A. Derivation of FIM from KLD

In this appendix we show that $\frac{1}{2}\Delta^\intercal F(\boldsymbol{\theta})\Delta$ is a second order Taylor approximation of $D_{\text{KL}}(p(\boldsymbol{\theta})\|p(\boldsymbol{\theta}+\Delta))$. First, let

$$\begin{aligned} g_q(\boldsymbol{\theta}) :=& D_{\text{KL}}(q\|p(\boldsymbol{\theta})) \\ =& \sum_{\omega\in\Omega} q(\omega)\ln\left(\frac{q(\omega)}{p(\omega|\boldsymbol{\theta})}\right). \end{aligned}$$

We begin by deriving equations for the Jacobian and Hessian of g_q at $\boldsymbol{\theta}$:

$$\begin{aligned} \frac{\partial g_q(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} =& \sum_{\omega\in\Omega} q(\omega)\frac{p(\omega|\boldsymbol{\theta})}{q(\omega)}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{q(\omega)}{p(\omega|\boldsymbol{\theta})}\right) \\ =& \sum_{\omega\in\Omega} q(\omega)\frac{p(\omega|\boldsymbol{\theta})}{q(\omega)}\left(\frac{-q(\omega)\frac{\partial p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}}{p(\omega|\boldsymbol{\theta})^2}\right) \\ =& \sum_{\omega\in\Omega} -\frac{q(\omega)}{p(\omega|\boldsymbol{\theta})}\frac{\partial p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}, \end{aligned} \tag{4}$$

and so:

$$\begin{aligned} \frac{\partial^2 g_q(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2} =& \frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\partial g_q(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right) \\ =& -\sum_{\omega\in\Omega} q(\omega)\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1}{p(\omega|\boldsymbol{\theta})}\frac{\partial p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right) \\ =& -\sum_{\omega\in\Omega}\frac{q(\omega)}{p(\omega|\boldsymbol{\theta})}\frac{\partial^2 p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2} \\ &+\sum_{\omega\in\Omega}\frac{q(\omega)}{p(\omega|\boldsymbol{\theta})^2}\frac{\partial p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^\intercal \\ =& -\sum_{\omega\in\Omega}\frac{q(\omega)}{p(\omega|\boldsymbol{\theta})}\frac{\partial^2 p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2} \\ &+\sum_{\omega\in\Omega} q(\omega)\frac{\partial\ln p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\ln p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^\intercal. \end{aligned} \tag{5}$$

Next we compute a second order Taylor expansion of $g_q(\boldsymbol{\theta}+\Delta)$ around $g_q(\boldsymbol{\theta})$:

$$\begin{aligned} g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta}+\Delta) \overset{\text{Taylor}_2}{\approx}& g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta}) + \Delta^\intercal\frac{\partial g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \\ &+\frac{1}{2}\Delta^\intercal\frac{\partial^2 g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\Delta. \end{aligned} \tag{6}$$

Notice that

$$g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta}) = D_{\text{KL}}(p(\boldsymbol{\theta})\|p(\boldsymbol{\theta})) = 0,$$

and by (4)

$$\begin{aligned} \Delta^\intercal\frac{\partial g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} =& -\Delta^\intercal\sum_{\omega\in\Omega}\frac{p(\omega|\boldsymbol{\theta})}{p(\omega|\boldsymbol{\theta})}\frac{\partial p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \\ =& -\Delta^\intercal\frac{\partial}{\partial\boldsymbol{\theta}}\left(\sum_{\omega\in\Omega} p(\omega|\boldsymbol{\theta})\right) \\ \overset{\text{(a)}}{=}& 0, \end{aligned}$$

where **(a)** holds because

$$\sum_{\omega\in\Omega} p(\omega|\boldsymbol{\theta}) = 1,$$

so

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left(\sum_{\omega\in\Omega} p(\omega|\boldsymbol{\theta})\right) = \frac{\partial 1}{\partial\boldsymbol{\theta}} = \mathbf{0}. \tag{7}$$

