Article

On the Fisher Metric of Conditional Probability Polytopes

Guido Montúfar [1,*], **Johannes Rauh** [1] **and Nihat Ay** [1,2,3]

[1] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, Leipzig 04103, Germany;
E-Mails: jrauh@mis.mpg.de (J.R.); nay@mis.mpg.de (N.A.)

[2] Department of Mathematics and Computer Science, Leipzig University, PF 10 09 20,
Leipzig 04009, Germany

[3] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

* Author to whom correspondence should be addressed; E-Mail: montufar@mis.mpg.de;
Tel.: +49-341-9959-521.

Abstract: We consider three different approaches to define natural Riemannian metrics on polytopes of stochastic matrices. First, we define a natural class of stochastic maps between these polytopes and give a metric characterization of Chentsov type in terms of invariance with respect to these maps. Second, we consider the Fisher metric defined on arbitrary polytopes through their embeddings as exponential families in the probability simplex. We show that these metrics can also be characterized by an invariance principle with respect to morphisms of exponential families. Third, we consider the Fisher metric resulting from embedding the polytope of stochastic matrices in a simplex of joint distributions by specifying a marginal distribution. All three approaches result in slight variations of products of Fisher metrics. This is consistent with the nature of polytopes of stochastic matrices, which are Cartesian products of probability simplices. The first approach yields a scaled product of Fisher metrics; the second, a product of Fisher metrics; and the third, a product of Fisher metrics scaled by the marginal distribution.

1. Introduction

The Riemannian structure of a function's domain has a crucial impact on the performance of gradient optimization methods, especially in the presence of plateaus and local maxima. The natural gradient [1] gives the steepest increase direction of functions on a Riemannian space. For example, artificial neural networks can often be trained by following some function's gradient on a space of probabilities. In this context, it has been observed that following the natural gradient with respect to the Fisher information metric, instead of the Euclidean metric, can significantly alleviate the plateau problem [1,2]. The Fisher information metric, which is also called Shahshahani metric [3] in biological contexts, is broadly recognized as the natural metric of probability spaces. An important argument was given by Chentsov [4], who showed that the Fisher information metric is the only metric on probability spaces for which certain natural statistical embeddings, called Markov morphisms, are isometries. More generally, Chentsov's theorem characterizes the Fisher metric and α-connections of statistical manifolds uniquely (up to a multiplicative constant) by requiring invariance with respect to Markov morphisms. Campbell [5] gave another proof that characterizes invariant metrics on the set of non-normalized positive measures, which restrict to the Fisher metric in the case of probability measures (up to a multiplicative constant). In this paper, we explore ways of defining distinguished Riemannian metrics on spaces of stochastic matrices.

In learning theory, when modeling the policy of a system, it is often preferred to consider stochastic matrices instead of joint probability distributions. For example, in robotics applications, policies are optimized over a parametric set of stochastic matrices by following the gradient of a reward function [6,7]. The set of stochastic matrices can be parametrized in many ways, e.g., in terms of feedforward neural networks, Boltzmann machines [8] or projections of exponential families [9]. The information geometry of policy models plays an important role in these applications and has been studied by Kakade [2], Peters and co-workers [10–12], and Bagnell and Schneider [13], among others. A stochastic matrix is a tuple of probability distributions, and therefore, the space of stochastic matrices is a Cartesian product of probability simplices. Accordingly, in applications, usually a product metric is considered, with the usual Fisher metric on each factor. On the other hand, Lebanon [14] takes an axiomatic approach, following the ideas of Chentsov and Campbell, and characterizes a class of invariant metrics of positive matrices that restricts to the product of Fisher metrics in the case of stochastic matrices. We will consider three different approaches discussed in the following.

In the first part, we take another look at Lebanon's approach for characterizing a distinguished metric on polytopes of stochastic matrices. However, since the maps considered by Lebanon do not map stochastic matrices to stochastic matrices, we will use different maps. We show that the product of Fisher metrics can be characterized by an invariance principle with respect to natural maps between stochastic matrices.

In the second part, we consider an approach that allows us to define Riemannian structures on arbitrary polytopes. Any polytope can be identified with an exponential family by using the coordinates of the polytope vertices as observables. The inverse of the moment map then defines an embedding of the polytope in a probability simplex. This embedding can be used to pull back geometric structures from the probability simplex to the polytope, including Riemannian metrics, affine connections, divergences, *etc*.

This approach has been considered in [9] as a way to define low-dimensional families of conditional probability distributions. More general embeddings can be defined by identifying each exponential family with a point configuration, $\mathbf{B}$, together with a weight function, ν. Given $\mathbf{B}$ and ν, the corresponding exponential family defines geometric structures on the set $(\text{conv}\,\mathbf{B})^\circ$, which is the relative interior of the convex support of the exponential family. Moreover, we can define natural morphisms between weighted point configurations as surjective maps between the point sets, which are compatible with the weight functions. As it turns out, the Fisher metric on $(\text{conv}\,\mathbf{B})^\circ$ can be characterized by invariance under these maps.

In the third part, we return to stochastic matrices. We study natural embeddings of conditional distributions in probability simplices as joint distributions with a fixed marginal. These embeddings define a Fisher metric equal to a weighted product of Fisher metrics. This result corresponds to the definitions commonly used in robotics applications.

All three approaches give very similar results. In all cases, the identified metric is a product metric. This is a sensible result, since the set of $k \times m$ stochastic matrices is a Cartesian product of probability simplices $\Delta_{m-1} \times \cdots \times \Delta_{m-1} = \Delta_{m-1}^k$, which suggests using the product metric of the Fisher metrics defined on the factor simplices, Δ_{m-1}. Indeed, this is the result obtained from our second approach. The first approach yields that same result with an additional scaling factor of $1/k$. Only when stochastic matrices of different sizes are compared, the two approaches differ. The third approach yields a product of Fisher metrics scaled by the marginal distribution that defines the embedding.

Which metric to use depends on the concrete problem and whether a natural marginal distribution is defined and known. In Section 7, we do a case study using a reward function that is given as an expectation value over a joint distribution. In this simple example, the weighted product metric gives the best asymptotic rate of convergence, under the assumption that the weights are optimally chosen. In Section 8, we sum up our findings.

The contents of the paper is organized as follows. Section 2 contains basic definitions around the Fisher metric and concepts of differential geometry. In Section 3, we discuss the theorems of Chentsov, Campbell and Lebanon, which characterize natural geometric structures on the probability simplex, on the set of positive measures and on the cone of positive matrices, respectively. In Section 4, we study metrics on polytopes of stochastic matrices, which are invariant under natural embeddings. In Section 5, we define a Riemannian structure for polytopes, which generalizes the Fisher information metric of probability simplices and conditional models in a natural way. In Section 6, we study a class of weighted product metrics. In Section 7, we study the gradient flow with respect to an expectation value. Section 8 contains concluding remarks. In Appendix A, we investigate restrictions on the parameters of the metrics characterized in Sections 3 and 4 that make them positive definite. Appendix B contains the proofs of the results from Section 4.

2. Preliminaries

We will consider the simplex of probability distributions on $[m] := \{1, \ldots, m\}$, $m \geq 2$, which is given by $\Delta_{m-1} := \{(p_i)_i \in \mathbb{R}^m \colon p_i \geq 0, \sum_i p_i = 1\}$. The relative interior of Δ_{m-1} consists of all strictly positive probability distributions on $[m]$, and will be denoted Δ_{m-1}°. This is a subset of $\mathbb{R}_+^m$,

the cone of strictly positive vectors. The set of $k \times m$ row-stochastic matrices is given by $\Delta_{m-1}^k :=$ $\{(K_{ij})_{ij} \in \mathbb{R}^{k \times m} : (K_{ij})_j \in \Delta_{m-1} \text{ for all } i \in [k]\}$ and is equal to the Cartesian product $\times_{i \in [k]} \Delta_{m-1}$. The relative interior $(\Delta_{m-1}^k)^\circ$ is a subset of $\mathbb{R}_+^{k \times m}$, the cone of strictly positive matrices.

Given two random variables X and Y taking values in the finite sets $[k]$ and $[m]$, respectively, the conditional probability distribution of Y given X is the stochastic matrix $K = (P(y|x))_{x \in [k], y \in [m]}$ with rows $(P(y|x))_{y \in [m]} \in \Delta_{m-1}$ for all $x \in [k]$. Therefore, the polytope of stochastic matrices Δ_{m-1}^k is called a conditional polytope.

The tangent space of $\mathbb{R}_+^n$ at a point $p \in \mathbb{R}_+^n$, denoted by $T_p\mathbb{R}_+^n$, is the real vector space spanned by the vectors $\partial_1, \ldots, \partial_n$ of partial derivatives with respect to the n components. The tangent space of Δ_{n-1}° at a point $p \in \Delta_{n-1}^\circ \subset \mathbb{R}_+^n$ is the subspace $T_p\Delta_{n-1}^\circ \subset T_p\mathbb{R}_+^n$ consisting of the vectors:

$$u = \sum_i u_i \partial_i \in T_p\mathbb{R}_+^n \quad \text{with} \quad \sum_i u_i = 0. \tag{1}$$

The Fisher metric on the positive probability simplex Δ_{n-1}° is the Riemannian metric given by:

$$g_p^{(n)}(u, v) = \sum_{i=1}^n \frac{u_i v_i}{p_i}, \quad \text{for all } u, v \in T_p\Delta_{n-1}^\circ. \tag{2}$$

The same formula (2) also defines a Riemannian metric on $\mathbb{R}_+^n$, which we will denote by the same symbol. This, however, is not the only way in which the Fisher metric can be extended from Δ_{n-1}° to $\mathbb{R}_+^n$. We will discuss other extensions in the next section (see Campbell's theorem, Theorem 3).

Consider a smoothly parametrized family of probability distributions $\mathcal{M} = \{(p(x; \theta))_{x \in [n]} : \theta \in \Omega\} \subseteq \Delta_{n-1}^\circ$, where $\Omega \subseteq \mathbb{R}^d$ is open. Then, $g^{(n)}$ induces a Riemannian metric on $\mathcal{M}$. Denote by $\partial_{\theta_i} = \frac{\partial}{\partial \theta_i}$ the tangent vector corresponding to the partial derivative with respect to θ_i, for all $i \in [d]$. Then, the Fisher matrix has coordinates:

$$g_\theta^{\mathcal{M}}(\partial_{\theta_i}, \partial_{\theta_j}) = \sum_{x \in [n]} p(x; \theta) \frac{\partial \log p(x; \theta)}{\partial \theta_i} \frac{\partial \log p(x; \theta)}{\partial \theta_j}, \quad \text{for all } i, j \in [d], \quad \text{for all } \theta \in \Omega. \tag{3}$$

Here, it is not necessary to assume that the parameters θ_i are independent. In particular, the dimension of $\mathcal{M}$ may be smaller than d, in which case the matrix is not positive definite. If the map $\Omega \to \mathcal{M}, \theta \mapsto p(\cdot; \theta)$ is an embedding (*i.e.*, a smooth injective map that is a diffeomorphism onto its image), then $g_\theta^{\mathcal{M}}$ defines a Riemannian metric on Ω, which corresponds to the pull-back of $g^{(n)}$.

Consider an embedding $f : \mathcal{E} \to \mathcal{E}'$. The pull-back of a metric g' on $\mathcal{E}'$ through f is defined as:

$$(f^* g')_p(u, v) := g'_{f(p)}(f_* u, f_* v), \quad \text{for all } u, v \in T_p\mathcal{E}, \tag{4}$$

where f_* denotes the push-forward of $T_p\mathcal{E}$ through f, which in coordinates is given by:

$$f_* : \quad T_p\mathcal{E} \to T_{f(p)}\mathcal{E}'; \quad \sum_i u_i \partial_{\theta_i} \mapsto \sum_j \sum_i u_i \frac{\partial f_j(p)}{\partial \theta_i} \partial_{\theta'_j}, \tag{5}$$

where $\{\partial_{\theta_i}\}_i$ spans $T_q\mathcal{E}$ and $\{\partial_{\theta'_j}\}_j$ spans $T_{f(p)}\mathcal{E}'$.

