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

Lecture 6: Variational representation, HCR and CR lower bounds.

Lecturer: Yihong Wu *Scribe: Georgios Rovatsos, Feb 11, 2016 [Ed. Apr 20]*

Last lecture we discussed systematic methods to find the best inequalities between different f-divergence via their joint range. We showed that examining the binary cases is sufficient to derive optimal inequalities. In this lecture we will further discuss lower bounds for statistical estimation using f-divergences.

Outline:

- Variational representation of f-divergences.
  - Convexity.
  - Lower semi-continuity.
- (Specializing to χ^2) Lower bounds for statistical estimation.
  - Hammersley-Chapman-Robbins (HCR) lower bound.
  - Cramér-Rao (CR) lower bound.
  - Bayesian Hammersley-Chapman-Robbins (HCR) lower bound.
  - Bayesian Cramér-Rao (CR) lower bound.

6.1 Variational representation of f-divergences

We begin with an example regarding the total variation metric.

Example 6.1 (Total variation)**.** Let $(\mathcal{X}, \mathcal{F})$ a measure space and P, Q two probability distributions. In previous lectures we saw how by choosing $f(x) = \frac{1}{2}|x-1|$ the f-divergence becomes the total variation metric. In particular, we saw that:

$$d_{\mathrm{TV}}(P,Q) = D_f(P\|Q) = \frac{1}{2}\int |P-Q| = \sup_{E\in\mathcal{F}} |P(E)-Q(E)| = \frac{1}{2}\sup_{\|f\|_\infty \le 1} |\mathbb{E}_P f(x) - \mathbb{E}_Q f(x)|.$$

It should be noted that the requirement of f to be convex in the definition of f-divergence is essential. In Euclidean spaces any convex function can be represented as the pointwise supremum of a family of affine functions and vice versa, every supremum of a family of affine functions produces a convex function. Take $f(x) = \frac{1}{2}|x-1|$ as an example. We see that it can be written as a pointwise supremum of $f_1(x) = \frac{1}{2}(x-1)$ and $f_2(x) = \frac{1}{2}(1-x)$. This remark can be used not only as a geometric interpretation of convex functions but as a definition of convexity. For f-divergences which are convex functions of probability measures, its variational representation amounts to writing it as a pointwise supremum of affine functions.

6.1.1 Convex conjugate

Let $f : (0, +\infty) \to \mathbb{R}$ be a convex function. The convex conjugate f^* of f is defined by:

$$f^*(y) = \sup_{x \in \mathbb{R}} [xy - f(x)]. \tag{6.1}$$

Two important properties of the convex conjugates are

1. f^* is also convex (which holds regardless of f being convex or not);
2. Biconjugation: $(f^*)^* = f$.

In particular, the definition of f^* yields the following (Young-Fenchel inequality)

$$f(x) \geq xy - f^*(y), \tag{6.2}$$

where the last inequality holds for any y.

Using the notion of convex conjugate, we obtain a variational representation of f-divergence in terms of the convex conjugate of f:[1]

$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right] = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))], \tag{6.3}$$

where g is such that both expectations are finite (of course). This representation is insightful for many reasons. For example, we get the following properties for free:

1. Convexity: First of all, note that $D_f(P\|Q)$ is expressed as a supremum of affine functions (since the expectation is a linear operation). As a result, we get that $(P, Q) \mapsto D_f(P\|Q)$ is convex, which was proved in previous lectures using different method.

2. Weak lower semicontinuity: We begin with an example. Assume $\{X_i\}$ are i.i.d. Rademachers (± 1). Then, by the central limit theorem we have that

   $$\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \xrightarrow{\mathrm{D}} \mathcal{N}(0, 1);$$

   however,

   $$D_f\left(\frac{P_{X_1+X_2+\ldots+X_n}}{\sqrt{n}} \middle\| \mathcal{N}(0,1)\right) \not\to 0,$$

   since the former distribution is discrete and the latter is continuous. Therefore the best we can hope for f-divergence is semicontinuity. Indeed, if $\mathcal{X}$ is a nice space (e.g., Euclidean space), in (6.3) we can restrict the function g to continuous bounded functions, in which case $D_f(P\|Q)$ is expressed as a supremum of weakly continuous functionals (note that $f^* \circ g$ is also continuous and bounded since f^* is continuous) and is hence weakly lower semi-continuous, i.e., for any sequence of distributions P_n and Q_n such that $P_n \xrightarrow{w} P$ and $Q_n \xrightarrow{w} Q$, we have

   $$\liminf_{n\to\infty} D_f(P_n\|Q_n) \geq D_f(P\|Q).$$

[1] Equivalently, one can take the supremum over all kernels $P_{Z|X}$ where Z is $\mathbb{R}$-valued.

