14 Information Geometry

As we developed, information divergence $D(\cdot\|\cdot)$ defined via

$$D(p\|q) = \sum_y p(y) \log \frac{p(y)}{q(y)}$$

provides a meaningful way to measure how different two distributions are with respect to the task of inference. In this section, we explore the space of distributions and see how interpreting information divergence as a kind of distance measure leads to a rich and useful geometry on this space, from which we can derive valuable intuition. In particular, we will see that in some important respects, information divergence behaves like the square of Euclidean distance.

We start with the concept of the probability simplex. Let y be a random variable defined over an alphabet $\mathcal{Y}$. We emphasize the case of finite alphabets, since this is sufficient to reveal the key concepts. However, the concepts naturally extend to countably and uncountably infinite alphabets as well.

Accordingly, without loss of generality, let the alphabet be $\mathcal{Y} = \{1, 2, \ldots, M\}$, where $M < \infty$. Then each possible distribution $p_{\mathsf{y}}(\cdot)$ for y is represented by a point in a M-dimensional space, with the coordinates corresponding to the probabilities of each of the M values. Since distributions are normalized, the collection of all possible models lies in an (affine) hyperplane of dimension $M-1$ in this space that intersects each of the M coordinate axes at unit distance from the origin. Moreover, because probabilities are nonnegative, this hyperplane is restricted to the first orthant. The case $M = 3$ is depicted in Fig. 1.

We use $\mathcal{P}^{\mathcal{Y}}$ to denote the collection of all possible distributions over the alphabet $\mathcal{Y}$, which is referred to as the *probability simplex*. Our development will focus on various subsets $\mathcal{P} \subset \mathcal{P}^{\mathcal{Y}}$ and their relationships. Before we begin our development, consider some simple but representative examples.

Example 1. Fix $p_1, p_2 \in \mathcal{P}^{\mathcal{Y}}$, and consider the set of distributions

$$\mathcal{P} = \left\{ p \in \mathcal{P}^{\mathcal{Y}} \colon p(y) = p_{\mathsf{y}}(y; x) = \frac{p_1(y)^x \, p_2(y)^{1-x}}{Z(x)}, \quad \text{for some } x \in [0,1] \right\},$$

where the partition function $Z(x)$ normalizes the distribution. This set of distributions, corresponding to weighted geometric mean of two base distributions, traces out a one-dimensional curve between p_1 and p_2 in the $(M-1)$-dimensional probability simplex $\mathcal{P}^{\mathcal{Y}}$. Recall that this corresponds to a one-dimensional exponential family.

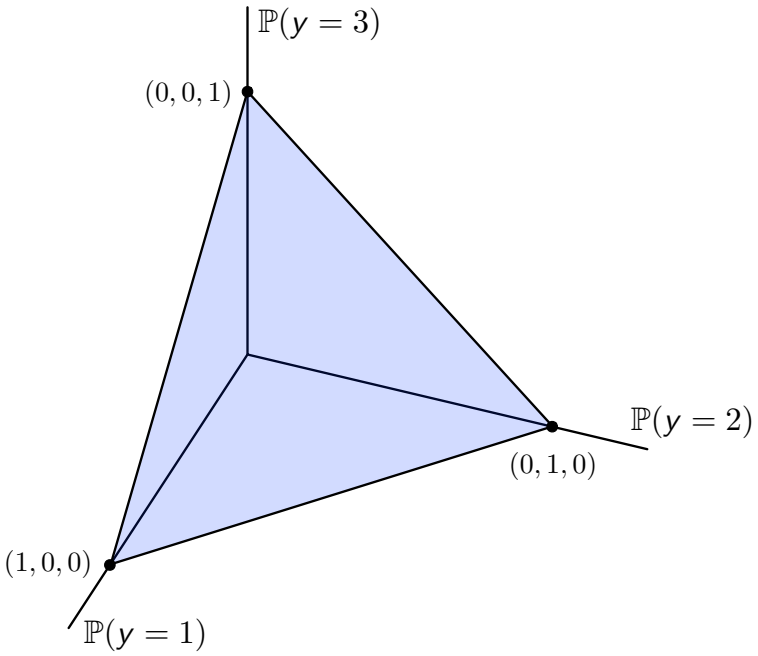

Figure 1: The probability simplex for distributions defined on the alphabet $\mathcal{Y} = \{1, 2, 3\}$.

Example 2. Again, fix $p_1, p_2 \in \mathcal{P}^{\mathcal{Y}}$, but now consider the set of distributions

$$\mathcal{P} = \big\{p \in \mathcal{P}^{\mathcal{Y}} \colon p(y) = p_{\mathsf{y}}(y; x) = x\, p_1(y) + (1 - x)\, p_2(y), \quad \text{for some } x \in [0, 1]\big\}.$$

In this case, the set of distributions, corresponding to weighted arithmetic means of two base distributions, traces out a line between p_1 and p_2 in the probability simplex $\mathcal{P}^{\mathcal{Y}}$. This example is an instance of a set of distributions that is convex.

Example 3. Fix a (nonconstant) function $t \colon \mathcal{Y} \mapsto \mathbb{R}$ and a constant $c \in \mathbb{R}$, and consider the set of distributions

$$\mathcal{P} = \big\{p \in \mathcal{P}^{\mathcal{Y}} \colon \mathbb{E}_p\left[t(\mathsf{y})\right] = c\big\}.$$

Since

$$\mathbb{E}_p\left[t(\mathsf{y})\right] = \sum_{y \in \mathcal{Y}} p(y)\, t(y) = c \tag{1}$$

is a linear constraint on the the points in the probability simplex, the subset $\mathcal{P}$ is a $(M - 2)$-dimensional affine subspace restricted to the simplex.[1] This example is an

[1]More precisely, this statement is true provided t and c such that (1) is satisfiable by some distribution. Otherwise, $\mathcal{P} = \varnothing$.

instance of what we will refer to as a *linear family* of distributions. Note that it is also a convex set.

As some specific examples, we could have, for numeric random variables, $t(y) = y$, which corresponds to a mean constraint, or $t(y) = y^2$, which corresonds to a second-moment constraint. Less obvious might be choices like $t(y) = \log q(y)$ for some fixed distribution $q \in \mathcal{P}^{\mathcal{Y}}$, which corresponds to the constraint

$$H(p) - D(p\|q) = c$$

or $t(y) = \mathbb{1}_{y=b}$ for some particular $b \in \mathcal{Y}$, which corresponds to a support constraint of the form $p(b) = 0$.

14.1 Divergence on the Boundary of the Simplex

Information divergence is a "warped" measure of distance by comparison with Euclidean distance. This warping, and the asymmetry of divergence, can be pronounced at the boundary of the simplex, as we now describe.

