**Information and Coding Theory** **Autumn 2022**

Lecture 14: November 10, 2022

Lecturer: Madhur Tulsiani

# 1 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 1.1.** *For any given real-valued functions $f_1, f_2, ..., f_k$ on $\mathcal{X}$ and $\alpha_1, \alpha_2, ..., \alpha_k \in \mathbb{R}$, the set*

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

*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 1.2.** *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 $\bigcup_{P \in \mathcal{L}} \mathsf{Supp}(P) = \mathcal{X}$. 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 $\mathrm{Supp}(P) \subseteq \mathrm{Supp}(P^*)$ 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 $p_t(a) \geq 0$ for all $t < 0$. We find a $\beta > 0$ such that for $t \in [-\beta, 0]$

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

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 $D(P||Q) = D(P||P^*) + D(P^*||Q)$. ■

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

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

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

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

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

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

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

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 1.2, 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 [WJ08]. A good reference for looking up the convex-duality based arguments above, is Chapter 5 of the excellent book by Boyd and Vandenberghe [BV04].

## 2 Matrix Scaling

We will consider an application of I-projections to a problem known as matrix scaling. Say we are given two nonnegative matrices $M, N \in \mathbb{R}_+^{n \times n}$ such that for all $i, j$, $M_{ij} = 0 \;\Leftrightarrow\; N_{ij} = 0$. The goal is to multiply (scale) each row $i$ of $M$ by a number $r_i$ and each column $j$ by $c_j$, such that the resulting matrix $M'$ has the same row and column sums as the target matrix $N$. Another way of stating this is that we want to find diagonal matrices $D_1$ and $D_2$ such that for $M' = D_1 M D_2$, we have

$$\sum_j M'_{ij} \;=\; \sum_j N_{ij} \quad \forall i \in [n] \qquad \text{and} \qquad \sum_i M'_{ij} \;=\; \sum_i N_{ij} \quad \forall j \in [n] \,.$$

We will show a special case when the goal is to scale $M$ so that the resulting matrix $M'$ is doubly stochastic i.e.,

$$\sum_j M'_{ij} \;=\; \sum_i M'_{ij} \;=\; 1 \quad \forall i, j \in [n] \,.$$

First, note that by a *global* scaling of $1/\sum_{i,j} M_{ij}$, we can assume that $\sum_{i,j} M_{ij} = 1$, and the goal is instead to scale it to have row and column sums equal to $1/n$ i.e.,

$$\sum_j M'_{ij} \;=\; \sum_i M'_{ij} \;=\; \frac{1}{n} \quad \forall i, j \in [n]\,.$$

We can now think of this as a problem of going from one distribution to another. Assume that $M_{ij} > 0$ for all $i, j$, and think of the target matrix $N$ with $N_{ij} = 1/n^2$ for all $i, j$. Since the entries of $M$ are positive and sum to 1, we can think of it as a probability distribution $Q$ with $\mathrm{Supp}(Q) = [n] \times [n]$ (where $q(i, j) = M_{ij}$). We consider the linear family of distributions on $[n] \times [n]$ (written as matrices) with the required row and column sums.

$$\mathcal{L} \;:=\; \left\{ P \;\mid\; \sum_j p(i,j) \;=\; \sum_i p(i,j) \;=\; \frac{1}{n} \quad \forall i, j \in [n] \right\}$$

Note that the above is a linear family as defined in the previous lecture, since we can consider functions $f_1, \ldots, f_n$ and $g_1, \ldots, g_n$ defined as

$$f_{i_0}(i,j) \;=\; \begin{cases} 1 & \text{if } i = i_0 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad g_{j_0}(i,j) \;=\; \begin{cases} 1 & \text{if } j = j_0 \\ 0 & \text{otherwise} \end{cases}.$$

Then, the above family can be written in terms of the expectations of the functions $f_i$ and $g_j$ for all $i, j \in [n]$. Moreover, we know from the previous lecture that the I-projection $P^*$ of $Q$ onto $\mathcal{L}$ is of the form

$$\begin{aligned} p^*(i,j) \;&=\; c_0 \cdot q(i,j) \cdot \exp\left( \sum_{i_0} \lambda_{i_0} \cdot f_{i_0}(i,j) + \sum_{j_0} \mu_{j_0} \cdot g_{j_0}(i,j) \right) \\ &=\; c_0 \cdot q(i,j) \cdot \exp\left(\lambda_i + \mu_j\right) \\ &=\; \left(\sqrt{c_0} \cdot \exp(\lambda_i)\right) \cdot M_{i,j} \cdot \left(\sqrt{c_0} \cdot \exp(\mu_j)\right)\,. \end{aligned}$$

Thus, we can define the diagonal matrices $D_1$ and $D_2$ as

$$(D_1)_{ii} \;=\; \sqrt{c_0} \cdot \exp(\lambda_i) \qquad \text{and} \qquad (D_2)_{jj} \;=\; \sqrt{c_0} \cdot \exp(\mu_j)\,.$$

We then have that the distribution $p^*$ given by the resulting matrix $M' = D_1 M D_2$, belongs to the linear family $\mathcal{L}$. Thus, the row and column sums of $M'$ are $1/n$. Combining this with another global scaling (replace $\sqrt{c_0}$ by $\sqrt{c_0 \cdot n}$) we can also get all the row and column sums to be 1 (i.e., make the matrix doubly stochastic).

**Exercise 2.1.** *Where did we use the fact that $M_{ij} > 0$ for all $i, j \in [n]$?*

**Exercise 2.2.** *Use this the above techniques to solve the matrix scaling problem for an arbitrary target matrix $N$ (assuming $M_{ij} = 0 \Leftrightarrow N_{ij} = 0$).*

Matrix scaling and its generalization, known as operator scaling have found a variety of applications in combinatorial optimization, complexity theory and analysis. Please take a look at the recent tutorial by Wigderson [Wig17] for an introduction to many of these connections.

## References


[BV04] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004. 3

[Wig17] Avi Wigderson, *Operator scaling: theory, applications and connections*, 2017, Tutorial given at CCC 2017. 5

[WJ08] Martin J Wainwright and Michael Irwin Jordan, *Graphical models, exponential families, and variational inference*, Now Publishers Inc, 2008. 3