Axiomatic Geometry of Conditional Models

Guy Lebanon

***Abstract*—We formulate and prove an axiomatic characterization of the Riemannian geometry underlying manifolds of conditional models. The characterization holds for both normalized and nonnormalized conditional models. In the normalized case, the characterization extends the derivation of the Fisher information by Čencov while in the nonnormalized case it extends Campbell's theorem. Due to the close connection between the conditional I-divergence and the product Fisher information metric, we provides a new axiomatic interpretation of the geometries underlying logistic regression and AdaBoost.**

***Index Terms*—Conditional probability estimation, congruent embedding by a Markov morphism, information geometry.**

I. Introduction

THE theory of information geometry presents a geometric interpretation of statistical properties and techniques. Among many examples of such properties and techniques are efficiency of estimators, robustness, maximum-likelihood estimation for exponential models, and hypothesis testing. The geometric properties of a space of statistical models $\{p(x;\theta) : \theta \in \Theta\}$ is studied using the mathematical framework of Riemannian geometry. In this framework, the geometry of a space Θ is specified by a local inner product $g_\theta(\cdot,\cdot)$, $\theta \in \Theta$ which translates into familiar concepts such as distance, curvature, and angles. An overview of the wide range of results in this field may be found in the monographs of Amari [1] and Kass and Voss [2].

A fundamental assumption in the information-geometric framework, is the choice of the Fisher information as the metric that underlies the geometry of probability distributions. The choice of the Fisher information metric may be motivated in several ways the strongest of which is Čencov's characterization theorem ([3, Lemma 11.3]). In his theorem, Čencov proves that the Fisher information metric is the only metric that is invariant under a family of probabilistically meaningful mappings termed congruent embeddings by a Markov morphism. Later, Campbell extended Čencov's result to include nonnormalized positive models [4].

The theorems of Čencov and Campbell are particularly interesting since Fisher information is pervasive in statistics and machine learning. It is the asymptotic variance of the maximum-likelihood estimators under some regularity conditions. Cramér and Rao used it to compute a lower bound on the variance of arbitrary unbiased estimators. In Bayesian statistics, it was used by Jeffreys to define noninformative prior. It is tightly connected to the Kullback–Leibler divergence which is the cornerstone of maximum-likelihood estimation for exponential models as well as various aspects of information theory.

While the geometric approach to statistical inference has attracted considerable attention, little research was conducted on the geometric approach to conditional inference. The characterization theorems of Čencov and Campbell no longer apply in this setting and the different ways of choosing a geometry for the space of conditional distributions, in contrast to the nonconditional case, are not supported by theoretical considerations.

In this paper, we extend the results of Čencov and Campbell to provide an axiomatic characterization of conditional information geometry. We derive the characterization theorem in the setting of nonnormalized conditional models from which the geometry for normalized models is obtained as a special case. In addition, we demonstrate a close connection between the characterized geometry and the conditional I-divergence which leads to a new axiomatic interpretation of the geometry underlying the primal problems of logistic regression and AdaBoost. This interpretation builds on the recently found connection between AdaBoost and constrained minimization of I-divergence [5].

Throughout the paper, we consider spaces of strictly positive conditional models where the sample spaces of the explanatory and response variable are finite. Moving to the infinite case presents some serious difficulties. The positivity constraint on the other hand does not play a crucial role and may by discarded at some notational cost.

The next section describes some relevant concepts from Riemannian geometry and is followed by a description of the manifolds of normalized and nonnormalized conditional models. Section IV describes a family of probabilistic mappings that will serve as the basis for the invariance requirement of the characterization theorem in Section V. Section VI applies the characterization result to logistic regression and AdaBoost and is followed by concluding remarks.

II. Relevant Concepts From Riemannian Geometry

In this section, we describe briefly relevant concepts from Riemannian geometry. For more details refer to any textbook discussing Riemannian geometry, for example [6], [7].

A homeomorphism $\phi : X \to Y$ is a bijection for which both ϕ and ϕ^{-1} are continuous. We then say that X and Y are homeomorphic. An n-dimensional topological manifold $\mathcal{M}$ is a topological subspace of $\mathbb{R}^m$, $m \geq n$, that is locally equivalent to $\mathbb{R}^n$, i.e., for every point $x \in \mathcal{M}$ there exists an open neighborhood U that is homeomorphic to $\mathbb{R}^n$. The above definition of a topological manifold makes use of an ambient Euclidean

Manuscript received July 29, 2004; revised December 17, 1004. The material in this paper was presented at Uncertainty in Artificial Intelligence 2004, Banff, AB, Canada, July 2004.

The author is with the Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: lebanon@cs.cmu.edu).

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2005.844060

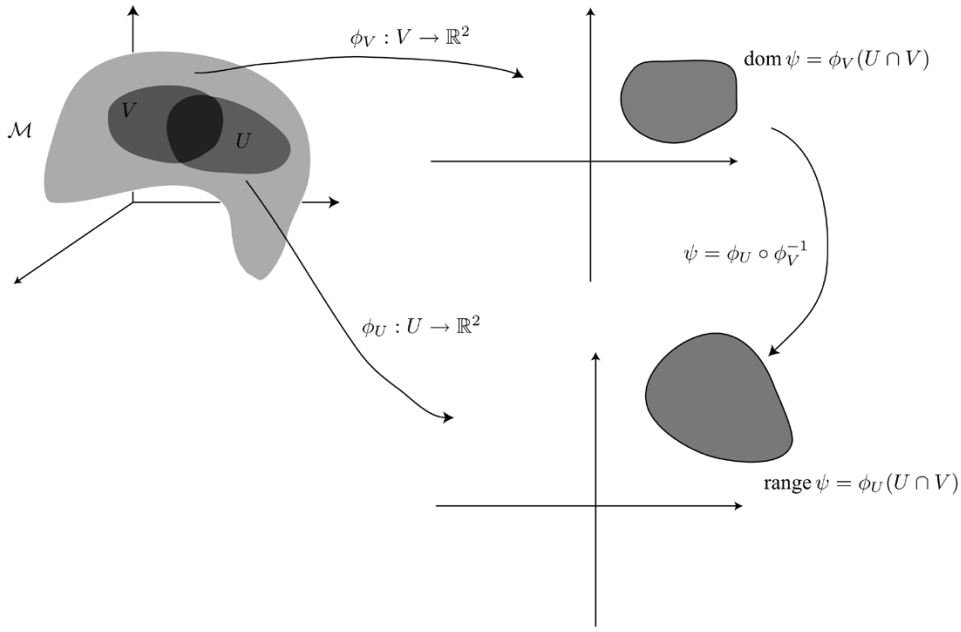

Fig. 1. Two neighborhoods U, V in a two-dimensional manifold $\mathcal{M}$, the coordinate charts ϕ_U, ϕ_V, and the transition function ψ between them.

space $\mathbb{R}^m$. While sufficient for our purposes, such a reference to $\mathbb{R}^m$ is not strictly necessary and may be discarded at the cost of certain topological assumptions [8]. Unless otherwise noted, for the remainder of this section we assume that all manifolds are of dimension n.

The local homeomorphisms in the above definition $\phi_U : U \subset \mathcal{M} \to \mathbb{R}^n$ are usually called charts. If for every pair of charts ϕ_U, ϕ_V the transition function defined by

$$\psi : \phi_V(U \cap V) \subset \mathbb{R}^n \to \mathbb{R}^n, \qquad \psi = \phi_U \circ \phi_V^{-1}$$

is a C^∞ differentiable map then $\mathcal{M}$ is called an n-differentiable manifold. The charts and transition function for a two-dimensional manifold are illustrated in Fig. 1.

Differentiable manifolds of dimensions 1 and 2 may be visualized as smooth curves and surfaces in Euclidean space. Examples of n-dimensional differentiable manifolds are the Euclidean space $\mathbb{R}^n$, the n-sphere

$$\mathbb{S}_n = \left\{ x \in \mathbb{R}^{n+1} : \sum_{i=1}^{n} x_i^2 = 1 \right\} \tag{1}$$

and the n-simplex

$$\mathbb{P}_n = \left\{ x \in \mathbb{R}^{n+1} : \forall\, i, x_i > 0, \sum_{i=1}^{n} x_i = 1 \right\}. \tag{2}$$

Using the charts, we can extend the definition of differentiable maps to real-valued functions on manifolds $f : \mathcal{M} \to \mathbb{R}$ and functions from one manifold to another $f : \mathcal{M} \to \mathcal{N}$. This extension is based on the known definition of differentiability of maps between Euclidean spaces $f' : \mathbb{R}^n \to \mathbb{R}^m$.

A continuous function $f : \mathcal{M} \to \mathbb{R}$ is said to be $C^\infty(\mathcal{M})$ differentiable if for every chart ϕ_U the function

$$f \circ \phi_U^{-1} : \mathbb{R}^n \to \mathbb{R}$$

is C^∞ differentiable. A continuous mapping between two differentiable manifolds $f : \mathcal{M} \to \mathcal{N}$ is said to be $C^\infty(\mathcal{M}, \mathcal{N})$ differentiable if

$$\forall\, r \in C^\infty(\mathcal{N}) \quad r \circ f \in C^\infty(\mathcal{M}).$$

A diffeomorphism between two manifolds $\mathcal{M}, \mathcal{N}$ is a bijection $f : \mathcal{M} \to \mathcal{N}$ such that

$$f \in C^\infty(\mathcal{M}, \mathcal{N}) \quad \text{and} \quad f^{-1} \in C^\infty(\mathcal{N}, \mathcal{M}).$$

For every point $x \in \mathcal{M}$, we define an n-dimensional vector space $T_x\mathcal{M}$ called the tangent space. The tangent space is equivalent to $\mathbb{R}^n$ and its members are vectors that act as directional derivatives on $C^\infty(\mathcal{M})$ differentiable functions. Intuitively, tangent spaces and tangent vectors are a generalization of the usual notions for smooth two-dimensional surfaces in an embedding $\mathbb{R}^3$. The technical definition, however, makes no use of an embedding space [6].