To begin, observe that a distribution q is on the boundary of the simplex, i.e., $q \in \partial(\mathcal{P}^{\mathcal{Y}})$, if and only if $q(y') = 0$ for some $y' \in \mathcal{Y}$. Phrased differently, p is in the interior of the simplex, i.e., $p \in \text{int}(\mathcal{P}^{\mathcal{Y}})$ if and only if p is strictly positive, i.e., $p(y) > 0$ for all $y \in \mathcal{Y}$. Moreover, the boundary corresponding to those distributions r for which $r(y') = 0$ is the simplex for the set of distributions over the alphabet $\mathcal{Y}\backslash\{y'\}$.

First, if both p and q are in the interior of the simplex, then clearly $D(p\|q) < \infty$ since all distributions are strictly positive.

In contrast, if q on the boundary of simplex but p is not, then the divergence of q from p is infinitely large. Specifically, if $p \in \text{int}(\mathcal{P}^{\mathcal{Y}})$ but $q \in \partial(\mathcal{P}^{\mathcal{Y}})$, then

$$D(p\|q) = \sum_y p(y) \log \frac{p(y)}{q(y)} = \infty,$$

since any term in the summation corresponding to $y = y'$ such that $q(y') = 0$ will be infinite since $p(y') > 0$.

However, the opposite is not true: if p is on the boundary but q is not, then the divergence of q from p is finite. If $p \in \partial(\mathcal{P}^{\mathcal{Y}})$ but $q \in \text{int}(\mathcal{P}^{\mathcal{Y}})$, then

$$D(p\|q) = \sum_y p(y) \log \frac{p(y)}{q(y)} < \infty,$$

since any term in the summation corresponding to $y = y'$ such that $p(y') = 0$ will be zero by our convention that $0 \log 0 \triangleq 0$.

To obtain additional results, note that we can decompose the boundary of the simplex into M components, each corresponding to distributions r such that $r(y') = 0$ for a different $y' \in \mathcal{Y}$.

If both p and q are on all the same boundary component(s) of the simplex, then $D(p\|q) < \infty$. Since this is the case where $p(y')$ and $q(y')$ are zero for exactly the same values of $y' \in \mathcal{Y}$, this corresponds to p and q being effectively defined over a smaller alphabet, so both lie on a lower dimensional simplex.

If both p and q lie on the boundary of the simplex, but q lies on at least one boundary component that p doesn't, then $D(p\|q) = \infty$. But if both p and q lie on the boundary of the simplex, but p lies on at least one boundary component that q doesn't, then $D(p\|q) < \infty$.

14.2 Information Projection and Pythagoras' Theorem

We first establish the notion of the projection of a distribution on the probability simplex, then obtain a version of Pythagoras' theorem for the space of distributions with the information divergence as distance.

Definition 1 (I-Projection)**.** *The* information projection *or* I-projection *of a distribution q onto a (nonempty and closed) set of distributions $\mathcal{P}$ is the distribution p^* such that*

$$p^* = \arg\min_{p \in \mathcal{P}} D(p\|q). \tag{2}$$

In general, p^* exists since $D(p\|q)$ is nonnegative and continuous in p, and $\mathcal{P}$ is nonempty and closed. In general, p^* may or may not be unique, depending on the properties of $\mathcal{P}$. However, when $\mathcal{P}$ is a convex set, as will be of primary interest to us, p^* is unique.

Theorem 1 (Pythagoras' Theorem, Information Version)**.** *Let q be any distribution, and let p^* be the I-projection of q onto a (nonempty) closed, convex set $\mathcal{P}$. Then*

$$D(p\|q) \geq D(p\|p^*) + D(p^*\|q) \quad \textit{for all } p \in \mathcal{P}. \tag{3}$$

Fig. 2 provides a graphical depiction of Theorem 1.

Proof. If $q \in \mathcal{P}$ then $p^* = q$ and (3) holds trivially. Next, we consider $q \notin \mathcal{P}$. Since $D(p\|q)$ is continuous and, as shown in Appendix 14.6, convex in p, p^* exists and is unique.

Let $p \in \mathcal{P}$ be an arbitrary distribution in $\mathcal{P}$. Then for $0 \leq \lambda \leq 1$, the distribution

$$p_\lambda = (1-\lambda)p^* + \lambda p \tag{4}$$

must also be in $\mathcal{P}$, by convexity of $\mathcal{P}$. Furthermore, $\lim_{\lambda \to 0} p_\lambda = p^*$.

Since $D(p^*\|q)$ is the minimum of $D(p_\lambda\|q)$ along the path from p^* to p parameterized by λ as it goes from 0 to 1, we must have

$$\frac{d}{d\lambda} D(p_\lambda\|q)\bigg|_{\lambda=0} \geq 0. \tag{5}$$

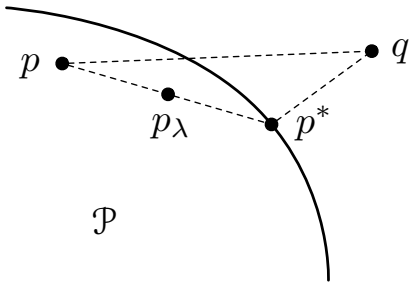

Figure 2: Divergence relationships in the information version of Pythagoras' theorem. The distribution p^* is the I-projection of the distribution q onto the set of distributions $\mathcal{P}$.

Otherwise, for some sufficiently small $\epsilon > 0$ we would have $D(p_\epsilon \| q) \leq D(p^* \| q)$.

Now we can evaluate the left-hand side of (5) by first noting that

$$\begin{aligned}
\frac{d}{d\lambda} D(p_\lambda \| q) &= \frac{d}{d\lambda} \sum_y p_\lambda(y) \log \frac{p_\lambda(y)}{q(y)} \\
&= \sum_y [p(y) - p^*(y)] \log \frac{p_\lambda(y)}{q(y)} + \sum_y p_\lambda(y) \cdot \frac{1}{p_\lambda(y)} [p(y) - p^*(y)] \\
&= \sum_y [p(y) - p^*(y)] \log \frac{p_\lambda(y)}{q(y)}. \qquad (6)
\end{aligned}$$

Setting $\lambda = 0$ in (6) and substituting into (5) yields

$$\begin{aligned}
0 &\leq \frac{d}{d\lambda} D(p_\lambda \| q) \Big|_{\lambda=0} \qquad (7) \\
&= \sum_y [p(y) - p^*(y)] \log \frac{p^*(y)}{q(y)} \\
&= \sum_y p(y) \log \left[\frac{p(y)}{q(y)} \frac{p^*(y)}{p(y)} \right] - \sum_y p^*(y) \log \frac{p^*(y)}{q(y)} \\
&= \sum_y p(y) \log \frac{p(y)}{q(y)} - \sum_y p(y) \log \frac{p(y)}{p^*(y)} - \sum_y p^*(y) \log \frac{p^*(y)}{q(y)} \\
&= D(p \| q) - D(p \| p^*) - D(p^* \| q), \qquad (8)
\end{aligned}$$

as desired. □

One significant consequence of Theorem 1 follows immediately from the results of Section 14.1.