An embedding $f : \mathcal{E} \to \mathcal{E}'$ of two Riemannian manifolds $(\mathcal{E}, g)$ and $(\mathcal{E}', g')$ is an isometry iff:

$$g_p(u, v) = (f^* g')_p(u, v), \quad \text{for all } p \in \mathcal{E} \text{ and } u, v \in T_p\mathcal{E}. \tag{6}$$

In this case, we say that the metric g is invariant with respect to f (and g').

3. The Results of Campbell and Lebanon

One of the theoretical motivations for using the Fisher metric is provided by Chentsov's characterization [4], which states that the Fisher metric is uniquely specified, up to a multiplicative constant, by an invariance principle under a class of stochastic maps, called Markov morphisms. Later, Campbell [5] considered the characterization problem on the space $\mathbb{R}^n_+$ instead of Δ°_{n-1}. This simplifies the computations, since $\mathbb{R}^n_+$ has a more symmetric parametrization.

Definition 1. Let $2 \leq m \leq n$. A (row) stochastic partition matrix (or just row-partition matrix) is a matrix $Q \in \mathbb{R}^{m \times n}$ of non-negative entries, which satisfies $\sum_{j \in A_{i'}} Q_{ij} = \delta_{ii'}$ for an m block partition $\{A_1, \ldots, A_m\}$ of $[n]$. The linear map defined by:

$$\mathbb{R}^m_+ \to \mathbb{R}^n_+; \quad p \mapsto p \cdot Q \tag{7}$$

is called a congruent embedding by a Markov mapping of $\mathbb{R}^m_+$ to $\mathbb{R}^n_+$ or just a Markov map, for short.

An example of a 3×5 row-partition matrix is:

$$Q = \begin{pmatrix} 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/3 & 0 & 2/3 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{8}$$

Markov maps preserve the 1-norm and restrict to embeddings $\Delta^\circ_{m-1} \to \Delta^\circ_{n-1}$.

Theorem 2 (Chentsov's theorem.)**.**

- *Let $g^{(m)}$ be a Riemannian metric on Δ°_{m-1} for $m \in \{2, 3, \ldots\}$. Let this sequence of metrics have the property that every congruent embedding by a Markov mapping is an isometry. Then, there is a constant $C > 0$ that satisfies:*

$$g_p^{(m)}(u, v) = C \sum_i \frac{u_i v_i}{p_i}. \tag{9}$$

- *Conversely, for any $C > 0$, the metrics given by Equation (9) define a sequence of Riemannian metrics under which every congruent embedding by a Markov mapping is an isometry.*

The main result in Campbell's work [5] is the following variant of Chentsov's theorem.

Theorem 3 (Campbell's theorem.)**.**

- *Let $g^{(m)}$ be a Riemannian metric on $\mathbb{R}^m_+$ for $m \in \{2, 3, \ldots\}$. Let this sequence of metrics have the property that every embedding by a Markov mapping is an isometry. Then:*

$$g_p^{(m)}(\partial_i, \partial_j) = A(|p|) + \delta_{ij} C(|p|) \frac{|p|}{p_i}, \tag{10}$$

- *where $|p| = \sum_{i=1}^m p_i$, δ_{ij} is the Kronecker delta, and A and C are C^∞ functions on $\mathbb{R}_+$ satisfying $C(\alpha) > 0$ and $A(\alpha) + C(\alpha) > 0$ for all $\alpha > 0$.*
- *Conversely, if A and C are C^∞ functions on $\mathbb{R}^+$ satisfying $C(\alpha) > 0$, $A(\alpha) + C(\alpha) > 0$ for all $\alpha > 0$, then Equation (10) defines a sequence of Riemannian metrics under which every embedding by a Markov mapping is an isometry.*

The metrics from Campbell's theorem also define metrics on the probability simplices Δ_{m-1}° for $m = 2, 3, \ldots$. Since the tangent vectors $v = \sum_i v_i \partial_i \in T_p \Delta_{m-1}^\circ$ satisfy $\sum_i v_i = 0$, for any two vectors $u, v \in T_p \Delta_{m-1}^\circ$, also $\sum_i \sum_j A u_i v_j = 0$ for any A. In this case, the choice of A is immaterial, and the metric becomes Chentsov's metric.

Remark 4. Observe that Chentsov's theorem is not a direct implication of Campbell's theorem. However, it can be deduced from it by the following arguments. Suppose that we have a family of Riemannian simplices $(\Delta_{m-1}^\circ, g^{(m)})$ for $m \in \{2, 3, \ldots\}$, and suppose that they are isometric with respect to Markov maps. If we can extend every $g^{(m)}$ to a Riemannian metric $\tilde{g}^{(m)}$ on $\mathbb{R}_+^m$ in such a way that the resulting spaces $(\mathbb{R}_+^m, \tilde{g}^{(m)})$ are still isometric with respect to Markov maps, then Campbell's theorem implies that $g^{(m)}$ is a multiple of the Fisher metric. Such metric extensions can be defined as follows. Consider the diffeomorphism:

$$\Delta_{m-1}^\circ \times \mathbb{R}_+ \cong \mathbb{R}_+^m, \quad (p, r) \mapsto r \cdot p. \tag{11}$$

Any tangent vector $u \in T_{(p,r)} \mathbb{R}_+^m$ can be written uniquely as $u = u_p + u_r \partial_r$, where u_p is tangent to $r \Delta_{m-1}^\circ$. Since each Markov map f preserves the one-norm $| \cdot |$, its push-forward f_* maps the tangent vector $\partial_r \in T_{(p,r)} \mathbb{R}_+^m$ to the corresponding tangent vector $\partial_r \in T_{f(p,r)} \mathbb{R}_+^m$; that is, $f_* u = f_* u_p + u_r \partial_r$. Therefore,

$$\tilde{g}_{(p,r)}^{(m)}(u, v) := g_p^{(m)}(u_p, v_p) + u_r v_r \tag{12}$$

is a metric on $\mathbb{R}_+^m$ that is invariant under f.

In what follows, we will focus on positive matrices. In order to define a natural Riemannian metric, we can use the identification $\mathbb{R}_+^{k \times m} \cong \mathbb{R}_+^{km}$ and apply Campbell's theorem. This leads to metrics of the form:

$$g_M^{(k,m)}(\partial_{ij}, \partial_{kl}) = A(|M|) + \delta_{ik} \delta_{jl} C(|M|) / M_{ij}, \tag{13}$$

where $\partial_{ij} = \frac{\partial}{\partial M_{ij}}$ and $|M| = \sum_{ij} M_{ij}$. However, a disadvantage of this approach is that the action of general Markov maps on $\mathbb{R}_+^{km}$ has no natural interpretation in terms of the matrix structure. Therefore, Lebanon [14] considered a special class of Markov maps defined as follows.

Definition 5. Consider a $k \times l$ row-partition matrix R and a collection of $m \times n$ row-partition matrices $Q = \{Q^{(1)}, \ldots, Q^{(k)}\}$. The map:

$$\mathbb{R}_+^{k \times m} \to \mathbb{R}_+^{l \times n}; \quad M \mapsto R^\top (M \otimes Q) \tag{14}$$

is called a congruent embedding by a Markov morphism of $\mathbb{R}_+^{k \times m}$ to $\mathbb{R}_+^{l \times n}$ in [15]. We will refer to such an embedding as a Lebanon map. Here, the row product $M \otimes Q$ is defined by:

$$(M \otimes Q)_{ab} = (M \cdot Q^{(a)})_{ab}, \quad \text{for all } a \in [k], b \in [n]; \tag{15}$$

that is, the a-th row of M is multiplied by the matrix $Q^{(a)}$.

In a Lebanon map, each row of the input matrix M is mapped by an individual Markov mapping $Q^{(i)}$, and each resulting row is copied and scaled by an entry of R. This kind of map preserves the sum of all matrix entries. Therefore, with the identification $\mathbb{R}_+^{k \times m} \cong \mathbb{R}_+^{km}$, each Lebanon map restricts to a

map $\Delta_{mk-1}^{\circ} \to \Delta_{nl-1}^{\circ}$. The set Δ_{mk-1}° can be identified with the set of joint distributions of two random variables. Lebanon maps can be regarded as special Markov maps that incorporate the product structure present in the set of joint probability distributions of a pair of random variables. In Section 4, we will give an interpretation of these maps.

Contrary to what is stated in [15], a Lebanon map does not map $(\Delta_{m-1}^{k})^{\circ}$ to $(\Delta_{n-1}^{l})^{\circ}$, unless $k = l$. Therefore, later, we will provide a characterization for the metrics on $(\Delta_{m-1}^{k})^{\circ}$ in terms of invariance under other maps (which are not Markov nor Lebanon maps).

The main result in Lebanon's work [15, Theorems 1 and 2] is the following.

Theorem 6 (Lebanon's theorem.)**.**

- *For each $k \geq 1, m \geq 2$, let $g^{(k,m)}$ be a Riemannian metric on $\mathbb{R}_{+}^{k \times m}$ in such a way that every Lebanon map is an isometry. Then:*

$$g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = A(|M|) + \delta_{ac}\left(\frac{B(|M|)}{|M_a|} + \delta_{bd}\frac{C(|M|)}{M_{ab}}\right) \tag{16}$$

  for some differentiable functions $A, B, C \in C^{\infty}(\mathbb{R}_+)$.
- *Conversely, let $\{(\mathbb{R}_{+}^{k \times m}, g^{(k,m)})\}$ be a sequence of Riemannian manifolds, with metrics $g^{(k,m)}$ of the form (16) for some $A, B, C \in C^{\infty}(\mathbb{R}_+)$. Then, every Lebanon map is an isometry.*

Lebanon does not study the question under which assumptions on $A, B, C \in C^{\infty}(\mathbb{R}_+)$ the formula (16) does indeed define a Riemannian metric. This question has the following simple answer, which we will prove in Appendix A:

Proposition 7. *The matrix (16) is positive definite if and only if $C(|M|) > 0$, $B(|M|) + C(|M|) > 0$ and $A(|M|) + B(|M|) + C(|M|) > 0$.*

The class of metrics (16) is larger than the class of metrics (13) derived in Campbell's theorem. The reason is that Campbell's metrics are invariant with respect to a larger class of embeddings.

The special case with $A(|M|) = 0$, $B(|M|) = 0$ and $C(|M|) = 1$ is called product Fisher metric,

$$g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = \delta_{ac}\delta_{bd}\frac{1}{M_{ab}}. \tag{17}$$

Furthermore, if we restrict to $(\Delta_{m-1}^{k})^{\circ}$, the functions A and B do not play any role. In this case $|M| = k$, and we obtain the scaled product Fisher metric:

$$g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = \delta_{ac}\delta_{bd}\frac{C(k)}{M_{ab}}, \tag{18}$$

where $C(k) : \mathbb{N} \to \mathbb{R}_+$ is a positive function. As mentioned before, Lebanon's theorem does not give a characterization of invariant metrics of stochastic matrices, since Lebanon maps do not preserve the stochasticity of the matrices. However, Lebanon maps are natural maps on the set Δ_{mk-1}° of positive joint distributions. In the same way as Chentsov's theorem can be derived from Campbell's theorem (see Remark 4), we obtain the following corollary:

Corollary 8.

- *Let $\{(\Delta_{km-1}^{\circ}, g^{(k,m)})\colon k \geq 1, m \geq 2\}$ be a double sequence of Riemannian manifolds with the property that every Lebanon map is an isometry. Then:*

$$g_P^{(k,m)}(u,v) = B \sum_a \sum_{b,c} \frac{u_{ab} u_{ac}}{|P_a|} + C \sum_a \sum_b \frac{u_{ab} v_{ab}}{P_{ab}}, \quad \textit{for each } P \in \Delta_{km-1}^{\circ}, \qquad (19)$$

  for some constants $B, C \in \mathbb{R}$ with $C > 0$ and $B + C > 0$, where $|P_a| = \sum_b P_{ab}$.
- *Conversely, let $\{(\Delta_{km-1}^{\circ}, g^{(k,m)})\}$ be a sequence of Riemannian manifolds with metrics $g^{(k,m)}$ of the form of Equation (19) for some $B, C \in \mathbb{R}$. Then, every Lebanon map is an isometry.*

Observe that these metrics agree with (a multiple of) the Fisher metric only if $B = 0$. The case $B = 0$ can also be characterized; note that Lebanon maps do not treat the two random variables symmetrically. Switching the two random variables corresponds to transposing the joint distribution matrix P. When exchanging the role of the two random variables, the Lebanon map becomes $P \mapsto (P^\top \otimes Q)^\top R$. We call such a map a dual Lebanon map. If we require invariance under both Lebanon maps and their duals in Theorem 6 or Corollary 8, the statements remain true with the additional restriction that $B = 0$ (as a function or constant, respectively).