In many cases, the manifold $\mathcal{M}$ is a submanifold of a larger manifold, often $\mathbb{R}^m$, $m \geq n$. For example, both $\mathbb{P}_n$ and $\mathbb{S}_n$ defined in (1), (2) are submanifolds of $\mathbb{R}^{n+1}$. In these cases, the tangent space of the submanifold $T_x\mathcal{M}$ is a vector subspace of $T_x\mathbb{R}^m \cong \mathbb{R}^m$ and we may represent tangent vectors $v \in T_x\mathcal{M}$ in the standard basis $\{\partial_i\}_{i=1}^m$ of the embedding tangent space $T_x\mathbb{R}^m$ as $v = \sum_{i=1}^m v_i \partial_i$. For example, for the simplex and the sphere we have

$$T_x\mathbb{P}_n = \left\{ v \in \mathbb{R}^n : \sum_{i=1}^{n} v_i = 0 \right\} \tag{3}$$

$$T_x\mathbb{S}_n = \left\{ v \in \mathbb{R}^n : \sum_{i=1}^{n} v_i x_i = 0 \right\}. \tag{4}$$

Fig. 2 illustrates the tangent spaces of the 2-simplex and the 2-sphere.

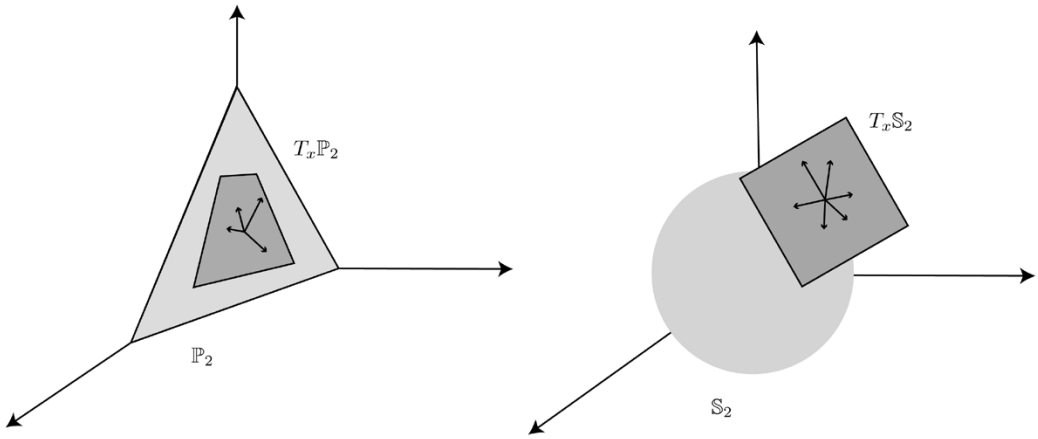

Fig. 2. Tangent spaces of the 2-simplex $T_x\mathbb{P}_2$ and the 2-sphere $T_x\mathbb{S}_2$.

The definition of topological manifolds is suited for notions such as continuity and convergence. The additional structure that differentiable manifolds possess is necessary to deal with differentiation. To define geometric quantities such as length, curvature, and angles, differentiable manifolds have to be augmented with a Riemannian metric.

A Riemannian manifold $(\mathcal{M}, g)$ is a differentiable manifold $\mathcal{M}$ equipped with a Riemannian metric g. The metric g is defined by a symmetric positive-definite inner product on the tangent spaces $T_x\mathcal{M}$ that is C^∞ differentiable in x

$$g_x(\cdot,\cdot) : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}, \qquad x \in \mathcal{M}.$$

Since for every $u, v \in T_x\mathcal{M}$

$$g_x(v,u) = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i u_j g_x(\partial_i, \partial_j)$$

g_x is completely described by $\{g_x(\partial_i, \partial_j) : 1 \le i, j \le n\}$—the set of inner products between the basis elements $\{\partial_i\}_{i=1}^n$ of $T_x\mathcal{M}$.

The metric enables us to define lengths of tangent vectors $v \in T_x\mathcal{M}$ by $\|v\| = \sqrt{g_x(v,v)}$ and lengths of curves $\gamma : [a,b] \to \mathcal{M}$ by $L(\gamma) = \int_a^b \|\dot\gamma(t)\| \mathrm{d}t$ where $\dot\gamma(t)$ is the velocity vector of the curve γ at time t. Using the above definition of lengths of curves, we can define the distance $d(x,y)$ between two points $x, y \in \mathcal{M}$ as

$$d(x,y) = \inf_{\gamma \in \Gamma(x,y)} \int_a^b \|\dot\gamma(t)\| \mathrm{d}t$$

where $\Gamma(x,y)$ is the set of piecewise-differentiable curves connecting x and y. The distance d, also called geodesic distance, satisfies the usual requirements of a distance and is compatible with the topological structure of $\mathcal{M}$ as a topological manifold.

Given two Riemannian manifolds $(\mathcal{M}, g)$, $(\mathcal{N}, h)$, and a diffeomorphism between them $f : \mathcal{M} \to \mathcal{N}$ we define the push-forward and pull-back maps below, which are of crucial importance to the characterization theorems of Section V.

Definition 1: The push-forward map $f_* : T_x\mathcal{M} \to T_{f(x)}\mathcal{N}$, associated with $f : \mathcal{M} \to \mathcal{N}$ is the vector f_*v that satisfies

$$v(r \circ f) = (f_*v)r, \qquad \forall\, r \in C^\infty(\mathcal{N}).$$

Intuitively, the push-forward transforms velocity vectors of curves γ to velocity vectors of transformed curves $f(\gamma)$.

Definition 2: Given $(\mathcal{N}, h)$ and a diffeomorphism $f : \mathcal{M} \to \mathcal{N}$ we define a metric f^*h on $\mathcal{M}$ called the pull-back metric by the relation

$$(f^*h)_x(u,v) = h_{f(x)}(f_*u, f_*v).$$

Definition 3: An isometry is a diffeomorphism $f : \mathcal{M} \to \mathcal{N}$ between two Riemannian manifolds $(\mathcal{M}, g)$, $(\mathcal{N}, h)$ for which the following condition holds:

$$g_x(u,v) = (f^*h)_x(u,v), \qquad \forall\, x \in \mathcal{M},\ \forall\, u, v \in T_x\mathcal{M}.$$

Isometries, as defined above, identify two Riemannian manifolds as identical in terms of their Riemannian structure. Accordingly, isometries preserve all the geometric properties including the geodesic distance function

$$d_g(x,y) = d_h(f(x), f(y)).$$

Note that the above definition of an isometry is defined through the local metric in contrast to the global definition of isometry in other branches of mathematical analysis.

We proceed in Section III to define the Fisher geometry on a manifold of distributions and to examine manifolds of conditional models.

III. Normalized and Nonnormalized Conditional Manifolds

Parametric inference in statistics is concerned with a parametric family of distributions $\{p(y;\theta) : \theta \in \Theta \subset \mathbb{R}^n\}$. If the parameter space Θ is a differentiable manifold and the mapping $\theta \mapsto p(y;\theta)$ is a diffeomorphism we can identify statistical models in the family as points on the manifold Θ. The Fisher information matrix $E\{ss^\top\}$ where s is the gradient of the log likelihood or the score $s = \nabla_\theta \log p(y;\theta)$ may be used to endow Θ with the following Riemannian metric:

$$g_\theta(u,v) = \sum_{i,j} u_i v_j \int p(y;\theta) \frac{\partial}{\partial\theta_i} \log p(y;\theta) \frac{\partial}{\partial\theta_j} \log p(y;\theta) \mathrm{d}y.$$

In the finite nonparametric setting, which is the topic of this paper, the event space $\mathcal{Y}$ is a finite set with $|\mathcal{Y}| = n$ and $\Theta = \mathbb{P}_{n-1}$ which represents the manifold of all positive probability models over $\mathcal{Y}$. The positivity constraint is necessary for $\Theta = \mathbb{P}_{n-1}$ to be a manifold. If zero probabilities are admitted, the appropriate framework for the parameter space $\Theta = \overline{\mathbb{P}_{n-1}}$ is a manifold with corners [7]. In the final section, we return to this topic and demonstrate how to extend the results in this paper to the space of nonnegative conditional models. The finiteness of $\mathcal{X}$ and $\mathcal{Y}$ is necessary for Θ to be a finite-dimensional manifold. Relaxing the finiteness assumption results in a manifold where each neighborhood is homeomorphic to an infinite-dimensional Hilbert space [9]. Such manifolds are called Frechet manifolds and are the topic of a branch of geometry called global analysis [10].

Considering $\mathbb{P}_{n-1}$ as a submanifold of $\mathbb{R}^n$, we represent tangent vectors $v \in T_\theta \mathbb{P}_{n-1}$ in the standard basis of $T_\theta \mathbb{R}^n$. As mentioned earlier (3), this results in the following representation of $v \in T_\theta \mathbb{P}_{n-1}$:

$$v = \sum_{i=1}^{n} v_i \partial_i \quad \text{subject to} \quad \sum_i v_i = 0.$$

Using this representation, the Fisher information metric becomes

$$g_\theta(u,v) = \sum_{i=1}^{n} \frac{u_i v_i}{\theta_i}, \qquad u, v \in T_\theta \mathbb{P}_{n-1}.$$

Note that the Fisher metric emphasizes coordinates that correspond to low probabilities. The fact that the metric $g_\theta(u,v) \to \infty$ when $\theta \to 0$ is not problematic since length of curves that involves integrals over g do converge. For more details on the Fisher information metric in the parametric and the nonparametric cases refer to [1].

Given two finite event sets $\mathcal{X}, \mathcal{Y}$ of sizes k and m, respectively, a conditional probability model $p(y|x)$ reduces to an element of $\mathbb{P}_{m-1}$ for each $x \in \mathcal{X}$. We may thus identify the space of conditional probability models associated with $\mathcal{X}$ and $\mathcal{Y}$ as the product space

$$\mathbb{P}_{m-1} \times \cdots \times \mathbb{P}_{m-1} = \mathbb{P}_{m-1}^k.$$

For our purposes, it will be more convenient to work with the more general case of positive nonnormalized conditional models. Dropping the normalization constraints $\sum_i p(y_i|x_j) = 1$ we obtain conditional models in the cone of $k \times m$ matrices with positive entries, denoted by $\mathbb{R}_+^{k\times m}$. Since a normalized conditional model is also a nonnormalized one, we can consider $\mathbb{P}_{m-1}^k$ to be a subset of $\mathbb{R}_+^{k\times m}$. Results obtained for nonnormalized models apply then to normalized models as a special case. In addition, some of the notation and formulation is simplified by working with nonnormalized models. By taking this approach, we follow the philosophy of [4] and [5].