Corollary 1 (Pythagoras' Corollary)**.** *The I-projection p^* of any q not on the boundary of $\mathcal{P}^{\mathcal{Y}}$ onto a linear family $\mathcal{P}$ cannot lie on a boundary component of $\mathcal{P}^{\mathcal{Y}}$ unless all of $\mathcal{P}$ is confined to that particular boundary component.*

Proof. Since $q \in \text{int}(\mathcal{P}^{\mathcal{Y}})$, we know that the left-hand side of (3) is finite, i.e., $D(p\|q) < \infty$. Hence, the first term on the right-hand side of (3) must be finite, i.e., $D(p\|p^*) < \infty$. But if $p^* \in \partial(\mathcal{P}^{\mathcal{Y}})$, then this means that p must lie on the same boundary component as p^*. Since $p \in \mathcal{P}$ is arbitrary, the corollary follows. □

In some important cases, the inequality in (3) holds with equality. To understand such scenarios, we need to develop the notion of a linear family.

14.3 Linear Families

I-projections of q in the simplex onto certain sets of distributions have special properties, as we develop in this section. We begin by defining set structure for which this is the case.

Definition 2 (Linear Family)**.** *A set of distributions $\mathcal{P} \subset \mathcal{P}^{\mathcal{Y}}$ is a* linear family *if for some $K < M$ there exist functions $\mathbf{t} = [t_1(\cdot), \ldots, t_K(\cdot)]^{\mathrm{T}}$ and constants $\bar{\mathbf{t}} = [\bar{t}_1, \ldots, \bar{t}_K]^{\mathrm{T}}$ such that*

$$\mathbb{E}_p\left[t_i(\boldsymbol{y})\right] = \bar{t}_i, \quad i = 1, \ldots, K \quad \textit{for all } p \in \mathcal{P}. \tag{9}$$

We can express (9) in matrix form as

$$\underbrace{\begin{bmatrix} t_1(1) - \bar{t}_1 & \cdots & t_1(M) - \bar{t}_1 \\ \vdots & \ddots & \vdots \\ t_K(1) - \bar{t}_K & \cdots & t_K(M) - \bar{t}_K \end{bmatrix}}_{\triangleq \tilde{\mathbf{T}}} \begin{bmatrix} p(1) \\ \vdots \\ p(M) \end{bmatrix} = \mathbf{0}.$$

Hence, $\mathcal{P}$ corresponds to the intersection of the null space of $\tilde{\mathbf{T}}$ with the simplex $\mathcal{P}^{\mathcal{Y}}$ If there is no such intersection, the linear family is degenerate: $\mathcal{P} = \varnothing$. Otherwise, the intersection is an affine subspace. We are interested only in latter case, so will restrict our attention to choices of $(\mathbf{t}, \bar{\mathbf{t}})$ for which $\mathcal{P} \neq \varnothing$.

For every linearly independent row in $\tilde{\mathbf{T}}$, the dimension of the null space is reduced by one relative to the dimension $M - 1$ of the original simplex. Since the number of linearly independent rows is the rank of $\tilde{\mathbf{T}}$, this means that

$$\dim \mathcal{P} = M - \text{rank}(\tilde{\mathbf{T}}) - 1. \tag{10}$$

When $\text{rank}(\tilde{\mathbf{T}}) = 0$, (9) imposes no constraints, and $\mathcal{P}$ is the entire simplex. When all the rows in $\mathbf{T}$ are linearly independent, $\text{rank}(\tilde{\mathbf{T}}) = K$ and $\mathcal{P}$ is an affine subspace

of dimension of $M - K - 1$. We say the representation $(\mathbf{t}, \bar{\mathbf{t}})$ of a linear family is *minimal* if $\tilde{\mathbf{T}}$ has linearly independent rows.

As an aside, note that the affine subspace defining this linear family is, equivalently, the set of solutions to

$$\underbrace{\begin{bmatrix} t_1(1) & \cdots & t_1(M) \\ \vdots & \ddots & \vdots \\ t_K(1) & \cdots & t_K(M) \\ 1 & \cdots & 1 \end{bmatrix}}_{\triangleq \mathbf{T}_+} \begin{bmatrix} p(1) \\ \vdots \\ p(M) \end{bmatrix} = \begin{bmatrix} \bar{t}_1 \\ \vdots \\ \bar{t}_K \\ 1 \end{bmatrix}, \tag{11}$$

where the last row is the simplex constraint, so we can also write

$$\dim \mathcal{P} = M - \operatorname{rank}(\mathbf{T}_+). \tag{12}$$

Eqs. (10) and (12) are consistent since $\operatorname{rank}(\mathbf{T}_+) = \operatorname{rank}(\tilde{\mathbf{T}}_+)$ with

$$\tilde{\mathbf{T}}_+ \triangleq \begin{bmatrix} t_1(1) - \bar{t}_1 & \cdots & t_1(M) - \bar{t}_1 \\ \vdots & \ddots & \vdots \\ t_K(1) - \bar{t}_K & \cdots & t_K(M) - \bar{t}_K \\ 1 & \cdots & 1 \end{bmatrix} \tag{13}$$

because (13) and $\mathbf{T}_+$ in (11) are related by row operations, and $\operatorname{rank}(\tilde{\mathbf{T}}_+) = \operatorname{rank}(\tilde{\mathbf{T}}) + 1$. To verify the latter, suppose this were not true. Then we must be able to write the last row of $\tilde{\mathbf{T}}_+$ as a linear combination of the others, i.e., there must exist $a_1, \ldots, a_K$ such that

$$\sum_{i=1}^{K} a_i (t_i(y) - \bar{t}_i) = 1, \quad \text{for all } y \in \mathcal{Y}. \tag{14}$$

But taking the expectation of the left hand side of (14) with respect to $p \in \mathcal{P}$ gives

$$\sum_y p(y) \sum_{i=1}^{K} a_i (t_i(y) - \bar{t}_i) = \sum_{i=1}^{K} a_i \sum_y p(y)(t_i(y) - \bar{t}_i) = \sum_{i=1}^{K} a_i \left(\mathbb{E}_p \left[t_i(y) \right] - \bar{t}_i \right) = 0,$$

regardless of the choice of $a_1, \ldots, a_K$, which is in contradiction with the right hand side of (14) being 1.

