Metric Learning for Text Documents

Guy Lebanon

Abstract—Many algorithms in machine learning rely on being given a good distance metric over the input space. Rather than using a default metric such as the Euclidean metric, it is desirable to obtain a metric based on the provided data. We consider the problem of learning a Riemannian metric associated with a given differentiable manifold and a set of points. Our approach to the problem involves choosing a metric from a parametric family that is based on maximizing the inverse volume of a given data set of points. From a statistical perspective, it is related to maximum likelihood under a model that assigns probabilities inversely proportional to the Riemannian volume element. We discuss in detail learning a metric on the multinomial simplex where the metric candidates are pull-back metrics of the Fisher information under a Lie group of transformations. When applied to text document classification the resulting geodesic distance resemble, but outperform, the tfidf cosine similarity measure.

Index Terms—Distance learning, text analysis, machine learning.

1 Introduction

Machine learning algorithms often require an embedding of data points into some space. Algorithms such as k-nearest neighbors and neural networks assume the embedding space to be $\mathbb{R}^n$, while SVM and other kernel methods embed the data in a Hilbert space through a kernel operation. Whatever the embedding space is, the notion of metric structure has to be carefully considered. For high-dimensional structured data such as text documents or images, it is hard to devise an appropriate metric by hand. This has led, in many cases, to the use of default metrics such as the pixel-wise Euclidean distance in images and the cosine similarity term frequency distance in text documents. These assumptions of default metrics is often used without justification by data or modeling arguments. We argue that, in the absence of direct evidence of Euclidean geometry, the metric structure should be inferred from the available data. The obtained metric may be useful in learning tasks such as classification and clustering through algorithms such as nearest neighbor and k-means. The learned metric d may also be useful for statistical modeling of the data through custom probability distribution such as $p(x) = Z^{-1} \exp(-d^2(x, \mu)/2\sigma^2)$.

Several attempts have recently been made to learn the metric structure of the embedding space from a given data set. Saul and Jordan [12] use geometrical arguments to learn optimal paths connecting two points in a space. Xing et al. [13] learn a global metric structure that is able to capture non-Euclidean geometry. The learned metric is global and not local as the resulting distances are invariant to translation of the data points. While an invariant metric may be desirable, in some cases, it is often not natural for compact or bounded manifolds. Lanckriet et al. [6] learn a kernel matrix that represents similarities between all pairs of the supplied data points. While such an approach does learn the kernel structure from data, the resulting Gram matrix does not generalize to unseen points.

Learning a Riemannian metric is also related to finding a lower dimensional representation of a data set. Work in this area includes linear methods such as principal component analysis and nonlinear methods such as spherical subfamily models [2], locally linear embedding [11], and curved multinomial subfamilies [3]. Once such a submanifold is found, distances $d(x, y)$ may be computed as the lengths of shortest paths on the submanifold connecting x and y. As shown in Section 3, this approach is a limiting case of learning a Riemannian metric for the high-dimensional embedding space.

Lower dimensional representations are useful for visualizing high-dimensional data. However, these methods assume strict conditions that are often violated in real-world, high-dimensional data. The obtained submanifold is tuned to the training data and new data points will likely lie outside the submanifold due to noise. It is necessary to specify some way of projecting the off-manifold points into the manifold. There is no notion of non-Euclidean geometry outside the submanifold and if the estimated submanifold does not fit current and future data perfectly, Euclidean projections are usually used.

Another source of difficulty is estimating the dimension of the submanifold. The dimension of the submanifold is notoriously hard to estimate for high-dimensional sparse data sets. Moreover, the data may have different lower dimensions in different locations or may lie on several disconnected submanifolds, thus violating the assumptions underlying the submanifold approach.

We propose an alternative approach to the metric learning problem. The obtained metric is local, thus capturing local variations within the space and is defined on the entire embedding space. A set of metric candidates is represented as a parametric family of transformations or, equivalently, as a parametric family of statistical models and the obtained metric is chosen from it based on some performance criterion. We examine the application of the metric learning techniques in the context of classification of text documents and images and provide experimental results for text classification.

In Section 3, we discuss our formulation of the Riemannian metric problem. Section 4 describes the set of

• *The author is with the Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907. E-mail: lebanon@stat.purdue.edu.*

Manuscript received 27 Jan. 2005; revised 15 Aug. 2005; accepted 18 Aug. 2005; published online 14 Feb. 2006.
Recommended for acceptance by A. Srivastava.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0059-0105.

metric candidates as pull-back metrics of a group of transformations followed by a discussion of the resulting generative model in Section 6. In Section 7, we apply the framework to text classification and report experimental results on the WebKB data. The appendix contains a review of relevant concepts from Riemannian geometry.

2 The Fisher Geometry

In this section, we describe some well-known results concerning the Fisher geometry of a space of probability distributions. The reader may want to consult the appendix at this point for a review of relevant concepts from Riemannian geometry. For more details on the Fisher geometry, refer to the monographs [5], [1].

Parametric inference in statistics is concerned with a parametric family of distributions $\{p(x;\theta) : \theta \in \Theta \subset \mathbb{R}^n\}$ over the event space $\mathcal{X}$. If the parameter space Θ is a differentiable manifold and the mapping $\theta \mapsto p(x;\theta)$ is a diffeomorphism, we can identify statistical models in the family as points on the manifold Θ. In this paper, we will mostly be concerned with the manifold of multinomial models[1]

$$\mathbb{P}_n = \left\{ \theta \in \mathbb{R}^{n+1} : \ \forall i \ \theta_i > 0, \ \sum_i \theta_i = 1 \right\}.$$

The manifold $\mathbb{P}_n$ is described as a subset of $\mathbb{R}^{n+1}$ despite the fact that it is an n-dimensional manifold. This notation leads to substantially simpler expressions later on. Notice that the above manifold contains a parameter vector θ of the multinomial distribution—that also happens to be a probability vector by itself. The simplex $\mathbb{P}$ is a submanifold of $\mathbb{R}^{n+1}$ and, as such, we can write its tangent vectors in the standard base of $\mathbb{R}^{n+1}$. Using this expression for tangent vectors, it is easy to identify the tangent space of the simplex as

$$T_\theta \mathbb{P}_n = \left\{ v \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} v_i = 0 \right\}.$$

Note that the above representation of $T_\theta \mathbb{P}$ does not depend on θ and is not unique. Since the tangent space $T_\theta \mathbb{P}_n$ is an n-dimensional vector space, we can express tangent vectors as vectors in $\mathbb{R}^{n+1}$ in many ways, each corresponding to a specific choice of a base.

The Fisher information matrix $E\{ss^\top\}$, where s is the gradient of the log-likelihood: $[s]_i = \partial \log p(x;\theta)/\partial\theta_i$, may be used to endow Θ with the following Riemannian metric

$$\begin{aligned} \mathcal{J}_\theta(u,v) &\stackrel{\text{def}}{=} \sum_{i,j} u_i v_j \int p(x;\theta) \frac{\partial}{\partial\theta_i} \log p(x;\theta) \frac{\partial}{\partial\theta_j} \log p(x;\theta) \mathrm{d}x \\ &= \sum_{i,j} u_i v_j \ E\left\{ \frac{\partial \log p(x;\theta)}{\partial\theta_i} \frac{\partial \log p(x;\theta)}{\partial\theta_j} \right\}, \end{aligned} \tag{1}$$

where the above integral is replaced with a sum if $\mathcal{X}$ is discrete. Note that, in this paper, we adopt the terminology of differential geometry: A symmetric, positive definite bilinear form (local inner product) is referred to as the metric, rather than the distance function $d(\cdot,\cdot)$. Consult Appendix A for further details.

Another important manifold that will appear in this paper is the positive sphere

$$\mathbb{S}_+^n = \left\{ \theta \in \mathbb{R}^{n+1} : \ \forall i \ \theta_i > 0, \ \sum_i \theta_i^2 = 1 \right\}.$$

Tangent vectors to the positive sphere, much like the simplex, may be written in the standard basis of $\mathbb{R}^{n+1}$ leading to the following identification of the tangent space

$$T_\theta \mathbb{S}_+^n = \left\{ v \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} v_i \theta_i = 0 \right\}.$$

Using the above expression for tangent vectors, the metric δ on $\mathbb{S}_+^n$ defined as $\delta_\theta(u,v) \stackrel{\text{def}}{=} \sum_{i=1}^{n+1} u_i v_i$ has the same functional form as the standard Euclidean inner product. Since this inner product characterizes Euclidean geometry, the local geometry of $(\mathbb{S}_+^n, \delta)$ is the Euclidean geometry, restricted to the sphere.

Fortunately, distances $d_{\mathcal{J}}(\theta,\eta)$ (see (13) for the definition of $d_{\mathcal{J}}$) on $(\mathbb{P}, \mathcal{J})$ have a closed form expression. The expression is obtained by noticing that

$$f : \mathbb{P} \to \mathbb{S}_+^n \quad f(\theta) = \left(\sqrt{\theta_1}, \ldots, \sqrt{\theta_{n+1}} \right)$$

is an isometry between $(\mathbb{P}, \mathcal{J})$, and $(\mathbb{S}_+^n, \delta)$, and noticing that $d_\delta(\theta,\eta)$ is given by the length of the great circle connecting the two points $d_\delta(\eta,\theta) = \arccos(\sum \eta_i \theta_i)$. It then follows that

$$d_{\mathcal{J}}(\theta,\eta) = d_\delta(f(\theta), f(\eta)) = \arccos\left(\sum_{i=1}^{n+1} \sqrt{\theta_i \eta_i} \right).$$

See Appendix A.3 for a definition of isometry in differential geometry. It is well-known that the transformation $f : \mathbb{P} \to \mathbb{S}_+^n$ is an isometry. A proof may be found at Section 4.1 of [7].