In the interest of simplicity, we will often use matrix notation instead of the standard probabilistic notation. A conditional model (either normalized or nonnormalized) is described by a positive matrix M such that $M_{ij} = p(y_j|x_i)$. Matrices that correspond to normalized models are (row) stochastic matrices. We denote tangent vectors to $\mathbb{R}_+^{k\times m}$ using the standard basis

$$T_M \mathbb{R}_+^{k\times m} = \operatorname{span}\{\partial_{ij} : i = 1, \ldots, k, j = 1, \ldots, m\}.$$

Tangent vectors to $\mathbb{P}_{m-1}^k$, when expressed using the basis of the embedding tangent space $T_M \mathbb{R}_+^{k\times m}$ are linear combinations of $\{\partial_{ij}\}$ such that the sum of the combination coefficients over each row are 0, e.g.,

$$\frac{1}{2}\partial_{11} + \frac{1}{2}\partial_{12} - \partial_{13} + \frac{1}{3}\partial_{21} - \frac{1}{3}\partial_{22} \in T_M \mathbb{P}_2^3.$$

The identification of the space of conditional models as a product of simplexes demonstrates the topological and differentiable structure. In particular, we do not assume that the metric has a product form. However, it is instructive to consider, as a special case, the product Fisher information metric on $\mathbb{P}_{n-1}^k$ and $\mathbb{R}_+^{k\times m}$. Using the above representation of tangent vectors it reduces to

$$g_M(u,v) = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{u_{ij} v_{ij}}{M_{ij}} \tag{5}$$

where $u, v \in T_M \mathbb{R}_+^{k\times m}$ or $u, v \in T_M \mathbb{P}_{m-1}^k$. A different way of expressing (5) is by specifying the values of the metric on pairs of basis elements

$$g_M(\partial_{ab}, \partial_{cd}) = \delta_{ac}\delta_{bd} \frac{1}{M_{ab}} \tag{6}$$

where $\delta_{ab} = 1$ if $a = b$ and 0 otherwise.

In the characterization theorem we will make use of the fact that $\mathbb{P}_{m-1}^k \cap \mathbb{Q}^{k\times m}$ and $\mathbb{R}_+^{k\times m} \cap \mathbb{Q}^{k\times m} = \mathbb{Q}_+^{k\times m}$ are dense in $\mathbb{P}_{m-1}^k$ and $\mathbb{R}_+^{k\times m}$, respectively. The set $\mathbb{Q}_+^{k\times m}$ of positive rational matrices is assumed to be the appropriate subset of $\mathbb{R}_+^{k\times m}$. Since continuous functions are uniquely characterized by their values on dense sets, it is enough to compute the metric for positive rational models $\mathbb{Q}_+^{k\times m}$. The value of the metric on nonrational models follows from its continuous extension to $\mathbb{R}_+^{k\times m}$.

In Section IV, we define a class of transformations called congruent embeddings by a Markov morphism. These transformations set the stage for the axioms in the characterization theorem of Section V.

IV. Congruent Embeddings by Markov Morphisms of Conditional Models

The characterization result of Section V is based on axioms that require geometric invariance through a set of transformations between conditional models. These transformations are a generalization of the transformations underlying Čencov's theorem. For consistency with the terminology of Čencov [3] and Campbell [4] we refer to these transformations as Congruent embeddings by Markov morphisms of conditional models.

Definition 4: Let $\mathcal{A} = \{A_1, \ldots, A_m\}$ be a set partition of $\{1, \ldots, n\}$ with $0 < m \leq n$. A matrix $Q \in \mathbb{R}^{m\times n}$ is called $\mathcal{A}$-stochastic if

$$\forall\, i \quad \sum_{j=1}^{n} Q_{ij} = 1 \quad \text{and} \quad Q_{ij} = \begin{cases} c_{ij} > 0, & j \in A_i \\ 0, & j \notin A_i. \end{cases}$$

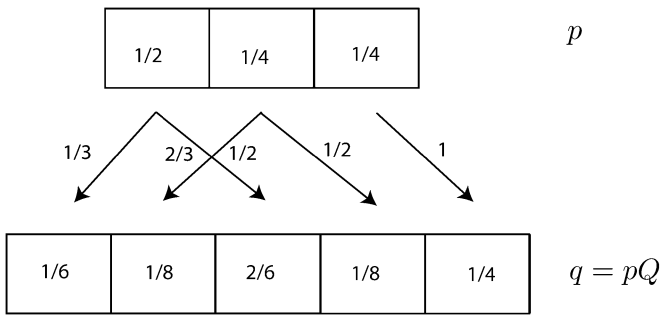

Fig. 3. Congruent embedding by a Markov morphism of $p = (1/2, 1/4, 1/4)$.

In other words, $\mathcal{A}$-stochastic matrices are stochastic matrices whose rows are concentrated on the sets of the partition $\mathcal{A}$.

For example, if $\mathcal{A} = \{\{1,3\},\{2,4\},\{5\}\}$ then the following matrix is $\mathcal{A}$-stochastic:

$$\begin{pmatrix} 1/3 & 0 & 2/3 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{7}$$

Obviously, the columns of any $\mathcal{A}$-stochastic matrix have precisely one nonzero element. If $m = n$ then an $\mathcal{A}$-stochastic matrix is a permutation matrix.

Multiplying a row probability vector $p \in \mathbb{R}_+^{1\times m}$ with an $\mathcal{A}$-stochastic matrix $Q \in \mathbb{R}^{m\times n}$ results in a row probability vector $q \in \mathbb{R}_+^{1\times n}$. The mapping $p \mapsto pQ$ has the following statistical interpretation. The event x_i is split into $|A_i|$ distinct events stochastically, with the splitting probabilities given by the i-row of Q. The new event space, denoted by $\mathcal{Z} = \{z_1, \ldots, z_n\}$, may be considered a refinement of $\mathcal{X} = \{x_1, \ldots, x_m\}$ (if $m < n$) and the model $q(z)$ is a consistent refinement of $p(x)$. For example, multiplying $p = (1/2, 1/4, 1/4)$ with the matrix Q in (7) yields

$$q = pQ = (1/6, 1/8, 2/6, 1/8, 1/4).$$

In this transformation, x_1 was split into $\{z_1, z_3\}$ with unequal probabilities, x_2 was split into $\{z_2, z_4\}$ with equal probabilities and x_3 was relabeled z_5 (Fig. 3)

The transformation $q \mapsto qQ$ is injective and therefore invertible. For example, the inverse transformation to Q in (7) is

$$\begin{aligned} p(x_1) &= q(z_1) + q(z_3) \\ p(x_2) &= q(z_2) + q(z_4) \\ p(x_3) &= q(z_5). \end{aligned}$$

The inverse transformation may be interpreted as extracting a sufficient statistic T from $\mathcal{Z}$. The sufficient statistic joins events in $\mathcal{Z}$ to create the event space $\mathcal{X}$, hence transforming models on $\mathcal{Z}$ to corresponding models on $\mathcal{X}$.

So far we have considered transformations of nonconditional models. The straightforward generalization to conditional models involves performing a similar transformation on the response space $\mathcal{Y}$ for every nonconditional model $p(\cdot|x_i)$ followed by transforming the explanatory space $\mathcal{X}$. It is formalized in the definitions below and illustrated in Fig. 4.

Definition 5: Let $M \in \mathbb{R}^{k\times m}$ and $Q = \{Q^{(i)}\}_{i=1}^k$ be a set of matrices in $\mathbb{R}^{m\times n}$. We define the row product $M \otimes Q \in \mathbb{R}^{k\times n}$ as

$$[M \otimes Q]_{ij} = \sum_{s=1}^{m} M_{is} Q_{sj}^{(i)} = [MQ^{(i)}]_{ij}. \tag{8}$$

In other words, the ith row of $M \otimes Q$ is the ith row of the matrix product $MQ^{(i)}$.

Definition 6: Let $\mathcal{B}$ be a k sized partition of $\{1, \ldots, l\}$ and $\{\mathcal{A}^{(i)}\}_{i=1}^k$ be a set of m sized partitions of $\{1, \ldots, n\}$. Furthermore, let $R \in \mathbb{R}_+^{k\times l}$ be a $\mathcal{B}$-stochastic matrix and $Q = \{Q^{(i)}\}_{i=1}^k$ a sequence of $\mathcal{A}^{(i)}$-stochastic matrices in $\mathbb{R}_+^{m\times n}$. Then the map

$$f : \mathbb{R}_+^{k\times m} \to \mathbb{R}_+^{l\times n} \quad f(M) = R^\top (M \otimes Q) \tag{9}$$

is termed a congruent embedding by a Markov morphism of $\mathbb{R}_+^{k\times m}$ into $\mathbb{R}_+^{l\times n}$ and the set of all such maps is denoted by $\mathfrak{F}_{k,m}^{l,n}$.

Congruent embeddings by a Markov morphism f are injective and if restricted to the space of normalized models $\mathbb{P}_{m-1}^k$ they produce a normalized model as well, i.e., $f(\mathbb{P}_{m-1}^k) \subset \mathbb{P}_{n-1}^l$.

The component-wise version of (9) is

$$[f(M)]_{ij} = \sum_{s=1}^{k} \sum_{t=1}^{m} R_{si} Q_{tj}^{(s)} M_{st} \tag{10}$$

with the above sum containing precisely one nonzero term since every column of $Q^{(s)}$ and R contains only one nonzero entry. The push-forward map $f_* : T_M \mathbb{R}_+^{k\times m} \to T_{f(M)} \mathbb{R}_+^{l\times n}$ associated with f is

$$f_*(\partial_{ab}) = \sum_{i=1}^{l} \sum_{j=1}^{n} R_{ai} Q_{bj}^{(a)} \partial'_{ij} \tag{11}$$

where $\{\partial_{ab}\}_{a,b}$ and $\{\partial'_{ij}\}_{ij}$ are the bases of $T_M \mathbb{R}_+^{k\times m}$ and $T_{f(M)} \mathbb{R}_+^{l\times n}$, respectively.

Using Definition 2 and (11), the pull-back of a metric g on $\mathbb{R}_+^{l\times n}$ through $f \in \mathfrak{F}_{k,m}^{l,n}$ is

$$\begin{aligned} &(f^* g)_M(\partial_{ab}, \partial_{cd}) \\ &\quad = g_{f(M)}(f_* \partial_{ab}, f_* \partial_{cd}) \\ &\quad = \sum_{i=1}^{l} \sum_{j=1}^{n} \sum_{s=1}^{l} \sum_{t=1}^{n} R_{ai} R_{cs} Q_{bj}^{(a)} Q_{dt}^{(c)} g_{f(M)}(\partial'_{ij}, \partial'_{st}). \end{aligned} \tag{12}$$

An important special case of a congruent embedding by a Markov morphism is specified by uniform $\mathcal{A}$-stochastic matrices defined next.