It is straightforward to verify that a linear family $\mathcal{P}$ is a convex set. Indeed, suppose $p_1, p_2 \in \mathcal{P}$, Then with $p(y) = \lambda\, p_1(y) + (1 - \lambda)\, p_2(y)$ for $\lambda \in [0, 1]$ and

$i = 1, \ldots, K$ we have

$$\begin{aligned}
\mathbb{E}_p\left[t_i(\mathsf{y})\right] &= \sum_y t_i(y)p(y) \\
&= \sum_y t_i(y)[\lambda\, p_1(y) + (1-\lambda)\, p_2(y)] \\
&= \lambda \sum_y t_i(y)p_1(y) + (1-\lambda) \sum_y t_i(y)p_2(y) \\
&= \lambda \bar{t}_i + (1-\lambda)\bar{t}_i \\
&= \bar{t}_i, \qquad (15)
\end{aligned}$$

so $p \in \mathcal{P}$. Note, too, that a linear family is a closed set.

However, linear families have an additional property that is relevant for us, which the following claim expresses.

Claim 1. *A linear family $\mathcal{L} \subset \mathcal{P}^{\mathcal{Y}}$ has the property that if $p_1, p_2 \in \mathcal{L}$ then for every $\lambda \in \mathbb{R}$ such that $p = \lambda\, p_1 + (1-\lambda)\, p_2 \in \mathcal{P}^{\mathcal{Y}}$, we have $p \in \mathcal{L}$.*

We emphasize that in Claim 1, λ is not constrained to the interval $[0, 1]$.

Proof. It suffices to recognize that the result (15) relies only on p being a distribution and not whether λ lies in the range $[0, 1]$. □

The inequality (3) in Pythagoras' Theorem holds with equality for the I-projection of a distribution q in the interior of the simplex onto a linear family. Specifically, we have the following.

Corollary 2 (Pythagorean Identity)**.** *Let $\mathcal{L}$ be a linear family of distributions defined on the alphabet $\mathcal{Y}$. Let q be an arbitrary distribution in the interior of the associated simplex $\mathcal{P}^{\mathcal{Y}}$. Then the I-projection of q onto $\mathcal{L}$, i.e.,*

$$p^* = \arg\min_{p \in \mathcal{L}} D(p\|q),$$

satisfies

$$D(p\|q) = D(p\|p^*) + D(p^*\|q), \quad \textit{for all } p \in \mathcal{L}. \qquad (16)$$

Proof. First, we show there must exist some $\lambda < 0$ such that

$$p_\lambda = \lambda p + (1-\lambda)p^*$$

is a distribution. By contradiction, suppose $p_\lambda(y') < 0$ at some $y' \in \mathcal{Y}$ for $\lambda < 0$ sufficiently close to 0. Then this would mean that $p(y') > 0$ but $p^*(y') = 0$, i.e., that p^* is on a boundary component that $p \in \mathcal{L}$ is not. However, by Corollary 1, we know this is not possible.

In turn, it follows that (7) in the proof of Theorem 1 holds with equality in this case; otherwise for some sufficiently small $\epsilon > 0$ we would have $D(p_{-\epsilon}\|q) \leq D(p^*\|q)$. In turn, (16) follows. □

Analogous to the case of Euclidean geometry, (16) is a statement about when a I-projection is, in an information geometric sense, orthogonal.

14.4 Orthogonal Families

Consider a linear family on the simplex $\mathcal{P}^{\mathcal{Y}}$ that has a minimal representation corresponding to a particular choice of $\mathbf{t} = [t_1, \ldots, t_K]^{\mathrm{T}}$. Given for any $p^* \in \mathcal{P}^{\mathcal{Y}}$, there is a unique choice of $\bar{\mathbf{t}} = [\bar{t}_1, \ldots, \bar{t}_K]^{\mathrm{T}}$ such that p^* is in this linear family; indeed, $\bar{\mathbf{t}} = \mathbb{E}_{p^*}[\mathbf{t}(\mathsf{y})]$.

Let $\mathcal{L}_{\mathbf{t}}(p^*)$ denote this linear family, where our notation emphasizes that p^* is in the family, but this parameterization is clearly not unique; indeed,

$$\mathcal{L}_{\mathbf{t}}(p^*) = \mathcal{L}_{\mathbf{t}}(p) \quad \text{for any } p \in \mathcal{L}_{\mathbf{t}}(p^*),$$

since for all such p we have $\mathbb{E}_p[\mathbf{t}(\mathsf{y})] = \bar{\mathbf{t}} = \mathbb{E}_{p^*}[\mathbf{t}(\mathsf{y})]$.

For this family, we can ask the question: what is the set of all distributions in the simplex whose I-projection onto $\mathcal{L}_{\mathbf{t}}(p^*)$ is p^*? These distributions then have the interpretation of being orthogonal, in the sense of information geometry, to this linear family at the point p^*.

The following theorem establishes our main result.

Theorem 2. *Let $p^* \in \mathcal{P}^{\mathcal{Y}}$ be an arbitrary distribution. Then p^* is the I-projection of a distribution $q \in \mathcal{P}^{\mathcal{Y}}$ onto the linear family*

$$\mathcal{L}_{\mathbf{t}}(p^*) = \left\{ p \in \mathcal{P}^{\mathcal{Y}} \colon \mathbb{E}_p[\mathbf{t}(\mathsf{y})] = \bar{\mathbf{t}} \triangleq \mathbb{E}_{p^*}[\mathbf{t}(\mathsf{y})] \right\}$$

if and only if q is in the exponential family

$$\mathcal{E}_{\mathbf{t}}(p^*) = \left\{ q \in \mathcal{P}^{\mathcal{Y}} \colon q(y) = p^*(y) \exp\left\{ \mathbf{x}^{\mathrm{T}} \mathbf{t}(y) - \alpha(\mathbf{x}) \right\}, \text{ for all } y \in \mathcal{Y}, \text{ some } \mathbf{x} \in \mathbb{R} \right\}. \tag{17}$$

These relationships are depicted graphically in Fig. 3. Note that $\mathcal{L}$ is an $(M-K)$-dimensional linear family, and $\mathcal{E}$ is a K-dimensional linear exponential family.

We refer to $\mathcal{E}_{\mathbf{t}}(p^*)$ defined in (17) as the *orthogonal family* to $\mathcal{L}_{\mathbf{t}}(p^*)$ at p^*. Note that our notation $\mathcal{E}_{\mathbf{t}}(p^*)$ emphasizes that this exponential family includes p^*, but this parameterization is also clearly not unique; indeed,

$$\mathcal{E}_{\mathbf{t}}(p^*) = \mathcal{E}_{\mathbf{t}}(q) \quad \text{for any } q \in \mathcal{E}_{\mathbf{t}}(p^*).$$

To see this, since $q \in \mathcal{E}_{\mathbf{t}}(p^*)$ there must exist $\mathbf{x}_0 \in \mathbb{R}^K$ such that

$$q(y) = p^*(y) \exp\left\{ \mathbf{x}_0^{\mathrm{T}} \mathbf{t}(y) - \alpha(\mathbf{x}_0) \right\}, \quad \text{for all } y \in \mathcal{Y}.$$

