/p} \lesssim \frac{w(K)}{p}.$$

Urysohn's lemma helps us characterize the bound of minimax risk. In our case, $K = B_{\|\cdot\|}$, then

$$w(K) = \mathbb{E}\left[\sup_{x \in K} \langle x, Z\rangle\right] = \mathbb{E}\left[\sup_{\|x\| \le 1} \langle x, Z\rangle\right] = \mathbb{E}\|Z\|_*,$$

which is in fact the expected *dual norm* of Z. From the lemma, we have $V_p^{1/p} \lesssim \frac{\mathbb{E}\|Z\|_*}{p}$. Therefore,

$$R^* \gtrsim \frac{1}{n} V_p^{-2/p} \gtrsim \frac{1}{n}\left(\frac{p}{\mathbb{E}\|Z\|_*}\right)^2.$$

13.2 Fano's method

Recall the inequality chain,

$$\inf_{P_{\hat\theta|\theta} : \mathbb{E}\|\hat\theta - \theta\| \le R_\pi^*} I(\theta; \hat\theta) \le I(\theta; \hat\theta) \le I(\theta; X) \le \text{capacity}.$$

In this section, we discuss Fano's method that reduces the LHS to multiple hypothesis testing problem, which is easier to compute.

The steps are followings:

1. (Discretize) Instead of Θ, consider a discrete subset $\tilde\Theta = \{\theta_1, \cdots, \theta_n\} \subset \Theta$. Points are picked to satisfy $\|\theta_i - \theta_j\| \ge \epsilon$ for all $i \ne j$. Figure 13.1 visualizes this discretization.

2. (Reduce to multiple hypothesis testing) Assume uniform prior such that $\theta \sim \pi = \mathrm{unif}(\{\theta_1, \cdots, \theta_n\})$ and let f be a quantizer that maps $\theta \in \Theta$ to $\theta_i \in \tilde\Theta$, the closest point to θ. Note that $f(\theta) = \theta$ because θ is drawn over $\tilde\Theta$. So by data processing inequality for $\theta - X - \hat\theta - f(\hat\theta)$,

$$I(\theta; \hat\theta) \ge I(\theta; f(\hat\theta)).$$

   Note that $I(\theta; f(\hat\theta))$ is a function of joint probability mass over discrete space $\tilde\Theta \times \tilde\Theta$.

   Let's see the error events $\{\theta \ne f(\hat\theta)\}$. Let say the true source is $\theta = \theta_k$. If error happens, it implies our estimate $\hat\theta$ closer to $\theta_j = f(\hat\theta)$ than θ_k for some j. In other words, if error happens,

$$\|\hat\theta - f(\hat\theta)\| \le \|\hat\theta - \theta_k\|.$$

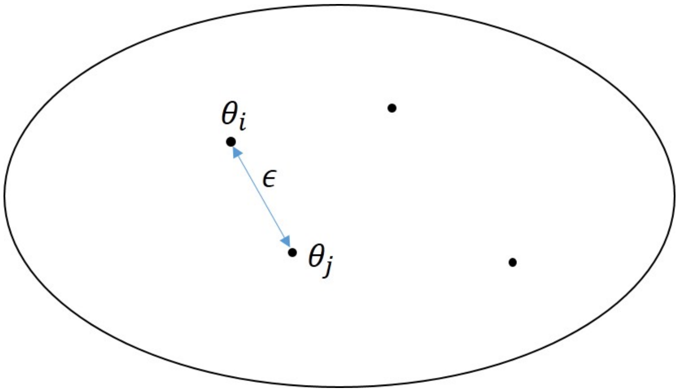

Figure 13.1: Discretization

So due to triangular inequality, the error event implies

$$\begin{aligned}\epsilon \le \|f(\hat\theta) - \theta_k\| &= \|f(\hat\theta) - \hat\theta + \hat\theta - \theta_k\| \le \|f(\hat\theta) - \hat\theta\| + \|\hat\theta - \theta_k\| \\ &\le 2\|\hat\theta - \theta_k\|, \\ \Rightarrow \quad \frac{\epsilon}{2} &\le \|\hat\theta - \theta_k\|.\end{aligned}$$