4. Invariance Metric Characterizations for Conditional Polytopes

According to Chentsov's theorem (Theorem 2), a natural metric on the probability simplex can be characterized by requiring the isometry of natural embeddings. Lebanon follows this axiomatic approach to characterize metrics on products of positive measures (Theorem 6). However, the maps considered by Lebanon dissolve the row-normalization of conditional distributions. In general, they do not map conditional polytopes to conditional polytopes. Therefore, we will consider a slight modification of Lebanon maps, in order to obtain maps between conditional polytopes.

4.1. Stochastic Embeddings of Conditional Polytopes

A matrix of conditional distributions $P(Y|X)$ in Δ_{m-1}^k can be regarded as the equivalence class of all joint probability distributions $P(X,Y) \in \Delta_{km-1}$ with conditional distribution $P(Y|X)$. Which Markov maps of probability simplices are compatible with this equivalence relation? The most obvious examples are permutations (relabelings) of the state spaces of X and Y.

In information theory, stochastic matrices are also viewed as channels. For any distribution of X, the stochastic matrix gives us a joint distribution of the pair (X, Y) and, hence, a marginal distribution of Y. If we input a distribution of X into the channel, the stochastic matrix determines what the distribution of the output Y will be.

Channels can be combined, provided the cardinalities of the state spaces fit together. If we take the output Y of the first channel $P(Y|X)$ and feed it into another channel $P(Y'|Y)$ then we obtain a combined channel $P(Y'|X)$. The composition of channels corresponds to ordinary matrix multiplication. If the first channel is described by the stochastic matrix K and the second channel by Q, then the combined channel is described by $K \cdot Q$. Observe that in this case, the joint distribution P

(considered as a normalized matrix $P \in \Delta_{km-1}$) is transformed similarly; that is, the joint distribution of the pair (X, Y') is given by $P \cdot Q$.

More general maps result from compositions where the choice of the second channel depends on the input of the first channel. In other words, we have a first channel that takes as input X and gives as output Y, and we have another channel that takes as input (X, Y) and gives as output Y'; we are interested in the resulting channel from X to Y'. The second channel can be described by a collection of stochastic matrices $Q = \{Q^{(i)}\}_i$. If K describes the first channel, then the combined channel is described by the row product $K \otimes Q$ (see Definition 5). Again, the joint distribution of (X, Y') arises in a similar way as $P \otimes Q$.

We can also consider transformations of the first random variable X. Suppose that we use X as the input to a channel described by a stochastic matrix R. In this case, the joint distribution of the output X' of the channel and Y is described by $R^\top X$. However, in general, there is not much that we can say about the conditional distribution of Y given X'. The result depends in an essential way on the original distribution of X. However, this is not true in the special case that the channel is "not mixing", that is, in the case that R is a stochastic partition matrix. In this case, the conditional distribution $P(Y|X')$ is described by $\overline{R}^\top K$, where $\overline{R}$ is the corresponding partition indicator matrix, where all non-zero entries of R are replaced by one. In other words, each state of X corresponds to several states of X', and the corresponding row of K is copied a corresponding number of times.

To sum up, if we combine the transformations due to Q and R, then the joint probability distribution transforms as $P \mapsto R^\top(P \otimes Q)$ and the conditional transforms as $K \mapsto \overline{R}^\top(K \otimes Q)$. In particular, for the joint distribution, we obtain the definition of a Lebanon map. Figure 1 illustrates the situation.

Figure 1. An interpretation for Lebanon maps and conditional embeddings. The variable X' is computed from X by R, and Y' is computed from X and Y by Q.

joint distributions: $P' = R^\top(P \otimes Q)$

conditional distributions: $K' = \overline{R}^\top(K \otimes Q)$
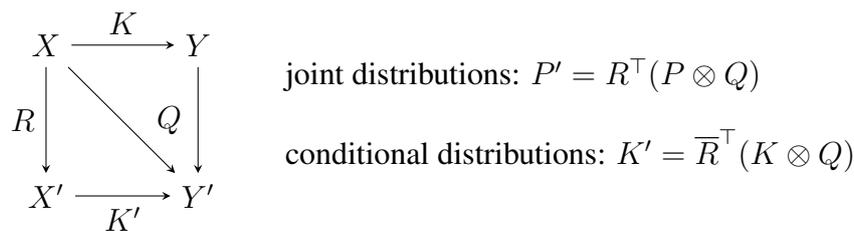
Finally, we will also consider the special case where the partition of R (and $\overline{R}$) is homogeneous, *i.e.*, such that all blocks have the same size. For example, this describes the case where there is a third random variable Z that is independent of Y given X. In this case, the conditional distribution satisfies $P(Y|X) = P(Y|X, Z)$, and R describes the conditional distribution of (X, Z) given X.

Definition 9. A (row) partition indicator matrix is a matrix $\overline{R} \in \{0, 1\}^{k \times l}$ that satisfies:

$$\overline{R}_{ij} = \begin{cases} 1, & \text{if } j \in A_i, \\ 0, & \text{else,} \end{cases} \tag{20}$$

for a k block partition $\{A_1, \ldots, A_k\}$ of $[l]$.

For example, the 3×5 partition indicator matrix corresponding to Equation (8) is:

$$\overline{R} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{21}$$

Definition 10. Consider a $k \times l$ partition indicator matrix $\overline{R}$ and a collection of $m \times n$ stochastic partition matrices $Q = \{Q^{(i)}\}_{i=1}^k$. We call the map:

$$f: \quad \mathbb{R}_+^{k \times m} \to \mathbb{R}_+^{l \times n}; \quad M \mapsto \overline{R}^\top (M \otimes Q) \tag{22}$$

a conditional embedding of $\mathbb{R}_+^{k \times m}$ in $\mathbb{R}_+^{l \times n}$. We denote the set of all such maps by $\hat{\mathcal{F}}_{k,m}^{l,n}$. If $\overline{R}$ is the partition indicator matrix of a homogeneous partition (with partition blocks of equal cardinality), then we call f a homogeneous conditional embedding. We denote the set of all such homogeneous conditional embeddings by $\mathcal{F}_{k,m}^{l,n}$ and assume that l is a multiple of k.

Conditional embeddings preserve the 1-norm of the matrix rows; that is, the elements of $\hat{\mathcal{F}}_{k,m}^{l,n}$ map $(\Delta_{m-1}^k)^\circ$ to $(\Delta_{n-1}^l)^\circ$. On the other hand, they do not preserve the 1-norm of the entire matrix. Conditional embeddings are Markov maps only when $k = l$, in which case they are also Lebanon maps.

4.2. Invariance Characterization

Considering the conditional embeddings discussed in the previous section, we obtain the following metric characterization.

Theorem 11.

- *Let $g^{(k,m)}$ denote a metric on $\mathbb{R}_+^{k \times m}$ for each $k \geq 1$ and $m \geq 2$. If every homogeneous conditional embedding $f \in \mathcal{F}_{k,m}^{l,n}$ is an isometry with respect to these metrics, then:*

$$g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = \frac{A}{k^2} + \delta_{ac}\left(k\frac{B}{k^2} + \delta_{bd}\frac{|M|}{M_{ab}}\frac{C}{k^2}\right), \quad \text{for all } M \in \mathbb{R}_+^{k \times m}, \tag{23}$$

  for some constants $A, B, C \in \mathbb{R}$, where $\partial_{ab} = \frac{\partial}{\partial M_{ab}}$ and $|M| = \sum_{ab} M_{ab}$.
- *Conversely, given the metrics defined by Equation (23) for any non-degenerate choice of constants $A, B, C \in \mathbb{R}$, each homogeneous conditional embedding $f \in \mathcal{F}_{k,m}^{l,n}$, $k \leq l$, $m \leq n$ is an isometry.*
- *Moreover, the tensors $g^{(k,m)}$ from Equation (23) are positive-definite for all $k \geq 1$ and $m \geq 2$ if and only if $C > 0$, $B + C > 0$ and $A + B + C > 0$.*

The proof of Theorem 11 is similar to the proof of the theorems of Chentsov, Campbell and Lebanon. Due to its technical nature, we defer it to Appendix B.

Now, for the restriction of the metric $g^{(k,m)}$ to $(\Delta_{m-1}^k)^\circ$, we have the following. In this case, $|M| = k$. Since tangent vectors $v = \sum_{ab} v_{ab}\partial_{ab} \in T_M(\Delta_{m-1}^k)^\circ$ satisfy $\sum_b v_{ab} = 0$ for all a, the constants A and B become immaterial, and the metric can be written as:

$$g_M^{(k,m)}(u, v) = \sum_{ab} \frac{|M| u_{ab} v_{ab}}{M_{ab}} \frac{C}{k^2} = \sum_{ab} \frac{u_{ab} v_{ab}}{M_{ab}} \frac{C}{k}, \quad \text{for all } u, v \in T_M(\Delta_{m-1}^k)^\circ. \tag{24}$$

This metric is a specialization of the metric (18) derived by Lebanon (Theorem 6).

The statement of Theorem 11 becomes false if we consider general conditional embeddings instead of homogeneous ones:

Theorem 12. *There is no family of metrics $g^{(k,m)}$ on $\mathbb{R}_+^{k \times m}$ (or on $(\Delta_{m-1}^k)^\circ$) for each $k \geq 1$ and $m \geq 2$, for which every conditional embedding $f \in \hat{\mathcal{F}}_{k,m}^{l,n}$ is an isometry.*

This negative result will become clearer from the perspective of Section 6: as we will show in Theorem 17, although there are no metrics that are invariant under all conditional embeddings, there are families of metrics (depending on a parameter, ρ) that transform covariantly (that is, in a well-defined manner) with respect to the conditional embeddings. We defer the proof of Theorem 12 to Appendix B.

5. The Fisher Metric on Polytopes and Point Configurations

In the previous section, we obtained distinguished Riemannian metrics on $\mathbb{R}_+^{k \times m}$ and $(\Delta_{m-1}^k)^\circ$ by postulating invariance under natural maps. In this section, we take another viewpoint based on general considerations about Riemannian metrics on arbitrary polytopes. This is achieved by embedding each polytope in a probability simplex as an exponential family. We first recall the necessary background. In Section 5.2, we then present our general results, and in Section 5.3, we discuss the special case of conditional polytopes.

5.1. Exponential Families and Polytopes

Let $\mathcal{X}$ be a finite set and $A \in \mathbb{R}^{d \times \mathcal{X}}$ a matrix with columns a_x indexed by $x \in \mathcal{X}$. It will be convenient to consider the rows A_i, $i \in [d]$ of A as functions $A_i : \mathcal{X} \to \mathbb{R}$. Finally, let $\nu : \mathcal{X} \to \mathbb{R}_+$. The exponential family $\mathcal{E}_{A,\nu}$ is the set of probability distributions on $\mathcal{X}$ given by:

$$p(x; \theta) = \exp(\theta^\top a_x + \log(\nu(x)) - \log(Z(\theta))), \quad \text{for all } x \in \mathcal{X}, \quad \text{for all } \theta \in \mathbb{R}^d, \qquad (25)$$

with the normalization function $Z(\theta) = \sum_{x' \in \mathcal{X}} \exp(\theta^\top a_{x'} + \log(\nu(x')))$. The functions A_i are called the observables and ν the reference measure of the exponential family. When the reference measure ν is constant, $\nu(x) = 1$ for all $x \in \mathcal{X}$, we omit the subscript and write $\mathcal{E}_A$.