Example 6.2 (Total variation)**.** By using $f(x) = \frac{1}{2}|x-1|$ in the formula of f-divergence we get the total variation metric given by

$$d_{TV}(P,Q) = \frac{1}{2}\int |P-Q|.$$

By using the definition of convex conjugate it is easy to see that

$$f^*(y) = \sup_x \left\{ xy - \frac{1}{2}|x-1| \right\} = \begin{cases} +\infty & \text{if } |y| > \frac{1}{2} \\ y & \text{if } |y| \leq \frac{1}{2} \end{cases}$$

Thus (6.3) gives

$$d_{TV}(P,Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))] = \sup_{g:|g|\leq\frac{1}{2}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)], \tag{6.4}$$

where in the last equality we restricted the supremum to functions bounded by $1/2$, since any other function would make the term inside the supremum equal to $-\infty$.

Example 6.3 (KL-divergence)**.** By using $f(x) = x\log x$ in the formula of f-divergence we get the KL-divergence

$$D(P\|Q) = \mathbb{E}_P\left[\log \frac{P}{Q}\right].$$

By using differentiation to find the supremum it is easy to see that $f^*(y) = e^{y-1}$. Plugging in the formula of f-divergence we get

$$D(P\|Q) = 1 + \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[e^{g(X)}]. \tag{6.5}$$

In comparison, the famous Donsker-Varadhan representation is

$$D(P||Q) = \sup_g \mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[e^{g(X)}], \tag{6.6}$$

which is stronger than (6.5) in the sense that for each g, the RHS of (6.6) is at least that of (6.5), since $\log(1+t) \leq t$.

Example 6.4 (χ^2-divergence)**.** By using $f(x) = (x-1)^2$ in the formula of f-divergence we get the χ^2-divergence

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

By using differentiation to find the supremum it is easy to see that $f^*(y) = y + \frac{y^2}{4}$. Hence

$$\chi^2(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q\left[g(X) + \frac{g^2(X)}{4}\right].$$

Finally by a change of variable $h(x) = \frac{1}{2}g(x) + 1$ we get

$$\chi^2(P\|Q) = \sup_{h:\mathcal{X}\to\mathbb{R}} 2\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h^2(X)] - 1. \tag{6.7}$$

This result is very important and will be used repeatedly for the derivation of the Hammersley-Chapman-Robbins (HCR) lower bound as well as their Bayesian version in the next section.

6.2 Hammersley-Chapman-Robbins (HCR) lower bound

In this section, we apply the variational representation for the χ^2-divergence to probability distributions P and Q on $\mathbb{R}$.[2] By limiting the choice of function h to affine functions, the equality (6.7) becomes an inequality. In particular, let $h(x) = ax + b$ and optimize over $a, b \in \mathbb{R}$, we have

$$\chi^2(P\|Q) \geq \sup_{a,b\in\mathbb{R}} \left\{ 2(a\mathbb{E}_P(X) + b) - \mathbb{E}_Q[(aX + b)^2] - 1 \right\} = \frac{(\mathbb{E}_P[X] - \mathbb{E}_Q[X])^2}{\mathrm{var}_Q(X)}. \tag{6.8}$$

Note: The inequality (6.8) can be interpreted as follows: On the left hand side of the inequality we have the χ^2-divergence, a measure of the dissimilarity between two distributions. Looking at the right hand side we see that if the two distributions are centered at very distant locations, then the right hand side will be large. Due to (6.8), this will lead to a bigger χ^2-divergence something that was in fact expected.