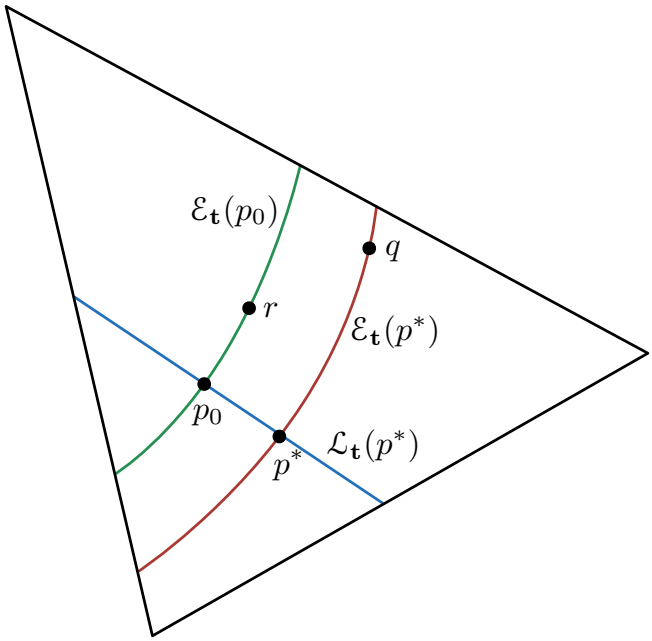

Figure 3: The distributions in the exponential family $\mathcal{E}_{\mathbf{t}}(p^*) = \mathcal{E}_{\mathbf{t}}(q)$ all have the same I-projection p^* onto the linear family $\mathcal{L}_{\mathbf{t}}(p^*) = \mathcal{L}_{\mathbf{t}}(p)$.

But then

$$\begin{aligned}
\mathcal{E}_{\mathbf{t}}(q) &= \left\{r\colon r(y) = q(y)\exp\left\{\mathbf{x}^{\mathrm{T}}\mathbf{t}(y) - \alpha(\mathbf{x})\right\}, \quad \text{for all } y \in \mathcal{Y}, \text{ some } \mathbf{x} \in \mathbb{R}^K\right\} \\
&= \left\{r\colon r(y) = p^*(y)\exp\left\{(\mathbf{x}+\mathbf{x}_0)^{\mathrm{T}}\mathbf{t}(y) - \alpha(\mathbf{x}) - \alpha(\mathbf{x}_0)\right\}, \right. \\
&\qquad\qquad \left. \text{for all } y \in \mathcal{Y}, \text{ some } \mathbf{x} \in \mathbb{R}^K\right\} \\
&= \left\{r\colon r(y) = p^*(y)\exp\left\{\tilde{\mathbf{x}}^{\mathrm{T}}\mathbf{t}(y) - \alpha(\tilde{\mathbf{x}})\right\}, \quad \text{for all } y \in \mathcal{Y}, \text{ some } \tilde{\mathbf{x}} \in \mathbb{R}^K\right\} \\
&= \mathcal{E}_{\mathbf{t}}(p^*).
\end{aligned}$$

Proof of Theorem 2. We begin by proving the "if" statement, i.e., that the I-projection of each $q \in \mathcal{E}_{\mathbf{t}}(p^*)$ is p^*. To see this, note that for $p \in \mathcal{L}_{\mathbf{t}}(p^*)$ we have

$$\begin{aligned}
D(p\|q) &= \sum_y p(y)\log\frac{p(y)}{q(y)} \\
&= \sum_y p(y)\log\frac{p(y)}{p^*(y)e^{\mathbf{x}^{\mathrm{T}}\mathbf{t}(y)-\alpha(\mathbf{x})}} \\
&= D(p\|p^*) - (\log e)\sum_y p(y)\left[\mathbf{x}^{\mathrm{T}}\mathbf{t}(y) - \alpha(\mathbf{x})\right] \\
&= D(p\|p^*) - (\log e)\left[\mathbf{x}^{\mathrm{T}}\,\mathbb{E}_p\left[\mathbf{t}(y)\right] - \alpha(\mathbf{x})\right] \\
&= D(p\|p^*) - (\log e)\left[\mathbf{x}^{\mathrm{T}}\,\mathbb{E}_{p^*}\left[\mathbf{t}(y)\right] - \alpha(\mathbf{x})\right], \qquad (18)
\end{aligned}$$

where the last quality in follows from the fact that $p \in \mathcal{L}_{\mathbf{t}}(p^*)$. But since the term in brackets in (18) does not depend on p, it follows from the nonnegativity of divergence that

$$\arg\min_{p \in \mathcal{L}_{\mathbf{t}}(p^*)} D(p\|q) = \arg\min_{p \in \mathcal{L}_{\mathbf{t}}(p^*)} D(p\|p^*) = p^*.$$

Next, we prove the "only if" statement, i.e., that there are no distributions other than those in $\mathcal{E}_{\mathbf{t}}(p^*)$ whose I-projections are p^*. As depicted in Fig. 4, consider any distribution $r \notin \mathcal{E}_{\mathbf{t}}(p^*)$. Then

$$r(y) \not\propto p^*(y) \exp\left\{\mathbf{x}^{\mathrm{T}}\mathbf{t}(y)\right\}, \quad \text{for all } y \in \mathcal{Y}, \text{ any } \mathbf{x} \in \mathbb{R}^K. \tag{19}$$

Let p_0 be the intersection of $\mathcal{E}_{\mathbf{t}}(r) = \mathcal{E}_{\mathbf{t}}(p_0)$ with $\mathcal{L}_{\mathbf{t}}(p^*) = \mathcal{L}_{\mathbf{t}}(p_0)$, so from the "if" part of the theorem proved above we know that p_0 is the I-projection of r onto $\mathcal{L}_{\mathbf{t}}(p^*)$. Now since $p_0 \in \mathcal{E}_{\mathbf{t}}(r)$, there exists $\mathbf{x}_0 \in \mathbb{R}^K$ such that

$$p_0(y) \propto r(y) \exp\left\{\mathbf{x}_0^{\mathrm{T}}\mathbf{t}(y)\right\}, \quad \text{for all } y \in \mathcal{Y}. \tag{20}$$

But then substituting (19) into (20) we see

$$p_0(y) \not\propto p^*(y) \exp\left\{(\mathbf{x}+\mathbf{x}_0)^{\mathrm{T}}\mathbf{t}(y)\right\}, \quad \text{for all } y \in \mathcal{Y}, \text{ any } \mathbf{x} \in \mathbb{R}^K,$$

which for the particular choice $\mathbf{x} = -\mathbf{x}_0$ means that $p_0 \neq p^*$. Thus, since the I-projection onto a linear family is unique, p^* cannot be the I-projection of r onto $\mathcal{L}_{\mathbf{t}}(p^*)$ since p_0 is. □