A direct calculation shows that the Fisher information matrix of $\mathcal{E}_{A,\nu}$ at a point $\theta \in \mathbb{R}^d$ has coordinates:

$$g_\theta^{\mathcal{E}_{A,\nu}}(\partial_{\theta_i}, \partial_{\theta_j}) = \text{cov}_\theta(A_i, A_j), \qquad \text{for all } i, j \in [d]. \qquad (26)$$

Here, cov_θ denotes the covariance computed with respect to the probability distribution $p(\cdot; \theta)$.

The convex support of $\mathcal{E}_{A,\nu}$ is defined as:

$$\text{conv } A := \text{conv}\{a_x : x \in \mathcal{X}\} = \left\{\mathbb{E}_p[A] : p \in \Delta_{|\mathcal{X}|-1}\right\} = \left\{\mathbb{E}_p[A] : p \in \overline{\mathcal{E}_{A,\nu}}\right\}, \qquad (27)$$

where $\text{conv } S$ is the set of all convex combinations of points in S. The moment map $\mu : p \in \Delta_{n-1} \mapsto A \cdot p \in \mathbb{R}^d$ restricts to a homeomorphism $\overline{\mathcal{E}_{A,\nu}} \to \text{conv } A$; see [16]. Here, $\overline{\mathcal{E}_{A,\nu}}$ denotes the Euclidean closure of $\mathcal{E}_{A,\nu}$. The inverse of μ will be denoted by $\mu^{-1} : \text{conv } A \to \overline{\mathcal{E}_{A,\nu}} \subseteq \Delta_{n-1}$. This gives a natural

embedding of the polytope $\operatorname{conv} A$ in the probability simplex $\Delta_{|\mathcal{X}|-1}$. Note that the convex support is independent of the reference measure ν. See [17] for more details.

5.2. Invariance Fisher Metric Characterizations for Polytopes

Let $\mathbf{P} \in \mathbb{R}^d$ be a polytope with n vertices $a_1, \ldots, a_n$. Let $A = (a_1, \ldots, a_n)$ be the matrix with columns $a_i \in \mathbb{R}^d$ for all $i \in [n]$. Then, $\mathcal{E}_A \subseteq \Delta_{n-1}^\circ$ is an exponential family with convex support $\mathbf{P}$. We will also denote this exponential family by $\mathcal{E}_\mathbf{P}$. We can use the inverse of the moment map, μ^{-1}, to pull back geometric structures on Δ_{n-1}° to the relative interior $\mathbf{P}^\circ$ of $\mathbf{P}$.

Definition 13. The Fisher metric on $\mathbf{P}^\circ$ is the pull-back of the Fisher metric on $\mathcal{E}_A \subseteq \Delta_{n-1}^\circ$ by μ^{-1}.

Some obvious questions are: Why is this a natural construction? Which maps between polytopes are isometries between their Fisher metrics? Can we find a characterization of Chentsov type for this metric?

Affine maps are natural maps between polytopes. However, in order to obtain isometries, we need to put some additional constraints. Consider two polytopes $\mathbf{P} \in \mathbb{R}^d$, $\mathbf{P}' \in \mathbb{R}^{d'}$ and an affine map $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ that satisfies $\phi(\mathbf{P}) \subseteq \mathbf{P}'$. A natural condition in the context of exponential families is that ϕ restricts to a bijection between the set $\operatorname{vert}(\mathbf{P})$ of vertices of $\mathbf{P}$ and the set $\operatorname{vert}(\mathbf{P}')$ of vertices of $\mathbf{P}'$. In this case, $\mathcal{E}_{\mathbf{P}'} \subseteq \mathcal{E}_\mathbf{P} \subseteq \Delta_{n-1}^\circ$. Moreover, the moment map μ' of $\mathbf{P}'$ factorizes through the moment map μ of $\mathbf{P}$: $\mu' = \phi \circ \mu$. Let $\phi^{-1} = \mu \circ \mu'^{-1}$. Then, the following diagram commutes:

$$
\begin{array}{ccc}
\mathbf{P}^\circ & \xrightarrow{\;\mu^{-1}\;} & \mathcal{E}_\mathbf{P} \\[2pt]
\phi^{-1} \Big\uparrow & & \Big\uparrow \searrow \\[2pt]
& & \Delta_{n-1}^\circ \\[2pt]
\mathbf{P}'^\circ & \xrightarrow[\;\mu'^{-1}\;]{} & \mathcal{E}_{\mathbf{P}'} \nearrow
\end{array}
\tag{28}
$$

It follows that ϕ^{-1} is an isometry from $\mathbf{P}'^\circ$ to its image in $\mathbf{P}^\circ$. Observe that the inverse moment map itself arises in this way: In the diagram (28), if $\mathbf{P}$ is equal to Δ_{n-1}, then the upper moment map μ^{-1} is the identity map, and ϕ^{-1} equals the inverse moment map μ'^{-1} of $\mathbf{P}'$.

The constraint of mapping vertices to vertices bijectively is very restrictive. In order to consider a larger class of affine maps, we need to generalize our construction from polytopes to weighted point configurations.

Definition 14. A weighted point configuration is a pair (A, ν) consisting of a matrix $A \in \mathbb{R}^{d \times n}$ with columns $a_1, \ldots, a_n$ and a positive weight function $\nu : \{1, \ldots, n\} \to \mathbb{R}_+$ assigning a weight to each column a_i. The pair (A, ν) defines the exponential family $\mathcal{E}_{A,\nu}$.

The (A, ν)-Fisher metric on $(\operatorname{conv} A)^\circ$ is the pull-back of the Fisher metric on Δ_{n-1}° through the inverse of the moment map.

We recover Definition 13 as follows. For a polytope $\mathbf{P}$, let A be the point configuration consisting of the vertices of $\mathbf{P}$. Moreover, let ν be a constant function. Then, $\mathcal{E}_\mathbf{P} = \mathcal{E}_{A,\nu}$, and the two definitions of the Fisher metric on $\mathbf{P}^\circ$ coincide.

The following are natural maps between weighted point configurations:

Definition 15. Let (A, ν), (A', ν') be two weighted point configurations with $A = (a_i)_i \in \mathbb{R}^{d \times n}$ and $A' = (a'_j)_j \in \mathbb{R}^{d' \times n'}$. A morphism $(A, \nu) \to (A', \nu')$ is a pair (ϕ, σ) consisting of an affine map $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ and a surjective map $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n'\}$ with $\phi(a_i) = a'_{\sigma(i)}$ and $\nu'(a'_j) = \alpha \sum_{i:\sigma(i)=j} \nu(a_i)$, where $\alpha > 0$ is a constant that does not depend on j.

Consider a morphism $(\phi, \sigma) : (A, \nu) \to (A', \nu')$. For each $j \in [n']$, let $A_j = \{i : \phi(a_i) = a'_j\}$. Then, $(A_1, \ldots, A_{n'})$ is a partition of $[n]$. Define a matrix $Q \in \mathbb{R}^{n' \times n}$ by:

$$
Q_{ji} = \begin{cases} \frac{\nu(i)}{\sum_{i' \in A_j} \nu(i')}, & \text{if } i \in A_j, \\ 0, & \text{else.} \end{cases}
\tag{29}
$$

Then, Q is a Markov mapping, and the following diagram commutes:

$$
\begin{array}{ccccc}
(\operatorname{conv} A)^\circ & \xrightarrow{\;\mu^{-1}\;} & \mathcal{E}_{A,\nu} & \lhook\joinrel\longrightarrow & \Delta^\circ_{n-1} \\
\Big\uparrow{\scriptstyle \phi^{-1}} & & \Big\updownarrow & & \Big\updownarrow{\scriptstyle Q} \\
(\operatorname{conv} A')^\circ & \xrightarrow[\;\mu'^{-1}\;]{} & \mathcal{E}_{A',\nu'} & \lhook\joinrel\longrightarrow & \Delta^\circ_{n'-1}
\end{array}
\tag{30}
$$

By Chentsov's theorem (Theorem 2), Q is an isometric embedding. It follows that ϕ^{-1} also induces an isometric embedding. This shows the first part of the following theorem:

Theorem 16.

- *Let $(\phi, \sigma) : (A, \nu) \to (A', \nu')$ be a morphism of weighted point configurations. Then, $\phi^{-1} : (\operatorname{conv} A')^\circ \to (\operatorname{conv} A)^\circ$ is an isometric embedding with respect to the Fisher metrics on $(\operatorname{conv} A)^\circ$ and $(\operatorname{conv} A')^\circ$.*

- *Let $g^{A,\nu}$ be a Riemannian metric on $(\operatorname{conv} A)^\circ$ for each weighted point configuration (A, ν). If every morphism $(\phi, \sigma) : (A, \nu) \to (A', \nu')$ of weighted point configurations induces an isometric embedding $\phi^{-1} : (\operatorname{conv} A')^\circ \to (\operatorname{conv} A)^\circ$, then there exists a constant $\alpha \in \mathbb{R}_+$ such that $g^{A,\nu}$ is equal to α times the (A, ν)-Fisher metric.*

Proof. The first statement follows from the discussion before the theorem. For the second statement, we show that under the given assumptions, all Markov maps are isometric embeddings. By Chentsov's theorem (Theorem 2), this implies that the metrics $g^{\mathbf{P}}$ agree with the Fisher metric whenever $\mathbf{P}$ is a simplex. The statement then follows from the two facts that the metric on $\mathbf{P}^\circ$ or $(\operatorname{conv} A)^\circ$ is the pull-back of the Fisher metric through the inverse of the moment map and that μ^{-1} is itself a morphism.

Observe that $\Delta_{n-1} = \operatorname{conv} I_n = \operatorname{conv}\{e_1, \ldots, e_n\}$ is a polytope, and Δ°_{n-1} is the corresponding exponential family. Consider a Markov embedding $Q : \Delta^\circ_{n'-1} \to \Delta^\circ_{n-1}, p \mapsto p \cdot Q$. Let $\nu(i) = \sum_j Q_{ji}$ be the value of the unique non-zero entry of Q in the i-th column. This defines a morphism and an embedding as follows:

Let A be the matrix that arises from Q by replacing each non-zero entry by one. We define ϕ as the linear map represented by the matrix A, and define $\sigma : [n] \to [n']$ by $\sigma(j) = i$ if and only if $a_j = e_i$, that is, $\sigma(j)$ indicates the row i in which the j-th column of A is non-zero. Then, (ϕ, σ) is a morphism

$(I_n, \nu) \to (I_{n'}, 1)$, and by assumption, the inverse ϕ^{-1} is an isometric embedding $\Delta^\circ_{n'-1} \to \Delta^\circ_{n-1}$. However, ϕ^{-1} is equal to the Markov map Q. This shows that all Markov maps are isometric embeddings, and so, by Chentsov's theorem, the statement holds true on the simplices. $\quad\square$

Theorem 16 defines a natural metric on $(\Delta^k_{m-1})^\circ$ that we want to discuss in more detail next.

5.3. Independence Models and Conditional Polytopes

Consider k random variables with finite state spaces $[n_1], \dots, [n_k]$. The independence model consists of all joint distributions $p \in \Delta_{\prod_{i \in [k]} n_i - 1}$ of these variables that factorize as:

$$p(x_1, \dots, x_k) = \prod_{i \in [k]} p_i(x_i), \quad \text{for all } x_1 \in [n_1], \dots, x_k \in [n_k], \tag{31}$$

where $p_i \in \Delta_{n_i-1}$ for all $i \in [k]$. Assuming fixed $n_1, \dots, n_k$, we denote the independence model by $\overline{\mathcal{E}_k}$. It is the Euclidean closure of an exponential family (with observables of the form δ_{iy_i}). The convex support of $\mathcal{E}_k$ is equal to the product of simplices $\mathbf{P}_k := \Delta_{n_1-1} \times \cdots \times \Delta_{n_k-1}$. The parametrization (31) corresponds to the inverse of the moment map.

