Riemannian Geometry and Statistical Machine Learning

Guy Lebanon

CMU-LTI-05-189

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Riemannian Geometry and Statistical Machine Learning

Doctoral Thesis

Guy Lebanon
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
lebanon@cs.cmu.edu

January 31, 2005

Abstract

Statistical machine learning algorithms deal with the problem of selecting an appropriate statistical model from a model space Θ based on a training set $\{x_i\}_{i=1}^N \subset \mathcal{X}$ or $\{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$. In doing so they either implicitly or explicitly make assumptions on the geometries of the model space Θ and the data space $\mathcal{X}$. Such assumptions are crucial to the success of the algorithms as different geometries are appropriate for different models and data spaces. By studying these assumptions we are able to develop new theoretical results that enhance our understanding of several popular learning algorithms. Furthermore, using geometrical reasoning we are able to adapt existing algorithms such as radial basis kernels and linear margin classifiers to non-Euclidean geometries. Such adaptation is shown to be useful when the data space does not exhibit Euclidean geometry. In particular, we focus in our experiments on the space of text documents that is naturally associated with the Fisher information metric on corresponding multinomial models.

Thesis Committee: John Lafferty (chair)
Geoffrey J. Gordon, Michael I. Jordan, Larry Wasserman

Acknowledgements

This thesis contains work that I did during the years 2001-2004 at Carnegie Mellon University. During that period I received a lot of help from faculty, students and friends. However, I feel that I should start by thanking my family: Alex, Anat and Eran Lebanon, for their support during my time at Technion and Carnegie Mellon. Similarly, I thank Alfred Bruckstein, Ran El-Yaniv, Michael Lindenbaum and Hava Siegelmann from the Technion for helping me getting started with research in computer science.

At Carnegie Mellon University I received help from a number of people, most importantly my advisor John Lafferty. John helped me in many ways. He provided excellent technical hands-on assistance, as well as help on high-level and strategic issues. Working with John was a very pleasant and educational experience. It fundamentally changed the way I do research and turned me into a better researcher. I also thank John for providing an excellent environment for research without distractions and for making himself available whenever I needed.

I thank my thesis committee members Geoffrey J. Gordon, Michael I. Jordan and Larry Wasserman for their helpful comments. I benefited from interactions with several graduate students at Carnegie Mellon University. Risi Kondor, Leonid Kontorovich, Luo Si, Jian Zhang and Xioajin Zhu provided helpful comments on different parts of the thesis. Despite not helping directly on topics related to the thesis, I benefited from interactions with Chris Meek at Microsoft Research, Yoram Singer at the Hebrew University and Joe Verducci and Douglas Critchlow at Ohio State University. These interactions improved my understanding of machine learning and helped me write a better thesis.

Finally, I want to thank Katharina Probst, my girlfriend for the past two years. Her help and support made my stay at Carnegie Mellon very pleasant, despite the many stressful factors in the life of a PhD student. I also thank her for patiently putting up with me during these years.

Contents

Mathematical Notation

Following is a list of the most frequent mathematical notations in the thesis.