Thus, the first two terms on the right side of (6) are zero, and thus:

$$g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta}+\Delta) \overset{\text{Taylor}_2}{\approx} \frac{1}{2}\Delta^\intercal\frac{\partial^2 g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\Delta. \tag{8}$$

Next we focus on the Hessian, (5), with $q = p(\boldsymbol{\theta})$:

$$\begin{aligned} \frac{\partial^2 g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2} =& \underbrace{-\sum_{\omega\in\Omega}\frac{p(\omega|\boldsymbol{\theta})}{p(\omega|\boldsymbol{\theta})}\frac{\partial^2 p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}}_{\overset{\text{(a)}}{=}0} \\ &+\sum_{\omega\in\Omega} p(\omega|\boldsymbol{\theta})\frac{\partial\ln p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\ln p(\omega|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^\intercal \\ =& F(\boldsymbol{\theta}), \end{aligned}$$

where **(a)** comes from taking the derivative of both sides of (7) with respect to $\boldsymbol{\theta}$. Substituting this into (8) we have that

$$g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta}+\Delta) \overset{\text{Taylor}_2}{\approx} \frac{1}{2}\Delta^\intercal F(\boldsymbol{\theta})\Delta.$$

B. Derivation of EIM from Energy Distance

In this section we show that $\Delta^\intercal\mathcal{E}(\boldsymbol{\theta})\Delta$ is a second order Taylor approximation of $D_{\text{E}}(p(\boldsymbol{\theta}), p(\boldsymbol{\theta}+\Delta))^2$. First, let

$$\begin{aligned} g_q(\boldsymbol{\theta}) :=& D_{\text{E}}(q, p(\boldsymbol{\theta})) \\ =& 2\sum_{\omega_1\in\Omega,\omega_2\in\Omega} q(\omega_1)p(\omega_2|\boldsymbol{\theta})d_q(\omega_1,\omega_2) \\ &-\sum_{\omega_1\in\Omega,\omega_2\in\Omega} p(\omega_1|\boldsymbol{\theta})p(\omega_2|\boldsymbol{\theta})d_q(\omega_1,\omega_2) \\ &-\sum_{\omega_1\in\Omega,\omega_2\in\Omega} q(\omega_1)q(\omega_2)d_q(\omega_1,\omega_2), \end{aligned}$$

where we use d_q to denote that d should be the distance metric at the distribution q. We begin by deriving an expression for the Jacobian of g_q at $\boldsymbol{\theta}$:

$$\begin{aligned}
\frac{\partial g_q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} =& \frac{\partial}{\partial \boldsymbol{\theta}} \Bigg(2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} q(\omega_1) p(\omega_2|\boldsymbol{\theta}) d_q(\omega_1, \omega_2) \\
& - \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} p(\omega_1|\boldsymbol{\theta}) p(\omega_2|\boldsymbol{\theta}) d_q(\omega_1, \omega_2) \\
& - \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} q(\omega_1) q(\omega_2) d_q(\omega_1, \omega_2) \Bigg) \\
=& 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) q(\omega_1) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
& - \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) \frac{\partial p(\omega_1|\boldsymbol{\theta}) p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
=& 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) q(\omega_1) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
& - \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) p(\omega_1|\boldsymbol{\theta}) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
& - \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) p(\omega_2|\boldsymbol{\theta}) \frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.
\end{aligned}$$

Notice that the last two lines are equal because d_q is symmetric—swap ω_1 and ω_2 in the last line, and you get the second to last line with $d_q(\omega_2, \omega_1) = d_q(\omega_1, \omega_2)$. So:

$$\begin{aligned}
\frac{\partial g_q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} =& 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) q(\omega_1) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
& - 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) p(\omega_1|\boldsymbol{\theta}) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
=& 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} (q(\omega_1) - p(\omega_1|\boldsymbol{\theta})).
\end{aligned} \tag{9}$$

