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

Lecture 11: Mutual Information Method

Lecturer: Yihong Wu *Scribe: Jaeho Lee, Mar 1, 2016 [Ed. Mar 9]*

Quick review: Assouad's lemma

In Assouad's lemma discussed in the last lecture, we made the following assumptions:

- Parameter space is a hypercube embedded in $\mathbb{R}^p$, i.e. $\Theta = \{\theta_0, \theta_1\}^p$.
- Loss function $l(\theta, \hat{\theta})$ is *separable*, i.e. $l(\theta, \hat{\theta}) = \sum_{i=1}^p l(\theta_i, \hat{\theta}_i)$, (e.g. Hamming, ℓ_2 squared.)
- and satisfies α-triangle inequality, i.e. $l(\theta_0, \theta_1) \le \alpha \left[l(\theta_0, \hat{\theta}) + l(\theta_1, \hat{\theta}) \right]$.

Letting $\pi \sim \text{Unif}(\Theta)$, we could proceed as:

$$\begin{aligned}
\mathcal{R}^* \ge \mathcal{R}^*_\pi = \inf_{\hat{\theta}} \sum_{i=1}^p \mathbb{E}_{\theta\sim\pi}[l(\theta_i, \hat{\theta}_i)] &= \sum_{i=1}^p \inf_{\hat{\theta}_i} \mathbb{E}_{\theta\sim\pi}[l(\theta_i, \hat{\theta}_i)] \\
&\overset{\text{Le Cam}}{\ge} \sum_{i=1}^p \frac{l(\theta_0, \theta_1)}{4\alpha} \left[1 - \text{TV}(P_{X|\theta_i=\theta_0}, P_{X|\theta_i=\theta_1}) \right] \\
&\overset{\text{convexity}}{\ge} \frac{p \cdot l(\theta_0, \theta_1)}{4\alpha} \left[1 - \max_{d_H(\theta,\theta')=1} \text{TV}(P_\theta, P_{\theta'}) \right].
\end{aligned}$$

where the last line could be thought of as a "deteriorated" version.

Example 11.1 (Gaussian Location Model)**.** As usual, let $Z \sim \mathcal{N}(0, I_p)$, $\Theta = \left\{0, \frac{1}{\sqrt{n}}\right\}^p$, and $l(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$ (which satisfies 2-triangle inequality). Then, Assouad's lemma gives us:

$$\mathcal{R}^* \ge \frac{p}{8n} \left[1 - \max_{d_H(\theta,\theta')=1} \text{TV}(P_\theta, P_{\theta'}) \right] = \frac{p}{8n} \left[1 - \text{TV}\left(\mathcal{N}\left(0, \frac{1}{n}\right), \mathcal{N}\left(\frac{1}{\sqrt{n}}, \frac{1}{n}\right) \right) \right] \simeq \frac{0.3}{8} \frac{p}{n}$$

which is not very good compared to $\frac{p}{n}$.

Along with the above example, the fact that the loss function is not always separable (e.g. ℓ_∞) necessitates a search a more versatile method. In this lecture, we discuss the "mutual information method" where the most important measure of information would be, of course, the mutual information $I(X;Y)$.

11.1 Mutual Information $I(X;Y)$

Recall that KL-divergence was defined using the function $f(x) = x \log x$:

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

Now, the mutual information can be defined using KL-divergence.

Definition 11.1 (Mutual Information)**.** Given a joint probability distribution P_{XY}, the mutual information between X and Y is defined as

$$I(X;Y) \triangleq D(P_{XY} \| P_X P_Y),$$

the distance between the original distribution and the hypothetical distribution assuming that X and Y are independent.

Mutual information has the following useful properties:

Proposition 11.1 (Properties of Mutual Information)**.** *Followings are true:*

1. $I(X;Y) = D(P_{Y|X} \| P_Y | P_X) = \mathbb{E}_{x \sim P_X}[D(P_{Y|X=x} \| P_Y)]$
2. *(Symmetry)* $I(X;Y) = I(Y;X)$.
3. *(Measure of dependency)* $I(X;Y) \geq 0$ *with equality iff* $X \perp\!\!\!\perp Y$.
4. *(I vs H: Y discrete)* $I(X;Y) = H(Y) - H(Y|X)$*, where* $H(Y)$ *denotes the Shannon entropy* $H(Y) \triangleq \sum_y P_Y(y) \log \frac{1}{P_Y(y)}$.
5. *(I vs h: Y continuous)* $I(X;Y) = h(Y) - h(Y|X)$*, where* $h(Y)$ *denotes the differential entropy* $h(Y) \triangleq \int f_Y(y) \log \frac{1}{f_Y(y)} dy$.