We can write any tangent vector $u \in T_{(p_1, \dots, p_k)} \mathbf{P}_k^\circ$ of this open product of simplices as a linear combination $u = \sum_{i \in [k]} \sum_{x_i \in [n_i]} u_{ix_i} \partial_{i,x_i}$, where $\sum_{x_i \in [n_i]} v_{ix_i} = 0$ for all $i \in [k]$. Given two such tangent vectors, the Fisher metric is given by:

$$g^{\mathbf{P}_k}_{(p_1, \dots, p_k)}(u, v) = \sum_{i \in [k]} \sum_{x_i \in [n_i]} \frac{u_{ix_i} v_{ix_i}}{p_i(x_i)}. \tag{32}$$

Just as the convex support of the independence model is the Cartesian product of probability simplices, the Fisher metric on the independence model is the product metric of the Fisher metrics on the probability simplices of the individual variables. If $n_1 = \cdots = n_k =: n$, then $\mathbf{P}_k = \Delta^k_{n-1}$ can be identified with the set of $k \times n$ stochastic matrices.

The Fisher metric on the product of simplices is equal to the product of the Fisher metrics on the factors. More generally, if $\mathbf{P} = \mathbf{Q}_1 \times \mathbf{Q}_2$ is a Cartesian product, then the Fisher metric on $\mathbf{P}^\circ$ is equal to the product of the Fisher metrics on $\mathbf{Q}_1^\circ$ and $\mathbf{Q}_2^\circ$. In fact, in this case, the inverse of the moment map of $\mathbf{P}$ can be expressed in terms of the two moment map inverses $\mu_1 : \mathbf{Q}_1 \to \overline{\mathcal{E}_{\mathbf{Q}_1}} \subseteq \Delta_{m_1-1}$ and $\mu_2 : \mathbf{Q}_2 \to \overline{\mathcal{E}_{\mathbf{Q}_2}} \subseteq \Delta_{m_2-1}$ and the moment map $\tilde{\mu}$ of the independence model $\Delta_{m_1-1} \times \Delta_{m_2-1}$, by:

$$\mu^{-1}(q_1, q_2) = \tilde{\mu}^{-1}(\mu_1^{-1}(q_1), \mu_2^{-1}(q_2)). \tag{33}$$

Therefore, the pull-back by μ^{-1} factorizes through the pull-back by $\tilde{\mu}^{-1}$, and since the independence model carries a product metric, the product of polytopes also carries a product metric.

Let us compare the metric $g_K^{(k,m)}$ from Equation (24), with the Fisher metric $g^{\mathbf{P}_k}_{(K_1, \dots, K_k)}$ from Equation (32) on the product of simplices $\mathbf{P}^\circ = (\Delta^k_{m-1})^\circ$. In both cases, the metric is a product metric; that is, it has the form:

$$g = g_1 + \cdots + g_k, \tag{34}$$

where g_i is a metric on the i-th factor Δ°_{m-1}. For $g_K^{\Delta^k_{m-1}}$, g_i is equal to the Fisher metric on Δ°_{m-1}. However, for $g_K^{(k,m)}$, g_i is equal to $1/k$ times the Fisher metric on Δ°_{m-1}. Since this factor only

depends on k, it only plays a role if stochastic matrices of different sizes are compared. The additional factor of $1/k$ can be interpreted as the uniform distribution on k elements. This is related to another more general class of Riemannian metrics that are used in applications; namely, given a function $K \in \Delta_{m-1}^k \to \rho^K \in \mathbb{R}_+^k$, it is common to use product metrics with g_i equal to $\rho^K(i)$ times the Fisher metric on Δ_{m-1}°. When K has the interpretation of a channel or when K describes the policy by which a system reacts to some sensor values, a natural possibility is to let ρ^K be the stationary distribution of the channel input or of the sensor values, respectively. We will discuss this approach in Section 6.

6. Weighted Product Metrics for Conditional Models

In this section, we consider metrics on spaces of stochastic matrices defined as weighted sums of the Fisher metrics on the spaces of the matrix rows, similar to Equation (34). This kind of metric was used initially by Amari [1] in order to define a natural gradient in the supervised learning context. Later, in the context of reinforcement learning, Kakade [2] defined a natural policy gradient based on this kind of metric, which has been further developed by Peters *et al.* [10]. Related applications within unsupervised learning have been pursued by Zahedi *et al.* [18].

Consider the following weighted product Fisher metric:

$$g_K^{\rho,m} = \sum_a \rho^K(a) g_{K_a}^{(m),a}, \quad \text{for all } K \in (\Delta_{m-1}^k)^\circ, \tag{35}$$

where $g_{K_a}^{(m),a}$ denotes the Fisher metric of Δ_{m-1}° at the a-th row of K and $\rho^K \in \Delta_{k-1}^\circ$ is a probability distribution over a associated with each $K \in (\Delta_{m-1}^k)^\circ$. For example, the distribution ρ^K could be the stationary distribution of sensor values observed by an agent when operating under a policy described by K.

In the following, we will try to illuminate the properties of polytope embeddings that yield the metric (35) as the pull-back of the Fisher information metric on a probability simplex. We will focus on the case that $\rho^K = \rho$ is independent of K.

There are two direct ways of embedding Δ_{n-1}^k in a probability simplex. In Section 5, we used the inverse of the moment map of an exponential family, possibly with some reference measure. This embedding is illustrated in the left panel of Figure 2. If we have given a fixed probability distribution $\rho \in \Delta_{k-1}^\circ$, there is a second natural embedding $\psi_\rho \colon \Delta_{m-1}^k \to \Delta_{k \cdot m-1}$ defined as follows:

$$\psi_\rho(K)(x,y) = \rho(x) K_{x,y} \quad \text{for all } x \in [k], y \in [m]. \tag{36}$$

If ρ is the distribution of a random variable X and $K \in \Delta_{m-1}^k$ is the stochastic matrix describing the conditional distribution of another variable Y given X, then $\psi_\rho(K)$ is the joint distribution of X and Y. Note that ψ_ρ is an affine embedding. See the right panel of Figure 2 for an illustration.

The pull-back of the Fisher metric on Δ_{km-1}° through ψ_ρ is given by:

$$\begin{aligned} g_{\psi_\rho(K)}^{(km)}(\psi_{\rho_*}u, \psi_{\rho_*}v) &= \sum_{a,b} \sum_{c,d} \sum_{i,j} \rho(i) K_{ij} u_{ab} \frac{\partial \log \rho(i) K_{ij}}{\partial K_{ab}} v_{cd} \frac{\partial \log \rho(i) K_{ij}}{\partial K_{cd}} \\ &= \sum_i \rho(i) \sum_j \frac{u_{ij} v_{ij}}{K_{ij}} = \sum_i \rho(i) g_{K_i}^i(u_i, v_i) = g_K^{\rho,m}(u,v). \end{aligned} \tag{37}$$

This recovers the weighted sum of Fisher metrics from Equation (35).

Figure 2. An illustration of different embeddings of the conditional polytope Δ_{m-1}^k in a probability simplex. The left panel shows an embedding in Δ_{m^k-1} by the inverse of the moment map μ of the independence model. The right panel shows an affine embedding in $\Delta_{k\cdot m-1}$ as a set of joint probability distributions for two different specifications of marginals.
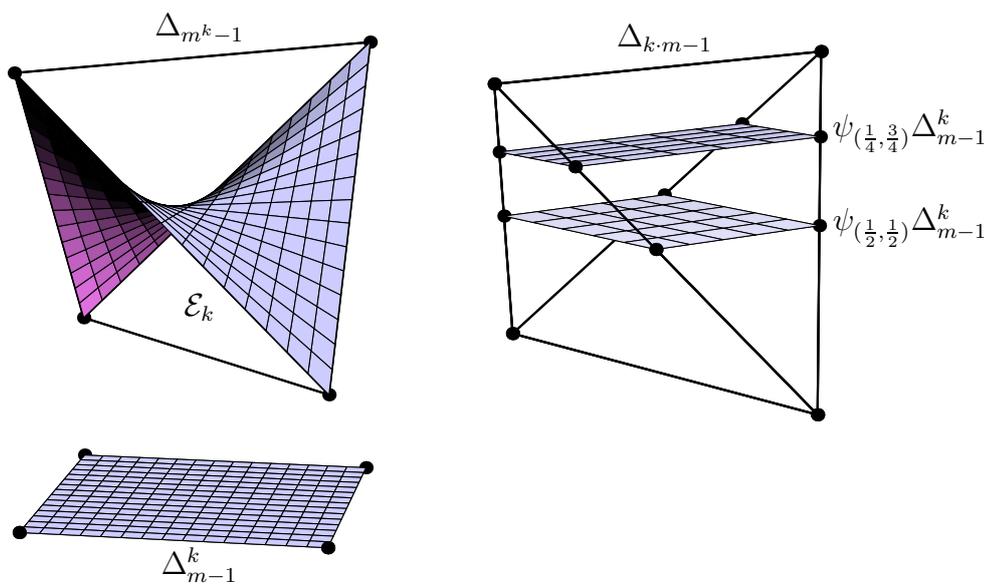
Are there natural maps that leave the metrics $g^{\rho,m}$ invariant? Let us reconsider the stochastic embeddings from Definition 10. Let $\overline{R}$ be a $k \times l$ indicator partition matrix and R a stochastic partition matrix with the same block structure as $\overline{R}$. Observe that to each indicator partition matrix $\overline{R}$ there are many compatible stochastic partition matrices R, but the indicator partition matrix $\overline{R}$ for any stochastic partition matrix R is unique. Furthermore, let $Q = \{Q^{(a)}\}_{a\in[k]}$ be a collection of stochastic partition matrices. The corresponding conditional embedding $\overline{f}$ maps $K \in \Delta_{m-1}^k$ to $\overline{f}(K) := \overline{R}^\top (K \otimes Q) \in \Delta_{n-1}^l$.

Let $\rho \in \Delta_{k-1}^\circ$. Suppose that K describes the conditional distribution of Y given X and that $\psi_\rho(K)$ describes the joint distribution of Y and X. As explained in Section 4.1, the matrix $f(P) := R^\top(P \otimes Q)$ describes the joint distribution of a pair of random variables (X', Y'), and the conditional distribution of Y' given X' is given by $\overline{f}(K)$. In this situation, the marginal distribution of X' is given by $\rho' = \rho R$. Therefore, the following diagram commutes:

$$
\begin{array}{ccc}
(\Delta_{m-1}^k)^\circ & \xrightarrow{\ \psi_\rho\ } & \Delta_{mk-1}^\circ \\
{\scriptstyle \overline{f}}\Big\downarrow & & \Big\downarrow{\scriptstyle f} \\
(\Delta_{n-1}^l)^\circ & \xrightarrow[\ \psi_{\rho'}\ ]{} & \Delta_{nl-1}^\circ
\end{array}
\tag{38}
$$

The preceding discussion implies the first statement of the following result:

Theorem 17.

- *For any $k \geq 1$ and $m \geq 2$ and any $\rho \in \Delta_{k-1}^\circ$, the Riemannian metric $g^{\rho,m}$ on $(\Delta_{m-1}^k)^\circ$ satisfies:*

$$g^{\rho,m} = \overline{f}^*(g^{\rho',n}), \quad for \ \rho' = \rho R, \tag{39}$$

   for any conditional embedding $\overline{f} : K \mapsto \overline{R}(K \otimes Q)$.

- *Conversely, suppose that for any $k \geq 1$ and $m \geq 2$ and any $\rho \in \Delta_{k-1}^\circ$, there is a Riemannian metric $g^{(\rho,m)}$ on $(\Delta_{m-1}^k)^\circ$, such that Equation (39) holds for all conditional embeddings, and suppose that $g^{(\rho,m)}$ depends continuously on ρ. Then, there is a constant $A > 0$ that satisfies $g^{(\rho,m)} = A g^{\rho,m}$.*

Proof. The first statement follows from the commutative diagram (38). For the second statement, denote by ρ^k the uniform distribution on a set of k elements. If $\overline{f} \colon K \mapsto \overline{R}(K \otimes Q)$ is a homogeneous conditional embedding of Δ_{m-1}^k in Δ_{n-1}^l, then $R = \frac{k}{l}\overline{R}$ is a stochastic partition matrix corresponding to the partition indicator matrix $\overline{R}$. Observe that $\rho^l = \rho^k R$. Therefore, the family of Riemannian metrics $g^{\rho^k,m}$ on Δ_{m-1}^k satisfies the assumptions of Theorem 11. Therefore, there is a constant $A > 0$ for which $g^{\rho^k,m}$ equals A/k times the product Fisher metric. This proves the statement for uniform distributions ρ.