Theorem 2 will have important implications and applications in later installments of the notes. However, for now, we simply observe that this theorem provides a straightforward procedure for computing the I-projection p^* of any distribution q onto a linear family specified by $\mathbf{t}(\cdot)$ and $\bar{\mathbf{t}}$. In particular, since p^* lies at the intersection of this linear family and the corresponding orthogonal family, we know

$$p^*(y) = q(y) \exp\left\{\mathbf{x}^{\mathrm{T}}\mathbf{t}(y) - \alpha(\mathbf{x})\right\}, \quad \text{for some } \mathbf{x} \in \mathbb{R}^K, \tag{21}$$

and

$$\sum_y \mathbf{t}(y)\, p^*(y) = \bar{\mathbf{t}}. \tag{22}$$

Substituting (21) into (22) we obtain

$$\sum_y \mathbf{t}(y)\, q(y) \exp\left\{\mathbf{x}^{\mathrm{T}}\mathbf{t}(y) - \alpha(\mathbf{x})\right\} = \bar{\mathbf{t}}.$$

which is a (nonlinear) system of K equations, which can be solved to obtain the K elements of the $\mathbf{x}$ vector corresponding to p^*.

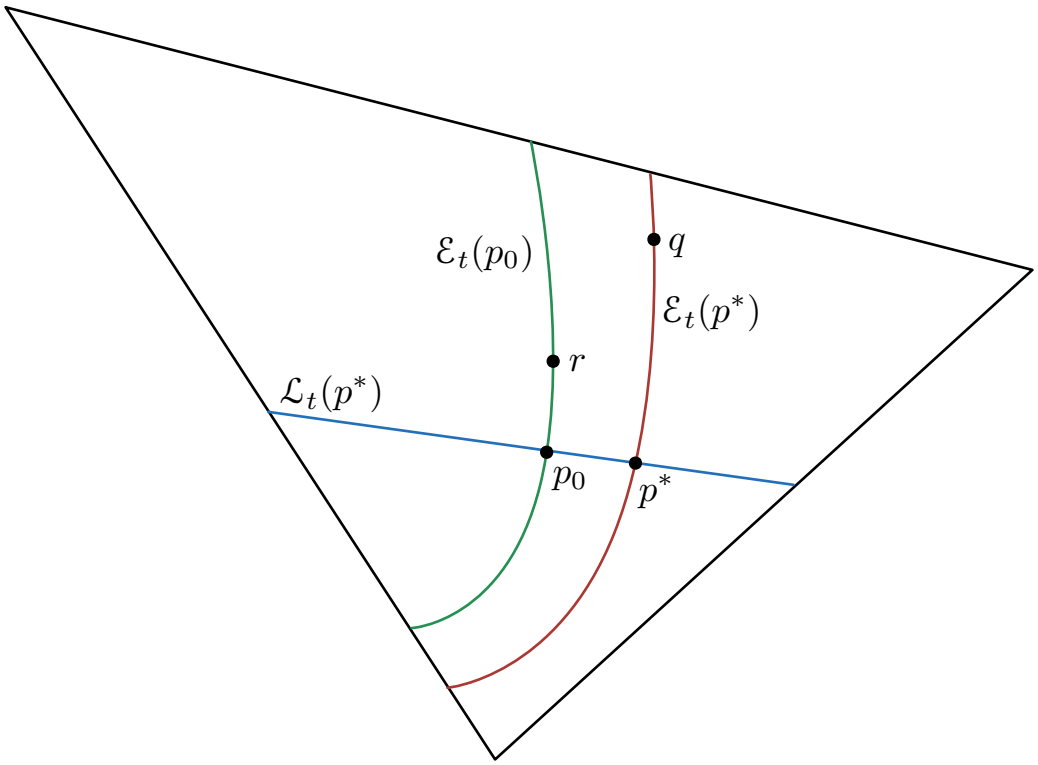

Figure 4: Orthogonality relationships in the simplex. The members of exponential family $\mathcal{E}_{\mathbf{t}}(p^*) = \mathcal{E}_{\mathbf{t}}(q)$, denoted by the red curve, all have an I-projection of p^* onto the linear family $\mathcal{L}_{\mathbf{t}}(p^*) = \mathcal{L}_{\mathbf{t}}(p_0)$, denoted by the blue line. The members of the exponential family $\mathcal{E}_{\mathbf{t}}(r) = \mathcal{E}_{\mathbf{t}}(p_0)$, denoted by the green curve, all have an I-projection of p_0 onto this linear family.

14.5 Local Information Geometry

As we have discussed, information geometry has some aspects of the structure of Euclidean geometry, and, in particular, information divergence behaves, in some respects, like the square of Euclidean distance.

In this section, we show that in fact in any sufficiently small neighborhood of the interior of the probability simplex for a finite alphabet $\mathcal{Y}$, information geometry is effectively Euclidean.

To see this, consider a neighborhood about a strictly positive distribution p_0 in the simplex. Let p denote an arbitrary distribution in this neighborhood, and let

$$\phi(y) \triangleq \frac{p(y) - p_0(y)}{\sqrt{2p_0(y)}}. \tag{23}$$

denote a normalized representation for p. Hence, we can write

$$p(y) = p_0(y) + \sqrt{2p_0(y)}\phi(y), \tag{24}$$

where we note the second term in (24) satisfies

$$\sum_y \sqrt{2p_0(y)}\phi(y) = 0. \tag{25}$$

Let the neighborhood be the ball

$$|\phi(y)| \leq B, \quad \text{for all } y.$$

Next, let ϕ_1 and ϕ_2 denote the normalized representations for two distributions p_1

and p_2, respectively, in this neighborhood. Then

$$\begin{aligned}
D(p_1\|p_2) &= \sum_y p_1(y)\log\frac{p_1(y)}{p_2(y)} \\
&= \sum_y \left(p_0(y)+\sqrt{2p_0(y)}\phi_1(y)\right)\log\frac{p_0(y)+\sqrt{2p_0(y)}\phi_1(y)}{p_0(y)+\sqrt{2p_0(y)}\phi_2(y)} \\
&= \sum_y \left(p_0(y)+\sqrt{2p_0(y)}\phi_1(y)\right)\left[\log\left(1+\frac{\sqrt{2}\phi_1(y)}{\sqrt{p_0(y)}}\right)-\log\left(1+\frac{\sqrt{2}\phi_2(y)}{\sqrt{p_0(y)}}\right)\right] \\
&= \sum_y \left(p_0(y)+\sqrt{2p_0(y)}\phi_1(y)\right)\left[-\frac{\sqrt{2}\phi_1(y)}{\sqrt{p_0(y)}}+\frac{\phi_1^2(y)}{p_0(y)}+\mathcal{O}(\phi_1^3(y))\right. \\
&\qquad \left.+\frac{\sqrt{2}\phi_2(y)}{\sqrt{p_0(y)}}-\frac{\phi_2^2(y)}{p_0(y)}+\mathcal{O}(\phi_2^3(y))\right], \phi_1(y),\phi_2(y)\to 0 \\
&= \underbrace{\sum_y \sqrt{2p_0(y)}\left(\phi_2(y)-\phi_1(y)\right)}_{D_1(\phi_1,\phi_2)} \\
&\qquad + \underbrace{\sum_y 2\phi_1(y)(\phi_2(y)-\phi_1(y))+(\phi_1^2(y)-\phi_2^2(y))}_{D_2(\phi_1,\phi_2)} \\
&\qquad\qquad + \underbrace{\sum_y \mathcal{O}(\phi_1(y)^3)+\mathcal{O}(\phi_2(y)^3)+\mathcal{O}(\phi_1(y)\phi_2(y)^2)}_{D_3(\phi_1,\phi_2)}
\end{aligned}$$

