Information and Coding Theory **Autumn 2014**

Lecture 10: October 30, 2014

Lecturer: Madhur Tulsiani Scribe: Min Jae Song

1 I-Projections and applications

We first review I-Projections from last lecture.

Definition 1.1 *Let Π be a closed convex set of distributions over U. In addition, assume that $Supp(P) \subseteq Supp(Q)$ for all $P \in \Pi$, and $Supp(Q) = U$. Then*

$$Proj_{\Pi}(Q) = \arg\min_{P \in \Pi} D(P||Q) = P^*$$

We proved in last lecture that P^* exists and is unique. It is immediate from definition that if $P \in \Pi$, then $D(P||Q) \geq D(P^*||Q)$. In fact, P^* tells us more. It also tells us how "far" P is away from Q in KL-divergence measure.

Theorem 1.2 *Let $P^* = Proj_{\Pi}(Q)$. Then, for all $P \in \Pi$,*

$$\begin{aligned} Supp(P) &\subseteq Supp(P^*) \\ D(P||Q) &\geq D(P||P^*) + D(P^*||Q) \end{aligned}$$

Proof: Define $P_t = tP + (1-t)P^*$, where $t \in [0,1]$. It is clear that $D(P_t||Q) - D(P^*||Q) \geq 0$. Then

$$\begin{aligned} 0 &\leq \frac{1}{t}[D(P_t||Q) - D(P^*||Q)] \\ &= \frac{d}{dt} D(P_t||Q)|_{t=t' \in [0,t]} \end{aligned}$$

by Mean Value Theorem. Since $t' \to 0$ as $t \to 0$,

$$\lim_{t \downarrow 0} \frac{d}{dt} D(P_t||Q) \geq 0$$

Now we compute $\frac{d}{dt} D(P_t||Q)$.

$$\frac{d}{dt} D(P_t||Q) = \sum_{a \in U} \frac{d}{dt} p_t(a) \log \frac{p_t(a)}{q(a)} + \sum_{a \in U} p_t(a) \frac{d}{dt} (\log p_t(a) - \log q(a))$$

Note that

$$\begin{aligned}\frac{d}{dt}p_t(a) &= p(a) - p^*(a)\\ \frac{d}{dt}\log p_t(a) &= \frac{1}{\ln 2}\frac{1}{p_t(a)}(p(a) - p^*(a))\end{aligned}$$

Using these facts, we have

$$\begin{aligned}\frac{d}{dt}D(P_t||Q) &= \sum_{a\in U}(p(a) - p^*(a))\log\frac{p_t(a)}{q(a)} + \sum_{a\in U}\frac{1}{\ln 2}(p(a) - p^*(a))\\ &= \sum_{a\in U}(p(a) - p^*(a))\log\frac{p_t(a)}{q(a)}\end{aligned}$$

Here, note that if $(\exists a)$ such that $p(a) > 0$ and $p^*(a) = 0$, then $\lim_{t\downarrow 0}\frac{d}{dt}D(P_t||Q) \to -\infty$, which contradicts the fact that $\frac{d}{dt}D(P_t||Q) \geq 0$. Hence, if $p(a) > 0$, then $p^*(a) > 0$ and therefore, $\text{Supp}(P) \subseteq \text{Supp}(P^*)$. This proves the first part of the theorem. Now we evaluate $\frac{d}{dt}D(P_t||Q)$ at $t = 0$.

$$\begin{aligned}\frac{d}{dt}D(P_t||Q)|_{t=0} &= \sum_{a\in U}p(a)\log\frac{p^*(a)}{q(a)} - p^*(a)\log\frac{p^*(a)}{q(a)}\\ &= \sum_{a\in U}p(a)\log\frac{p^*(a)}{q(a)}\frac{p(a)}{p(a)} - D(P^*||Q)\\ &= \sum_{a\in U}p(a)\log\frac{p(a)}{q(a)} - \sum_{a\in U}p(a)\log\frac{p(a)}{p^*(a)} - D(P^*||Q)\\ &= D(P||Q) - D(P||P^*) - D(P^*||Q) \geq 0\end{aligned}$$

Hence, $D(P||Q) \geq D(P||P^*) + D(P^*||Q)$. ■

The following example will show that the inequality in Theorem 1.2 can be strict.

Example 1.3 *Let $U = \{0, 1\}$ and $\Pi = \{P : P(1) \leq 1/2\}$. Choose $\varepsilon < 1/10$ and define the distribution Q as*

$$Q = \begin{cases}1 & \textit{with prob. } 1 - \varepsilon\\ 0 & \textit{with prob. } \varepsilon\end{cases}$$

Exercise 1.4 *Show that*

$$P^* = \begin{cases}1 & \textit{with prob. } 1/2\\ 0 & \textit{with prob. } 1/2\end{cases}$$

Exercise 1.5 *Show that $D(P||Q) > D(P||P^*) + D(P^*||Q)$ for the above example.*

Now we examine cases where the inequality in Theorem 1.2 is actually an equality.