A general distribution $\rho \in \Delta_{k-1}^\circ$ can be approximated by a distribution with rational probabilities. Since $g^{(\rho,m)}$ is assumed to be continuous, it suffices to prove the statement for rational ρ. In this case, there exists a stochastic partition matrix R for which $\rho' := \rho R$ is a uniform distribution, and so, $g^{(\rho',n)}$ is of the desired form. Equation (39) shows that $g^{(\rho,m)}$ is also of the desired form. $\square$

7. Gradient Fields and Replicator Equations

In this section, we use gradient fields in order to compare Riemannian metrics on the space $(\Delta_{n-1}^k)^\circ$.

7.1. Replicator Equations

We start with gradient fields on the simplex Δ_{n-1}°. A Riemannian metric g on Δ_{n-1}° allows us to consider gradient fields of differentiable functions $F \colon \Delta_{n-1}^\circ \to \mathbb{R}$. To be more precise, consider the differential $d_p F : T_p \Delta_{n-1}^\circ \to \mathbb{R}$ of F in p. It is a linear form on $T_p \Delta_{n-1}^\circ$, which maps each tangent vector u to $d_p F(u) = \frac{\partial F}{\partial u}(p) \in \mathbb{R}$. Using the map $u \mapsto g_p(u, \cdot)$, this linear form can be identified with a tangent vector in $T_p \Delta_{n-1}^\circ$, which we denote by $\mathrm{grad}_p F$. If we choose the Fisher metric $g^{(n)}$ as the Riemannian metric, we obtain the gradient in the following way. First consider a differentiable extension of F to the positive cone $\mathbb{R}_+^n$, which we will denote by the same symbol F. With the partial derivatives $\partial_i F$ of F, the Fisher gradient of F on the simplex Δ_{n-1}° is given as:

$$(\mathrm{grad}_p F)_i \; = \; p_i \left(\partial_i F(p) - \sum_{j=1}^n p_j \, \partial_j F(p) \right), \qquad i \in [n]. \tag{40}$$

Note that the expression on the right-hand side of Equation (40) does not depend on the particular differentiable extension of F to $\mathbb{R}_+^n$. The corresponding differential equation is well known in theoretical biology as the replicator equation; see [19,20].

$$\dot{p}_i \; = \; p_i \left(\partial_i F(p) - \sum_{j=1}^n p_j \, \partial_j F(p) \right), \qquad i \in [n]. \tag{41}$$

We now apply this gradient formula to functions that have the structure of an expectation value. Given real numbers F_i, $i \in [n]$, referred to as fitness values, we consider the mean fitness:

$$\bar{F}(p) := \sum_{i=1}^{n} p_i F_i. \tag{42}$$

Replacing the p_i by any positive real numbers leads to a differentiable extension of F, also denoted by F. Obviously, we have $\partial_i F = F_i$, which leads to the following replicator equation:

$$\dot{p}_i = p_i \left(F_i - \bar{F}(p) \right), \qquad i \in [n]. \tag{43}$$

This equation has the solution:

$$p_i(t) = \frac{p_i(0) e^{t F_i}}{\sum_{j=1}^{n} p_j(0) e^{t F_i}}, \qquad i \in [n]. \tag{44}$$

Clearly, the mean fitness will increase along this solution of the gradient field. The rate of increase can be easily calculated:

$$\frac{d}{dt} \bar{F}\big(p(t)\big) = \sum_{i=1}^{n} \dot{p}_i(t) F_i = \sum_{i=1}^{n} p_i \left(F_i - \bar{F}(p) \right) F_i = \sum_{i=1}^{n} p_i \left(F_i - \bar{F}(p) \right)^2 = \text{var}_p(F) > 0. \tag{45}$$

As limit points of this solution, we obtain:

$$\lim_{t \to -\infty} p_i(t) = \begin{cases} \frac{p_i(0)}{\sum_{j \in \text{argmin} F} p_j(0)}, & \text{if } i \in \text{argmin} F \\ 0 & , \quad \text{otherwise} \end{cases} , \qquad i \in [n], \tag{46}$$

and:

$$\lim_{t \to +\infty} p_i(t) = \begin{cases} \frac{p_i(0)}{\sum_{j \in \text{argmax} F} p_j(0)}, & \text{if } i \in \text{argmax} F \\ 0 & , \quad \text{otherwise} \end{cases} , \qquad i \in [n]. \tag{47}$$

7.2. Extension of the Replicator Equations to Stochastic Matrices

Now, we come to the corresponding considerations of gradient fields in the context of stochastic matrices $K \in (\Delta_{n-1}^k)^\circ$. We consider a function:

$$K \mapsto F(K) = F(K_{11}, \ldots, K_{1n}; K_{21}, \ldots, K_{2n}; \ldots; K_{k1}, \ldots, K_{kn}). \tag{48}$$

One way to deal with this is to consider for each $i \in [k]$ the corresponding replicator equation:

$$\dot{K}_{ij} = K_{ij} \left(\partial_{ij} F(K) - \sum_{j'=1}^{n} K_{ij'} \, \partial_{ij'} F(K) \right), \qquad j \in [n]. \tag{49}$$

Obviously, this is the gradient field that one obtains by using the product Fisher metric on $(\Delta_{n-1}^k)^\circ$ (Equation (17)):

$$g_K^{(k,m)}(u, v) = \sum_{ij} \frac{1}{K_{ij}} u_{ij} v_{ij}. \tag{50}$$

If we replace the metric by the weighted product Fisher metric considered by Kakade (Equation (35)),

$$g_K^{\rho,m}(u,v) \;=\; \sum_{ij} \frac{\rho_i}{K_{ij}}\, u_{ij} v_{ij}, \tag{51}$$

then we obtain

$$\dot{K}_{ij} \;=\; \frac{K_{ij}}{\rho_i}\left(\partial_{ij} F(K) - \sum_{j'=1}^{n} K_{ij'}\, \partial_{ij'} F(K) \right), \qquad j \in [n]. \tag{52}$$

7.3. The Example of Mean Fitness

Next, we want to study how the gradient flows with respect to different metrics compare. We restrict to the class of metrics $g^{\rho,m}$ (Equation (35)), where $\rho \in \Delta_k^\circ$ is a probability distribution. In principle, one could drop the normalization condition $\sum_i \rho_i = 1$ and allow arbitrary coefficients ρ_i. However, it is clear that the rate of convergence can always be increased by scaling all values ρ_i with a common positive factor. Therefore, some normalization condition is needed for ρ.

With a probability distribution $p \in \Delta_{k-1}^\circ$ and fitness values F_{ij}, let us consider again the example of an expectation value function:

$$\bar{F}(K) \;=\; \sum_{i=1}^{k} p_i \sum_{j=1}^{n} K_{ij}\, F_{ij}. \tag{53}$$

With $\partial_{ij}\bar{F}(\pi) = p_i\, F_{ij}$, this leads to:

$$\dot{K}_{ij} \;=\; \frac{p_i}{\rho_i}\, K_{ij}\left(F_{ij} - \sum_{j'=1}^{n} K_{ij'}\, F_{ij'} \right), \qquad j \in [n]. \tag{54}$$

The corresponding solutions are given by:

$$K_{ij}(t) \;=\; \frac{K_{ij}(0)\, e^{t\frac{p_i}{\rho_i} F_{ij}}}{\sum_{j'=1}^{n} K_{ij'}(0)\, e^{t\frac{p_i}{\rho_i} F_{ij'}}}, \qquad i \in [n]. \tag{55}$$

Since $\operatorname{argmax}(\frac{p_i}{\rho_i} F_{i\cdot})$ and $\operatorname{argmin}(\frac{p_i}{\rho_i} F_{i\cdot})$ are independent of $\rho_i > 0$, the limit points are given independently of the chosen ρ as:

$$\lim_{t\to-\infty} K_{ij}(t) \;=\; \begin{cases} \frac{K_{ij}(0)}{\sum_{j'\in\operatorname{argmin} F_{i\cdot}} K_{ij'}(0)}, & \text{if } j \in \operatorname{argmin} F_{i\cdot} \\ 0 & , \quad \text{otherwise} \end{cases}, \qquad i \in [n], \tag{56}$$

and:

$$\lim_{t\to+\infty} K_{ij}(t) \;=\; \begin{cases} \frac{K_{ij}(0)}{\sum_{j'\in\operatorname{argmax} F_{i\cdot}} K_{ij'}(0)}, & \text{if } j \in \operatorname{argmax} F_{i\cdot} \\ 0 & , \quad \text{otherwise} \end{cases}, \qquad i \in [n]. \tag{57}$$

This is consistent with the fact that the critical points of gradient fields are independent of the chosen Riemannian metric. However, the speed of convergence does depend on the metric:

For each i, let $G_i = \max_j F_{ij}$ and $g_i = \max_{j \notin \operatorname{argmax}(F_{ij})} F_{ij}$ be the largest and second-largest values in the i-th row of F_{ij}, respectively. Then, as: $t \to \infty$,

$$K_{ij}(t) \;=\; \begin{cases} 1 - O(\exp(-\frac{p_i}{\rho_i}(G_i - g_i)t), & \text{if } i \in \operatorname{argmax} F_{i\cdot} \\ O(\exp(-\frac{p_i}{\rho_i}(G_i - F_{ij})t) & , \quad \text{otherwise} \end{cases} \tag{58}$$

Therefore,

$$\bar{F}(K(t)) = \sum_i p_i \sum_{j \in \text{argmax } F_{i\cdot}} F_{ij} + O\left(\exp(-\frac{p_i}{\rho_i}(G_i - g_i)t)\right)$$

$$= \sum_i p_i \sum_{j \in \text{argmax } F_{i\cdot}} F_{ij} + O\left(\exp(-\inf_i \left\{\frac{p_i}{\rho_i}(G_i - g_i)\right\} t)\right). \quad (59)$$

Thus, in the long run, the rate of convergence is given by $\inf_i\{\frac{p_i}{\rho_i}(G_i - g_i)\}$, which depends on the parameter ρ of the metric. As a result, in this case study, the optimal choice of ρ_i, *i.e.*, with the largest convergence rate, can be computed if the numbers G_i and g_i are known.

Consider, for example, the case that the differences $G_i - g_i$ are of comparable sizes for all i. Then, we need to find the choice of ρ that maximizes $\inf_i\{\frac{p_i}{\rho_i}\}$. Clearly, $\inf_i\{\frac{p_i}{\rho_i}\} \leq 1$ (since there is always an index i with $p_i \leq \rho_i$). Equality is attained for the choice $\rho_i = p_i$. Thus, we recover the choice of Kakade.

8. Conclusions

So, which Riemannian metric should one use in practice on the set of stochastic matrices, $(\Delta_{m-1}^k)^{\circ}$? The results provided in this manuscript give different answers, depending on the approach. In all cases, the characterized Riemannian metrics are products of Fisher metrics with suitable factor weights. Theorem 11 suggests to use a factor weight proportional to $1/k$, and Theorem 16 suggests to use a constant weight independent of k. In many cases, it is possible to work within a single conditional polytope $(\Delta_{m-1}^k)^{\circ}$ and a fixed k, and then, these two results are basically equivalent. On the other hand, Theorem 17 gives an answer that allows arbitrary factor weights ρ.

Which metric performs best obviously depends on the concrete application. The first observation is that in order to use the metric $g^{\rho,m}$ of Theorem 17, it is necessary to know ρ. If the problem at hand suggests a natural marginal distribution ρ, then it is natural to make use of it and choose the metric $g^{\rho,m}$. Even if ρ is not known at the beginning, a learning system might try to learn it to improve its performance.

On the other hand, there may be situations where there is no natural choice of the weights ρ. Observe that ρ breaks the symmetry of permuting the rows of a stochastic matrix. This is also expressed by the structural difference between Theorems 11 and 16 on the one side and Theorem 17 on the other. While the first two theorems provide an invariance metric characterization, Theorem 17 provides a "covariance" classification; that is, the metrics $g^{\rho,m}$ are not invariant under conditional embeddings, but they transform in a controlled manner. This again illustrates that the choice of a metric should depend on which mappings are natural to consider, e.g., which mappings describe the symmetries of a given problem.

