Information and Coding Theory **Autumn 2017**

Lecture 10: October 30, 2017

Lecturer: Madhur Tulsiani

1 I-Projections and applications

In this lecture, we will talk more about finding the distribution in a set Π that minimizes $D(P\|Q)$ for a fixed distribution Q. We encountered this when discussing Sanov's theorem and will not discuss its properties in some detail. When Q is the uniform distribution on U. Then we also have,

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

Hence, in this case 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 can be thought of as picking P^* in the set of distributions Π, subject to the least amount of additional assumptions. This is sometimes called the *Maximum Entropy Principle*. In this lecture we will charcterize the distributions obtained by minimizing Kl-divergence (or maximizing entropy).

For closed convex set Π, such a P is called the I-projection of Q onto Π.

Definition 1.1 *Let Π be a closed convex set of distributions over U. In addition, assume that $\mathrm{Supp}(Q) = U$. Then*

$$\mathsf{Proj}_{\Pi}(Q) \;:=\; \arg\min_{P\in\Pi} D(P\|Q) \;=\; P^*$$

Note that the assumption $\mathrm{Supp}(Q) = U$ above is without loss of generality since $D(P\|Q) = \infty$ for any P such that $\mathrm{Supp}(P) \not\subseteq \mathrm{Supp}(Q)$. Use the (strict) convexity of KL-divergence to check the following.

Exercise 1.2 *For a closed, convex set Π, the projection $P^* = \mathsf{Proj}_{\Pi}(Q)$ 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 1.3 *Let $P^* = \mathsf{Proj}_{\Pi}(Q)$. Then, for all $P \in \Pi$,*

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

Proof: Define $P_t = tP + (1-t)P^*$, where $t \in [0,1]$. By minimality of P^*, it is clear that $D(P_t||Q) - D(P^*||Q) \geq 0$. By the mean value theorem, we also have that

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

Since $t' \to 0$ as $t \to 0$, we get

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

We now 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, $D(P||Q) \geq D(P||P^*) + D(P^*||Q)$. ∎

Consider the following example, which shows that the inequality can in fact be strict.

Exercise 1.4 *Let $U = \{0, 1\}$ and $\Pi = \{P : P(1) \leq 1/2\}$. Let Q be defined as*

$$Q = \begin{cases} 1 & \text{with prob. } 3/4 \\ 0 & \text{with prob. } 1/4 \end{cases}$$

1. *Show that*
$$P^* = \begin{cases} 1 & \text{with prob. } 1/2 \\ 0 & \text{with prob. } 1/2 \end{cases}$$
2. *Show that $D(P||Q) > D(P||P^*) + D(P^*||Q)$ for the above example.*

Next, we show how to compute and characterize I-projections for some special sets of distributions.

1.1 Linear families and I-projections

Definition 1.5 *For any given functions $f_1, f_2, ..., f_k$ on U and $\alpha_1, \alpha_2, ..., \alpha_k \in \mathbb{R}$, the set*

$$\mathcal{L} = \left\{ P \;\middle|\; \sum_{a \in U} P(a) \cdot f_i(a) \;=\; \mathop{\mathbb{E}}_{a \sim P}[f_i(a)] \;=\; \alpha_i,\ \forall i \in [k] \right\}$$

is called a linear family *of distributions.*

We show that for linear families, the inequality proved above, is in fact tight. Moreover, the projection P^* lies in the interior of the polytope defining $\mathcal{L}$.

Lemma 1.6 *Let $\mathcal{L}$ be a linear family given by*

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

and $\bigcup_{P \in \mathcal{L}} \mathsf{Supp}(P) = U$. Let $P^* = \mathsf{Proj}_{\mathcal{L}}(Q)$. Then, for all $P \in \mathcal{L}$

1. *There exists $\beta > 0$ such that for $t \in [-\beta, 0]$, $P_t = tP + (1-t)P^* \in \mathcal{L}$.*
2. $D(P\|Q) \;=\; D(P\|P^*) + D(P^*\|Q)$

Then the I-Projection P^* of Q onto $\mathcal{L}$ satisfies the Pythagorean identity