Definition 7: An $\mathcal{A}$-stochastic matrix is called uniform if every row has the same number of nonzero elements and if all its positive entries are identical.

For example, the following matrix is a uniform $\mathcal{A}$-stochastic matrix for $\mathcal{A} = \{\{1,3\},\{2,4\},\{5,6\}\}$:

$$\begin{pmatrix} 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix}.$$

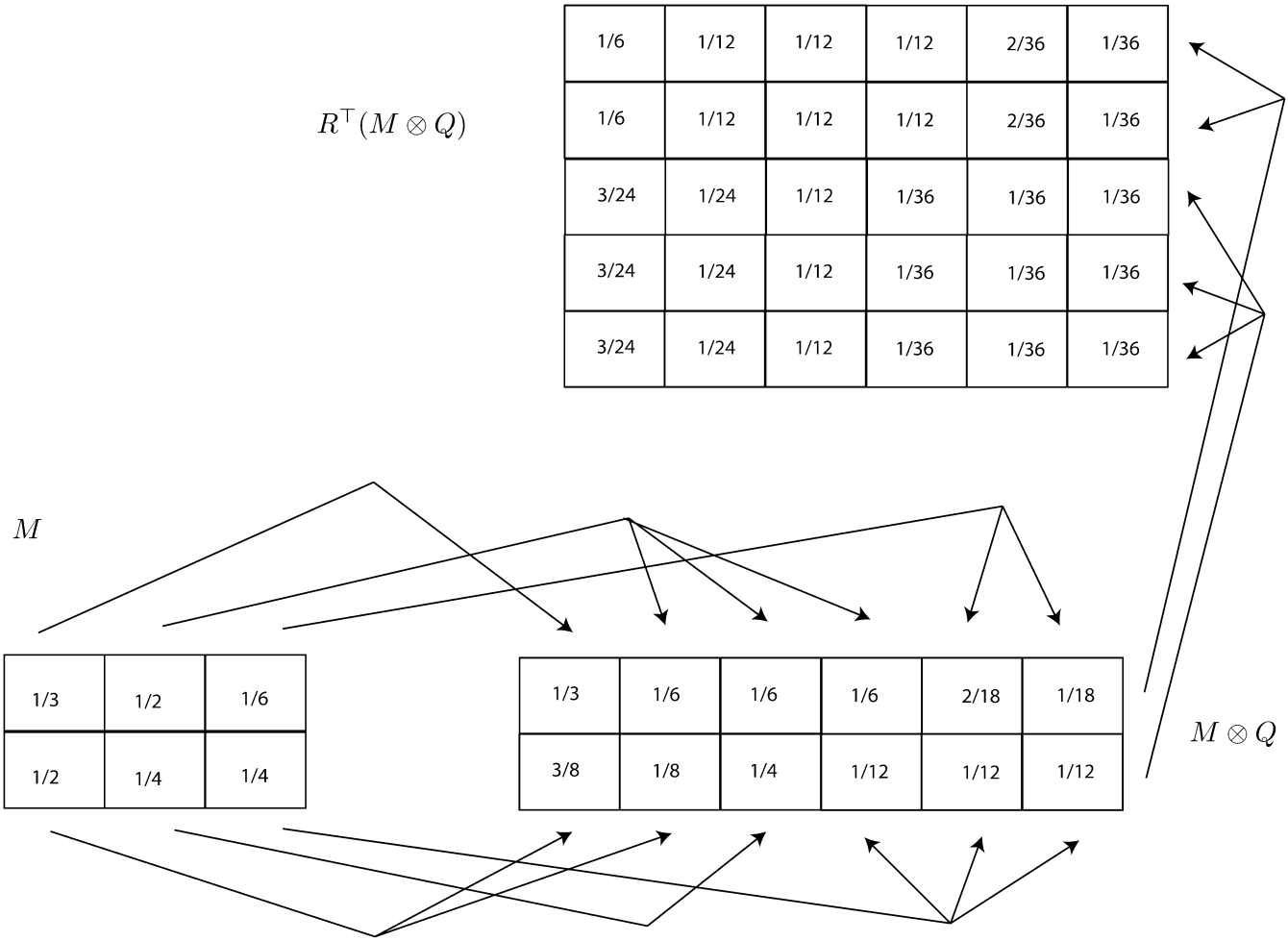

Fig. 4. Congruent embedding by a Markov morphism of $\mathbb{R}_+^{2\times 3}$ into $\mathbb{R}_+^{5\times 6}$.

We proceed in Section V to state and prove the characterization theorem.

V. A Characterization of Metrics on Conditional Manifolds

As mentioned in the previous section, congruent embeddings by a Markov morphism have a strong probabilistic interpretation. Such maps transform conditional models to other conditional models in a manner consistent with changing the granularity of the event spaces. Moving to a finer or coarser description of the event space should not have an effect on the models if such a move may be expressed as a sufficient statistic. It makes sense then to require that the geometry of a space of conditional models be invariant under such transformations. Such geometrical invariance is obtained by requiring maps $f \in \mathfrak{F}_{k,m}^{l,n}$ to be isometries. The main results of the paper are Theorems 1 and 2 below followed by Corollary 1. The proof of Theorem 1 bears some similarity to the proof of Campbell's theorem [4] which in turn is related to the proof technique used in Khinchin's characterization of the entropy [11]. Throughout the paper, we avoid Čencov's style of using category theory and use only standard techniques in differential geometry.

A. Three Useful Transformations

Before we turn to the characterization theorem, we show that congruent embeddings by a Markov morphisms are norm preserving and examine three special cases that will be useful later.

We denote by M_i the ith row of the matrix M and by $|\cdot|$ the L^1 norm applied to vectors or matrices

$$|v| = \sum_i |v_i| \quad |M| = \sum_i |M_i| = \sum_{ij} |M_{ij}|.$$

Proposition 1: Maps in $\mathfrak{F}_{k,m}^{l,n}$ are norm preserving

$$|M| = |f(M)| \quad \forall\, f \in \mathfrak{F}_{k,m}^{l,n}, \qquad \forall\, M \in \mathbb{R}_+^{k\times m}.$$

Proof: Multiplying a positive row vector v by an $\mathcal{A}$-stochastic matrix T is norm preserving

$$|vT| = \sum_i [vT]_i = \sum_j v_j \sum_i T_{ji} = \sum_j v_j = |v|.$$

As a result, $|[MQ^{(i)}]_i| = |M_i|$ for any positive matrix M and hence,

$$|M| = \sum_i |M_i| = \sum_i |[MQ^{(i)}]_i| = |M \otimes Q|.$$

A map $f \in \mathfrak{F}_{k,m}^{l,n}$ is norm preserving since

$$\begin{aligned}|M| &= |M \otimes Q| = |(M \otimes Q)^\top| = |(M \otimes Q)^\top R| \\ &= |R^\top (M \otimes Q)| = |f(M)|. \end{aligned} \qquad \square$$

We denote the symmetric group of permutations over k letters by $\mathfrak{S}_k$. The first transformation $\mathfrak{h}_\sigma^\Pi \in \mathfrak{F}_{k,m}^{k,m}$, parameterized by $\sigma \in \mathfrak{S}_k$ and

$$\Pi = (\pi^{(1)}, \ldots, \pi^{(k)}), \qquad \pi^{(i)} \in \mathfrak{S}_m,$$

is defined by $Q^{(i)}$ being the permutation matrix that corresponds to $\pi^{(i)}$ and R being the permutation matrix that corresponds to σ. The push-forward is

$$\mathfrak{h}_{\sigma *}^\Pi(\partial_{ab}) = \partial'_{\sigma(a)\pi^{(a)}(b)} \tag{13}$$

and requiring $\mathfrak{h}_\sigma^\Pi$ to be an isometry from $(\mathbb{R}_+^{k\times m}, g)$ to itself amounts to

$$g_M(\partial_{ab}, \partial_{cd}) = g_{\mathfrak{h}_\sigma^\Pi(M)}(\partial_{\sigma(a)\pi^{(a)}(b)}, \partial_{\sigma(c)\pi^{(c)}(d)}) \tag{14}$$

for all $M \in \mathbb{R}_+^{k\times m}$ and for every pair of basis vectors ∂_{ab}, ∂_{cd} in $T_M \mathbb{R}_+^{k\times m}$.

The usefulness of $\mathfrak{h}_\sigma^\Pi$ stems in part from the following proposition.

Proposition 2: Given $\partial_{a_1 b_1}, \partial_{a_2 b_2}, \partial_{c_1 d_1}, \partial_{c_2 d_2}$ with $a_1 \neq c_1$ and $a_2 \neq c_2$ there exists σ, Π such that

$$\mathfrak{h}_{\sigma *}^\Pi(\partial_{a_1 b_1}) = \partial_{a_2 b_2} \quad \mathfrak{h}_{\sigma *}^\Pi(\partial_{c_1 d_1}) = \partial_{c_2 d_2}. \tag{15}$$

Proof: The desired map may be obtained by selecting Π, σ such that $\sigma(a_1) = a_2, \sigma(c_1) = c_2$ and $\pi^{(a_1)}(b_1) = b_2$, $\pi^{(c_1)}(d_1) = d_2$. $\square$

The second transformation $\mathfrak{r}_{zw} \in \mathfrak{F}_{k,m}^{kz,mw}$, parameterized by $z, w \in \mathbb{N}$, is defined by $Q^{(1)} = \cdots = Q^{(k)} \in \mathbb{R}^{m\times mw}$ and $R \in \mathbb{R}^{k\times kz}$ being uniform matrices (in the sense of Definition 7). Note that each row of $Q^{(i)}$ has precisely w nonzero entries of value $1/w$ and each row of R has precisely z nonzero entries of value $1/z$. The exact forms of $\{Q^{(i)}\}$ and R are immaterial for our purposes and any uniform matrices of the above sizes will suffice. By (11) the push-forward is

$$\mathfrak{r}_{zw *}(\partial_{st}) = \frac{1}{zw} \sum_{i=1}^{z} \sum_{j=1}^{w} \partial'_{\pi(i)\sigma(j)}$$

for some permutations π, σ that depend on s, t and the precise shape of $\{Q^{(i)}\}$ and R. The pull-back of g is

$$\begin{aligned}(\mathfrak{r}_{zw}^* g)_M(\partial_{ab}, \partial_{cd}) = \frac{1}{(zw)^2} \sum_{i=1}^{z} \sum_{j=1}^{w} \sum_{s=1}^{z} \sum_{t=1}^{w} g_{\mathfrak{r}_{zw}(M)} \\ \times (\partial'_{\pi_1(i),\sigma_1(j)}, \partial'_{\pi_2(s),\sigma_2(t)}) \end{aligned} \tag{16}$$

again, for some permutations π_1, π_2, σ_1, σ_2.