Next we compute the Hessian of g_q at $\boldsymbol{\theta}$:

$$\begin{aligned}
\frac{\partial^2 g_q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} =& \frac{\partial}{\partial \boldsymbol{\theta}} \left(\frac{\partial g_q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \\
=& \frac{\partial}{\partial \boldsymbol{\theta}} \Bigg(2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} q(\omega_1) \Bigg) \\
& \frac{\partial}{\partial \boldsymbol{\theta}} \Bigg(- 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} p(\omega_1|\boldsymbol{\theta}) \Bigg) \\
=& 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) q(\omega_1) \frac{\partial^2 p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \\
& - 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) \frac{\partial}{\partial \boldsymbol{\theta}} \left(\frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} p(\omega_1|\boldsymbol{\theta}) \right) \\
=& 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) q(\omega_1) \frac{\partial^2 p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \\
& - 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial p(\omega_1|\boldsymbol{\theta})^\intercal}{\partial \boldsymbol{\theta}} \\
& - 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) p(\omega_1|\boldsymbol{\theta}) \frac{\partial^2 p(\omega_2|\boldsymbol{\theta})}{\partial^2 \boldsymbol{\theta})} \\
=& 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_q(\omega_1, \omega_2) (q(\omega_1) - p(\omega_1|\boldsymbol{\theta})) \frac{\partial^2 p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \\
& - 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_{p(\boldsymbol{\theta})}(\omega_1, \omega_2) \frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial p(\omega_2|\boldsymbol{\theta})^\intercal}{\partial \boldsymbol{\theta}}.
\end{aligned} \tag{10}$$

Next we compute a second order Taylor expansion of $g_q(\boldsymbol{\theta} + \Delta)$ around $g_q(\boldsymbol{\theta})$:

$$\begin{aligned}
g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta} + \Delta) \overset{\text{Taylor}_2}{\approx} & g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta}) + \Delta^\intercal \frac{\partial g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
& + \frac{1}{2} \Delta^\intercal \frac{\partial^2 g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \Delta.
\end{aligned} \tag{11}$$

Notice that

$$g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta}) = D_{\mathrm{E}}(p(\boldsymbol{\theta}), p(\boldsymbol{\theta})) = 0,$$

and by (9)

$$\begin{aligned}
\Delta^\intercal \frac{\partial g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} =& - \Delta^\intercal 2 \sum_{\omega_1 \in \Omega, \omega_2 \in \Omega} d_{p(\boldsymbol{\theta})}(\omega_1, \omega_2) \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
& \times (p(\omega_1|\boldsymbol{\theta}) - p(\omega_1|\boldsymbol{\theta})) \\
=& 0.
\end{aligned}$$

The first two terms on the right side of (11) are zero, and thus:

$$g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta} + \Delta) \overset{\text{Taylor}_2}{\approx} \frac{1}{2} \Delta^\intercal \frac{\partial^2 g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \Delta. \tag{12}$$

Next we focus on the Hessian, (10), with $q = p(\boldsymbol{\theta})$:

$$\begin{aligned}\frac{\partial^2 g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} =& 2\sum_{\omega_1\in\Omega,\omega_2\in\Omega} d_{p(\boldsymbol{\theta})}(\omega_1,\omega_2)\underbrace{(p(\omega_1|\boldsymbol{\theta}) - p(\omega_1|\boldsymbol{\theta}))}_{=0}\frac{\partial^2 p(\omega_2|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\\ &-2\sum_{\omega_1\in\Omega,\omega_2\in\Omega} d_{p(\boldsymbol{\theta})}(\omega_1,\omega_2)\frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^\intercal\\ =& 2\mathcal{E}(\boldsymbol{\theta}).\end{aligned}$$

Substituting this into (12) we have that

$$g_{p(\boldsymbol{\theta})}(\boldsymbol{\theta}+\Delta) \overset{\text{Taylor}_2}{\approx} \Delta^\intercal\mathcal{E}(\boldsymbol{\theta})\Delta.$$