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

Proof: Recall that $\text{Supp}(P) \subseteq \text{Supp}(P^*)$ and $p_t(a) = t \cdot p(a) + (1-t) \cdot p^*(a)$. Since the conditions defining $\mathcal{L}$ are linear, we have that for *all* $t \in \mathbb{R}$ and all $i \in [k]$

$$\sum_{a \in U} p_t(a) \cdot f_i(a) \;=\; t \cdot \sum_{a \in U} p(a) \cdot f_i(a) + (1-t) \cdot \sum_{a \in U} p^*(a) \cdot f_i(a) \;=\; \alpha_i$$

However, we may not have $p_t(a) \geq 0$ for all $t < 0$. We find a $\beta > 0$ such that for $t \in [-\beta, 0]$

$$p_t(a) \;\geq\; 0 \quad \Leftrightarrow \quad t(p(a) - p^*(a)) \;\geq\; -p^*(a)$$

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 $\text{Supp}(P^*) \supseteq \cup_{P \in \mathcal{L}} \text{Supp}(P)$.

The above implies that $\frac{d}{dt} D(P_t || Q)|_{t=0} = 0$ by the minimality of P^*, which in turn implies the equality $D(P||Q) = D(P||P^*) + D(P^*||Q)$. ∎

The above can also be used to show that the I-projection onto $\mathcal{L}$ is of a special form. To describe this, we define the following 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*

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

is called an exponential family *of distributions.*

We will show that $P^* = \text{Proj}_{\mathcal{L}}(Q) \in \mathcal{E}_Q(f_1, ..., f_k)$. We prove this for a linear family defined by a single constraint. The proof for families with multiple constraints is identical.

Let $f : U \to \mathbb{R}$ and let $\mathcal{L}$ be defined as

$$\mathcal{L} = \left\{ P \;\mid\; \sum_{a \in U} P(a) \cdot f(a) \;=\; \mathop{\mathbb{E}}_{a \sim P}[f(a)] \;=\; \alpha \right\}$$

The projection P^* is the optimal solution to the convex program

$$\begin{aligned} \text{minimize} \quad & D(P \| Q) \\ \text{subject to} \quad & \sum_{a \in U} P(a) \cdot f(a) \;=\; \alpha \\ & \sum_{a \in U} P(a) \;=\; 1 \\ & P(a) \;\geq\; 0 \quad \forall a \in U\,. \end{aligned}$$

For $\lambda_0, \lambda_1 \in \mathbb{R}$, we write the Lagrangian as

$$\Lambda(P; \lambda_0, \lambda_1) \;=\; D(P\|Q) + \lambda_0 \cdot \left(\sum_a P(a) - 1 \right) + \lambda_1 \cdot \left(\sum_a P(a) \cdot f(a) - \alpha \right) .$$

The problem above can be written in terms of the Lagrangian as

$$\inf_{P \geq 0} \sup_{\lambda_0, \lambda_1 \in \mathbb{R}} \Lambda(P; \lambda_0, \lambda_1) .$$

From the lemma above, we know that P^* lies in the relative interior of the polytope defining $\mathcal{L}$. Then, strong duality holds for the above program and we can write

$$\inf_{P \geq 0} \sup_{\lambda_0, \lambda_1 \in \mathbb{R}} \Lambda(P; \lambda_0, \lambda_1) \;=\; \sup_{\lambda_0, \lambda_1 \in \mathbb{R}} \inf_{P \geq 0} \Lambda(P; \lambda_0, \lambda_1) .$$

We now characterize the form of the optimal solution by considering the second (dual) program. For a given value of λ_0, λ_1, we can find the optimal solution P^* by setting the derivative of $\Lambda(P; \lambda_0, \lambda_1)$ with respect to $P(a)$ to zero, for every $a \in U$. This gives

$$\log \left(\frac{p^*(a)}{q(a)} \right) + \frac{1}{\ln 2} + \lambda_0 + \lambda_1 \cdot f(a) \;=\; 0$$

Thus, we have for all $a \in U$

$$p^*(a) \;=\; q(a) \cdot 2^{-\lambda_0 - \lambda_1 \cdot f(a)} .$$

The proof for linear families defined by multiple constraints is identical. The above also shows that maximum entropy distributions subject to linear constraints, always belong to an exponential family.