We will often express rational conditional models $M \in \mathbb{Q}_+^{k\times m}$ as

$$M = \frac{1}{z}\tilde{M}, \qquad \tilde{M} \in \mathbb{N}^{k\times m},\ z \in \mathbb{N}$$

where $\mathbb{N}$ is the set of natural numbers. Given a rational model M, the third mapping

$$\mathfrak{y}_M \in \mathfrak{F}_{k,m}^{|\tilde{M}|, \prod_i |\tilde{M}_i|}, \qquad \text{where } M = \frac{1}{z}\tilde{M} \in \mathbb{Q}_+^{k\times m}$$

is associated with $Q^{(i)} \in \mathbb{R}^{m\times \prod_s |\tilde{M}_s|}$ and $R \in \mathbb{R}^{k\times |\tilde{M}|}$ which are defined as follows. The i-row of $R \in \mathbb{R}^{k\times |\tilde{M}|}$ is required to have $|\tilde{M}_i|$ nonzero elements of value $|\tilde{M}_i|^{-1}$. Since the number of columns equals the number of positive entries, it is possible to arrange the entries such that each columns will have precisely one positive entry. R then is an $\mathcal{A}$-stochastic matrix for some partition $\mathcal{A}$. The jth row of $Q^{(i)} \in \mathbb{R}^{m\times \prod_s |\tilde{M}_s|}$ is required to have $\tilde{M}_{ij} \prod_{s\neq i} |\tilde{M}_s|$ nonzero elements of value $(\tilde{M}_{ij} \prod_{s\neq i} |M_s|)^{-1}$. Again, the number of positive entries

$$\sum_j \tilde{M}_{ij} \prod_{s\neq i} |\tilde{M}_s| = \prod_s |\tilde{M}_s|$$

is equal to the number of columns and hence, $Q^{(i)}$ is a legal $\mathcal{A}$ stochastic matrix for some $\mathcal{A}$. Note that the number of positive entries, and also columns of $Q^{(i)}$ does not depend on i hence, $\{Q^{(i)}\}$ are of the same size. The exact forms of $\{Q^{(i)}\}$ and R do not matter for our purposes as long as the above restriction and the requirements for $\mathcal{A}$-stochasticity apply (Definition 6).

The usefulness of $\mathfrak{y}_M$ comes from the fact that it transforms rational models M into a constant matrix.

Proposition 3: For $M = \frac{1}{z}\tilde{M} \in \mathbb{Q}_+^{k\times m}$

$$\mathfrak{y}_M(M) = \left(z \prod_s |\tilde{M}_s| \right)^{-1} \mathbf{1}$$

where $\mathbf{1}$ is a matrix of ones of size $|\tilde{M}| \times \prod_s |\tilde{M}_s|$.

Proof: $[M \otimes Q]_i$ is a row vector of size $\prod_s |\tilde{M}_s|$ whose elements are

$$\begin{aligned}[M \otimes Q]_{ij} &= [MQ^{(i)}]_{ij} = \frac{1}{z}\tilde{M}_{ir} \frac{1}{\tilde{M}_{ir} \prod_{s\neq i} |\tilde{M}_s|} \\ &= \left(z \prod_{s\neq i} |\tilde{M}_s| \right)^{-1} \end{aligned}$$

for some r that depends on i, j. Multiplying on the left by R results in

$$\begin{aligned}[R^\top (M \otimes Q)]_{ij} &= R_{ri}[M \otimes Q]_{rj} = \frac{1}{|\tilde{M}_r|} \frac{1}{z \prod_{s\neq r} |\tilde{M}_s|} \\ &= \left(z \prod_s |\tilde{M}_s| \right)^{-1} \end{aligned}$$

for some r that depends on i, j. $\square$

A straightforward calculation using (11) and the definition of $\mathfrak{y}_{M*}$ above shows that the push-forward of $\mathfrak{y}_M$ is

$$\mathfrak{y}_{M*}(\partial_{ab}) = \frac{\sum_{i=1}^{|\tilde{M}_a|} \sum_{j=1}^{\tilde{M}_{ab}} \prod_{l\neq a} |\tilde{M}_l|}{\tilde{M}_{ab} \prod_i |\tilde{M}_i|} \partial'_{\pi(i)\sigma(j)} \tag{17}$$

for some permutations π, σ that depend on M, s, t. Substituting (17) in (12) gives the pull-back

$$(\mathfrak{y}_M^* g)_M(\partial_{ab}, \partial_{cd}) = \frac{\sum_i \sum_s \sum_j \sum_t g_{\mathfrak{y}_M(M)}(\partial_{\pi_1(i)\sigma_1(j)}, \partial_{\pi_2(s)\sigma_2(t)})}{\tilde{M}_{ab}\tilde{M}_{cd}\prod_s |\tilde{M}_s|^2} \quad (18)$$

where the first two summations are over $1, \ldots, |\tilde{M}_a|$ and $1, \ldots, |\tilde{M}_c|$ and the last two summations are over

$$1, \ldots, \tilde{M}_{ab}\prod_{l\neq a}|\tilde{M}_l| \quad \text{and} \quad 1, \ldots, \tilde{M}_{cd}\prod_{l\neq c}|\tilde{M}_l|.$$

B. The Characterization Theorem

Theorems 1 and 2 in this subsection are the main result of the paper.

Theorem 1: Let $\{(\mathbb{R}_+^{k\times m}, g^{(k,m)}) : k \geq 1, m \geq 2\}$ be a sequence of Riemannian manifolds with the property that every congruent embedding by a Markov morphism is an isometry. Then

$$g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = A(|M|) + \delta_{ac}\left(\frac{|M|}{|M_a|}B(|M|) + \delta_{bd}\frac{|M|}{M_{ab}}C(|M|)\right) \quad (19)$$

for some differentiable functions $A, B, C \in C^\infty(\mathbb{R}_+)$.

Proof: The following proof uses the isometry requirement to obtain restrictions on $g_M^{(k,m)}(\partial_{ab}, \partial_{cd})$ first for $a \neq c$, followed by the case of $a = c$, $b \neq d$, and, finally, for the case $a = c$, $b = d$. In each of these cases, we first characterize the metric at constant matrices U and then compute it for rational models M by pulling back the metric at U through $\mathfrak{y}_M$. The value of the metric at nonrational models follows from the rational case by the denseness of $\mathbb{Q}_+^{k\times m}$ in $\mathbb{R}_+^{k\times m}$ and the continuity of the metric.

Part I: $g_M^{(k,m)}(\partial_{ab}, \partial_{cd})$ for $a \neq c$.

We start by computing the metric at constant matrices U. Given $\partial_{a_1b_1}, \partial_{c_1d_1}, a_1 \neq c_1$ and $\partial_{a_2b_2}, \partial_{c_2d_2}, a_2 \neq c_2$, we can use Proposition 2 and (14) to pull back through a corresponding $\mathfrak{h}_\sigma^\Pi$ to obtain

$$\begin{aligned} g_U^{(k,m)}(\partial_{a_1b_1}, \partial_{c_1d_1}) &= g_{\mathfrak{h}_\sigma^\Pi(U)}^{(k,m)}(\partial_{a_2b_2}, \partial_{c_2d_2}) \\ &= g_U^{(k,m)}(\partial_{a_2b_2}, \partial_{c_2d_2}). \end{aligned} \quad (20)$$

Since (20) holds for all a_1, a_2, b_1, b_2 with $a_1 \neq c_1$, $a_2, \neq c_2$ we have that $g_U^{(k,m)}(\partial_{ab}, \partial_{cd}), a \neq c$ depends only on k, m, and $|U|$ and we denote it temporarily by $\hat{A}(k, m, |U|)$.

A key observation, illustrated in Fig. 5, is the fact that pushing forward $\partial_{a,b}$, $\partial_{c,d}$ for $a \neq c$ through any $f \in \mathfrak{F}_{k,m}^{l,n}$ results in two sets of basis vectors whose pairs have disjoint rows. As a result, in the pull-back (12), all the terms in the sum represent metrics between two basis vectors with different rows.

As a result of the above observation, in computing the pull back $g^{(kz,mw)}$ through $\mathfrak{r}_{zw}$ (16) we have a sum of z^2w^2 metrics between vectors of disjoint rows

$$\begin{aligned} \hat{A}(k, m, |U|) &= g_U^{(k,m)}(\partial_{ab}, \partial_{cd}) \\ &= \frac{(zw)^2}{(zw)^2}\hat{A}(kz, mw, |\mathfrak{r}_{zw}(U)|) \\ &= \hat{A}(kz, mw, |U|) \end{aligned} \quad (21)$$

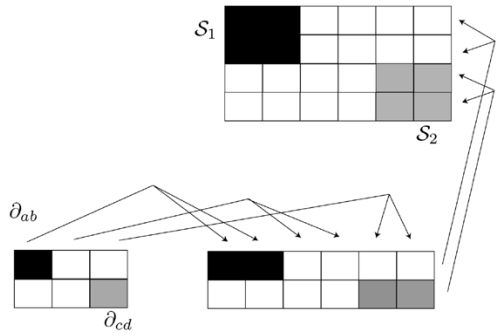

Fig. 5. Pushing forward ∂_{ab}, ∂_{cd} for $a \neq c$ through any $f \in \mathfrak{F}_{k,m}^{l,n}$ results in two sets of basis vectors $\mathcal{S}_1$ (black) and $\mathcal{S}_2$ (gray) for which every pair of vectors $\{(v, u) : v \in \mathcal{S}_1, u \in \mathcal{S}_2\}$ are in disjoint rows.

since $\mathfrak{r}_{zw}(U)$ is a constant matrix with the same norm as U. Equation (21) holds for any $z, w \in \mathbb{N}$ and hence, $g_U^{(k,m)}(\partial_{ab}, \partial_{cd})$ does not depend on k, m and we write

$$g_U^{(k,m)}(\partial_{ab}, \partial_{cd}) = A(|U|), \qquad \text{for some } A \in C^\infty(\mathbb{R}_+).$$

We turn now to computing $g_M^{(k,m)}(\partial_{ab}, \partial_{cd}), a \neq c$ for rational models $M = \frac{1}{z}\tilde{M}$. Pulling back through $\mathfrak{y}_M$ according to (18) we have

$$\begin{aligned} g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) &= \frac{\tilde{M}_{ab}\tilde{M}_{cd}\prod_s |\tilde{M}_s|^2}{\tilde{M}_{ab}\tilde{M}_{cd}\prod_s |\tilde{M}_s|^2}A(|\mathfrak{y}_M(M)|) \\ &= A(|M|). \end{aligned} \quad (22)$$

Again, we made use of the fact that in the pull-back (18) all the terms in the sum are metrics between vectors of different rows.