Definition 1.6 *For any given functions $f_1, f_2, ..., f_k$ on U and $\alpha_1, \alpha_2, ..., \alpha_k \in \mathbb{R}$, the set*

$$\mathcal{L} = \{P : \sum_{a \in U} p(a) f_i(a) = \alpha_i,\ i \in [k]\}$$

is called a **linear family** *of distributions.*

Definition 1.7 *Let Q be a given distribution. For any given functions $g_1, g_2, ..., g_k$ on U and $\lambda_1, \lambda_2, ..., \lambda_k \in \mathbb{R}$, the set*

$$\mathcal{E}_Q = \{P : p(a) = c \cdot Q(a) \exp(\sum_{i=1}^{k} \lambda_i g_i(a))\}$$

is called an **exponential family** *of distributions.*

Theorem 1.8 *Let $\mathcal{L}$ be a linear family given by*

$$\mathcal{L} = \{P : \sum_{a \in U} p(a) f_i(a) = \alpha_i,\ i \in [k]\}$$

and $\bigcup_{P \in \mathcal{L}} supp(P) = U$. Then the I-Projection P^* of Q onto $\mathcal{L}$ satisfies the Pythagorean identity

$$D(P||Q) = D(P||P^*) + D(P^*||Q)$$

Moreover, $P^* = Proj_{\mathcal{L}}(Q) \in \mathcal{E}_Q(f_1, ..., f_k)$.

Proof: We will only prove the Pythagorean identity. It suffices to prove that for all $P \in \mathcal{L}$ there exists a small $\beta > 0$ such that for $t \in [-\beta, 0]$, $P_t = tP + (1-t)P^* \in \mathcal{L}$. This is because if this were true, then $\frac{d}{dt} D(P_t||Q)|_{t=0} = 0$ by the minimality of P^*, which in turn implies the equality $D(P||Q) = D(P||P^*) + D(P^*||Q)$.

Now we find β for a given $P \in \mathcal{L}$. Recall that $supp(P) \subseteq supp(P^*)$ and $p_t(a) = tp(a) + (1-t)p^*(a)$. Since $p_t(a) \geq 0$, we want to show that for $t \in [-\beta, 0]$

$$t(p(a) - p^*(a)) \geq -p^*(a)$$

Note that above inequality clearly holds if $p(a) - p^*(a) < 0$. Now choose β such that

$$\beta = \min_{a: p(a) - p^*(a) > 0} \left\{ \frac{p^*(a)}{p(a) - p^*(a)} \right\}$$

Notice that $\beta > 0$ since if $p^*(a) = 0$, then $p(a) = 0$. Hence, $P_t \in \mathcal{L}$ for $t \in [-\beta, 0]$ and therefore, $D(P||Q) = D(P||P^*) + D(P^*||Q)$. ■

Let Q be a uniform distribution on U. Then,

$$D(P||Q) = \log |U| - H(P)$$

Hence, P^* is a distribution that maximizes entropy. In general, when the given information does not uniquely determine a distribution, we choose P^* that maximizes entropy. This is because P^*, being the projection of Q onto the set of distributions Π, is subject to the least amount of additional assumptions. This is known as the **Maximum Entropy Principle**.

2 Parameter Estimation

Let P_θ be a distribution which depends on the parameter θ. For example,

$$P_\theta = \begin{cases} 1 & \text{with prob. } \theta \\ 0 & \text{with prob. } 1-\theta \end{cases}$$

We want to estimate the true parameter θ. So we design estimators $T : U \to \mathbb{R}$.

Definition 2.1 *An estimator T is* **unbiased** *if*

$$\mathop{\mathbb{E}}_{x \sim P_\theta^n}[T(x) - \theta] = 0$$

Example 2.2 *Let $x = (x_1, ..., x_n) \sim P_\theta^n$ for a distribution P_θ which is 1 with probability θ and 0 otherwise. Then the following estimators are all unbiased.*

$$\begin{aligned} T_1 &= x_1 \\ T_2 &= \frac{x_1 + x_2 + x_{17}}{3} \\ T_3 &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

Definition 2.3 *The* **variance** *of an estimator T is defined as*

$$\mathsf{Var}(T) = \mathop{\mathbb{E}}_{x \sim P_\theta^n}[(T - \mathbb{E}\, T)^2]$$

Ideally, we would like an unbiased estimator with small variance. However, there is a limit to how small the variance can be if we have an unbiased estimator. This is given by the Cramér-Rao lower bound. To state this inequality, we will first need a few definitions.

Definition 2.4 *Score function $v_\theta : U \to \mathbb{R}$ is defined as*

$$v_\theta(a) = \frac{d}{d\theta} \ln p_\theta(a) = \frac{1}{p_\theta(a)} \frac{d}{d\theta} p_\theta(a)$$

Note that a larger score is given to perturbations in small probabilities.

Definition 2.5 *Fisher information $J(\theta)$ is defined as*

$$J(\theta) = \mathop{\mathbb{E}}_{x \sim P_\theta^n}[v_\theta(x)^2]$$

Theorem 2.6 (Cramér-Rao) *For any unbiased estimator T estimating the true parameter θ,*

$$\mathsf{Var}(T) \geq \frac{1}{J(\theta)}$$

Proof: Just two lines. By Cauchy-Schwarz,

$$\begin{aligned}(\mathbb{E}[(v_\theta - \mathbb{E}\, v_\theta)(T - \mathbb{E}\, T)])^2 &\leq \mathbb{E}[(v_\theta - \mathbb{E}\, v_\theta)^2]\, \mathbb{E}[(T - \mathbb{E}\, T)^2] \\ &= J(\theta)\, \mathsf{Var}(T)\end{aligned}$$

■

Exercise 2.7 *Show that*

$$\mathbb{E}[(v_\theta - \mathbb{E}\, v_\theta)(T - \mathbb{E}\, T)] = 1$$