C. Proof of Theorem 1

Let $|\Omega| = m$, and let D be a $m\times m$ distance matrix where $D_{ij} := d_{p(\boldsymbol{\theta})}(\omega_i,\omega_j)$. Let M be an $m\times n$ matrix where the i^{th} row is $\frac{\partial p(\omega_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$. The EIM can then be written as:

$$\mathcal{E}(\boldsymbol{\theta}) = -M^\intercal DM.$$

Recall from (7) that $\sum_{i=1}^m \frac{\partial p(\omega_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = 0$. This means that each column of M must sum to zero, and thus that for any $\mathbf{x}\in\mathbb{R}^n$, $M\mathbf{x}$ is a vector whose entries also sum to 0. Hence, if D is conditionally negative definite then $\mathcal{E}(\boldsymbol{\theta})$ is negative semidefinite since

$$\mathbf{x}^\intercal\mathcal{E}(\boldsymbol{\theta})\mathbf{x} = -\mathbf{x}^\intercal M^\intercal DM\mathbf{x} \overset{\text{(a)}}{\geq} 0,$$

for all $\mathbf{x}$, where **(a)** holds from the definition of conditionally positive semidefinite matrices.

D. Discussion of CND Distances

Conditionally negative definite distances are related to Euclidean distances, as shown by Schoenberg (1938).

Corollary 1. *Assume $|\Omega| < \infty$ and $\sqrt{d_{p(\boldsymbol{\theta})}}$ is a metric and Euclidean embeddable, that is, there exists a mapping ϕ from Ω to a Euclidean space with distance d' so that $\sqrt{d_{p(\boldsymbol{\theta})}(\omega_1,\omega_2)} = d'(\phi(\omega_1),\phi(\omega_2))$. Then $\mathcal{E}(\boldsymbol{\theta})$ is positive semidefinite. For $|\Omega|\leq 4$, every distance is Euclidean embeddable, therefore $\sqrt{d_{p(\boldsymbol{\theta})}}$ being a metric is sufficient in this case.*

The corollary follows directly from Theorem 1 and the work by Schoenberg (1938) and Rao (1984).

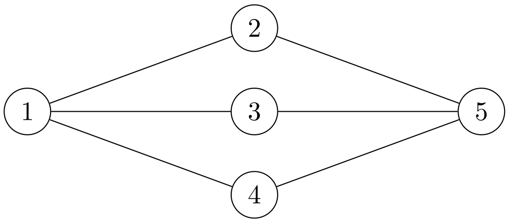

Figure 5: The distance defined by the length of the shortest path between two nodes is not conditionally negative definite.

We now provide an example of a distance metric d that is *not* conditionally negative definite. We define a distance d over the set $\Omega = \{1,2,3,4,5\}$ by the number of edges in the shortest path between two nodes in the graph depicted in Figure 5. For example, the distance between 1 and 5 is $d(1,5) = 2$, while $d(1,2) = 1$ and $d(1,1) = 0$. One can easily verify by enumeration that d is actually a distance: it satisfies $d(w_1,w_2)\geq 0$, and $d(w_1,w_2) = 0 \Leftrightarrow w_1 = w_2$, and the triangle inequality. The distance matrix, D, of d is

$$D = \begin{bmatrix} 0 & 1 & 1 & 1 & 2\\ 1 & 0 & 2 & 2 & 1\\ 1 & 2 & 0 & 2 & 1\\ 1 & 2 & 2 & 0 & 1\\ 2 & 1 & 1 & 1 & 0\end{bmatrix}.$$

The vector

$$x = \begin{bmatrix} -3\\ 2\\ 2\\ 2\\ -3\end{bmatrix}$$

satisfies $\sum_{i=1}^5 x_i = 0$ and gives $x^\intercal Dx = 12$. Hence, d is not conditionally negative semidefinite.