3 The Metric Learning Problem

The metric learning problem may be formulated as follows: Given a differentiable manifold $\mathcal{M}$ and a data set $D = \{x_1, \ldots, x_N\} \subset \mathcal{M}$, select a Riemannian metric g from a set of metric candidates $\mathcal{G}$. As in statistical inference, $\mathcal{G}$ may be a parametric family $\mathcal{G} = \{g^\lambda : \lambda \in \Lambda \subset \mathbb{R}^k\}$ or as in nonparametric statistics a less constrained set of candidates. We focus on the parametric approach, as we believe it to generally perform better for high-dimensional sparse data such as text documents. We use a superscript for the parameter g^λ since the subscript of the metric is reserved for its value at a particular point of the manifold (see Appendix A.3).

Let $\{e_i\}_i$ represent a basis of the tangent space $T_x\mathcal{M}$. The volume element of g at x is defined as $\mathrm{dvol}\ g(x) \stackrel{\text{def}}{=} \sqrt{\det G(x)}$, where $G(x)$ is the matrix whose entries are $[G(x)]_{ij} = g_x(e_i, e_j)$. Note that $\det G(x) > 0$ since $G(x)$ is positive definite. Intuitively, the volume element $\mathrm{dvol}\ g(x)$ summarizes the "size" of the metric g at x in one scalar (it is originally a bilinear form or a matrix). Similarly, the inverse volume element measures the "smallness" of the metric at x. Paths crossing areas with high inverse volume will tend to be shorter than paths over an area with high inverse volume.

1. The parameters θ are required to be positive for the simplex to be a manifold, rather than a manifold with corners. This is a technical issue and does not influence possible applications.

The size of the metric at a data set $D = \{x_1, \ldots, x_N\}$ may be measured as the product of the inverse volume elements at the points x_i. One problem is that the above quantity is unbounded. This can be demonstrated using basic properties of determinants $\mathrm{dvol}\,(cg(x)) = c^{n/2}\mathrm{dvol}\,(g(x))$. A simple solution is to enforce the total volume to be constant through normalizing. We therefore propose to choose the metric based on the following objective function

$$\mathcal{O}(g, D) = \prod_{i=1}^{N} \frac{(\mathrm{dvol}\; g(x_i))^{-1}}{\int_{\mathcal{M}} (\mathrm{dvol}\; g(x))^{-1}\, \mathrm{d}x}. \tag{2}$$

Maximizing the inverse volume in (2) will result in shorter curves across densely populated regions of $\mathcal{M}$. As a result, the geodesics will tend to pass through densely populated regions. This agrees with the intuition that distances between data points should be measured on the lower dimensional data submanifold, thus capturing the intrinsic geometrical structure of the data. Note that the normalized inverse volume element may be seen as a probability distribution over the manifold and maximizing $\mathcal{O}(g, D)$ may be considered as a maximum-likelihood problem. The normalization in $\mathcal{O}$ is necessary for the same reason it is necessary in probabilities. We are not interested in the total mass but in local variations of it.

If $\mathcal{G}$ is completely unconstrained, the metric maximizing the above criterion will have a volume element tending to 0 at the data points and $+\infty$ everywhere else. Such a solution is analogous to estimating a distribution by an impulse train at the data points and 0 elsewhere (the empirical distribution). As in statistics, we avoid this degenerate solution by restricting the set of candidates $\mathcal{G}$ to a constrained set of smooth functions.

The case of extracting a low-dimensional submanifold (or linear subspace) may be recovered from the above framework if $g \in \mathcal{G}$ is equal to the metric inherited from the embedding Euclidean space across a submanifold and tending to $+\infty$ outside. In this case, distances between two points on the submanifold will be measured as the shortest curve on the submanifold using the Euclidean length element.

If $\mathcal{G}$ is a parametric family of metrics $\mathcal{G} = \{g^\lambda : \lambda \in \Lambda\}$, the log of the objective function $\mathcal{O}$ is equivalent to the log likelihood of the data $\ell(\lambda)$ under the model

$$p(x; \lambda) = \frac{1}{Z}(\mathrm{dvol}\; g(x))^{-1}.$$

As a side note, if $g = \mathcal{J}$ the above model is the inverse of Jeffreys' prior $p(x) \propto \mathrm{dvol}\; \mathcal{J}(x)$ a widely studied distribution in Bayesian statistics. However, in the case of Jeffreys' prior, the metric is known in advance and there is no need for parameter estimation. For prior work on connecting volume elements and densities on manifolds, refer to [10].

Specifying the family of metrics $\mathcal{G}$ is not an intuitive task. Metrics are specified in terms of a local inner product and it may be difficult to understand the implications of a specific choice on the resulting distances. Instead of specifying a parametric family of metrics as discussed in the previous section, we specify a parametric family of transformations $\{F_\lambda : \lambda \in \Lambda\}$. The resulting set of metric candidates will be the pull-back metrics $\mathcal{G} = \{F_\lambda^* \mathcal{J} : \lambda \in \Lambda\}$ of the Fisher information metric $\mathcal{J}$ (See Appendix A.3 for the definition of the pull-back metric $F^* g$ with respect to a transformation F and a metric g). Since the metrics are pull-back metrics of the Fisher information for the multinomial distribution, a closed form expression for the distance $d_{F_\lambda^* \mathcal{J}}(x, y)$ is readily available (see Appendix A.3).

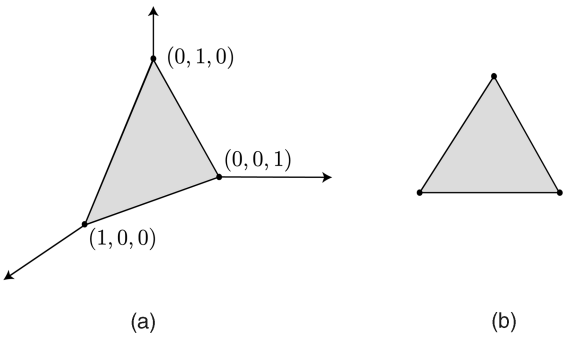


Fig. 1. The 2-simplex $\mathbb{P}_2$ may be visualized as (a) a surface in $\mathbb{R}^3$ or (b) as a triangle in $\mathbb{R}^2$.

Denoting the metric inherited from the embedding Euclidean space by δ, we define f to be a flattening transformation if $f : (\mathcal{M}, g) \to (\mathcal{N}, \delta)$ is an isometry. In this case, distances on the manifold $(\mathcal{M}, g) = (\mathcal{M}, f^*\delta)$ may be measured as the shortest Euclidean path on the manifold $\mathcal{N}$ between the transformed points. Such a computation is often simpler than the original distance computation for an arbitrary metric. A flattening transformation f, thus takes a locally distorted space and converts it into a subset of $\mathbb{R}^n$ equipped with the local Euclidean metric $\delta(u, v) = \sum_i u_i v_i$.

In the next sections, we work out in detail an implementation of the above framework in which the manifold $\mathcal{M}$ is the multinomial simplex $\mathbb{P}_n$.

4 A Parametric Class of Metrics

Consider the following family of diffeomorphisms $F_\lambda : \mathbb{P}_n \to \mathbb{P}_n$

$$F_\lambda(x) \stackrel{\mathrm{def}}{=} \left(\frac{x_1 \lambda_1}{\langle x, \lambda \rangle}, \ldots, \frac{x_{n+1} \lambda_{n+1}}{\langle x, \lambda \rangle} \right), \quad \lambda \in \mathbb{P}_n.$$

The family F_λ is a Lie group of transformations under composition whose parametric space is $\Lambda = \mathbb{P}$. The identity element is $(\frac{1}{n+1}, \ldots, \frac{1}{n+1})$ and the inverse of F_λ is $(F_\lambda)^{-1} = F_\eta$, where

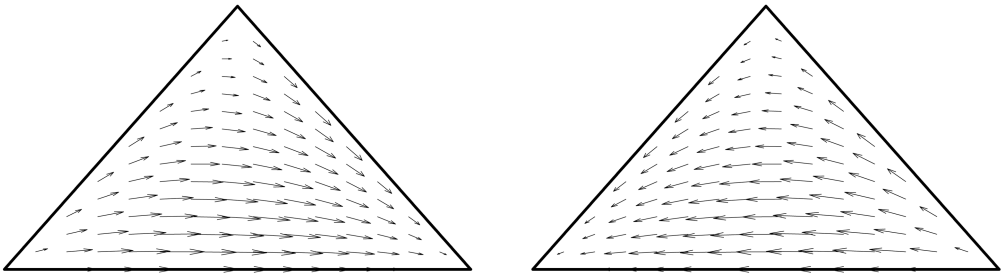
$$\eta_i = \frac{1/\lambda_i}{\sum_k 1/\lambda_k}.$$

The above transformation group acts on $x \in \mathbb{P}_n$ by increasing the components of x with high λ_i values while remaining in the simplex. Fig. 1 illustrates how to visualize $\mathbb{P}_2$ in two dimensions and Fig. 2 illustrates the above action in $\mathbb{P}_2$.