Finally, since $\mathbb{Q}_+^{k\times m}$ is dense in $\mathbb{R}_+^{k\times m}$ and $g_M^{(k,m)}$ is continuous in M, (22) holds for all models in $\mathbb{R}_+^{k\times m}$.

Part II: $g_M^{(k,m)}(\partial_{ab}, \partial_{cd})$ for $a = c$, $b \neq d$.

As before, we start with constant matrices U. Given ∂_{a_1,b_1}, ∂_{c_1,d_1} with $a_1 = c_1, b_1 \neq d_1$ and $\partial_{a_2,b_2}, \partial_{c_2,d_2}$ with $a_2 = c_2$, $b_2 \neq d_2$ we can pull back through $\mathfrak{h}_\sigma^\Pi$ with

$$\sigma(a_1) = a_2, \pi^{(a_1)}(b_1) = b_2$$

and $\pi^{(a_1)}(d_1) = d_2$ to obtain

$$\begin{aligned} g_U^{(k,m)}(\partial_{a_1b_1}, \partial_{c_1d_1}) &= g_{\mathfrak{h}_\sigma^\Pi(U)}^{(k,m)}(\partial_{a_2b_2}, \partial_{c_2d_2}) \\ &= g_U^{(k,m)}(\partial_{a_2b_2}, \partial_{c_2d_2}). \end{aligned}$$

It follows that $g_U^{(k,m)}(\partial_{ab}, \partial_{ad})$ depends only on k, m, $|U|$ and we temporarily denote

$$g_U^{(k,m)}(\partial_{ab}, \partial_{ad}) = \hat{B}(k, m, |U|).$$

As in Part I, we stop to make an important observation, illustrated in Fig. 6. Assume that f_* pushes forward $\partial_{a,b}$ to a set of vectors $\mathcal{S}_1$ organized in z rows and w_1 columns and $\partial_{a,d}, b \neq d$ to a set of vectors $\mathcal{S}_2$ organized in z rows and w_2 columns. Then, counting the pairs of vectors $\mathcal{S}_1 \times \mathcal{S}_2$, we obtain zw_1w_2 pairs of vectors that have the same rows but different columns and $zw_1(z-1)w_2$ pairs of vectors that have different rows and different columns.

Applying the above observation to the push-forward of $\mathfrak{r}_{k,m}^{kz,mw}$ we have among the set of pairs $\mathcal{S}_1 \times \mathcal{S}_2$, zw^2 pairs of vectors with the same rows but different columns and $zw^2(z-1)$ pairs of vectors with different rows and different columns.

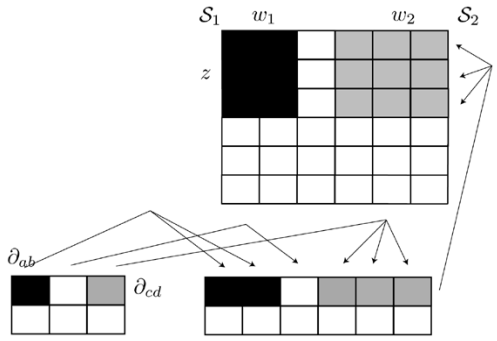

Fig. 6. Let f_* push forward ∂_{ab} to a set of vectors $\mathcal{S}_1$ (black) organized in z rows and w_1 columns and $\partial_{ab}, b \neq d$ to a set of vectors (gray) $\mathcal{S}_2$ organized in z rows and w_2 columns. Then counting the pairs of vectors $\mathcal{S}_1 \times \mathcal{S}_2$ we obtain $z\,w_1\,w_2$ pairs of vectors that have the same rows but different columns and $zw_1(z-1)w_2$ pairs of vectors that have different rows and different columns.

Pulling back through $\mathfrak{r}_{zw}$ according to (16) and the above observation we obtain

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

where the first term corresponds to the zw^2 pairs of vectors with the same rows but different columns and the second term corresponds to the $zw^2(z-1)$ pairs of vectors with different rows and different columns.

Rearranging and dividing by k results in

$$\frac{\hat{B}(k,m,|U|) - A(|U|)}{k} = \frac{\hat{B}(kz,mw,|U|) - A(|U|)}{kz}.$$

It follows that the above quantity is independent of k, m and we write

$$\frac{\hat{B}(k,m,|U|) - A(|U|)}{k} = B(|U|)$$

for some $B \in C^\infty(\mathbb{R}_+)$ which after rearrangement gives us

$$g_U^{(k,m)}(\partial_{ab},\partial_{ad}) = A(|U|) + kB(|U|). \tag{23}$$

We compute next the metric for positive rational matrices $M = \frac{1}{z}\tilde{M}$ by pulling back through $\mathfrak{y}_M$. We use again the observation in Fig. 6, but now with $z = |\tilde{M}_a|$, $w_1 = \tilde{M}_{ab}\prod_{l\neq a}|\tilde{M}_l|$, and $w_2 = \tilde{M}_{ad}\prod_{l\neq a}|\tilde{M}_l|$. Using (18) the pull-back through $\mathfrak{y}_M$ is

$$\begin{aligned}&g_M^{(k,m)}(\partial_{ab},\partial_{ad})\\ &= \frac{|\tilde{M}_a|\tilde{M}_{ab}\prod_{l\neq a}|\tilde{M}_l|(|\tilde{M}_a|-1)\tilde{M}_{ad}\prod_{l\neq a}|\tilde{M}_l|}{\tilde{M}_{ab}\tilde{M}_{ad}\prod_i|\tilde{M}_i|^2}A(|M|)\\ &\quad + \frac{|\tilde{M}_a|\tilde{M}_{ab}\tilde{M}_{ad}\prod_{l\neq a}|\tilde{M}_l|^2}{\tilde{M}_{ab}\tilde{M}_{ad}\prod_i|\tilde{M}_i|^2}\Big(A(|M|)+B(|M|)\sum_i|\tilde{M}_i|\Big)\\ &= \frac{|\tilde{M}_a|-1}{|\tilde{M}_a|}A(|M|) + \frac{1}{|\tilde{M}_a|}\Big(A(|M|)+B(|M|)\sum_i|\tilde{M}_i|\Big)\\ &= A(|M|)+\frac{|\tilde{M}|}{|\tilde{M}_a|}B(|M|) = A(|M|)+\frac{|M|}{|M_a|}B(|M|).\end{aligned} \tag{24}$$

The first term in the sums in (24) corresponds to the $zw_1(z-1)w_2$ pairs of vectors that have different rows and different columns and the second term corresponds to the zw_1w_2 pairs of vectors that have different columns but the same row.

As previously, by denseness of $\mathbb{Q}_+^{k\times m}$ in $\mathbb{R}_+^{k\times m}$ and continuity of $g^{(k,m)}$ (24) holds for all $M \in \mathbb{R}_+^{k\times m}$.

Part III: $g_M^{(k,m)}(\partial_{ab},\partial_{cd})$ for $a = c$, $b = d$.

As before, we start by computing the metric for constant matrices U. Given a_1, b_1, a_2, b_2 we pull back through $\mathfrak{h}_\sigma^\Pi$ with

$$\sigma(a_1) = a_2, \pi^{(a_1)}(b_1) = b_2$$

to obtain

$$g_U^{(k,m)}(\partial_{a_1b_1},\partial_{a_1b_1}) = g_U^{(k,m)}(\partial_{a_2b_2},\partial_{a_2b_2}).$$

It follows that $g_U^{(k,m)}(\partial_{ab},\partial_{ab})$ does not depend on a, b, and we temporarily denote

$$g_U^{(k,m)}(\partial_{ab},\partial_{ab}) = \hat{C}(k,m,|U|).$$

In the present case, pushing forward two identical vectors $\partial_{a,b}, \partial_{a,b}$ by a congruent embedding f results in two identical sets of vectors $\mathcal{S}, \mathcal{S}$ that we assume are organized in z rows and k columns. Counting the pairs in $\mathcal{S}\times\mathcal{S}$ we obtain zw pairs of identical vectors, $zw(w-1)$ pairs of vectors of identical rows but different columns and $zw^2(z-1)$ pairs of vectors of different rows and columns. These three sets of pairs allow us to organize the terms in the pull-back summation (12) into the three cases under considerations.