Example 11.2 (Additive noise: binary)**.** Let $Y = X \oplus Z$, where $X \sim \text{Bern}(\delta)$, $Z \sim \text{Bern}(\epsilon)$, $X \perp\!\!\!\perp Z$, and $\oplus$ denotes the XOR operation (binary addition). Then,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) = H(Y) - H(X \oplus Z|X) = H(Y) - H(Z) \\ &= H(\text{Bern}(\delta \star \epsilon)) - H(\text{Bern}(\epsilon)) = h(\delta \star \epsilon) - h(\epsilon). \end{aligned}$$

where $h(t) = t \log \frac{1}{t} + (1-t) \log \frac{1}{1-t}$ is the binary entropy function (not differential entropy!) and the convolution operation $\star$ on $[0,1]$ for Bernoulli random variables is $\delta \star \epsilon = \delta\bar{\epsilon} + \bar{\delta}\epsilon$.

Example 11.3 (Additive noise: Gaussian)**.** Let $Y = X + Z$ where $X \sim \mathcal{N}(0, S)$, $Z \sim \mathcal{N}(0,1)$, $X \perp\!\!\!\perp Z$. Then,

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) = h(Y) - h(Z) \\ &= h(\mathcal{N}(0, 1+S)) - h(\mathcal{N}(0,1)) = \frac{1}{2} \log(1+S). \end{aligned}$$

Alternatively, we could do

$$I(X;Y) = D(P_{Y|X} \| P_Y | P_X) = \mathbb{E}_{x \sim \mathcal{N}(0,S)}[D(\mathcal{N}(x,1) \| \mathcal{N}(0, 1+S))].$$

to arrive the same conclusion.

Like f-divergence, the mutual information has a very useful property when applied on Markov chains: the data processing inequality. In fact, the data processing inequality of mutual information is a direct consequence of that of KL-divergence.

Theorem 11.1 (Data processing inequality for M.I.). *Let $X \to Y \to Z$ forms a Markov chain. Then,*

$$I(X;Z) \le I(X;Y).$$

Proof. For the same kernel $P_{Z|Y}$, we have $P_{Y|X=x} \xrightarrow{P_{Z|Y}} P_{Z|X=x}$ for each x and similarly $P_Y \xrightarrow{P_{Z|Y}} P_Z$. Hence applying the data processing inequality for KL divergence yield

$$I(X;Z) = D(P_{Z|X}\|P_Z|P_X) \le D(P_{Y|X}\|P_Y|P_X) = I(X;Y).$$

□

Remark 11.1. For the longer Markov chain $W \to X \to Y \to Z$, we have $I(W;Z) \le I(X;Y)$.

Remark 11.2. For other f-divergences, we can define $I_f(X;Y) \triangleq D_f(P_{Y|X}\|P_Y|P_X)$ which naturally satisfies the data processing inequality on Markov chain.

For a detailed explanation on the materials presented in this section, please refer to [PW15, Ch.2.1-2.2] or [CT06].

11.2 Mutual information method: minimax lower bound

Here's the main idea of the mutual information method: As usual, we are trying to estimate the parameter θ distributed by some prior π, using the estimator $\hat{\theta}$ using the experiment X as its input. In other words, we have a Markov chain $\theta \to X \to \hat{\theta}$.

Then we can upper-bound the mutual information between θ and $\hat{\theta}$ as follows:

$$I(\theta,\hat{\theta}) \le I(\theta;X) \le \sup_{\pi\in\mathcal{M}(\Theta)} I(\theta;X),$$

where the first inequality is due to the data processing inequality of mutual information. The second inequality could be used to drop the assumption that we know the prior π, and is useful when the data X does not provide enough information about θ.

For the lower bound, we have the following:

$$I(\theta,\hat{\theta}) \ge \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}l(\theta,\hat{\theta})\le\mathcal{R}^*_\pi} I(\theta;\hat{\theta}),$$

for any 'good' $\hat{\theta}$ that satisfies $\mathbb{E}l(\theta,\hat{\theta}) \le \mathcal{R}^*_\pi$. This could be interpreted as a minimum amount of information required for an estimation task.

Also notice the followings:

- This line of inequalities is very similar to what we call joint-source channel coding, with the capacity-like upper bound and rate-distortion-like lower bound.
- Only the lower bound is related to the loss function.
- The problem might include a choosing of the prior to make our life easier.