We will consider the pull-back metrics of the Fisher information $\mathcal{J}$ through the above transformation group as our parametric family of metrics $\mathcal{G} = \{F_\lambda^* \mathcal{J} : \lambda \in \mathbb{P}_n\}$. Note that since the Fisher information itself is a pullback metric from the sphere under the square root transformation, we have that $F_\lambda^* \mathcal{J}$ is also the pull-back metric of $(\mathbb{S}_+^n, \delta)$ through the transformation

$$\hat{F}_\lambda(x) \stackrel{\mathrm{def}}{=} \left(\sqrt{\frac{x_1 \lambda_1}{\langle x, \lambda \rangle}}, \ldots, \sqrt{\frac{x_{n+1} \lambda_{n+1}}{\langle x, \lambda \rangle}} \right), \quad \lambda \in \mathbb{P}_n.$$

As a result of the above observation we have the following closed form for the geodesic distance under $F_\lambda^* \mathcal{J}$



(a) (b)

Fig. 2. F_λ acting on $\mathbb{P}_2$ for (a) $\lambda = (\frac{2}{10}, \frac{5}{10}, \frac{3}{10})$ and (b) F_λ^{-1} acting on $\mathbb{P}_2$. The arrows indicate the mapping that transforms x to (a) $F_\lambda(x)$ or (b) $F_\lambda^{-1}(x)$.

$$d_{F_\lambda^*\mathcal{J}}(x,y) = \operatorname{acos}\left(\sum_{i=1}^{n+1}\sqrt{\frac{x_i\lambda_i}{\langle x,\lambda\rangle}\frac{y_i\lambda_i}{\langle y,\lambda\rangle}}\right) = \operatorname{acos}\left(\sum_{i=1}^{n+1}\lambda_i\frac{\sqrt{x_iy_i}}{\sqrt{\langle x,\lambda\rangle\langle y,\lambda\rangle}}\right). \tag{3}$$

The above distance is surprisingly similar to the tfidf cosine similarity measure [4]. The differences are the square root, the normalization and the choice of non-idf λ parameters in (3).

5 Computing the Volume Element of $F_\lambda^*\mathcal{J}$

To apply the framework described in Section 3 to the metric $F_\lambda^*\mathcal{J}$, we need to compute the volume element given by the square root of the determinant of the Gram matrix of $F_\lambda^*\mathcal{J}$. This is done in several stages. First, the Gram matrix G is computed, then some useful lemmas concerning matrix determinants are proven and, finally, we compute $\det G$.

5.1 Computing the Gram Matrix G

We start by computing the Gram matrix $[G]_{ij} = F_\lambda^*\mathcal{J}(\partial_i, \partial_j)$, where $\{\partial_i\}_{i=1}^n$ is a basis for $T_\theta\mathbb{P}_n$ given by the rows of the matrix

$$U = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & 0 & \ddots & 0 & -1 \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \in \mathbb{R}^{n\times n+1} \tag{4}$$

and then proceed by computing $\det G$ in Propositions 1 and 2 below. Note that since the determinant is invariant under change of basis, we are free to select the convenient base expressed by the rows of (4).

Proposition 1. *The matrix $[G]_{ij} = F_\lambda^*\mathcal{J}(\partial_i, \partial_j)$ is given by*

$$G = JJ^\top = U(D - \lambda\alpha^\top)(D - \lambda\alpha^\top)^\top U^\top, \tag{5}$$

where $D \in \mathbb{R}^{n+1\times n+1}$ is a diagonal matrix whose entries are

$$[D]_{ii} = \sqrt{\frac{\lambda_i}{x_i}}\frac{1}{2\sqrt{\langle x,\lambda\rangle}}$$

and α is a column vector given by

$$[\alpha]_i = \sqrt{\frac{\lambda_i}{x_i}}\frac{x_i}{2\langle x,\lambda\rangle^{3/2}}.$$

Note that all vectors are treated as column vectors and for $\lambda, \alpha \in \mathbb{R}^{n+1}$, $\lambda\alpha^\top \in \mathbb{R}^{n+1\times n+1}$ is the outer product matrix $[\lambda\alpha^\top]_{ij} = \lambda_i\alpha_j$.

Proof. The jth component of the vector $\hat{F}_{\lambda *}v$ is

$$[\hat{F}_{\lambda *}v]_j = \frac{\mathrm{d}}{\mathrm{d}t}\sqrt{\frac{(x_j + tv_j)\lambda_j}{\langle x + tv,\lambda\rangle}}\bigg|_{t=0} = \frac{1}{2}\frac{v_j\lambda_j}{\sqrt{x_j\lambda_j}\sqrt{\langle x,\lambda\rangle}} - \frac{1}{2}\frac{\langle v,\lambda\rangle\sqrt{x_j\lambda_j}}{\langle x,\lambda\rangle^{3/2}}.$$

Taking the rows of U to be the basis $\{\partial_i\}_{i=1}^n$ for $T_x\mathbb{P}_n$ we have, for $i = 1,\ldots,n$ and $j = 1,\ldots,n+1$,

$$[\hat{F}_{\lambda *}\partial_i]_j = \frac{\lambda_j[\partial_i]_j}{2\sqrt{x_j\lambda_j}\sqrt{\langle x,\lambda\rangle}} - \frac{\sqrt{x_j\lambda_j}}{2\langle x,\lambda\rangle^{3/2}}\partial_i,\lambda = \frac{\delta_{j,i} - \delta_{j,n+1}}{2\sqrt{\langle x,\lambda\rangle}}\sqrt{\frac{\lambda_j}{x_j}} - \frac{\lambda_i - \lambda_{n+1}}{2\langle x,\lambda\rangle^{3/2}}\sqrt{\frac{\lambda_j}{x_j}}x_j.$$

If we define $J \in \mathbb{R}^{n\times n+1}$ to be the matrix whose rows are $\{\hat{F}_*\partial_i\}_{i=1}^n$, we have $J = U(D - \lambda\alpha^\top)$.

Since the metric $F_\lambda^*\mathcal{J}$ is the pullback of δ through $\hat{F}_\lambda$, we have $[G]_{ij} = \langle\hat{F}_{\lambda *}\partial_i\hat{F}_{\lambda *}\partial_j\rangle$ and $G = JJ^\top = U(D - \lambda\alpha^\top)(D - \lambda\alpha^\top)^\top U^\top$. □

Before we turn to computing the determinant of the matrix G above, we prove Lemmas 1 and 2 below that will prove to be useful in computing $\det G$.

5.2 Some Useful Lemmas Concerning Matrix Determinants

The determinant of a matrix $\det A \in \mathbb{R}^{n\times n}$ may be seen as a function of the rows of A, $\{A_i\}_{i=1}^n$

$$f : \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R} \qquad f(A_1,\ldots,A_n) = \det A.$$

The multilinearity property of the determinant means that the function f above is linear in each of its components

$$\forall j = 1,\ldots,n \quad f(A_1,\ldots,A_{j-1},A_j + B_j,A_{j+1},\ldots,A_n) = f(A_1,\ldots,A_{j-1},A_j,A_{j+1},\ldots,A_n) + f(A_1,\ldots,A_{j-1},B_j,A_{j+1},\ldots,A_n).$$

Lemma 1. *Let $D \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $D_{11} = 0$ and $\mathbf{1}$ a matrix of ones. Then,*

$$det(D - \mathbf{1}) = -\prod_{i=2}^m D_{ii}.$$

Proof. Subtract the first row from all the other rows to obtain

$$\begin{pmatrix} -1 & -1 & \cdots & -1 \\ 0 & D_{22} & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & D_{mm} \end{pmatrix}.$$

Now, compute the determinant by the cofactor expansion along the first column to obtain

$$\det(D - \mathbf{1}) = (-1)\prod_{j=2}^{m} D_{jj} + 0 + 0 + \cdots + 0.$$

□

Lemma 2. *Let $D \in \mathbb{R}^{n\times n}$ be a diagonal matrix and $\mathbf{1}$ a matrix of ones. Then,*

$$det(D - \mathbf{1}) = \prod_{i=1}^{m} D_{ii} - \sum_{i=1}^{m}\prod_{j\neq i} D_{jj}.$$

Proof. Using the multilinearity property of the determinant, we separate the first row of $D - \mathbf{1}$ as $(D_{11}, 0, \ldots, 0) + (-1, \ldots, -1)$. The determinant $\det D - \mathbf{1}$ then becomes $\det A + \det B$, where A is $D - \mathbf{1}$ with the first row replaced by $(D_{11}, 0, \ldots, 0)$ and B is the $D - \mathbf{1}$ with the first row replaced by a vector of -1.

Using Lemma 1, we have $\det B = -\prod_{j=2}^{n} D_{jj}$. The determinant $\det A$ may be expanded along the first row resulting in $\det A = D_{11}M_{11}$, where M_{11} is the minor resulting from deleting the first row and the first column. Note that M_{11} is the determinant of a matrix similar to $D - \mathbf{1}$ but of size $n - 1 \times n - 1$.

Repeating recursively the above multilinearity argument, we have

$$\det(D - \mathbf{1}) =$$
$$-\prod_{j=2}^{n} D_{jj} + D_{11}\left(-\prod_{j=3}^{n} D_{jj} + D_{22}\left(-\prod_{j=4}^{n} D_{jj} + D_{33}\right.\right.$$
$$\left.\left.\left(-\prod_{j=5}^{n} D_{jj} + D_{44}(\cdots)\right)\right)\right) = \prod_{i=1}^{n} D_{ii} - \sum_{i=1}^{n}\prod_{j\neq i} D_{jj}.$$

□

5.3 Computing det *G*

Proposition 2. *The determinant of the Gram matrix G of the metric $F_\lambda^* \mathcal{J}$ is*

$$det\, G \propto \frac{\prod_{i=1}^{n+1}(\lambda_i/x_i)}{\langle x, \lambda\rangle^{n+1}}. \tag{6}$$

Proof. We will factor G into a product of square matrices and compute $\det G$ as the product of the determinants of each factor.