Pulling back through $\mathfrak{r}_{zw}$ (16) we obtain

$$\begin{aligned}\hat{C}(k,m,|U|) &= \frac{zw\hat{C}(kz,mw,|U|)}{(zw)^2} + \frac{z(z-1)w^2A(|U|)}{(zw)^2}\\ &\quad + \frac{zw(w-1)(A(|U|)+kzB(|U|))}{(zw)^2}\\ &= \frac{\hat{C}(kz,mw,|U|)}{zw} + \Big(1-\frac{1}{zw}\Big)A(|U|)\\ &\quad + \Big(k-\frac{zk}{zw}\Big)B(|U|)\end{aligned}$$

which after rearrangement and dividing by km gives

$$\begin{aligned}&\frac{\hat{C}(k,m,|U|) - A(|U|) - kB(|U|)}{km}\\ &\qquad = \frac{\hat{C}(kz,mw,|U|) - A(|U|) - kzB(|U|)}{kzmw}.\end{aligned} \tag{25}$$

It follows that the left-hand side of (25) equals a function $C(|U|)$ for some $C \in C^\infty(\mathbb{R}_+)$ independent of k and m resulting in

$$g_U^{(k,m)}(\partial_{ab},\partial_{ab}) = A(|U|) + kB(|U|) + kmC(|U|).$$

Finally, we compute $g_M^{(k,m)}(\partial_{ab},\partial_{ab})$ for positive rational matrices $M = \frac{1}{z}\tilde{M}$. Pulling back through $\mathfrak{y}_M$ (18) and using the above division of $\mathcal{S}\times\mathcal{S}$ with $z = \tilde{M}_a$, $w = \tilde{M}_{ab}\prod_{l\neq a}|\tilde{M}_l|$ we obtain

$$\begin{aligned}&g_M^{(k,m)}(\partial_{ab},\partial_{ab})\\ &= \frac{|\tilde{M}_a|-1}{|\tilde{M}_a|}A(|M|)\\ &\quad + \Big(\frac{1}{|\tilde{M}_a|} - \frac{1}{\tilde{M}_{ab}\prod_i|\tilde{M}_i|}\Big)\Big(A(|M|)+B(|M|)\sum_i|\tilde{M}_i|\Big)\end{aligned}$$

$$+\frac{A(|M|)+B(|M|)\sum_i|\tilde{M}_i|+C(|M|)\prod_j|\tilde{M}_j|\sum_i|\tilde{M}_i|}{\tilde{M}_{ab}\prod_i|\tilde{M}_i|}$$
$$=A(|M|)+\frac{|\tilde{M}|}{|\tilde{M}_a|}B(|M|)+\frac{|\tilde{M}|}{\tilde{M}_{ab}}C(|M|)$$
$$=A(|M|)+\frac{|M|}{|M_a|}B(|M|)+\frac{|M|}{M_{ab}}C(|M|). \quad (26)$$

Since the positive rational matrices are dense in $\mathbb{R}_+^{k\times m}$ and the metric $g_M^{(k,m)}$ is continuous in M, (26) holds for all models $M\in\mathbb{R}_+^{k\times m}$. □

The following theorem is the converse of Theorem 1.

Theorem 2: Let $\{(\mathbb{R}_+^{k\times m}, g^{(k,m)})\}$ be a sequence of Riemannian manifolds, with the metrics $g^{(k,m)}$ given by

$$g_M^{(k,m)}(\partial_{ab},\partial_{cd})$$
$$=A(|M|)+\delta_{ac}\left(\frac{|M|}{|M_a|}B(|M|)+\delta_{bd}\frac{|M|}{M_{ab}}C(|M|)\right) \quad (27)$$

for some $A,B,C\in C^\infty(\mathbb{R}_+)$. Then every congruent embedding by a Markov morphism is an isometry.

Proof: To prove the theorem we need to show that

$$\forall\, M\in\mathbb{R}_+^{k\times m},\quad \forall\, f\in\mathfrak{F}_{k,m}^{l,n},\quad \forall\, u,v\in T_M\mathbb{R}_+^{k\times m}$$
$$g_M^{(k,m)}(u,v)=g_{f(M)}^{(l,n)}(f_*u,f_*v). \quad (28)$$

Considering arbitrary $M\in\mathbb{R}_+^{k\times m}$ and $f\in\mathfrak{F}_{k,m}^{l,n}$ we have by (12)

$$g_{f(M)}^{(l,n)}(f_*\partial_{ab},f_*\partial_{cd})$$
$$=\sum_{i=1}^{l}\sum_{j=1}^{n}\sum_{s=1}^{l}\sum_{t=1}^{n}R_{ai}R_{cs}Q_{bj}^{(a)}Q_{dt}^{(c)}g_{f(M)}^{(l,n)}(\partial'_{ij},\partial'_{st}). \quad (29)$$

For $a\neq c$, using the metric form of (27), the right-hand side of (29) reduces to

$$A(|f(M)|)\sum_{i=1}^{l}\sum_{j=1}^{n}\sum_{s=1}^{l}\sum_{t=1}^{n}R_{ai}R_{cs}Q_{bj}^{(a)}Q_{dt}^{(c)}$$
$$=A(|f(M)|)=A(|M|)=g_M^{(k,m)}(\partial_{ab},\partial_{cd})$$

since R and $Q^{(i)}$ are stochastic matrices.

Similarly, for $a=c, b\neq d$, the right hand side of (29) reduces to

$$A(|f(M)|)\sum_{i=1}^{l}\sum_{j=1}^{n}\sum_{s=1}^{l}\sum_{t=1}^{n}R_{ai}R_{cs}Q_{bj}^{(a)}Q_{dt}^{(c)}$$
$$+B(|f(M)|)\sum_i\frac{|f(M)|}{|[f(M)]_i|}R_{ai}^2\sum_j\sum_t Q_{bj}^{(a)}Q_{dt}^{(a)}$$
$$=A(|M|)+B(|M|)|M|\sum_i\frac{R_{ai}^2}{|[f(M)]_i|}. \quad (30)$$

Recall from (10) that

$$[f(M)]_{ij}=\sum_{s=1}^{k}\sum_{t=1}^{m}R_{si}Q_{tj}^{(s)}M_{st}.$$

Summing over j we obtain

$$|[f(M)]_i|=\sum_{s=1}^{k}R_{si}\sum_{t=1}^{m}M_{st}\sum_j Q_{tj}^{(s)}=\sum_{s=1}^{k}R_{si}|M_s|. \quad (31)$$

Since every column of R has precisely one nonzero element, it follows from (31) that R_{ai} is either 0 or $\frac{|[f(M)]_i|}{|M_a|}$ which turns (30) into

$$g_{f(M)}^{(l,n)}(f_*\partial_{ab},f_*\partial_{ad})=A(|M|)+B(|M|)|M|\sum_{i:R_{ai}\neq0}\frac{R_{ai}}{|M_a|}$$
$$=A(|M|)+B(|M|)\frac{|M|}{|M_a|}$$
$$=g_M^{(k,m)}(\partial_{ab},\partial_{ad}).$$

Finally, for the case $a=c, b=d$ the right-hand side of (29) becomes

$$A(|M|)+B(|M|)\frac{|M|}{|M_a|}+C(|M|)|M|\sum_{i=1}^{l}\sum_{j=1}^{n}\frac{(R_{ai}Q_{bj}^{(a)})^2}{[f(M)]_{ij}}.$$

Since in the double sum of (10)

$$[f(M)]_{ij}=\sum_{s=1}^{k}\sum_{t=1}^{m}R_{si}Q_{tj}^{(s)}M_{st}$$

there is a unique positive element, $R_{ai}Q_{bj}^{(a)}$ is either $[f(M)]_{ij}/M_{ab}$ or 0. It follows then that (29) equals

$$g_{f(M)}^{(l,n)}(f_*\partial_{ab},f_*\partial_{ab})=A(|M|)+B(|M|)\frac{|M|}{|M|_a}+C(|M|)|M|$$
$$\times\sum_{i:R_{ai}\neq0}\sum_{j:Q_{bj}^{(a)}\neq0}\frac{R_{ai}Q_{bj}^{(a)}}{M_{ab}}$$
$$=A(|M|)+B(|M|)\frac{|M|}{|M_a|}+C(|M|)\frac{|M|}{M_{ab}}$$
$$=g_M^{(k,m)}(\partial_{ab},\partial_{ab}).$$

We have shown that for arbitrary $M\in\mathbb{R}_+^{k\times m}$ and $f\in\mathfrak{F}_{k,m}^{l,n}$

$$g_M^{(k,m)}(\partial_{ab},\partial_{cd})=g_{f(M)}^{(l,n)}(f_*\partial_{ab},f_*\partial_{cd})$$

for each pair of tangent basis vectors ∂_{ab}, ∂_{cd} and hence the condition in (28) holds, thus proving that

$$f:\left(\mathbb{R}_+^{(k,m)},g^{(k,m)}\right)\to\left(\mathbb{R}_+^{(l,n)},g^{(l,n)}\right)$$

is an isometry. □

C. Normalized Conditional Models

A stronger statement can be made in the case of normalized conditional models. In this case, it turns out that the choices of A and B are immaterial and (19) reduces to the product Fisher information, scaled by a constant that represents the choice of the function C. The following corollary specializes the characterization theorem to the normalized manifolds $\mathbb{P}_{m-1}^k$.

Corollary 1: In the case of the manifold of normalized conditional models, (19) in Theorem 1 reduces to the product Fisher information metric up to a multiplicative constant.

Proof: For $u, v \in T_M \mathbb{P}^k_{m-1}$ expressed in the coordinates of the embedding tangent space $T_M \mathbb{R}^{k\times m}_+$

$$u = \sum_{ij} u_{ij}\partial_{ij} \quad v = \sum_{ij} v_{ij}\partial_{ij}$$

we have

$$\begin{aligned} g^{(k,m)}_M(u,v) &= \left(\sum_{ij} u_{ij}\right)\left(\sum_{ij} v_{ij}\right) A(|M|) \\ &\quad + \sum_i \left(\sum_j u_{ij}\right)\left(\sum_j v_{ij}\right)\frac{|M|}{|M_i|}B(|M|) \\ &\quad + \sum_{ij} u_{ij}v_{ij}\frac{|M|C(|M|)}{M_{ij}} = kC(k)\sum_{ij}\frac{u_{ij}v_{ij}}{M_{ij}} \end{aligned}$$

since $|M| = k$ and for $v \in T_M \mathbb{P}^k_{m-1}$ we have $\sum_j v_{ij} = 0$ for all i. We see that the choice of A and B is immaterial and the resulting metric is precisely the product Fisher information metric up to a multiplicative constant $kC(k)$, that corresponds to the choice of C. □

VI. A Geometric Interpretation of Logistic Regression and AdaBoost

In this section, we use the close relationship between the product Fisher information metric and conditional I-divergence to study the geometry implicitly assumed by logistic regression and AdaBoost.