The reason that the variance with respect to the Q distribution appears in the denominator is to quantify how different the two means are *relatively*. Indeed, the standard deviation must appear as a normalizing factor because the LHS is a numerical number. Also, the bound only involves the variance under Q not P, which is consistent with the asymmetry of χ^2-divergence.

Using (6.7) we now derive the HCR lower bound on the variance of an estimator (possibly randomized). To this end, assume that data $X \sim P_\theta$, where $\theta \in \Theta \subset \mathbb{R}$. We use quadratic cost to quantify the difference between the real and the predicted parameter, i.e., $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Then the risk of estimator $\hat{\theta}$ when the real parameter is θ is given by $R_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\theta - \hat{\theta})^2]$. Now, fix $\theta \in \Theta$. For any other $\theta' \in \Theta$ we will use (6.8) with $Q = P_\theta$ and $P = P_{\theta'}$. As a result we have that

$$\chi^2(P_{\theta'}\|P_\theta) \geq \chi^2(P_{\hat{\theta}}\|Q_{\hat{\theta}}) \geq \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\mathrm{var}_\theta(\hat{\theta})}$$

Where the first inequality arises by using the data processing inequality and the second inequality by (6.8). Finally, by swapping the denominator with the left hand side and taking the supremum over all $\theta' \neq \theta$, and since $\mathrm{var}_\theta(\hat{\theta})$ is not a function of θ', we derive the final result.

Theorem 6.1 (Hammersley-Chapman-Robbins (HCR) lower bound)**.** *For the quadratic loss, any estimator $\hat{\theta}$ satisfies*

$$R_\theta(\hat{\theta}) \geq \mathrm{var}_\theta(\hat{\theta}) \geq \sup_{\theta'\neq\theta} \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\chi^2(P_{\theta'}\|P_\theta)}, \quad \forall\theta \in \Theta. \tag{6.9}$$

6.3 Cramér-Rao (CR) lower bound

We now derive the Cramér-Rao lower bound as a consequence of the HCR lower bound. To this end, we restrict the problem to unbiased estimators, where an estimator $\hat{\theta}$ is said to be unbiased if $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Theta$. Then by applying the HCR lower bound we have that

$$\mathrm{var}_\theta(\hat{\theta}) \geq \sup_{\theta'\neq\theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'}\|P_\theta)} \geq \lim_{\theta'\to\theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'}\|P_\theta)}. \tag{6.10}$$

[2]This can always be assumed by allowing the likelihood ratio function $\frac{dP}{dQ}$ which is a sufficient statistic.

Here, we bypass the supremum by sending θ' to θ. However, when $\theta' \to \theta$ both the numerator and denominator will go to zero. Doing this, we hope that the denominator will go to zero quadratically as the numerator does. Remember that

$$\chi^2(P_{\theta'} \| P_\theta) = \int \frac{(P_\theta - P_{\theta'})^2}{P_\theta}.$$

Then by using the Taylor expansion for P_θ around θ' we get that

$$P_\theta - P_{\theta'} = (\theta - \theta')\frac{dP_\theta}{d\theta} + o[(\theta - \theta')^2],$$

for θ near θ'. Combining the above while ignoring the little-o terms we get that

$$\chi^2(P_{\theta'} \| P_\theta) = (\theta - \theta')^2 \int \frac{(\frac{dP_\theta}{d\theta})^2}{P_\theta}.$$

Plugging back in (**??**) we get the CR lower bound.

Theorem 6.2 (Cramér-Rao (CR) lower bound)**.** *For any unbiased estimator $\hat{\theta}$ and any $\theta \in \Theta$*

$$\mathrm{var}_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)},$$

where $I(\theta)$ is the Fisher information given by

$$I(\theta) = \int \frac{(\frac{dP_\theta}{d\theta})^2}{P_\theta}.$$

An intuitive interpretation of $I(\theta)$ is that it is a measure of the information the data contains for the estimation of the parameter when its true value is θ.