By factoring a diagonal matrix Λ, $[\Lambda]_{ii} = \sqrt{\frac{\lambda_i}{x_i}}\frac{1}{2\sqrt{\langle x,\lambda\rangle}}$ from $D - \lambda\alpha^\top$, we have

$$J = U\left(I - \frac{\lambda x^\top}{\langle x, \lambda\rangle}\right)\Lambda \tag{7}$$

$$G = U\left(I - \frac{\lambda x^\top}{\langle x, \lambda\rangle}\right)\Lambda^2\left(I - \frac{\lambda x^\top}{\langle x, \lambda\rangle}\right)^\top U^\top. \tag{8}$$

Note that $G = JJ^\top$ is not the desired decomposition since J is not a square matrix.

We proceed by studying the eigenvalues and eigenvectors of $I - \frac{\lambda x^\top}{\langle x,\lambda\rangle}$ in order to simplify (8) via an eigenvalue decomposition. First, note that, if (v, μ) is an eigenvector-eigenvalue pair of $\frac{\lambda x^\top}{\langle x,\lambda\rangle}$, then $(v, 1 - \mu)$ is an eigenvector-eigenvalue pair of $I - \frac{\lambda x^\top}{\langle x,\lambda\rangle}$. Next, note that vectors v such that $x^\top v = 0$ are eigenvectors of $\frac{\lambda x^\top}{\langle x,\lambda\rangle}$ with eigenvalue 0. Hence, they are also eigenvectors of $I - \frac{\lambda x^\top}{\langle x,\lambda\rangle}$ with eigenvalue 1. There are n such independent vectors $v_1, \ldots, v_n$. Since $\mathrm{trace}(I - \frac{\lambda x^\top}{\langle x,\lambda\rangle}) = n$, the sum of the eigenvalues is also n and we may conclude that the last of the $n + 1$ eigenvalues is 0.

The eigenvectors of $I - \frac{\lambda x^\top}{\langle x,\lambda\rangle}$ may be written in several ways. One possibility is as the columns of the following matrix

$$V = \begin{pmatrix} -\frac{x_2}{x_1} & -\frac{x_3}{x_1} & \cdots & -\frac{x_{n+1}}{x_1} & \lambda_1 \\ 1 & 0 & \cdots & 0 & \lambda_2 \\ 0 & 1 & \cdots & 0 & \lambda_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \lambda_{n+1} \end{pmatrix} \in \mathbb{R}^{n+1\times n+1},$$

where the first n columns are the eigenvectors that correspond to unit eigenvalues and the last eigenvector corresponds to a 0 eigenvalue.

Using the above eigenvector decomposition, we have $I - \frac{\lambda x^\top}{\langle x,\lambda\rangle} = V\tilde{I}V^{-1}$ and $\tilde{I}$ is a diagonal matrix containing all the eigenvalues. Since the diagonal of $\tilde{I}$ is $(1, 1, \ldots, 1, 0)$, we may write $I - \frac{\lambda x^\top}{\langle x,\lambda\rangle} = V^{|n}V^{-1|n}$, where $V^{|n} \in \mathbb{R}^{n+1\times n}$ is V with the last column removed and $V^{-1|n} \in \mathbb{R}^{n\times n+1}$ is V^{-1} with the last row removed.

We have then,

$$\begin{aligned} \det G &= \det\Big(U(V^{|n}V^{-1|n})\Lambda^2(V^{-1|n\top}V^{|n\top})U^\top\Big) \\ &= \det\Big((UV^{|n})(V^{-1|n}\Lambda^2V^{-1|n\top})(V^{|n\top}U^\top)\Big) \\ &= \Big(\det(UV^{|n})\Big)^2\ \det\Big(V^{-1|n}\Lambda^2V^{-1|n\top}\Big). \end{aligned}$$

Noting that

$$UV^{|n} = \begin{pmatrix} -\frac{x_2}{x_1} & -\frac{x_3}{x_1} & \cdots & -\frac{x_n}{x_1} & -\frac{x_{n+1}}{x_1} - 1 \\ 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \in \mathbb{R}^{n\times n},$$

we factor $1/x_1$ from the first row and add columns $2, \ldots, n$ to column 1, thus obtaining

$$\begin{pmatrix} -\sum_{i=1}^{n+1} x_i & -x_3 & \cdots & -x_n & -x_{n+1} - x_1 \\ 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix}.$$

Computing the determinant by minor expansion of the first column, we obtain

$$\det(UV^{|n})^2 = \left(\frac{1}{x_1}\sum_{i=1}^{n+1} x_i\right)^2 = \frac{1}{x_1^2}. \tag{9}$$

We proceed by computing $\det V^{-1|n}\Lambda^2 V^{-1|n\top}$.

The inverse of V, as may be easily verified is,

$$V^{-1} = \frac{1}{\langle x,\lambda\rangle} \begin{pmatrix} -x_1\lambda_2 & \langle x,\lambda\rangle - x_2\lambda_2 & -x_3\lambda_2 & \cdots & -x_{n+1}\lambda_2 \\ -x_1\lambda_3 & -x_2\lambda_3 & \langle x,\lambda\rangle - x_3\lambda_3 & \cdots & -x_{n+1}\lambda_3 \\ \vdots & \vdots & & \ddots & \\ -x_1\lambda_{n+1} & -x_2\lambda_{n+1} & \cdots & \cdots & \langle x,\lambda\rangle - x_{n+1}\lambda_{n+1} \\ x_1\lambda_1 & x_2\lambda_1 & \cdots & \cdots & x_{n+1}\lambda_1 \end{pmatrix}.$$

Removing the last row gives

$$V^{-1|n} = \frac{1}{\langle x,\lambda\rangle} \begin{pmatrix} -x_1\lambda_2 & \langle x,\lambda\rangle - x_2\lambda_2 & -x_3\lambda_2 & \cdots & -x_{n+1}\lambda_2 \\ -x_1\lambda_3 & -x_2\lambda_3 & \langle x,\lambda\rangle - x_3\lambda_3 & \cdots & -x_{n+1}\lambda_3 \\ \vdots & \vdots & & \ddots & \\ -x_1\lambda_{n+1} & -x_2\lambda_{n+1} & \cdots & \cdots & \langle x,\lambda\rangle - x_{n+1}\lambda_{n+1} \end{pmatrix},$$

$$= \frac{1}{\langle x,\lambda\rangle} P \begin{pmatrix} -x_1 & \langle x,\lambda\rangle/\lambda_2 - x_2 & -x_3 & \cdots & -x_{n+1} \\ -x_1 & -x_2 & \langle x,\lambda\rangle/\lambda_3 - x_3 & \cdots & -x_{n+1} \\ \vdots & \vdots & & \ddots & \\ -x_1 & -x_2 & \cdots & \cdots & \langle x,\lambda\rangle/\lambda_{n+1} - x_{n+1} \end{pmatrix},$$

where

$$P = \begin{pmatrix} \lambda_2 & 0 & \cdots & 0 \\ 0 & \lambda_3 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_{n+1} \end{pmatrix}.$$

$[V^{-1|n}\Lambda^2 V^{-1|n\top}]_{ij}$ is the scalar product of the i and j rows of the following matrix

$$V^{-1|n}\Lambda = \frac{1}{2}\langle x,\lambda\rangle^{-3/2} P \begin{pmatrix} -\sqrt{x_1\lambda_1} & \frac{\langle x,\lambda\rangle}{\sqrt{x_2\lambda_2}} - \sqrt{x_2\lambda_2} & -\sqrt{x_3\lambda_3} & \cdots & -\sqrt{x_{n+1}\lambda_{n+1}} \\ -\sqrt{x_1\lambda_1} & -\sqrt{x_2\lambda_2} & \frac{\langle x,\lambda\rangle}{\sqrt{x_3\lambda_3}} - \sqrt{x_3\lambda_3} & \cdots & -\sqrt{x_{n+1}\lambda_{n+1}} \\ \vdots & \vdots & & \ddots & \\ -\sqrt{x_1\lambda_1} & -\sqrt{x_2\lambda_2} & \cdots & \cdots & \frac{\langle x,\lambda\rangle}{\sqrt{x_{n+1}\lambda_{n+1}}} - \sqrt{x_{n+1}\lambda_{n+1}} \end{pmatrix}.$$

We therefore have

$$V^{-1|n}\Lambda^2 V^{-1|n\top} = \frac{1}{4}\langle x,\lambda\rangle^{-2} PQP,$$

where

$$Q = \begin{pmatrix} \frac{\langle x,\lambda\rangle}{x_2\lambda_2} - 1 & -1 & \cdots & -1 \\ -1 & \frac{\langle x,\lambda\rangle}{x_3\lambda_3} - 1 & \cdots & -1 \\ \vdots & & \ddots & \vdots \\ -1 & -1 & -1 & \frac{\langle x,\lambda\rangle}{x_{n+1}\lambda_{n+1}} - 1 \end{pmatrix}.$$

As a consequence of Lemma 2, we have

$$\det Q = x_1\lambda_1 \frac{\langle x,\lambda\rangle^n}{\prod_{i=1}^{n+1} x_i\lambda_i} - x_1\lambda_1 \frac{\langle x,\lambda\rangle^{n-1}\sum_{j=2}^{n+1} x_j\lambda_j}{\prod_{i=1}^{n+1} x_i\lambda_i} = x_1^2\lambda_1^2 \frac{\langle x,\lambda\rangle^{n-1}}{\prod_{i=1}^{n+1} x_i\lambda_i}.$$

and

$$\det V^{-1|n}\Lambda^2 V^{-1|n\top} = (1/4)^n \langle x,\lambda\rangle^{-2n}\left(\prod_{i=2}^{n+1}\lambda_i\right)$$

$$x_1^2\lambda_1^2 \frac{\langle x,\lambda\rangle^{n-1}}{\prod_{i=1}^{n+1} x_i\lambda_i}\left(\prod_{i=2}^{n+1}\lambda_i\right) = \frac{x_1^2\langle x,\lambda\rangle^{n-1}}{4^n\langle x,\lambda\rangle^{2n}}\prod_{i=1}^{n+1}\frac{\lambda_i}{x_i}.$$

This proves the proposition since multiplying $\det V^{-1|n}\Lambda^2 V_n^{-1|n\top}$ above by (9) gives (6). □

Propositions 1 and 2 reveal the form of the objective function $\mathcal{O}(g, D)$. Fig. 3 displays the inverse volume element on $\mathbb{P}_1$ with the corresponding geodesic distance from the left corner of $\mathbb{P}_1$. In the next section, we describe a maximum-likelihood estimation problem that is equivalent to maximizing $\mathcal{O}(g, D)$ and study its properties.