For example, consider a utility function of the form $F = \sum_i \rho_i \sum_j K_{ij} F_{ij}$. Row permutations do not leave $g^{\rho,m}$ invariant (for a general ρ), but they are not symmetries of the utility function F, either, and hence, they are not very natural mappings to consider. However, row permutations transform the metric $g^{\rho,m}$ and the utility function in a controlled manner; in such a way that the two transformations match. Therefore, in this case, it is natural to use $g^{\rho,m}$. On the other hand, when studying problems that are symmetric under all row permutations, it is more natural to use the invariant metric $g^{(k,m)}$.

Appendix

A. Conditions for Positive Definiteness

Equation (16) in Lebanon's Theorem 6 defines a Riemannian metrics whenever it defines a positive-definite quadratic form. The next proposition gives sufficient and necessary conditions for which this is the case.

Proposition 18. *For each pair $k \geq 1$ and $m \geq 2$, consider the tensor on $\mathbb{R}_+^{k \times m}$ defined by:*

$$g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = A(|M|) + \delta_{ac}\left(\frac{B(|M|)}{|M_a|} + \delta_{bd}\frac{C(|M|)}{M_{ab}}\right) \tag{A1}$$

for some differentiable functions $A, B, C \in C^\infty(\mathbb{R}_+)$. The tensor $g^{(k,m)}$ defines a Riemannian metric for all k and m if and only if $C(\alpha) > 0$, $B(\alpha) + C(\alpha) > 0$ and $A(\alpha) + B(\alpha) + C(\alpha) > 0$ for all $\alpha \in \mathbb{R}_+$.

Proof. The tensors are Riemannian metrics when:

$$g_M^{(k,m)}(V) = A(|M|)(\sum_{ab} V_{ab})^2 + B(|M|)\sum_a \frac{|M|}{|M_a|}(\sum_b V_{ab})^2 + C(|M|)\sum_{ab}\frac{|M|}{M_{ab}}V_{ab}^2 \tag{A2}$$

is strictly positive for all non-zero $V \in \mathbb{R}^{k \times m}$, for all $M \in \mathbb{R}_+^{k \times m}$.

We can derive necessary conditions on the functions A, B, C from some basic observations. Choosing $V = \partial_{ab}$ in Equation (A2) shows that $A(|M|) + \frac{|M|}{|M_a|}B(|M|) + \frac{|M|}{M_{ab}}C(|M|)$ has to be positive for all $a \in [k], b \in [m]$, for all $M \in \mathbb{R}_+^{k \times m}$. Since M_{ab} can be arbitrarily small for fixed $|M|$ and $|M_a|$, we see that C has to be non-negative. Since we can choose $|M_a| \approx M_{ab} \ll |M|$ for a fixed $|M|$, we find that $B + C$ has to be non-negative. Further, since we can choose $M_{ab} \approx |M_a| \approx |M|$ for a given $|M|$, we find that $A + B + C$ has to be non-negative. This shows that the quadratic form is positive definite only if $C \geq 0$, $B + C \geq 0$, $A + B + C \geq 0$. Since the cone of positive definite matrices is open, these inequalities have to be strictly satisfied. In the following, we study sufficient conditions.

For any given $M \in \mathbb{R}_+^{k \times m}$, we can write Equation (A2) as a product $V^\top G V$, for all $V \in \mathbb{R}^{km}$, where $G = G_A + G_B + G_C \in \mathbb{R}^{km \times km}$ is the sum of a matrix G_A with all entries equal to $A(|M|)$, a block diagonal matrix G_B whose a-th block has all entries equal to $\frac{|M|}{|M_a|}B(|M|)$, and a diagonal matrix G_C with diagonal entries equal to $\frac{|M|}{M_{ab}}C(|M|)$. The matrix G is obviously symmetric, and by Sylvester's criterion, it is positive definite iff all its leading principal minors are positive. We can evaluate the minors using Sylvester's determinant theorem. That theorem states that for any invertible $m \times m$ matrix X, an $m \times n$ matrix Y and an $n \times m$ matrix Z, one has the equality $\det(X + YZ) = \det(X)\det(I_n + ZX^{-1}Y)$.

Let us consider a leading square block G', consisting of all entries $G_{ab,cd}$ of G with row-index pairs (a, b) satisfying $b \in [m]$ for all $a < a'$ and $b \leq b'$ for $a = a'$ for some $a' \leq k$ and $b' \leq m$; and the same restriction for the column index pairs. The corresponding block $G'_A + G'_B$ can be written as the rank-a' matrix YZ, with Y consisting of columns $\mathbb{1}_a$ for all $a \leq a'$ and Z consisting of rows $A + \mathbb{1}_a\frac{|M|}{|M_a|}B$ for all $a \leq a'$. Hence, the determinant of G' is equal to:

$$\det(G') = \det(G'_C) \cdot \det(I_{a'} + ZG'^{-1}_C Y). \tag{A3}$$

Since G'_C is diagonal, the first term is just:

$$\det(G'_C) = \left(\prod_{a<a'} \prod_b \frac{|M|}{M_{ab}} C \right) \left(\prod_{b\leq b'} \frac{|M|}{M_{a'b}} C \right). \tag{A4}$$

The matrix in the second term of Equation (A3) is given by:

$$I_{a'} + ZG'^{-1}_C Y =$$

$$\frac{1}{C} \begin{pmatrix} C+B & & & \\ & \ddots & & \\ & & C+B & \\ & & & C+\frac{\sum_{b\leq b'} M_{a'b}}{|M_{a'}|}B \end{pmatrix} + \frac{1}{C} \begin{pmatrix} \frac{|M_1|}{|M|}A & \cdots & \frac{|M_{a'-1}|}{|M|}A & \frac{\sum_{b\leq b'} M_{a'b}}{|M|}A \\ \vdots & & \vdots & \vdots \\ \frac{|M_1|}{|M|}A & \cdots & \frac{|M_{a'-1}|}{|M|}A & \frac{\sum_{b\leq b'} M_{a'b}}{|M|}A \end{pmatrix}. \tag{A5}$$

By Sylvester's determinant theorem, we have:

$$\det(I_{a'} + ZG'^{-1}_C Y) = C^{-a'}(C+B)^{a'-1}(C + \frac{\sum_{b\leq b'} M_{a'b}}{|M_{a'}|}B)(1 + \sum_{a<a'} \frac{\frac{|M_a|}{|M|}A}{C+B} + \frac{\frac{\sum_{b\leq b'} M_{a'b}}{|M|}A}{C+\frac{\sum_{b\leq b'} M_{a'b}}{|M_{a'}|}B})$$

$$= \left(\prod_a \frac{C+B_a}{C} \right) \left(1 + \sum_a \frac{A_a}{C+B_a} \right), \tag{A6}$$

where $A_a = \frac{|M_a|}{|M|}A$ for $a < a'$ and $A_{a'} = \frac{\sum_{b\leq b'} M_{a'b}}{|M|}A$, and $B_a = B$ for $a < a'$ and $B_{a'} = \frac{\sum_{b\leq b'} M_{a'b}}{|M_{a'}|}B$.

This shows that the matrix G is positive definite for all M if and only if $C > 0$, $C + B > 0$ and $\left(1 + \sum_{a\leq a'} \frac{A_a}{C+B_a} \right) > 0$ for all a' and b'. The latter inequality is satisfied whenever $A + B + C > 0$. This completes the proof. $\square$

B. Proofs of the Invariance Characterization

The following lemma follows directly from the definition and contains all the technical details we need for the proofs.

Lemma 19. *The push-forward* $f_*: T_M \mathbb{R}^{k\times m}_+ \to T_{f(M)} \mathbb{R}^{l\times n}_+$ *of a map* $f \in \hat{\mathcal{F}}^{l,n}_{k,m}$ *is given by:*

$$f_*(\partial_{ab}) = \sum_{i=1}^l \sum_{j=1}^n \overline{R}_{ai} Q^{(a)}_{bj} \partial'_{ij}, \tag{A7}$$

and the pull-back of a metric $g^{(l,n)}$ *on* $\mathbb{R}^{l\times n}_+$ *through* f *is given by:*

$$(f^* g^{(l,n)})_M(\partial_{ab}, \partial_{cd}) = g^{(l,n)}_{f(M)}(f_*\partial_{ab}, f_*\partial_{cd}) = \sum_{i=1}^l \sum_{j=1}^n \sum_{s=1}^l \sum_{t=1}^n \overline{R}_{ai}\overline{R}_{cs} Q^{(a)}_{bj} Q^{(c)}_{dt} g^{(l,n)}_{f(M)}(\partial'_{ij}, \partial'_{st}). \tag{A8}$$

Proof of Theorem 11. We follow the strategy of [5,14]. The idea is to consider subclasses of maps from the class $\mathcal{F}^{l,n}_{k,m}$ and to evaluate their push-forward and pull-back maps together with the isometry requirement. This yields restrictions on the possible metrics, eventually fully characterizing them.

First. Consider the maps $h_{\pi,\sigma} \in \mathcal{F}_{k,m}^{k,m}$, resulting from permutation matrices $Q^{(a)} = P_{\pi^a}$, $\pi^a \colon [m] \to [m]$ for all $a \in [k]$, and $\overline{R} = P_\sigma$, $\sigma \colon [k] \to [k]$. Requiring isometry yields:

$$
(h_{\pi,\sigma})_*(\partial_{ab}) = \partial'_{\sigma(a)\,\pi^a(b)} \tag{A9}
$$

$$
g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = g_{h_{\pi,\sigma}(M)}^{(k,m)}(\partial_{\sigma(a)\,\pi^{(a)}(b)}, \partial_{\sigma(c)\,\pi^{(c)}(d)}). \tag{A10}
$$

Second. Consider the maps $r_{zw} \in \mathcal{F}_{k,m}^{kz,mw}$ defined by $Q^{(1)} = \cdots = Q^{(k)} \in \mathbb{R}^{m \times mw}$ and $\overline{R} \in \mathbb{R}^{k \times kz}$ being uniform. In this case, for some permutations π and σ,

$$
(r_{zw})_*(\partial_{ab}) = \frac{1}{w} \sum_{i=1}^{z} \sum_{j=1}^{w} \partial'_{\sigma^{(a)}(i)\,\pi^{(b)}(j)} \tag{A11}
$$

$$
(r_{zw}{}^* g^{(kz,mw)})_M(\partial_{ab}, \partial_{cd}) = \frac{1}{w^2} \sum_{i=1}^{z} \sum_{j=1}^{w} \sum_{s=1}^{z} \sum_{t=1}^{w} g_{r_{zw}(M)}^{(kz,mw)}(\partial'_{\sigma^{(a)}(i)\,\pi^{(b)}(j)}, \partial'_{\sigma^{(c)}(s)\,\pi^{(d)}(t)}). \tag{A12}
$$

Third. For a rational matrix $M = \frac{1}{Z}\tilde{M}$ with $\tilde{M} \in \mathbb{N}^{k \times m}$ and row-sum $|\tilde{M}_a| = N \in \mathbb{N}$ for all $a \in [k]$, consider the map $v_M \in \mathcal{F}_{k,m}^{zk,N}$ that maps M to a constant matrix. In this case, $\overline{R} \in \mathbb{R}^{k \times kz}$ and $Q^{(a)}$ has the b-th row with $|\tilde{M}_{ab}|$ entries with value $\frac{1}{|\tilde{M}_{ab}|}$ at positions $\pi^{(ab)}([\tilde{M}_{ab}]) \subseteq [N]$, and:

$$
(v_M)_*(\partial_{ab}) = \frac{1}{\tilde{M}_{ab}} \sum_{i=1}^{k} \sum_{j=1}^{\tilde{M}_{ab}} \partial'_{\sigma^{(a)}(i)\,\pi^{(ab)}(j)} \tag{A13}
$$

$$
(v_M{}^* g^{(kz,N)})_M(\partial_{ab}, \partial_{cd}) = \frac{1}{\tilde{M}_{ab}} \frac{1}{\tilde{M}_{cd}} \sum_{i=1}^{z} \sum_{j=1}^{\tilde{M}_{ab}} \sum_{s=1}^{z} \sum_{t=1}^{\tilde{M}_{cd}} g_{v_M(M)}^{(kz,N)}(\partial'_{\sigma^{(a)}(i)\,\pi^{(ab)}(j)}, \partial'_{\sigma^{(c)}(s)\,\pi^{(cd)}(t)}). \tag{A14}
$$

Step 1: $a \neq c$. Consider a constant matrix $M = U$. Then:

$$
g_U^{(k,m)}(\partial_{a_1 b_1}, \partial_{c_1 d_1}) = g_{h_{\pi,\sigma}(U)}^{(k,m)}(\partial_{a_2 b_2}, \partial_{c_2 d_2}) = g_U^{(k,m)}(\partial_{a_2 b_2}, \partial_{c_2 d_2}). \tag{A15}
$$

This implies that $g_U^{(k,m)}(\partial_{ab}, \partial_{cd}) = \hat{A}(k,m)$ when $a \neq c$.