But

$$D_1(\phi_1,\phi_2)=0$$

due to (25),

$$D_2(\phi_1,\phi_2)=\sum_y |\phi_1(y)-\phi_2(y)|^2,$$

which is $\mathcal{O}(B^2)$, and

$$D_3(\phi_1,\phi_2)=\mathcal{O}(B^3),$$

whence

$$D(p_1\|p_2)=\sum_y |\phi_1(y)-\phi_2(y)|^2\,(1+o(1)),\quad \text{as } B\to 0 \tag{26}$$

Finally, if for $\mathcal{Y}=\{1,\ldots,M\}$ we express ϕ in vector form via

$$\boldsymbol{\phi}=\begin{bmatrix}\phi(1) & \cdots & \phi(M)\end{bmatrix}^{\mathrm{T}}, \tag{27}$$

we can write (26) as

$$D(p_1\|p_2)\cong\|\boldsymbol{\phi}_1-\boldsymbol{\phi}_2\|^2, \tag{28}$$

where the approximation is accurate for sufficiently small neighborhoods. Hence, locally, information divergence behaves exactly like the square of Euclidean distance.[2]

Further insight is obtained by examing the structure of a linear family $\mathcal{L}_{\mathbf{t}}(p^*)$ and its corresponding orthogonal exponential family $\mathcal{E}_{\mathbf{t}}(p^*)$ in a neighborhood of their intersection p^*.

Consider first the linear family, which are the distributions p that satisfy

$$\mathbb{E}_p\left[t_i(y)\right] = \bar{t}_i, \quad i = 1, \ldots, K. \tag{29}$$

Via (23) we can express (29) as

$$\sum_y \left(p^*(y) + \sqrt{2p^*(y)}\phi_{\mathcal{L}}(y)\right)\left(t_i(y) - \bar{t}_i\right) = \sum_y \sqrt{2p^*(y)}(t_i(y) - \bar{t}_i)\phi_{\mathcal{L}}(y) = 0 \tag{30}$$

for $i = 1, \ldots, K$, where

$$\phi_{\mathcal{L}}(y) = \frac{p(y) - p^*(y)}{\sqrt{2p^*(y)}}.$$

In turn, since (30) expresses the relation

$$\underbrace{\begin{bmatrix} (t_1(1) - \bar{t}_1) & \cdots & (t_1(M) - \bar{t}_1) \\ \vdots & \ddots & \vdots \\ (t_K(1) - \bar{t}_K) & \cdots & (t_K(M) - \bar{t}_K) \end{bmatrix}}_{=\tilde{\mathbf{T}}} \underbrace{\operatorname{diag}\left(\sqrt{p^*(1)}, \ldots, \sqrt{p^*(M)}\right)}_{\triangleq \mathbf{\Pi}} \boldsymbol{\phi}_{\mathcal{L}} = \mathbf{0}.$$

Additionally, (25) implies $\boldsymbol{\phi}_{\mathcal{L}}$ must satisfy

$$\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \mathbf{\Pi}\boldsymbol{\phi}_{\mathcal{L}} = 0, \tag{31}$$

Hence, it follows that $\mathcal{L}_{\mathbf{t}}(p^*)$ corresponds to the null space of $\tilde{\mathbf{T}}_+\mathbf{\Pi}$, where

$$\tilde{\mathbf{T}}_+ = \begin{bmatrix} (t_1(1) - \bar{t}_1) & \cdots & (t_1(M) - \bar{t}_1) \\ \vdots & \ddots & \vdots \\ (t_K(1) - \bar{t}_K) & \cdots & (t_K(M) - \bar{t}_K) \\ 1 & \ldots & 1 \end{bmatrix},$$

which has the dimension of the null space of $\tilde{\mathbf{T}}_+$ since $\mathbf{\Pi}$ is invertible. But applying simple row operations we see that the null space $\tilde{\mathbf{T}}_+$ is the null space of

$$\mathbf{T}_+ = \begin{bmatrix} t_1(1) & \cdots & t_1(M) \\ \vdots & \ddots & \vdots \\ t_K(1) & \cdots & t_K(M) \\ 1 & \ldots & 1 \end{bmatrix}.$$

[2]Phrased differently, information geometry is locally isomorphic to Euclidean geometry.

Equivalently, $\mathcal{L}_{\mathbf{t}}(p^*)$ is the subspace corresponding to the intersection of the null space of $\tilde{\mathbf{T}}\mathbf{\Pi}$ with the hyperplane corresponding to (31). Hence, when the $t_1(\cdot), \ldots, t_K(\cdot)$ and the constant function $t_0(\cdot) = 1$ are linearly independent, this means that $\mathcal{L}_{\mathbf{t}}(p^*)$ is a subspace of dimension $M - K - 1$.

Next we note that linear exponential families can be locally approximated as linear families. For example, consider the exponential family $\mathcal{E}_{\mathbf{t}}(p^*)$ in a neighborhood of p^*. By a Taylor series expansion we have, for sufficiently small $\|\mathbf{x}\|$,

$$\begin{aligned} p_{\mathsf{y}}(y;\mathbf{x}) &= p^*(y)\exp\left\{\sum_{i=1}^{K} x_i t_i(y) - \alpha(\mathbf{x})\right\} \\ &\cong p^*(y)\left[1 + \sum_{i=1}^{K} x_i t_i(y) - \alpha(\mathbf{x})\right] \qquad (32) \\ &\cong p^*(y)\left[1 + \sum_{i=1}^{K} x_i t_i(y) - \sum_{i=1}^{K} x_i \left(\frac{\partial}{\partial x_i}\alpha(\mathbf{x})\Big|_{\mathbf{x}=0}\right)\right] \qquad (33) \\ &= p^*(y)\left[1 + \sum_{i=1}^{K} x_i (t_i(y) - \bar{t}_i)\right], \qquad (34) \end{aligned}$$

where to obtain (32) we have used the approximation $e^u \cong 1+u$ valid for u sufficiently small u, where to obtain (33) we have used the Taylor series approximation

$$\alpha(\mathbf{x}) \cong \alpha(\mathbf{0}) + \left(\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}\alpha(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{0}}\right)\mathbf{x}, \quad \text{valid for sufficiently small } \|\mathbf{x}\|,$$

with $\alpha(\mathbf{0}) = 0$ since $p_{\mathsf{y}}(y;\mathbf{0}) = p^*(y)$, and where to obtain (34) we have used that

$$\frac{\partial}{\partial x_i}\alpha(\mathbf{x})\Big|_{\mathbf{x}=0} = \mathbb{E}_{p^*}\left[t_i(\mathsf{y})\right].$$

Note that (34) is nonnegative for sufficient small $\mathbf{x}$, and normalized, as required.