E. Proof of Theorem 2

We have

$$\begin{aligned}\mathcal{E}(\boldsymbol{\theta}) &= -\sum_{\omega_1\in\Omega,\omega_2\in\Omega} d_{p(\boldsymbol{\theta})}(\omega_1,\omega_2)\frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^\intercal\\ &= -\sum_{\omega_1\in\Omega,\omega_2\in\Omega}\left(\frac{\mathbf{1}_{(\omega_1\neq\omega_2)}}{2p(\omega_1|\boldsymbol{\theta})} + \frac{\mathbf{1}_{(\omega_1\neq\omega_2)}}{2p(\omega_2|\boldsymbol{\theta})}\right)\frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^\intercal\\ &= -\sum_{\omega_1\in\Omega}\frac{1}{2p(\omega_1|\boldsymbol{\theta})}\frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\sum_{\omega_2\neq\omega_1}\frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^\intercal\\ &\quad-\sum_{\omega_2\in\Omega}\frac{1}{2p(\omega_2|\boldsymbol{\theta})}\left(\sum_{\omega_1\neq\omega_2}\frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^\intercal.\end{aligned}$$

By (7) we have that:

$$\sum_{i\neq j}\frac{\partial p(i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = -\frac{\partial p(j|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}},$$

and so

$$\begin{aligned}\mathcal{E}(\boldsymbol{\theta}) =& \sum_{\omega_1 \in \Omega} \frac{1}{2p(\omega_1|\boldsymbol{\theta})} \frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial p(\omega_1|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^\intercal \\ &+ \sum_{\omega_2 \in \Omega} \frac{1}{2p(\omega_2|\boldsymbol{\theta})} \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^\intercal \\ =& \sum_{\omega \in \Omega} \frac{1}{p(\omega|\boldsymbol{\theta})} \frac{\partial p(\omega|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial p(\omega|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^\intercal \\ =& \sum_{\omega \in \Omega} p(\omega|\boldsymbol{\theta}) \frac{\partial \ln p(\omega|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\omega|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^\intercal \\ =& F(\boldsymbol{\theta}).\end{aligned}$$

F. Proof of Theorem 3

Here we present a formal definition of what it means for an update direction to be covariant before proving Theorem 3, which states that the energetic natural gradient is a covariant update direction. Intuitively, an update is covariant if the direction of an update in the space of probability distributions does not depend on the parametrization of the space of probability distributions. We provide a (possibly unintuitive) formal definition below, which comes from the work of Dabney & Thomas (2014, Lemma 1).

Definition 1 (Congruency of PPMs)**.** *We say that two PPMs, p with parameters $\boldsymbol{\theta} \in \mathbb{R}^n$ and q with parameters $\boldsymbol{\phi} \in \mathbb{R}^n$, are* **congruent** *if there exists a continuous function $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ such that for all $\boldsymbol{\theta}$:*

$$p(\boldsymbol{\theta}) = q(\Phi(\boldsymbol{\theta})),$$

and the Jacobian of Φ is full rank.

Definition 2 (Covariant Update)**.** *The update direction $\widetilde{\nabla}$ is* **covariant** *if, for all congruent PPMs, p and q, and all $\boldsymbol{\theta} \in \mathbb{R}^n$:*

$$\widetilde{\nabla}(f \circ q)(\Phi(\boldsymbol{\theta})) = \frac{\partial \Phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \widetilde{\nabla}(f \circ p)(\boldsymbol{\theta}). \tag{13}$$

We now prove that the Energetic natural gradient is a covariant update direction. Our proof is similar to that of Dabney & Thomas (2014), who show that a broad class of natural gradient algorithms (not including the energetic natural gradient) are covariant. First notice that by the chain rule:

$$\frac{\partial \ln q(\omega|\Phi(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{\partial \Phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln q(\omega|\Phi(\boldsymbol{\theta}))}{\partial \Phi(\boldsymbol{\theta})},$$

and so, since the Jacobian of $\Phi(\boldsymbol{\theta})$ is full rank:

$$\frac{\partial \Phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^{-1} \frac{\partial \ln q(\omega|\Phi(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{\partial \ln q(\omega|\Phi(\boldsymbol{\theta}))}{\partial \Phi(\boldsymbol{\theta})}. \tag{14}$$