6 An Inverse Volume Probabilistic Model

Using Proposition 2, we have that the objective function $\mathcal{O}(g, D)$ may be regarded as a likelihood function under the model

$$p(x;\lambda) = \frac{1}{Z}\langle x,\lambda\rangle^{\frac{n+1}{2}}\prod_{i=1}^{n+1} x_i^{1/2} \quad x,\lambda\in\mathbb{P}_n, \tag{10}$$

where $Z = \int_{\mathbb{P}_n}\langle x,\lambda\rangle^{\frac{n+1}{2}}\prod_{i=1}^{n+1} x_i^{1/2}\mathrm{d}x$. The loglikelihood function for model (10) is given by

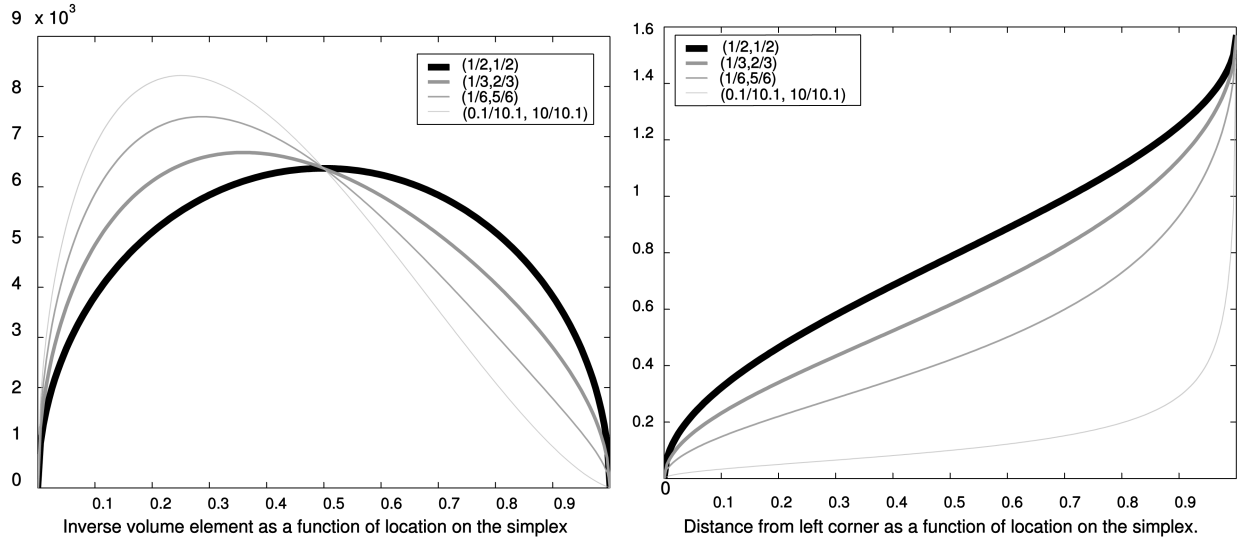
$$\ell(\lambda;x) = \frac{n+1}{2}\log(\langle x,\lambda\rangle) - \log\int_{\mathbb{P}_n}\langle x,\lambda\rangle^{\frac{n+1}{2}}\prod_{i=1}^{n+1}\sqrt{x_i}\ \mathrm{d}x.$$

The Hessian matrix $H(x,\lambda)$ of the log-likelihood function may be written as

$$[H(x,\lambda)]_{ij} = -k\frac{x_i}{\langle x,\lambda\rangle}\frac{x_j}{\langle x,\lambda\rangle} - (k^2-k)L\left(\frac{x_i}{\langle x,\lambda\rangle}\frac{x_j}{\langle x,\lambda\rangle}\right) + k^2 L\left(\frac{x_i}{\langle x,\lambda\rangle}\right)L\left(\frac{x_j}{\langle x,\lambda\rangle}\right),$$

where $k = \frac{n+1}{2}$ and L is the positive linear functional

$$Lf = \frac{\int_{\mathbb{P}_n}\langle x,\lambda\rangle^{\frac{n+1}{2}}\prod_{l=1}^{n+1}\sqrt{x_l}\ f(x,\lambda)\ \mathrm{d}x}{\int_{\mathbb{P}_n}\langle x,\lambda\rangle^{\frac{n+1}{2}}\prod_{l=1}^{n+1}\sqrt{x_l}\mathrm{d}x}.$$



(a) (b)

Fig. 3. (a) The inverse volume element $1/\sqrt{\det G(x)}$ as a function of $x \in \mathbb{P}_1$ and (b) the geodesic distance $d(x, 0)$ from the left corner as a function $x \in \mathbb{P}_1$. Different plots represent different metric parameters $\lambda \in \{(1/2, 1/2), (1/3, 2/3), (1/6, 5/6), (0.0099, 0.9901)\}$.

Note that the matrix given by $LH(x, \lambda) = [LH_{ij}(x, \lambda)]$ is negative definite due to its covariance-like form. In other words, for every value of λ, $H(x, \lambda)$ is negative definite on average, with respect to the model $p(x; \lambda)$. While not as strong as negative definite, this property indicates a favorable condition for maximization.

6.1 Computing the Normalization Term

We describe an efficient way to compute the normalization term Z through the use of dynamic programming and Fast Fourier Transform (FFT).

Assuming that $n = 2k - 1$ for some $k \in \mathbb{N}$, we have

$$Z = \int_{\mathbb{P}_n} \langle x, \lambda \rangle^k \prod_{i=1}^{n+1} x_i^{1/2} \mathrm{d}x = \sum_{a_1 + \cdots + a_{n+1} = k : a_i \geq 0} \frac{k!}{a_1! \cdots a_{n+1}!} \prod_{j=1}^{n+1} \lambda_j^{a_j}$$

$$\int_{\mathbb{P}_n} \prod_{j=1}^{n+1} x_j^{a_j + \frac{1}{2}} \mathrm{d}x \propto \sum_{a_1 + \cdots + a_{n+1} = k : a_i \geq 0} \prod_{j=1}^{n+1} \frac{\Gamma(a_j + 3/2)}{\Gamma(a_j + 1)} \lambda_j^{a_j}.$$

The following proposition and its proof describe a way to compute the summation in Z in $O(n^2 \log n)$ time.

Proposition 3. *The normalization term for model (10) may be computed in $O(n^2 \log n)$ time complexity.*

Proof. Using the notation $c_m = \frac{\Gamma(m+3/2)}{\Gamma(m+1)}$ the summation in Z may be expressed as

$$Z \propto \sum_{a_1=0}^{k} c_{a_1} \lambda_1^{a_1} \sum_{a_2=0}^{k-a_1} c_{a_2} \lambda_2^{a_2} \cdots \sum_{a_n=0}^{k - \sum_{j=1}^{n-1} a_j} c_{a_n} \lambda_n^{a_n} c_{k - \sum_{j=1}^{n} a_j} \lambda_{n+1}^{k - \sum_{j=1}^{n} a_j}. \quad (11)$$

A trivial dynamic program can compute (11) in $O(n^3)$ complexity.

However, each of the single subscript sums in (11) is, in fact, a linear convolution operation. By defining

$$B_{ij} = \sum_{a_i=0}^{j} c_{a_i} \lambda_i^{a_i} \cdots \sum_{a_n=0}^{j - \sum_{l=i}^{n-1} a_l} c_{a_n} \lambda_n^{a_n} c_{j - \sum_{l=i}^{n} a_l} \lambda_{n+1}^{j - \sum_{l=i}^{n} a_l},$$

we have $Z = B_{1k}$ and the recurrence relation $B_{ij} = \sum_{m=0}^{j} c_m \lambda_i^m B_{i+1,j-m}$ which is the linear convolution of $\{B_{i+1,j}\}_{j=0}^{k}$ with the vector $\{c_j \lambda_i^j\}_{j=0}^{k}$. By performing the convolution in the frequency domain (i.e., multiplying the FFT of the vectors and then computing the inverse FFT), filling in each row of the table B_{ij} for $i = 0, \ldots, n+1, j = 0, \ldots, k$ takes $O(n \log n)$ complexity leading to a total of $O(n^2 \log n)$ complexity. □

The computation method described in the proof may be used to compute the partial derivative of Z, resulting in $O(n^3 \log n)$ computation for the gradient. By careful dynamic programming, the gradient vector may be computed in $O(n^2 \log n)$ time complexity as well.