Using the second type of map, we get:

$$
\hat{A}(k,m) = \frac{z^2 w^2}{w^2} \hat{A}(kz, mw), \tag{A16}
$$

which implies $g_U^{(k,m)}(\partial_{ab}, \partial_{cd}) = \frac{A}{k^2}$, when $a \neq c$. Considering a rational matrix M and the map v_M yields:

$$
g_M^{(k,m)}(\partial_{ab}, \partial_{c,d}) = \frac{A}{k^2}. \tag{A17}
$$

Step 2: $b \neq d$. By similar arguments as in Part 1, $g_U^{(k,m)}(\partial_{ab}, \partial_{ad}) = \hat{B}(k,m)$. Evaluating the map r_{zw} yields:

$$
\hat{B}(k,m) = \frac{zw^2}{w^2} \hat{B}(kz, mw) + \frac{z(z-1)w^2}{w^2} \frac{A}{(kz)^2}
$$

$$
= z\hat{B}(kz, mw) + \frac{z-1}{z} \frac{A}{k^2}, \tag{A18}
$$

and therefore,

$$\frac{1}{z}\left(\hat{B}(k,m) - \frac{A}{k^2}\right) = \hat{B}(kz, mw) - \frac{A}{(kz)^2}, \tag{A19}$$

which implies that $\left(\hat{B}(k,m) - \frac{A}{k^2}\right)$ is independent of m and scales with the inverse of k, such that it can be written as $\frac{B}{k}$. Rearranging the terms yields $g_U^{(k,m)}(\partial_{ab}, \partial_{ad}) = \frac{A}{k^2} + \frac{B}{k}$, for $b \neq d$.

For a rational matrix M, the pull-back through v_M shows then:

$$g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = z\frac{\tilde{M}_{ab}\tilde{M}_{ad}}{\tilde{M}_{ab}\tilde{M}_{ad}}\left(\frac{A}{(kz)^2} + \frac{B}{kz}\right) + \frac{z(z-1)\tilde{M}_{ab}\tilde{M}_{ad}}{\tilde{M}_{ab}\tilde{M}_{ad}}\frac{A}{(kz)^2} = \frac{A}{k^2} + \frac{B}{k}. \tag{A20}$$

Step 3: $a = c$ and $b = d$. In this case, $g_U^{(k,m)}(\partial_{a_1 b_1}, \partial_{a_1 b_1}) = g_U^{(k,m)}(\partial_{a_2 b_2}, \partial_{a_2 b_2}) = \hat{C}(k,m)$, and:

$$\hat{C}(k,m) = \frac{1}{w^2}zw\hat{C}(kz, mw) + \frac{1}{w^2}zw(w-1)\left(\frac{A}{(kz)^2} + \frac{B}{kz}\right) + \frac{1}{w^2}zw^2 z(z-1)\frac{A}{(kz)^2}$$

$$= \frac{z}{w}\hat{C}(kz, mw) + (1 - \frac{1}{zw})\frac{A}{k^2} + (1 - \frac{z}{zw})\frac{B}{k}, \tag{A21}$$

which implies:

$$\frac{k}{m}\left(\hat{C}(k,m) - \frac{A}{k^2} - \frac{B}{k}\right) = \frac{kz}{mw}\left(\tilde{C}(kz, mw) - \frac{A}{(kz)^2} - \frac{B}{kz}\right), \tag{A22}$$

such that the left-hand side is a constant C, and $g_U^{(k,m)}(\partial_{ab}, \partial_{ab}) = \frac{A}{k^2} + \frac{B}{k} + \frac{m}{k}C$. Now, for a rational matrix M, pulling back through v_M gives:

$$g_M^{(k,m)}(\partial_{ab}, \partial_{ab}) = \frac{1}{\tilde{M}_{ab}^2}\tilde{M}_{ab}\left(\frac{A}{k^2} + \frac{B}{k} + \frac{|\tilde{M}_a|}{k}C\right) + \frac{1}{\tilde{M}_{ab}^2}\tilde{M}_{ab}(\tilde{M}_{ab} - 1)\left(\frac{A}{k^2} + \frac{B}{k}\right) + 0$$

$$= \frac{A}{k^2} + \frac{B}{k} + \frac{|\tilde{M}_a|}{\tilde{M}_{ab}k}C$$

$$= \frac{A}{k^2} + k\frac{B}{k^2} + \frac{|M|}{M_{ab}}\frac{C}{k^2}. \tag{A23}$$

Summarizing, we found:

$$g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = \frac{A}{k^2} + \delta_{ac}\left(k\frac{B}{k^2} + \delta_{bd}\frac{|M|}{M_{ab}}\frac{C}{k^2}\right), \tag{A24}$$

which proves the first statement. The second statement follows by plugging Equation (23) into Equation (A8). Finally, the statement about the positive-definiteness is a direct consequence of Proposition 7. □

Proof of Theorem 12. Suppose, contrary to the claim, that a family of metrics $g_M^{(k,m)}$ exists, which is invariant with respect to any conditional embedding. By Theorem 11, these metrics are of the form of Equation (23). To prove the claim, we only need to show that A, B and C vanish. In the following, we study conditional embeddings where Q consists of identity matrices and evaluate the isometry requirement $(f^* g^{(l,n)})_M(\partial_{ab}, \partial_{cd}) = g_M^{(k,m)}(\partial_{ab}, \partial_{cd})$.

Step 1: In the case $a \neq c$, we obtain from the invariance requirement and Equation (A8), that:

$$\frac{A}{k^2} = |\overline{R}_a||\overline{R}_c|\frac{A}{l^2}. \tag{A25}$$

Observe that:

$$\frac{1}{k}\sum_{i=1}^{k}|\overline{R}_i| = \frac{1}{k}|\overline{R}| = \frac{l}{k}. \tag{A26}$$

In fact, $|\overline{R}_i|$ is the cardinality of the i-th block of the partition belonging to $\overline{R}$. Therefore, if we choose $\overline{R}$ to be the partition indicator matrix of a partition that is not homogeneous and in which $|\overline{R}_a| > l/k$ and $|\overline{R}_c| > l/k$, then Equation (A25) implies that $A = 0$.

Step 2: In the case $a = c$ and $b \neq d$, we obtain from invariance and Equation (A8), that:

$$\frac{B}{k} = \sum_{i=1}^{l}\sum_{s=1}^{l}\overline{R}_{ai}\overline{R}_{as}\delta_{is}\frac{B}{l} = |\overline{R}_a|\frac{B}{l}. \tag{A27}$$

Again, we may chose $\overline{R}_a$ in such a way that $|\overline{R}_a| \neq \frac{k}{l}$ and find that $B = 0$.

Step 3: Finally, in the case $a = c$ and $b = d$, we obtain from invariance and Equation (A8), that:

$$\frac{C|M|}{k^2 M_{ab}} = \sum_{i=1}^{l}\sum_{s=1}^{l}\overline{R}_{ai}\overline{R}_{as}\delta_{i,s}\frac{C|\overline{R}^\top M|}{l^2(\overline{R}^\top M)_{ib}} = |\overline{R}_a|\frac{C|\overline{R}^\top M|}{l^2 M_{ab}}. \tag{A28}$$

If we chose $\overline{R}_a$, such that $|\overline{R}_a| \neq \frac{|M|}{|\overline{R}^\top M|}$, then we see that $C = 0$. Therefore, $g^{(k,m)}$ is the zero-tensor, which is not a metric. $\square$

Acknowledgments

Author Contributions

All authors contributed to the design of the research. The research was carried out by all authors, with main contributions by Guido Montúfar and Johannes Rauh. The manuscript was written by Guido Montúfar, Johannes Rauh and Nihat Ay. All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interests.

References

1. Amari, S. Natural gradient works efficiently in learning. *Neur. Comput.* **1998**, *10*, 251–276.

2. Kakade, S. A Natural Policy Gradient. *Advances in Neural Information Processing Systems 14*; MIT Press: Cambridge, MA, USA, 2001; pp. 1531–1538.

3. Shahshahani, S. *A New Mathematical Framework for the Study of Linkage and Selection*; American Mathematical Society: Providence, RI, USA, 1979.

4. Chentsov, N. *Statistical Decision Rules and Optimal Inference*; American Mathematical Society: Providence, RI, USA, 1982.

5. Campbell, L. An extended Čencov characterization of the information metric. *Proc. Am. Math. Soc.* **1986**, *98*, 135–141.

6. Sutton, R.S.; McAllester, D.; Singh, S.; Mansour, Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation. *In Advances in Neural Information Processing Systems 12*; MIT Press: Cambridge, MA, USA, 2000; pp. 1057–1063.

7. Marbach, P.; Tsitsiklis, J. Simulation-based optimization of Markov reward processes. *IEEE Trans. Autom. Control* **2001**, *46*, 191–209.

8. Montúfar, G.; Ay, N.; Zahedi, K. Expressive power of conditional restricted boltzmann machines for sensorimotor control. **2014**, arXiv:1402.3346.

9. Ay, N.; Montúfar, G.; Rauh, J. Selection Criteria for Neuromanifolds of Stochastic Dynamics. In *Advances in Cognitive Neurodynamics (III)*; Yamaguchi, Y., Ed.; Springer-Verlag: Dordrecht, The Netherlands 2013; pp. 147–154.

10. Peters, J.; Schaal, S. Natural Actor-Critic. *Neurocomputing* **2008**, *71*, 1180–1190.

11. Peters, J.; Schaal, S. Policy Gradient Methods for Robotics. In Proceedings of the IEEE International Conference on Intelligent Robotics Systems (IROS 2006), Beijing, China, 9–15 October 2006.

12. Peters, J.; Vijayakumar, S.; Schaal, S. Reinforcement learning for humanoid robotics. In Proceedings of the third IEEE-RAS international conference on humanoid robots, Karlsruhe, Germany, 29–30 September 2003; pp. 1–20.

13. Bagnell, J.A.; Schneider, J. Covariant policy search. In Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9–15, 2003; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2003; pp. 1019–1024.

14. Lebanon, G. Axiomatic geometry of conditional models. *IEEE Trans. Inform. Theor.* **2005**, *51*, 1283–1294.

15. Lebanon, G. An Extended Čencov-Campbell Characterization of Conditional Information Geometry. In Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence (UAI 04), Banff, AL, Canada, 7–11 July 2004; Chickering, D.M., Halpern, J.Y., Eds.; AUAI Press: Arlington, VA, USA, 2004; pp. 341–345.

16. Barndorff-Nielsen, O. *Information and Exponential Families: In statistical Theory*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1978.

17. Brown, L.D. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*; Institute of Mathematical Statistics: Hayward, CA, USA, 1986.

18. Zahedi, K.; Ay, N.; Der, R. Higher coordination with less control—A result of informaion maximiation in the sensorimotor loop. *Adapt. Behav.* **2010**, *18*.

19. Hofbauer, J.; Sigmund, K. *Evolutionary Games and Population Dynamics*; Cambridge University Press: Cambridge, United Kingdom, 1998.

20. Ay, N.; Erb, I. On a notion of linear replicator equations. *J. Dyn. Differ. Equ.* **2005**, *17*, 427–451.