

## Lecture 11: February 11, 2025

Lecturer: Madhur Tulsiani

## 1 I-Projections and applications

We will now talk more about finding a distribution in a set  $\Pi$  that minimizes  $D(P\|Q)$  for a fixed distribution  $Q$ . We encountered this when discussing Sanov's theorem and hypothesis testing, and will now discuss its properties in some detail. When  $Q$  is the uniform distribution on  $\mathcal{X}$ . Then we also have,

$$D(P\|Q) = \log |\mathcal{X}| - H(P)$$

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  $\Pi$ , subject to the least amount of additional assumptions. This is sometimes called the *Maximum Entropy Principle*. In this lecture, we will characterize the distributions obtained by minimizing KL-divergence (or maximizing entropy).

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

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

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

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

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

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

**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  $t' \rightarrow 0$  as  $t \rightarrow 0$ , we get

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

We now compute  $\frac{d}{dt}D(P_t||Q)$ .

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

Note that

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

Using these facts, we have

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

Here, note that if  $(\exists x)$  such that  $p(x) > 0$  and  $p^*(x) = 0$ , then  $\lim_{t \downarrow 0} \frac{d}{dt}D(P_t||Q) \rightarrow -\infty$ , which contradicts the fact that  $\frac{d}{dt}D(P_t||Q) \geq 0$ . Hence, if  $p(x) > 0$ , then  $p^*(x) > 0$  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  $t = 0$ .

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

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  $\mathcal{X} = \{0, 1\}$  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.

## 2 Linear families and I-projections

Building on the previous lecture, we will show how to compute and characterize I-projections for some special sets of distributions.

**Definition 2.1.** For any given real-valued functions  $f_1, f_2, \dots, f_k$  on  $\mathcal{X}$  and  $\alpha_1, \alpha_2, \dots, \alpha_k \in \mathbb{R}$ , the set

$$\mathcal{L} = \left\{ P \mid \sum_{x \in \mathcal{X}} p(x) \cdot f_i(x) = \mathbb{E}_{x \sim P} [f_i(x)] = \alpha_i, \forall i \in [k] \right\}$$

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 2.2.** Let  $\mathcal{L}$  be a linear family given by

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

and  $\cup_{P \in \mathcal{L}} \text{Supp}(P) = \mathcal{X}$ . Let  $P^* = \text{Proj}_{\mathcal{L}}(Q)$ . 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  $p_t(x) = t \cdot p(x) + (1-t) \cdot p^*(x)$ . Since the conditions defining  $\mathcal{L}$  are linear, we have that for all  $t \in \mathbb{R}$  and all  $i \in [k]$

$$\sum_{x \in \mathcal{X}} p_t(x) \cdot f_i(x) = t \cdot \sum_{x \in \mathcal{X}} p(x) \cdot f_i(x) + (1-t) \cdot \sum_{x \in \mathcal{X}} p^*(x) \cdot f_i(x) = \alpha_i$$

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

$$p_t(x) \geq 0 \Leftrightarrow t(p(x) - p^*(x)) \geq -p^*(x)$$

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

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

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

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 2.3.** Let  $Q$  be a given distribution. For any given functions  $g_1, g_2, \dots, g_k$  on  $\mathcal{X}$ , the set

$$\mathcal{E}_Q(g_1, \dots, g_k) := \left\{ P \mid \exists \lambda_1, \dots, \lambda_k \in \mathbb{R} \forall x \in \mathcal{X}, \quad p(x) = c \cdot q(x) \cdot \exp \left( \sum_{i=1}^k \lambda_i g_i(x) \right) \right\}$$

is called an exponential family of distributions.

We will show that  $P^* = \text{Proj}_{\mathcal{L}}(Q) \in \mathcal{E}_Q(f_1, \dots, f_k)$ . We prove this for a linear family defined by a single constraint. The proof for families with multiple constraints is identical. Let  $f : \mathcal{X} \rightarrow \mathbb{R}$  and let  $\mathcal{L}$  be defined as

$$\mathcal{L} = \left\{ P \mid \sum_{x \in \mathcal{X}} p(x) \cdot f(x) = \mathbb{E}_{x \sim P} [f(x)] = \alpha \right\}$$

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

$$\begin{aligned} & \text{minimize} && D(P \| Q) \\ & \text{subject to} && \sum_{x \in \mathcal{X}} p(x) \cdot f(x) = \alpha \\ & && \sum_{x \in \mathcal{X}} p(x) = 1 \\ & && p(x) \geq 0 \quad \forall x \in \mathcal{X}. \end{aligned}$$

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

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

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 [Lemma 2.2](#), 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  $\lambda_0, \lambda_1$ , we can find the optimal solution  $P^*$  by setting the derivative of  $\Lambda(P; \lambda_0, \lambda_1)$  with respect to  $p(x)$  to zero, for every  $x \in \mathcal{X}$ . This gives

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

Thus, we have for all  $a \in \mathcal{X}$

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

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. Exponential families have many interesting applications, and more material on these can be found in the survey by Jordan and Wainwright [?]. A good reference for looking up the convex-duality based arguments above, is Chapter 5 of the excellent book by Boyd and Vandenberghe [?].

## References

- [BV04] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.
- [WJ08] Martin J Wainwright and Michael Irwin Jordan, *Graphical models, exponential families, and variational inference*, Now Publishers Inc, 2008.