7 Applications

7.1 Text Classification

In this section, we describe applying the metric learning framework to document classification and report some results on the WebKB data set. We map documents to the simplex by multinomial MLE or MAP estimation. This mapping results in the well-known term-frequency (tf) representation where the multinomial model entries are the frequencies of the different terms in the document.

It is a well-known fact that less common terms across the text corpus tend to provide more discriminative information than the most common terms. In the extreme case, stopwords like `the`, `or`, and `of` are often severely down-weighted or removed from the representation. Geometrically, this means that we would like the geodesics to pass through corners of the simplex that correspond to sparsely occurring words, in contrast to densely populated simplex

idf	Estimated λ
tiff romano potra	disobedience seat alr
anitescu papeli theo	seizure refuse delegated
echo chimera trestle	soverigns territory
schlatter xiyong	mobocracy stabbed
⋮	⋮
at department with	will course system
this by office course	you page research with
are an from system	that by are at this
programming be last	home from office or as

Fig. 4. Comparison of top and bottom valued parameters for idf and model (10). The words are sorted by their idf or λ values. The data set is the faculty versus student Web page classification task from WebKB data set. Note that the least scored terms are similar for the two methods while the top scored terms are completely disjoint.

corners such as the ones that correspond to the stop-words above. To account for this in our framework, we use the metric $F_\lambda^* \mathcal{J} = (F_\theta^{-1})^* \mathcal{J}$, where θ is the MLE under model (10) obtained by a gradient descent, modified to work in $\mathbb{P}_n$, with early stopping procedure. In other words, we are pulling back the Fisher information metric through the inverse to the transformation that maximizes the normalized inverse volume of D. As a result, geodesics will tend to pass through sparsely populated regions emphasizing differences in dimensions that correspond to rare words.

The standard tfidf representation of a document consists of multiplying the tf parameter by an idf component

$$idf_k = \log \frac{N}{\#\text{documents that word } k \text{ appears in}}.$$

Given the tfidf representation of two documents, their cosine similarity is simply the scalar product between the two normalized tfidf representations. Despite its simplicity the tfidf representation leads to some of the best results in text classification (e.g., [4]) and information retrieval and is a natural candidate for a baseline comparison due to its similarity to the geodesic expression.

A comparison of the top and bottom terms between the metric learning and idf scores is shown in Fig. 4. Note that both methods rank similar words at the bottom. These are the most common words such as `this`, `at`, etc., that often carry little or no information for classification purposes. The top words, however, are completely different for the two schemes. Note the tendency of idf to give high scores to rare proper nouns while the metric learning method gives high scores for rare common nouns. This difference may be explained by the fact that idf considers appearance of words in documents as a binary event while the metric learning looks at the number of appearances of a term in each document through the documents representation as term frequencies. As a result, the total number of appearances of each term in the corpus is taken into account rather than the number of documents it appears in. Rare proper nouns such as the high scoring idf terms in Fig. 4 appear several times in a single Web page. As a result, these words will score higher with the idf scheme but lower with the metric learning scheme.

In Fig. 5, the rank-value plot for the estimated λ values and idf is shown on a log-log scale. The x axis represents different words that are sorted by increasing parameter value and the y axis represents the λ or idf value. An experimental observation is that the idf scores show a stronger linear trend in the log-log scale than the λ values.

To measure performance in classification we compared the testing error of a nearest neighbor classifier under two different distances. We compared geodesic distance under the learned metric with tfidf cosine similarity. Fig. 6 displays test-set error rates as a function of the training set size. The error rates were averaged over 30 experiments with random sampling of a fixed size training set. According to Fig. 6, the learned metric outperforms the standard tfidf measure by a considerable amount.

7.2 Image Classification

Images are typically represented as a two-dimensional array of pixels taking values in some bounded continuous range, e.g., $\Theta = (0,1)^{100\times 100} \cong (0,1)^{10,000}$. A metric g on the

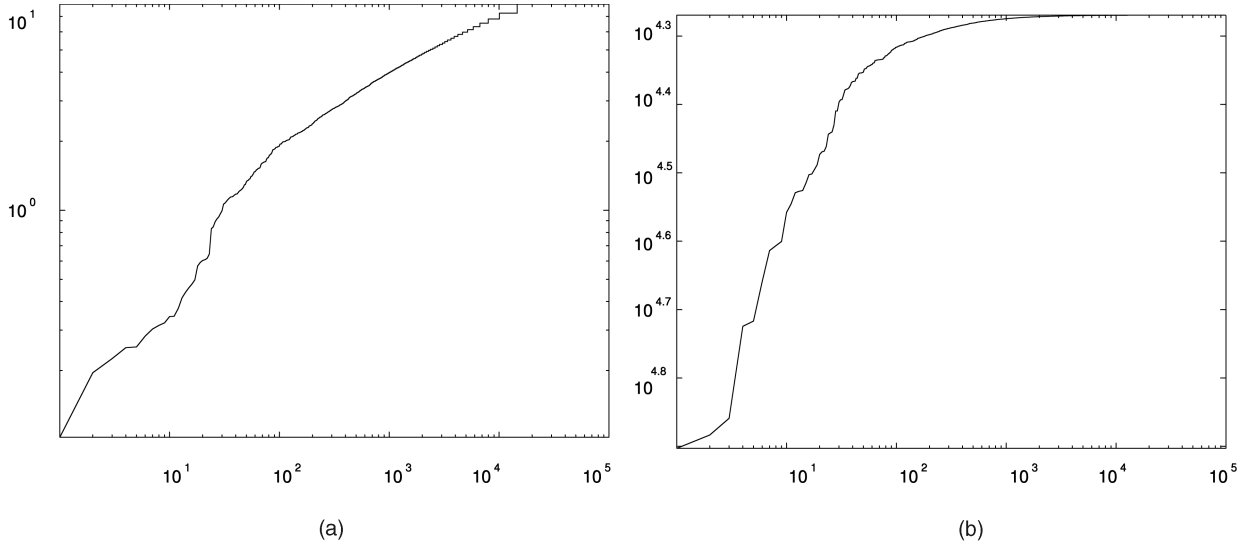


Fig. 5. (a) Log-log plots for sorted idf values and (b) the sorted λ values of the learned metric. The task is the same as in Fig. 4.

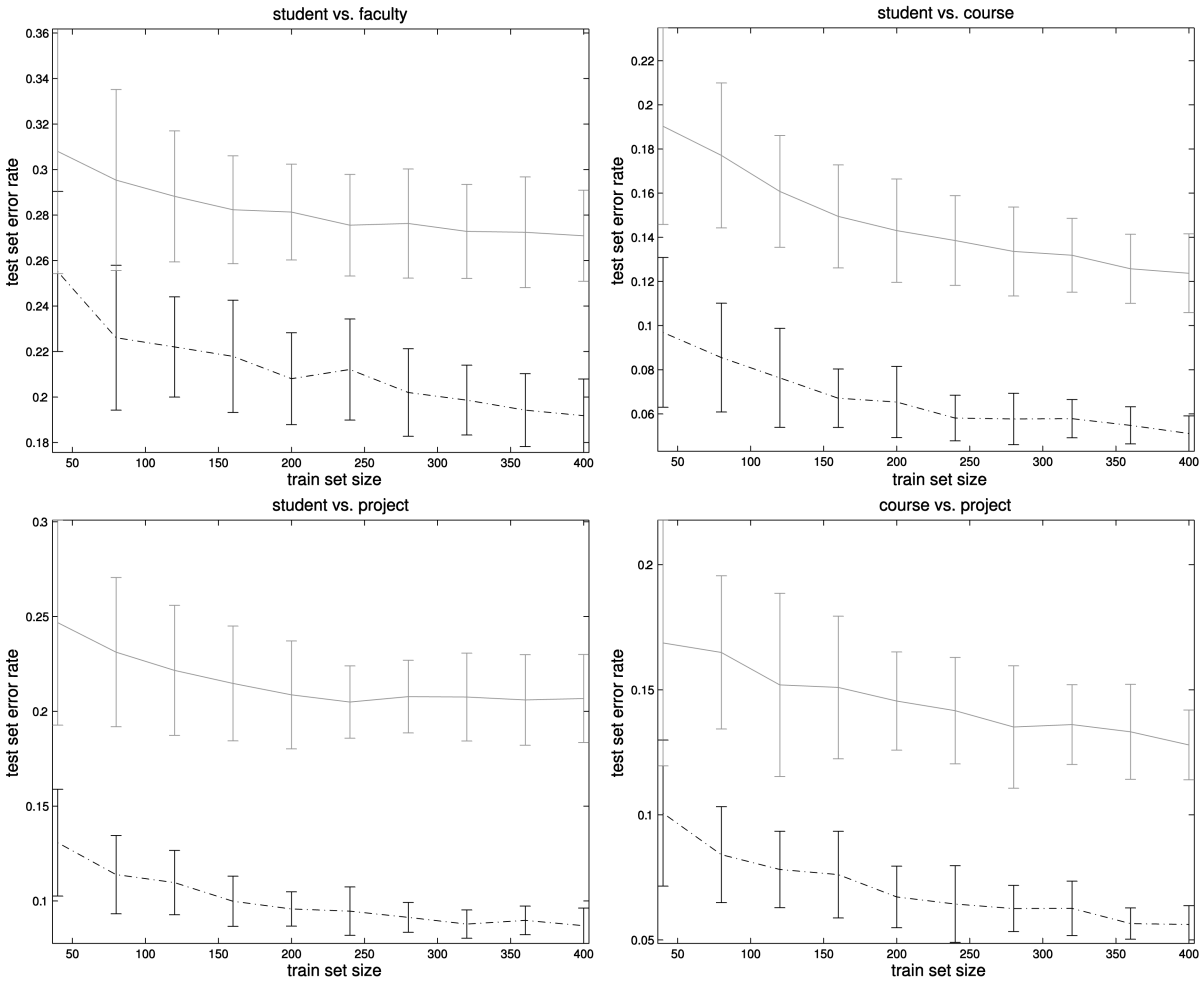


Fig. 6. Test set error rate for nearest neighbor classifier on WebKB binary tasks. Distances were computed by geodesic for the learned Riemannian metric (dashed) and tfidf with cosine similarity (solid). The plots are averaged over a set of 30 random samplings of training sets of the specified sizes, evenly divided between positive and negative examples. Error bars represent one standard deviation.

resulting manifold Θ is specified by defining its values for every pair of basis tangent vectors at each point in Θ

$$g_\theta(e_{ij}, e_{kl}) \quad i,k = 1,\ldots,n \quad j,l = 1,\ldots,m \quad \forall \theta \in \Theta.$$

The value $g_\theta(e_{ij}, e_{kl})$ may be interpreted as the cost of increasing the brightness of pixels (i,j) and (k,l) simultaneously in the image θ.