Using (23) with $p(\cdot) = p_{\mathsf{y}}(\cdot;\mathbf{x})$ as approximated by (34) and $p_0 = p^*$, we have

$$\phi_{\mathcal{E}}(y) = \frac{p_{\mathsf{y}}(y;\mathbf{x}) - p^*(y)}{\sqrt{2p^*(y)}} = \sqrt{\frac{p^*(y)}{2}}\sum_{i=1}^{K} x_i(t_i(y) - \bar{t}_i),$$

which we can equivalently write in the form

$$\boldsymbol{\phi}_{\mathcal{E}} = \frac{1}{\sqrt{2}}\underbrace{\operatorname{diag}\left(\sqrt{p^*(1)}, \ldots, \sqrt{p^*(M)}\right)}_{=\mathbf{\Pi}}\underbrace{\begin{bmatrix} (t_1(1) - \bar{t}_1) & \cdots & (t_K(1) - \bar{t}_K) \\ \vdots & \ddots & \vdots \\ (t_1(M) - \bar{t}_1) & \cdots & (t_K(M) - \bar{t}_K) \end{bmatrix}}_{=\tilde{\mathbf{T}}^{\mathrm{T}}}\begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix}, \qquad (35)$$

from which we see that in a neighborhood of p^*, the exponential family $\mathcal{E}_{\mathbf{t}}(p^*)$ corresponds to the column space of $(\tilde{\mathbf{T}}\boldsymbol{\Pi})^{\mathrm{T}}$, which since $\boldsymbol{\Pi}$ is invertible has the dimension of the column space of $\tilde{\mathbf{T}}^{\mathrm{T}}$.

But the column space of a matrix is the orthogonal complement of the null space of its transpose, so within the hyperplane defined by (31), and in a neighborhood of their intersection, $\mathcal{E}_{\mathbf{t}}(p^*)$ is orthogonal to $\mathcal{L}_{\mathbf{t}}(p^*)$ in a Euclidean sense, and moreover it is consists of *all* the distributions in the neighborhood that are orthogonal. In turn, this implies that when the $t_1(\cdot), \ldots, t_K(\cdot)$ and the constant function $t_0(\cdot) = 1$ are linearly independent, this means that $\mathcal{E}_{\mathbf{t}}(p^*)$ is a subspace of dimension K, since we saw in this case that $\mathcal{L}_{\mathbf{t}}(p^*)$ is a subspace of dimension $M - K - 1$.

In addition, this orthogonality manifests itself through the local version of Pythagoras' identity (16). In particular, combining (16) with (26) we obtain

$$\|\boldsymbol{\phi}_{\mathcal{L}} - \boldsymbol{\phi}_{\mathcal{E}}\|^2 = \|\boldsymbol{\phi}_{\mathcal{L}}\|^2 + \|\boldsymbol{\phi}_{\mathcal{E}}\|^2,$$

which is the Euclidean version of Pythagoras' identity and obviously expresses the relation $\boldsymbol{\phi}_{\mathcal{L}}^{\mathrm{T}}\boldsymbol{\phi}_{\mathcal{E}} = 0$.

14.6 Appendix: Convexity of Divergence

The convexity of divergence is most easily shown using the following inequality.

Lemma 1 (Log-Sum Inequality)**.** *For any $u_1, \ldots, u_K \geq 0$ and $v_1, \ldots, v_K \geq 0$, we have*

$$\sum_{i=1}^{K} u_i \log \frac{u_i}{v_i} \geq \left(\sum_{i=1}^{K} u_i \right) \log \frac{\sum_{i=1}^{K} u_i}{\sum_{i=1}^{K} v_i} \tag{36}$$

with equality if and only if $u_i \propto v_i$.

Proof. Due to our convention that $0 \log 0 \triangleq 0$, any terms in (36) for which $u_i = 0$ contribute zero to the left or right hand sides, so can be omitted from consideration. Likewise, if $v_i = 0$ for some i the left hand side of (36) is infinite so the claim is true. Hence it remains only to consider the case where all u_i and v_i are strictly positive.

Let $f(t) \triangleq t \log t$ and note this function is strictly convex, since $f''(t) = \log(e)/t > 0$ for $t > 0$. Hence, Jensen's inequality implies that

$$\sum_{i=1}^{K} \xi_i f(t_i) \geq f\left(\sum_{i=1}^{K} \xi_i t_i \right)$$

for ξ_i such that $\xi_i \geq 0$ and $\sum_{i=1}^{K} \xi_i = 1$, with equality if and only if the t_i are all identical. Choosing

$$\xi_i = \frac{v_i}{\sum_{j=1}^{K} v_j} \qquad \text{and} \qquad t_i = \frac{u_i}{v_i}$$

we obtain (36) and the condition for equality. □

The convexity of $D(p\|q)$ in p and q the follows from

$$\begin{aligned} & D(\lambda p_1 + (1-\lambda)p_2) \parallel \lambda q_1 + (1-\lambda)q_2) \\ & \quad = \sum_y [\lambda p_1(y) + (1-\lambda)p_2(y)] \log \frac{\lambda p_1(y) + (1-\lambda)p_2(y)}{\lambda q_1(y) + (1-\lambda)q_2(y)} \\ & \quad \leq \sum_y \left[\lambda p_1(y) \log \frac{\lambda p_1(y)}{\lambda q_1(y)} + (1-\lambda)p_2(y) \log \frac{(1-\lambda)p_2(y)}{(1-\lambda)q_2(y)} \right] \qquad (37) \\ & \quad = \lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2), \end{aligned}$$

which holds for all $0 \leq \lambda \leq 1$, and where to obtain (37) we have used (36) with $K = 2$, $u_1 = \lambda p_1(y)$, $u_2 = (1-\lambda)p_2(y)$, $v_1 = \lambda q_1(y)$, and $v_2 = (1-\lambda)q_2(y)$.

As a final comment, divergence being convex in (p, q) immediately implies it is convex in p and q individually. It suffices to set $q_1 = q_2 \triangleq q$ to see the former, and $p_1 = p_2 = p$ to see the latter.

14.7 Further reading

For additional perspectives and further detail on information geometry, see the tutorial "Information Theory and Statistics" by Csiszar and Shields, as well as the text by Cover and Thomas.