Logistic regression is a popular technique for conditional inference, usually represented by the following normalized conditional model:

$$p(1|x;\theta) = \frac{1}{Z}e^{\sum_i x_i\theta_i}, \qquad x,\theta \in \mathbb{R}^n,\ \mathcal{Y} = \{-1,1\}$$

where Z is the normalization factor [12]. A more general form [5] that is appropriate for $2 \le |\mathcal{Y}| < \infty$ is

$$p(y|x;\theta) = \frac{1}{Z}e^{\sum_i \theta_i f_i(x,y)}, \qquad x,\theta \in \mathbb{R}^n,\ y \in \mathcal{Y} \tag{32}$$

where $f_i : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ are arbitrary feature functions. The model (32) is a conditional exponential model and the parameters θ are normally obtained by maximum-likelihood estimation for a training set $\{(x_j, y_j)\}_{j=1}^N$

$$\arg\max_\theta \sum_{j=1}^N \sum_i \theta_i f_i(x_j,y_j) - \sum_{j=1}^N \log \sum_{y'\in\mathcal{Y}} e^{\sum_i \theta_i f_i(x_j,y')}. \tag{33}$$

AdaBoost is a linear classifier, usually viewed as an incremental ensemble method that combines weak learners [13]. The incremental rule that AdaBoost uses to select the weight vector θ is known to greedily minimize the exponential loss

$$\arg\min_\theta \sum_j \sum_{y\ne y_j} e^{\sum_i \theta_i(f_i(x_j,y) - f_i(x_j,y_j))} \tag{34}$$

associated with a nonnormalized model

$$p(y|x;\theta) = e^{\sum_i \theta_i f_i(x,y)}, \qquad x,\theta \in \mathbb{R}^n,\ y \in \mathcal{Y}.$$

By moving to the convex primal problems that correspond to maximum likelihood for logistic regression (33) and minimum exponential loss for AdaBoost (34), a close connection between the two algorithms appear [5]. Both problems select a model that minimizes the I-divergence

$$\begin{aligned} &D_r(p,q) \\ &\quad = \sum_x r(x)\sum_y \left(p(y|x)\log\frac{p(y|x)}{q(y|x)} - p(y|x) + q(y|x)\right) \end{aligned} \tag{35}$$

to a uniform distribution q where r is the empirical distribution over the training set $r(x) = \frac{1}{N}\sum_{i=1}^N \delta_{x,x_i}$.

The minimization is constrained by expectation equations with the addition of normalization constraints for logistic regression. The I-divergence above applies to nonnormalized conditional models and reduces to the conditional Kullback–Leibler divergence for normalized models. The conditional form above (35) is a generalization of the nonnormalized divergence for probability measures studied by Csiszár [14].

Assuming $\epsilon = q - p \to 0$, we may approximate $D_r(p,q) = D_r(p, p+\epsilon)$ by a second-order Taylor approximation around $\epsilon = 0$

$$\begin{aligned} D_r(p,q) &\approx D_r(p,p) + \sum_{xy} \frac{\partial D(p,p+\epsilon)}{\partial\epsilon(y,x)}\Big|_{\epsilon=0}\epsilon(y,x) \\ &\quad + \frac{1}{2}\sum_{x_1y_1}\sum_{x_2y_2}\frac{\partial^2 D(p,p+\epsilon)}{\partial\epsilon(y_1,x_1)\partial\epsilon(y_2,x_2)}\Big|_{\epsilon=0}\epsilon(y_1,x_1)\epsilon(y_2,x_2). \end{aligned}$$

The first-order terms

$$\frac{\partial D_r(p,p+\epsilon)}{\partial\epsilon(y_1,x_1)} = r(x_1)\left(1 - \frac{p(y_1|x_1)}{p(y_1|x_1)+\epsilon(y_1,x_1)}\right) \tag{36}$$

zero out for $\epsilon = 0$. The second-order terms

$$\frac{\partial^2 D_r(p,p+\epsilon)}{\partial\epsilon(y_1,x_1)\partial\epsilon(y_2,x_2)} = \frac{\delta_{y_1y_2}\delta_{x_1x_2}r(x_1)p(y_1|x_1)}{(p(y_1|x_1)+\epsilon(y_1,x_1))^2}$$

at $\epsilon = 0$ are $\delta_{y_1y_2}\delta_{x_1x_2}\frac{r(x_1)}{p(y_1|x_1)}$. Substituting these expressions in the Taylor approximation gives

$$D_r(p,p+\epsilon) \approx \frac{1}{2}\sum_{xy}\frac{r(x)\epsilon^2(y,x)}{p(y|x)} = \frac{1}{2}\sum_{xy}\frac{(r(x)\epsilon(y,x))^2}{r(x)p(y|x)}$$

which is the squared length of $r(x)\epsilon(y,x) \in T_{r(x)p(y|x)}\mathbb{R}^{k\times m}_+$ under the metric (19) for the choices $A(|M|) = B(|M|) = 0$ and $C(|M|) = 1/(2|M|)$.

The I-divergence $D_r(p,q)$ which both logistic regression and AdaBoost minimize is then approximately the squared geodesic distance between the conditional models $r(x)p(y|x)$ and $r(x)q(y|x)$ under a metric (19) with the above choices of A, B, C. The fact that the models $r(x)p(y|x)$ and $r(x)q(y|x)$ are not strictly positive is not problematic, since by the continuity of the metric, Theorems 1 and 2 pertaining to $\mathbb{R}^{k\times m}_+$ apply also to its closure $\overline{\mathbb{R}^{k\times m}_+}$—the set of all nonnegative conditional models.

The preceding result is not restricted to logistic regression and AdaBoost. It carries over to any conditional modeling technique that is based on maximum entropy or minimum Kullback–Leibler divergence.

VII. Discussion

We formulated and proved an axiomatic characterization of a family of metrics, the simplest of which is the product Fisher information metric in the conditional setting for both normalized

and nonnormalized models. This result is a strict generalization of Campbell's and Čencov's theorems. For the case $k = 1$, Theorems 1 and 2 reduce to Campbell's theorem [4] and Corollary 1 reduces to Čencov's theorem ([3, Lemma 11.3]).

In contrast to Čencov's and Campbell's theorems, we do not make any reference to a joint distribution and our analysis is strictly discriminative. If one is willing to consider a joint distribution it may be possible to derive a geometry on the space of conditional models $p(y|x)$ from Campbell's geometry on the space of joint models $p(x, y)$. Such a derivation may be based on the observation that the conditional manifold is a quotient manifold of the joint manifold. If such a derivation is carried over, it is likely that the derived metric would be different from the metric characterized in this paper.

As mentioned in Section III, the proper framework for considering nonnegative models is a manifold with corners [7]. The theorem stated here carries over by the continuity of the metric from the manifold of positive models to its closure. Extension to infinite $\mathcal{X}$ or $\mathcal{Y}$ poses considerable difficulty. For a brief discussion of infinite dimensional manifolds representing densities see [1, pp. 44-45].

The characterized metric (19) has three additive components. The first one represents a component that is independent of the tangent vectors, but depends on the norm of the model at which it is evaluated. Such a dependency may be used to produce the effect of giving higher importance to large models, that represent more confidence. The second term is nonzero if the two tangent vectors represent increases in the current model along $p(\cdot|x_a)$. In this case, the term depends not only on the norm of the model but also on $|M_a| = \sum_j p(y_j|x_a)$. This may be useful in dealing with nonnormalized conditional models whose values along the different rows $p(\cdot|x_i)$ are not on the same numeric scale. Such scale variance may represent different importance in the predictions made, when conditioning on different x_i. The last component represents the essence of the Fisher information quantity. It scales up with low values $p(y_j|x_i)$ to represent a kind of space stretching, or distance enhancing when we are dealing with points close to the boundary. It captures a similar effect as the log likelihood of increased importance given to near-zero erroneous predictions.

Using the characterization theorem, we give for the first time a differential geometric interpretation of logistic regression and AdaBoost whose metric is characterized by natural invariance properties. Such a geometry applies not only to the above models, but to any algorithmic technique that is based on maximum conditional entropy principles.

Despite the relationship between the I-divergence $D_r(p, q)$ and the geodesic distance $d(pr, qr)$, there are some important differences. The geodesic distance not only enjoys the symmetry and triangle inequality properties, but is also bounded. In contrast, the I-divergence grows to infinity—a fact that causes it to be extremely nonrobust. Indeed, in the statistical literature, the maximum-likelihood estimator is often replaced by more robust estimators, among them the minimum Hellinger distance estimator [15], [16]. Interestingly, the Hellinger distance is extremely similar to the geodesic distance under the Fisher information metric. It is likely that new techniques in conditional inference that are based on minimum geodesic distance in the primal space, will perform better than maximum entropy or conditional exponential models.

Another interesting aspect is that maximum entropy or conditional exponential models may be interpreted as transforming models p into rp where r is the empirical distribution of the training set. This makes sense since two models rp, rq become identical over x_i that do not appear in the training set, and indeed the lack of reference data makes such an embedding workable. It is conceivable, however, to consider embeddings $p \mapsto rp$ using distributions r different from the empirical training data distribution. Different x_i may have different importance associated with their prediction $p(\cdot|x_i)$ and some labels y_i may be known to be corrupted by noise with a distribution that depends on i.

ACKNOWLEDGMENT

The author wishes to thank John Lafferty for helpful discussions.

REFERENCES

[1] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Providence, RI: Amer. Math. Soc., 2000.
[2] R. E. Kass and P. W. Voss, *Geometrical Foudnation of Asymptotic Inference*. New York: Wiley, 1997.
[3] N. N. Čencov, *Statistical Decision Rules and Optimal Inference*. Providence, RI: Amer. Math. Soc., 1982.
[4] L. L. Campbell, "An extended Čencov characterization of the information metric," *Proc. Amer. Math. Soc.*, vol. 98, no. 1, pp. 135–141, 1986.
[5] G. Lebanon and J. Lafferty, "Boosting and maximum likelihood for exponential models," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2001, vol. 14.
[6] M. Spivak, *A Comprehensive Introduction to Differential Geometry*. Houston, TX: Publish or Perish, 1975, vol. 1–5.
[7] J. M. Lee, *Introduction to Smooth Manifolds*. New York: Springer-Verlag, 2002.
[8] ——, *Introduction to Topological Manifolds*. New York: Springer-Verlag, 2000.
[9] J. Lafferty, "The density manifold and configuration space quantization," *Trans. Amer. Math. Soc.*, vol. 305, no. 2, pp. 699–741, 1988.
[10] S. Lang, *Fundamentals of Differential Geometry*. New York: Springer-Verlag, 1999.
[11] A. I. Khinchin, *Mathematical Foundations of Information Theory*. New York: Dover, 1957.
[12] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 1989.
[13] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Proc. MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, CA, 2002.
[14] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, vol. 19, no. 4, pp. 2032–2066, 1991.
[15] R. Beran, "Minimum hellinger distance estimates for parameteric models," *Ann. Statist.*, vol. 5, no. 3, pp. 445–463, 1977.
[16] B. G. Lindsay, "Efficiency versus robustness: The case for minimum hellinger distance and related methods," *Ann. Statist.*, vol. 22, no. 2, pp. 1081–1114, 1994.