Hence,

$$\begin{aligned} P_e \triangleq \Pr(\theta_k \neq f(\hat\theta)) \le \Pr\left(\|\hat\theta - \theta_k\| \ge \frac{\epsilon}{2}\right) \le \frac{\mathbb{E}\|\hat\theta - \theta_k\|}{\epsilon/2} \le \frac{R_\pi^*}{\epsilon/2} = \frac{2R_\pi^*}{\epsilon} \\ \Rightarrow \inf_{P_{\hat\theta|\theta}:\mathbb{E}\|\hat\theta - \theta\| \le R_\pi^*} I(\theta;\hat\theta) \ge \inf_{P_e \le \frac{2R_\pi^*}{\epsilon}} I(\theta;\hat\theta) \ge \inf_{P_e \le \frac{2R_\pi^*}{\epsilon}} I(\theta; f(\hat\theta)).\end{aligned}$$

So, we reduce the LHS to a multiple hypothesis test problem where $\theta, f(\hat\theta)$ are both discrete.

3. (Apply Fano's inequality) Recall the data processing inequality for KL divergence by Figure 13.2. Here our processor is $1\{\theta \neq \hat\theta\}$, and we can further lower bound as

$$\begin{aligned} I(\theta; f(\hat\theta)) &= D\left(P_{\theta, f(\hat\theta)} \| P_\theta P_{f(\hat\theta)}\right) \\ &\ge D\left(\text{Bern}(P_e) \| \text{Bern}\left(1 - \frac{1}{n}\right)\right) \\ &= P_e \log \frac{P_e}{1 - \frac{1}{n}} + (1 - P_e) \log \frac{1 - P_e}{\frac{1}{n}} \\ &= -h(P_e) + \log n - P_e \log(n-1) \\ &\ge -\log 2 + \log n - P_e \log n, \\ \Rightarrow P_e &\ge 1 - \frac{I(\theta; f(\hat\theta)) + \log 2}{\log n},\end{aligned}$$

where $h(\cdot)$ is a binary entropy function. So finally we reach the bound

$$\begin{aligned} \frac{2R^*}{\epsilon} \ge \frac{2R_\pi^*}{\epsilon} \ge P_e \ge 1 - \frac{I(\theta; f(\hat\theta)) + \log 2}{\log n} \\ \Rightarrow \quad R^* \ge \frac{\epsilon}{2}\left(1 - \frac{I(\theta; f(\hat\theta)) + \log 2}{\log n}\right).\end{aligned}$$

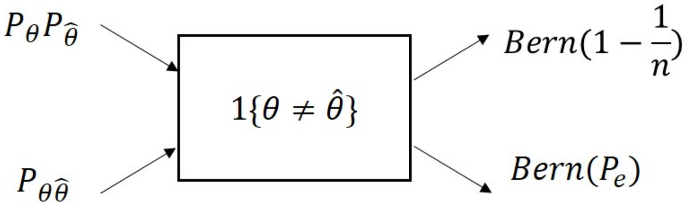

Figure 13.2: Data processing kernel for Fano's inequality

Note: The situation of the Fano's inequality in class is that

1. θ uniformly takes M values.
2. Markov chain $\theta - X - \hat\theta$ holds.

Then, the Fano's inequality says that

$$\begin{aligned} I(\theta; X) &\geq -\log 2 + \log M - P_e \log(M-1) \\ &\geq -\log 2 + (1-P_e)\log M, \\ \Rightarrow P_e &\geq 1 - \frac{I(\theta;X)+\log 2}{\log M}. \end{aligned}$$

The Fano's inequality intuitively means that when the mutual information is fixed, P_e cannot be less than a certain value. On the other hand, when P_e is fixed, the mutual information must be greater than a certain value.

Note: We can also use the Fano inequality as following:

$$I(\theta; X) \geq \min_{P_e \leq \frac{2R_\pi^*}{\epsilon}} I(\theta; X),$$

and similarly as above,

$$\begin{aligned} \frac{2R^*}{\epsilon} \geq \frac{2R_\pi^*}{\epsilon} &\geq P_e \geq 1 - \frac{I(\theta;X)+\log 2}{\log n} \\ \Rightarrow \quad R^* &\geq \frac{\epsilon}{2}\left(1 - \frac{I(\theta;X)+\log 2}{\log M}\right). \end{aligned}$$

Note: If the loss function is $\|\cdot\|^2$,

$$\min_{P_e \leq \frac{2R_\pi^*}{\epsilon}} I(\theta; X) \quad \Rightarrow \quad R^* \geq \left(\frac{\epsilon}{2}\right)^2 \left(1 - \frac{I(\theta;X)+\log 2}{\log M}\right).$$