Now consider $\mathcal{E}(\Phi(\boldsymbol{\theta}))$, where we write $\boldsymbol{\phi}$ as shorthand for $\Phi(\boldsymbol{\theta})$. Below, ... denotes that a long line was split onto two lines.

$$\begin{aligned}&\mathcal{E}(\boldsymbol{\phi}) \\ &= -\mathbf{E}_{\substack{\omega_1 \sim q(\boldsymbol{\phi}) \\ \omega_2 \sim q(\boldsymbol{\phi})}} \left[d_{q(\boldsymbol{\phi})}(\omega_1, \omega_2) \frac{\partial \ln q(\omega_1|\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \frac{\partial \ln q(\omega_2|\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}^\intercal \right] \\ &\overset{\text{(a)}}{=} -\mathbf{E}_{\substack{\omega_1 \sim q(\boldsymbol{\phi}) \\ \omega_2 \sim q(\boldsymbol{\phi})}} \Bigg[d_{q(\boldsymbol{\phi})}(\omega_1, \omega_2) \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \frac{\partial \ln q(\omega_1|\boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \\ &\quad \ldots \frac{\partial \ln q(\omega_2|\boldsymbol{\phi})}{\partial \boldsymbol{\theta}}^\intercal \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \right)^\intercal \Bigg] \\ &= -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \mathbf{E}_{\substack{\omega_1 \sim q(\boldsymbol{\phi}) \\ \omega_2 \sim q(\boldsymbol{\phi})}} \Bigg[d_{q(\boldsymbol{\phi})}(\omega_1, \omega_2) \frac{\partial \ln q(\omega_1|\boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \\ &\quad \ldots \frac{\partial \ln q(\omega_2|\boldsymbol{\phi})}{\partial \boldsymbol{\theta}}^\intercal \Bigg] \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \right)^\intercal,\end{aligned}$$

where **(a)** comes from (14). Since $q(\boldsymbol{\phi}) = p(\boldsymbol{\theta})$ we have that:

$$\begin{aligned}&\mathcal{E}(\boldsymbol{\phi}) \\ &= -\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \mathbf{E}_{\substack{\omega_1 \sim p(\boldsymbol{\theta}) \\ \omega_2 \sim p(\boldsymbol{\theta})}} \left[d_{p(\boldsymbol{\theta})}(\omega_1, \omega_2) \frac{\partial \ln p(\omega_1|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\omega_2|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^\intercal \right] \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \right)^\intercal \\ &= \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \mathcal{E}(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \right)^\intercal.\end{aligned}$$

So, we have that the left side of (13) is:

$$\mathcal{E}(\boldsymbol{\phi})^+ \frac{\partial (f \circ q)(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \left[\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \mathcal{E}(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \right)^\intercal \right]^+ \frac{\partial (f \circ q)(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}. \tag{15}$$

We can use the chain rule as before to show that

$$\frac{\partial (f \circ q)(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \frac{\partial (f \circ q)(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}},$$

and so continuing (15) we have that

$$\begin{aligned}\mathcal{E}(\boldsymbol{\phi})^+ \frac{\partial (f \circ q)(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} &= \left[\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \mathcal{E}(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \right)^\intercal \right]^+ \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \frac{\partial (f \circ q)(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \\ &\overset{\text{(a)}}{=} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta})^+ \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} \frac{\partial (f \circ q)(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \\ &= \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta})^+ \frac{\partial (f \circ q)(\boldsymbol{\phi})}{\partial \boldsymbol{\theta}} \\ &= \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}} \widetilde{\nabla}(f \circ p)(\boldsymbol{\theta}),\end{aligned}$$

where **(a)** comes from the assumption that $\partial \boldsymbol{\phi} / \partial \boldsymbol{\theta}$ has full rank, and so $[\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}^{-1} A]^+ = A^+ \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}$ for any matrix A. We therefore have that (13) holds for the energetic natural gradient.