11.3 Extremization of the mutual information

A good news is that we have the convexity and concavity of the mutual information at hand, which could help us find the infimum and supremum of the mutual information. In specific, we have the following property cf. [PW15, p.28]:

Proposition 11.2 (Convexity and Concavity of mutual information)**.** *Consider the notation* $I(P_X, P_{Y|X}) = I(X;Y)$*. Then*

- *For fixed* $P_{Y|X}$*,* $P_X \mapsto I(P_X, P_{Y|X})$ *is concave.*
- *For fixed* P_X*,* $P_{Y|X} \mapsto I(P_X, P_{Y|X})$ *is convex.*

The upper bound, or the maximization part, is the following task: given $P_{Y|X}$, we want to find $\max_{P_X \in \mathcal{P}} I(X;Y)$ where $\mathcal{P}$ is a convex set.

Example 11.4 (GLM, upper bound)**.** Again let $Y = X + Z$, where $Z \sim \mathcal{N}(0, I_p)$ and $X \perp\!\!\!\perp Z$. However, in this case we do not know the prior distribution of X. Rather, we consider a convex set of priors $\mathcal{P} = \{P_X : \mathbb{E}\|X\|_2^2 \le p \cdot S\}$, the signals with constrained average per-dimension power. Then, by the well-known formula for Gaussian channel capacity cf. [PW15, p.33]

$$\max_{P_X \in \mathcal{P}} I(X; X+Z) = \frac{p}{2}\log(1+S).$$

The lower bound, or the minimization part, is: given P_X, we want to find $\min_{P_{Y|X} \in \mathcal{P}} I(X;Y)$.

Example 11.5 (GLM, lower bound)**.** We are only assuming that $X \sim \mathcal{N}(0, S \cdot I_p)$. In the case of the squared distortion, it is known that [PW15, p.33]

$$\min_{P_{Y|X}:\|Y-X\|^2 \le p\cdot\epsilon} I(X;Y) = \begin{cases} \frac{p}{2}\log\left(\frac{S}{\epsilon}\right) & \cdots \quad \epsilon < S \\ 0 & \cdots \quad \text{otherwise.} \end{cases}$$

For non-Gaussian cases, it is in general difficult to find the bounds exactly, and in the following lectures we would discuss the further bounding on both bounds. But before that, we provide several more examples.

Example 11.6 (Bernoulli, lower bound)**.** Let $X \sim \mathrm{Bern}(\delta)^{\otimes p}$. Then:

$$\min_{P_{Y|X}:\mathbb{E}d_H(X,Y) \le p\cdot\epsilon} I(X;Y) = p\left[h(\delta) - h(\epsilon)\right] \cdots \epsilon < \delta < \frac{1}{2}.$$

Example 11.7 (p-dim, n-sample GLM, quadratic loss, combining bounds)**.** Let $\theta \sim \mathcal{N}(0, S \cdot I_p)$, and $\theta \to X \to \hat{\theta}$ hold. Following the usual assumptions, we have $P_{X|\theta} \sim \mathcal{N}(\theta, \frac{1}{n} I_p)$. Then, from the upper bound we know

$$I(\theta, \hat{\theta}) \le I(\theta; X) = \frac{p}{2}\log(1 + S \cdot n).$$

From the lower bound, we have:

$$I(\theta, \hat{\theta}) \ge \min_{P_{\hat{\theta}|\theta}:\|\hat{\theta}-\theta\|_2^2 \le \mathcal{R}_\pi^*} I(\theta;\hat{\theta}) = \frac{p}{2}\log\frac{S}{\mathcal{R}_\pi^*/p}.$$

Combining, we surprisingly have

$$\mathcal{R}_\pi^* \geq \frac{S \cdot p}{1 + S \cdot n}$$

which becomes $\mathcal{R}^* \geq \frac{p}{n}$ as $S \to \infty$.

Note: Statistical estimation task could be represented as a Markov chain $\theta \to X \to \hat{\theta}$ where we do not have a control over $P_{X|\theta}$. On the other hand, communication problem is a Markov chain $\theta \to C \to X \to \hat{\theta}$ where $P_{X|C}$ is not under our control but we can use the "encoder" $P_{C|\theta}$.

11.4 Coming next

Starting from the next lecture, we discuss various methods to further upper and lower bound $I(\theta, \hat{\theta})$. In specific:

- Fano's method (or Pyramid method) is again about transforming the estimation into testing, thereby forming the Markov chain $\theta \to X \to \hat{\theta} \to \hat{\theta}_{\text{test}}$, and investigating the value $\min I(\theta; \hat{\theta}_{\text{test}})$.
- Mutual information would be view as an information radius, and we would use the fact that radius is upper bounded by diameter, which would be more easily characterized.

References

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA, 2006.

[PW15] Y. Polyanskiy and Y. Wu. Lecture notes on information theory. Feb 2015. http://www.ifp.illinois.edu/~yihongwu/teaching/itlectures.pdf.