A reasonable restriction is to constrain g to a local diagonal form $g_\theta(e_{ij}, e_{kl}) = \delta_{ik}\delta_{jl} f(N(\theta_{ij}))$, where f is some function and $N(\theta_{ij})$ is a neighborhood of the pixel θ_{ij}. Using the above intuition, this means that the cost depends only on the neighborhood of the pixel and there is no pairwise interaction when simultaneously changing the values of two pixels. The volume element, in this case, is easily computed to be $\mathrm{dvol}g(\theta) = \prod_{ij} \sqrt{f(N(\theta_{ij}))}$. The parametric family of metrics reduces to a selection of a parametric family of functions $\{f_\lambda : \lambda \in \Lambda\}$.

The learned metric would then capture local properties of images in the training collection. For example, the metric learned for face images would be different from the metric learned for outdoors scene images. We leave the precise specification of f_λ and experimental results for future work.

8 Summary

We have proposed a new framework for the metric learning problem that enables robust learning of a local metric for high-dimensional sparse data. This is achieved by restricting the set of metric candidates to a parametric family and selecting a metric based on maximizing the inverse volume element.

In the case of learning a metric on the multinomial simplex, the metric candidates are taken to be pull-back metrics of the Fisher information under a continuous group of transformations. Since the geometries are isometric to the positive sphere equipped with the metric inherited from the Euclidean space, the geodesic distances are easily computed. Furthermore, the geometries are easily visualized and are shown to be of a form similar to the popular tfidf distances. The optimization problem, which may be cast as a maximum-likelihood problem, selects a specific geometry that is similar to tfidf, yet posseses qualitative differences that enable it to outperform tfidf in text classification.

The framework proposed in this section is quite general and may be employed in other domains. The key component is the specification of the set of metric candidates by flattening transformations and the ability to compute a closed form expression for their volume elements.

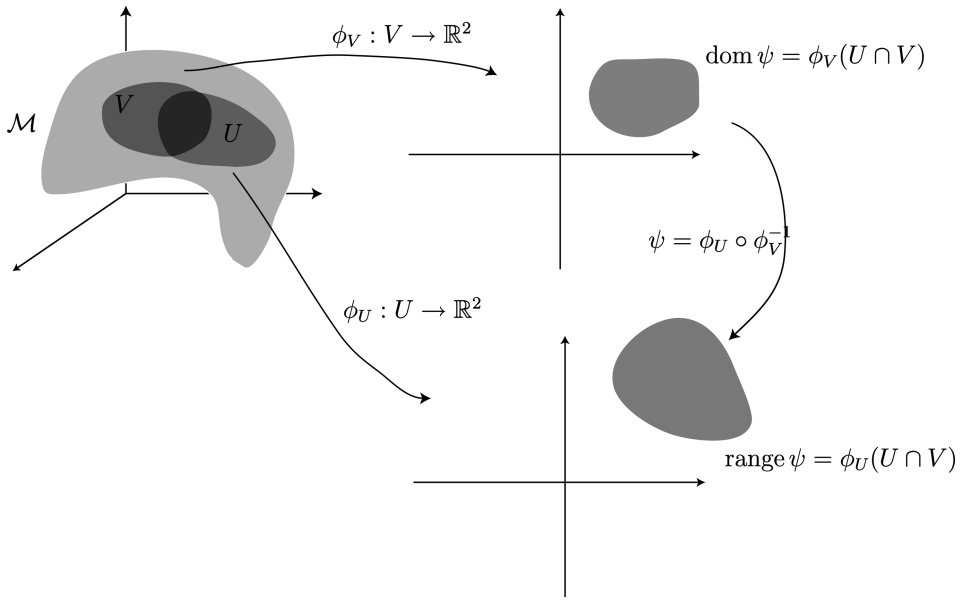


Fig. 7. Two neighborhoods U, V in a two-dimensional manifold $\mathcal{M}$, the coordinate charts ϕ_U, ϕ_V, and the transition function ψ between them.

APPENDIX A
REVIEW OF RIEMANNIAN GEOMETRY

In this section, we describe concepts from Riemannian geometry that are relevant to this paper. For more details, refer to any textbook discussing Riemannian geometry, e.g., [9].

Riemannian manifolds are built out of three layers of structure. The topological layer is suitable for treating topological notions such as continuity and convergence. The differentiable layer allows extending the notion of differentiability to the manifold and the Riemannian layer defines rigid geometrical quantities such as distances, angles, and curvature on the manifold. In accordance with this philosophy, we start below with the definition of topological manifold and quickly proceed to defining differentiable manifolds and Riemannian manifolds.

A.1 Topological and Differentiable Manifolds

A homeomorphism between two topological spaces X and Y is a bijection $\phi : X \to Y$ for which both ϕ and ϕ^{-1} are continuous. We then say that X and Y are homeomorphic and essentially equivalent from a topological perspective. 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 $U \subset \mathcal{M}$ that is homeomorphic to $\mathbb{R}^n$. The local homeomorphisms in the above definition $\phi_U : U \subset \mathcal{M} \to \mathbb{R}^n$ are usually called charts. Note that this definition of a topological manifold makes use of an ambient Euclidean space $\mathbb{R}^m$ (a Euclidean space such that the manifold is its topological subspace). 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[2] [8]. Unless otherwise noted, for the remainder of this section, we assume that all manifolds are of dimension n.

We are now in a position to introduce the differentiable structure. First, recall that a mapping between two open sets of Euclidean spaces $f : U \subset \mathbb{R}^k \to V \subset \mathbb{R}^l$ is infinitely differentiable, denoted by $f \in C^\infty(\mathbb{R}^k, \mathbb{R}^l)$ if f has continuous partial derivatives of all orders. If for every pair of charts ϕ_U, ϕ_V, the transition function defined by

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

(when $U \cap V \neq \emptyset$) is a $C^\infty(\mathbb{R}^n, \mathbb{R}^n)$ differentiable map then $\mathcal{M}$ is called an n-dimensional differentiable manifold. The charts and transition function for a two-dimensional manifold are illustrated in Fig. 7.

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 $\mathbb{S}^n = \{x \in \mathbb{R}^{n+1} : \sum x_i^2 = 1\}$ its positive orthant $\mathbb{S}_+^n = \{x \in \mathbb{R}^{n+1} : \sum x_i^2 = 1,\ \forall i\ x_i > 0\}$, and the n-simplex $\mathbb{P}_n = \{x \in \mathbb{R}^{n+1} : \sum x_i = 1,\ \forall i\ x_i > 0\}$.

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 $f : \mathcal{M} \to \mathcal{N}$. A continuous function $f : \mathcal{M} \to \mathbb{R}$ is said to be $C^\infty(\mathcal{M}, \mathbb{R})$ differentiable if for every chart ϕ_U the function $f \circ \phi_U^{-1} \in C^\infty(\mathbb{R}^n, \mathbb{R})$. 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

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

A diffeomorphism between two manifolds $\mathcal{M}, \mathcal{N}$ is a bijection $f : \mathcal{M} \to \mathcal{N}$ such that $f \in C^\infty(\mathcal{M}, \mathcal{N})$ and $f^{-1} \in C^\infty(\mathcal{N}, \mathcal{M})$.

A.2 The Tangent Space

For every point $x \in \mathcal{M}$, we define an n-dimensional real vector space $T_x\mathcal{M}$, isomorphic to $\mathbb{R}^n$, called the tangent space. The elements of the tangent space, the tangent vectors

2. The general definition, that uses the Hausdorff and second countability properties, is equivalent to the ambient Euclidean space definition by Whitney's embedding theorem. Nevertheless, it is considerably more elegant to do away with the excess baggage of an ambient space.