Example 6.5 (GLM)**.** Let $\theta \in \mathbb{R}$ and $X \sim P_\theta = \mathcal{N}(\theta, 1)$. Define the standard normal distribution by $\Phi(x)$. Note that $P_\theta(x) = \Phi(x - \theta)$. Next we calculate the Fisher information. By shifting x to θ, note that

$$I(\theta) = \int \frac{(\frac{dP_\theta(x)}{d\theta})^2}{P_\theta(x)} dx = \int \frac{(\frac{d}{d\theta}\Phi(x-\theta))^2}{\Phi(x-\theta)} dx = I(0).$$

Thus, $I(\theta) = I(0)$ for all $\theta \in \Theta$. In general, in any case where we have the model $X = \theta + Noise$, where the noise is standard normal (location model) we have that the fisher information is the same everywhere.

Remark

Another useful way of seeing the Fisher information is the following:

$$I(\theta) = \int \frac{(\frac{\partial P_\theta(x)}{\partial\theta})^2}{P_\theta(x)} \partial x = \mathbb{E}_\theta\left[\left(\frac{\frac{\partial P_\theta(X)}{\partial\theta}}{P_\theta(X)}\right)^2\right] = \mathbb{E}_\theta\left[\left(\frac{\partial \log P_\theta(X)}{\partial\theta}\right)^2\right] = \mathrm{var}_\theta\left[\frac{\partial \log P_\theta(X)}{\partial\theta}\right],$$

where the last equality holds after noticing that

$$\mathbb{E}_\theta\left[\frac{\partial \log P_\theta(X)}{\partial\theta}\right] = 0.$$

6.4 Biased estimators

Many times restricting ourselves to unbiased estimators proves to be very limiting. As a result, biased estimators need to be considered. Then it is useful to see how the HCR bound can be applied in this case. Define the bias of an estimator $\hat{\theta}$ by $b(\theta) = \mathbb{E}_\theta[\hat{\theta}] - \theta$. Assuming the risk function is quadratic it is easy to see that for a biased estimator by directly using HCR then

$$R_\theta(\hat{\theta}) = b^2(\theta) + \mathrm{var}_\theta(\hat{\theta}) \geq b^2(\theta) + \sup_{\theta' \neq \theta} \frac{(b(\theta') + \theta' - b(\theta) - \theta)^2}{\chi^2(P_{\theta'} \| P_\theta)}.$$

By using the same taylor expansion trick and assuming that $b(\theta)$ is differentiable we finally get that for an estimator $\hat{\theta}$ and any $\theta \in \Theta$

$$R_\theta(\hat{\theta}) \geq b^2(\theta) + \frac{(1 + b'(\theta))^2}{I(\theta)}.$$

Using this inequality we can find a lower bound on the worst case mini-max risk. In particular, we have that

$$R^* = \inf_{\hat{\theta}} \sup_\theta R_\theta(\hat{\theta}) \geq \inf_b \left[\sup_\theta \left(b^2(\theta) + \frac{(1 + b'(\theta))^2}{I(\theta)} \right) \right],$$

where in the last inequality we also used the fact that the choice of the estimator affects our quantity only through the bias.

6.5 Bayesian CR lower bound

Previously in this lecture we used the HCR bound to derive the CR lower bound. In order to derive the Bayesian version of the CR lower bound a similar approach can be used: first prove the Bayesian HCR and then derive the Bayesian CR lower bound as a result.

Theorem 6.3 (Bayesian CR lower bound)**.** *Assume that the loss function is quadratic, i.e., $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Also, for any estimator $\hat{\theta}$ (possibly randomized), and for any prior $\pi \in M(\Theta)$ define the Bayes risk of $\hat{\theta}$ by $R_\pi(\hat{\theta}) = \int R_\theta(\hat{\theta})\pi(d\theta) = \int \mathbb{E}_\theta[(\hat{\theta} - \theta)^2]\pi(d\theta)$. Then we have that*

$$R^*_\pi = \inf_{\hat{\theta}} R_\pi(\hat{\theta}) \geq \frac{1}{\mathbb{E}_{\theta \sim \pi}[I(\theta)] + I(\pi)},$$

where $I(\pi)$ the Fisher information of π, i.e.,

$$I(\pi) = \int \frac{(\pi'(\theta))^2}{\pi(\theta)} d\theta.$$