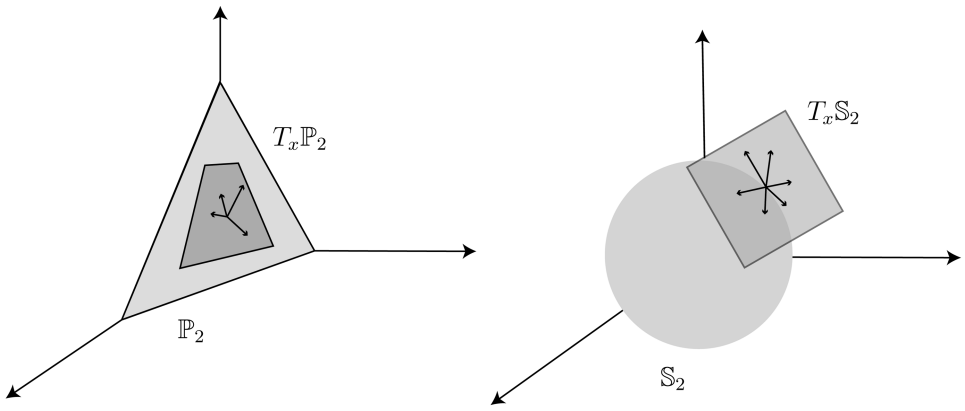


Fig. 8. Tangent spaces of the 2-simplex $T_x\mathbb{P}_2$ and the 2-sphere $T_x\mathbb{S}_2$.

$v \in T_x\mathcal{M}$, are usually defined as directional derivatives at x operating on $C^\infty(\mathcal{M}, \mathbb{R})$ differentiable functions or as equivalence classes of curves having the same velocity vectors at x. Intuitively, tangent vectors and tangent spaces are a generalization of geometric tangent vectors and spaces for smooth curves and two-dimensional surfaces in the ambient $\mathbb{R}^3$. For an n-dimensional manifold $\mathcal{M}$ embedded in an ambient $\mathbb{R}^m$ the tangent space $T_x\mathcal{M}$ is a copy of $\mathbb{R}^n$ translated so that its origin is positioned at x. See Fig. 8 for an illustration of this concept for two-dimensional manifolds in $\mathbb{R}^3$.

In many cases, the manifold $\mathcal{M}$ is a submanifold of an m-dimensional manifold $\mathcal{N}$, $m \geq n$. Considering $\mathcal{M}$ and its ambient space $\mathbb{R}^m, m \geq n$ is one special case of this phenomenon. For example, both $\mathbb{P}_n$ and $\mathbb{S}^n$ 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 $T_x\mathcal{N} \cong \mathbb{R}^m$ 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 (see Fig. 8)

$$\begin{aligned} T_x\mathbb{P}_n &= \left\{ v \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} v_i = 0 \right\} \\ T_x\mathbb{S}_n &= \left\{ v \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} v_i x_i = 0 \right\}. \end{aligned} \tag{12}$$

A C^∞ vector field X on $\mathcal{M}$ is a smooth assignment of tangent vectors to each point of $\mathcal{M}$. We denote the set of vector fields on $\mathcal{M}$ as $\mathcal{X}(\mathcal{M})$ and X_p is the value of the vector field X at $p \in \mathcal{M}$. Given a function $f \in C^\infty(\mathcal{M}, \mathbb{R})$, we define the action of $X \in \mathcal{X}(\mathcal{M})$ on f as

$$Xf \in C^\infty(\mathcal{M}, \mathbb{R}) \qquad (Xf)(p) = X_p(f)$$

in accordance with our definition of tangent vectors as directional derivatives of functions.

A.3 Riemannian Manifolds

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 local inner product on tangent vectors

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

that is symmetric, bilinear, positive definite, and C^∞ differentiable in x. By the bilinearity of the inner product g, for every $u, v \in T_x\mathcal{M}$

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

and g_x is completely described by $\{g_x(\partial_i, \partial_j) : 1 \leq i, j \leq n\}$ —the set of inner products between the basis elements $\{\partial_i\}_{i=1}^n$ of $T_x\mathcal{M}$. The Gram matrix $[G(x)]_{ij} = g_x(\partial_i, \partial_j)$ is a symmetric and positive definite matrix that completely describe the metric g_x.

The metric enables us to define lengths of tangent vectors $v \in T_x\mathcal{M}$ by $\sqrt{g_x(v, v)}$ and lengths of curves $\gamma : [a, b] \to \mathcal{M}$ by

$$L(\gamma) = \int_a^b \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\mathrm{d}t,$$

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 $d_g(x, y)$ between two points $x, y \in \mathcal{M}$ as

$$d_g(x, y) = \inf_{\gamma \in \Gamma(x,y)} \int_a^b \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\mathrm{d}t, \tag{13}$$

where $\Gamma(x, y)$ is the set of piecewise differentiable curves connecting x and y. The distance d_g is called geodesic distance and the minimal curve achieving it is called a geodesic curve.[3] 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:

Definition 1. *The push-forward map $f_* : T_x\mathcal{M} \to T_{f(x)}\mathcal{N}$, associated with the diffeomorphism $f : \mathcal{M} \to \mathcal{N}$ is the mapping that satisfies $v(r \circ f) = (f_*v)r$, $\forall r \in C^\infty(\mathcal{N}, \mathbb{R})$ and $\forall v \in T_x\mathcal{M}$.*

The push-forward is none other than a coordinate free version of the Jacobian matrix J or the total derivative operator associated with the local chart representation of f.

3. It is also common to define geodesics as curves satisfying certain differential equations. The above definition, however, is more intuitive and appropriate for our needs.

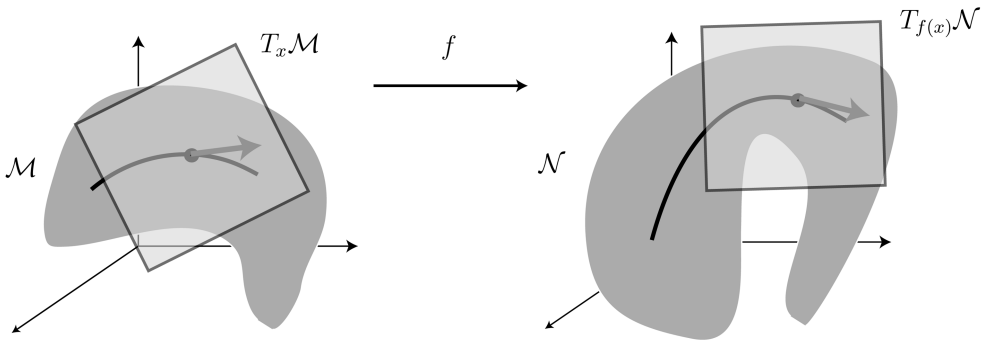


Fig. 9. The map $f : \mathcal{M} \to \mathcal{N}$ defines a push forward map $f_* : T_x\mathcal{M} \to T_{f(x)}\mathcal{N}$ that transforms velocity vectors of curves to velocity vectors of the transformed curves.

In other words, if we define the coordinate version of $f : \mathcal{M} \to \mathcal{N}$

$$\tilde{f} = \phi \circ f \circ \psi^{-1} : \mathbb{R}^n \to \mathbb{R}^m,$$

where ϕ, ψ are local charts of $\mathcal{N}, \mathcal{M}$ then the push-forward map is

$$f_* u = Ju = \sum_i \left(\sum_j \frac{\partial \tilde{f}_i}{\partial x_j} u_j \right) e_i,$$

where J is the Jacobian of $\tilde{f}$ and $\tilde{f}_i$ is the i-component function of $\tilde{f} : \mathbb{R}^m \to \mathbb{R}^n$. Intuitively, as illustrated in Fig. 9, 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 $g_x(u, v) = (f^*h)_x(u, v) \quad \forall x \in \mathcal{M}, \quad \forall u, v \in T_x\mathcal{M}$.*

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.

REFERENCES

[1] S. Amari and H. Nagaoka, *Methods of Information Geometry.* Am. Math. Soc., 2000.
[2] A. Gous, "Exponential and Spherical Subfamily Models," Stanford Univ., 1998.
[3] K. Hall and T. Hofmann, "Learning Curved Multinomial Subfamilies for Natural Language Processing and Information Retrieval," *Proc. 17th Int'l Conf. Machine Learning,* P. Langley, ed., pp. 351-358, 2000.
[4] T. Joachims, "The Maximum Margin Approach to Learning Text Classifiers Methods, Theory and Algorithms," PhD thesis, Dortmund Univ., 2000.
[5] R.E. Kass and P.W. Voss, *Geometrical Foudnation of Asymptotic Inference.* John Wiley and Sons, Inc., 1997.
[6] G.R.G. Lanckriet, P. Bartlett, N. Cristianini, L. ElGhaoui, and M.I. Jordan, "Learning the Kernel Matrix with Semidefinite Programming," *J. Machine Learning Research,* vol. 5, pp. 27-72, 2004.
[7] G. Lebanon, "Riemannian Geometry and Statistical Machine Learning," Technical Report CMU-LTI-05-189, Carengie Mellon Univ., 2005.
[8] J.M. Lee, *Introduction to Topological Manifolds.* Springer, 2000.
[9] J.M. Lee, *Introduction to Smooth Manifolds.* Springer, 2002.
[10] M.K. Murray and J.W. Rice, *Differential Geometry and Statistics.* CRC Press, 1993.
[11] S. Roweis and L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science,* vol. 290, p. 2323, 2000.
[12] L.K. Saul and M.I. Jordan, "A Variational Principle for Model-Based Interpolation," *Advances in Neural Information Processing Systems 9,* M.C. Mozer, M.I. Jordan, and T. Petsche, eds., 1997.
[13] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russel, "Distance Metric Learning with Applications to Clustering with Side Information," *Advances in Neural Information Processing Systems, 15,* S. Becker, S. Thrun and K. Obermayer, eds., pp. 505-512, 2003.

Guy Lebanon received the bachelor and master's degrees from Technion—Israel Institute of Technology and the PhD degree from Carnegie Mellon University. He is an assistant professor in the Department of Statistics and School of Electrical and Computer Engineering at Purdue University. His main research area is the theory and applications of statistical machine learning.