$\mathcal{X}$	Set/manifold of data points
Θ	Set/manifold of statistical models
$X \times Y$	Cartesian product of sets/manifold
X^k	The Cartesian product of X with itself k times
$\|\cdot\|$	The Euclidean norm
$\langle \cdot, \cdot \rangle$	Euclidean dot product between two vectors
$p(x\,;\theta)$	A probability model for x parameterized by θ
$p(y\|x\,;\theta)$	A probability model for y, conditioned on x and parameterized by θ
$D(\cdot,\cdot), D_r(\cdot,\cdot)$	I-divergence between two models
$T_x\mathcal{M}$	The tangent space to the manifold $\mathcal{M}$ at $x \in \mathcal{M}$
$g_x(u,v)$	A Riemannian metric at x associated with the tangent vectors $u, v \in T_x\mathcal{M}$
$\{\partial_i\}_{i=1}^n, \{e_i\}_{i=1}^n$	The standard basis associated with the vector space $T_x\mathcal{M} \cong \mathbb{R}^n$
$G(x)$	The Gram matrix of the metric g, $[G(x)]_{ij} = g_x(e_i, e_j)$
$\mathcal{J}_\theta(u,v)$	The Fisher information metric at the model θ associated with the vectors u, v
$\delta_x(u,v)$	The induced Euclidean local metric $\delta_x(u,v) = \langle u, v \rangle$
$\delta_{x,y}$	Kronecker's delta $\delta_{x,y} = 1$ if $x = y$ and 0 otherwise
$\iota : A \to X$	The inclusion map $\iota(x) = x$ from $A \subset X$ to X.
$\mathbb{N}, \mathbb{Q}, \mathbb{R}$	The natural, rational and real numbers respectively
$\mathbb{R}_+$	The set of positive real numbers
$\mathbb{R}^{k\times m}$	The set of real $k \times m$ matrices
$[A]_i$	The i-row of the matrix A
$\{\partial_{ab}\}_{a,b=1}^{k,m}$	The standard basis associated with $T_x\mathbb{R}^{k\times m}$
$\overline{X}$	The topological closure of X
$\mathbb{H}^n$	The upper half plane $\{x \in \mathbb{R}^n : x_n \geq 0\}$
$\mathbb{S}^n$	The n-sphere $\mathbb{S}^n = \{x \in \mathbb{R}^{n+1} : \sum_i x_i^2 = 1\}$
$\mathbb{S}_+^n$	The positive orthant of the n-sphere $\mathbb{S}_+^n = \{x \in \mathbb{R}_+^{n+1} : \sum_i x_i^2 = 1\}$
$\mathbb{P}_n$	The n-simplex $\mathbb{P}_n = \{x \in \mathbb{R}_+^{n+1} : \sum_i x_i = 1\}$
$f \circ g$	Function composition $f \circ g(x) = f(g(x))$
$C^\infty(\mathcal{M}, \mathcal{N})$	The set of infinitely differentiable functions from $\mathcal{M}$ to $\mathcal{N}$
$f_* u$	The push-forward map $f_* : T_x\mathcal{M} \to T_{f(x)}\mathcal{N}$ of u associated with $f : \mathcal{M} \to \mathcal{N}$
$f^* g$	The pull-back metric on $\mathcal{M}$ associated with $(\mathcal{N}, g)$ and $f : \mathcal{M} \to \mathcal{N}$
$d_g(\cdot,\cdot), d(\cdot,\cdot)$	The geodesic distance associated with the metric g
$\operatorname{dvol} g(\theta)$	The volume element of the metric g_θ, $\operatorname{dvol} g(\theta) = \sqrt{\det g_\theta} = \sqrt{\det G(\theta)}$
$\nabla_X Y$	The covariant derivative of the vector field Y in the direction of the vector field X associated with the connection ∇

$\ell(\theta)$	The log-likelihood function
$\mathcal{E}(\theta)$	The AdaBoost.M2 exponential loss function
$\tilde{p}$	Empirical distribution associated with a training set $D \subset \mathcal{X} \times \mathcal{Y}$
$\hat{u}$	The L^2 normalized version of the vector u
$d(x, S)$	The distance from a point to a set $d(x, S) = \inf_{y \in S} d(x, y)$

1 Introduction

There are two fundamental spaces in machine learning. The first space $\mathcal{X}$ consists of data points and the second space Θ consists of possible learning models. In statistical learning, Θ is usually a space of statistical models, $\{p(x\,;\theta) : \theta \in \Theta\}$ in the generative case or $\{p(y\,|\,x\,;\theta) : \theta \in \Theta\}$ in the discriminative case. The space Θ can be either a low dimensional parametric space as in parametric statistics, or the space of all possible models as in non-parametric statistics.

Learning algorithms select a model $\theta \in \Theta$ based on a training sample $\{x_i\}_{i=1}^n \subset \mathcal{X}$ in the generative case or $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ in the discriminative case. In doing so they, either implicitly or explicitly, make certain assumptions about the geometries of $\mathcal{X}$ and Θ. In the supervised case, we focus on the classification setting, where $\mathcal{Y} = \{y_1, \ldots, y_k\}$ is a finite set of unordered classes. By this we mean that the space

$$\mathcal{X} \times \mathcal{Y} = \mathcal{X} \times \cdots \times \mathcal{X} = \mathcal{X}^k$$

has the product geometry over $\mathcal{X}$. This is a common assumption that makes sense in many practical situations, where there is no clear relation or hierarchy between the classes. As a result, we will mostly ignore the role of $\mathcal{Y}$ and restrict our study of data spaces to $\mathcal{X}$.

Data and model spaces are rarely Euclidean spaces. In data space, there is rarely any meaning to adding or subtracting two data points or multiplying a data point by a real scalar. For example, most representations of text documents as vectors are non-negative and multiplying them by a negative scalar does not yield another document. Similarly, images I are usually represented as matrices whose entries are in some bounded interval of the real line $I \in [a, b]^{k \times m}$ and there is no meaning to matrices with values outside that range that are obtained by addition or scalar multiplication. The situation is similar in model spaces. For example, typical parametric families such as normal, exponential, Dirichlet or multinomial, as well as the set of non-negative, normalized distributions are not Euclidean spaces.

In addition to the fact that data and model spaces are rarely $\mathbb{R}^n$ as topological spaces, the geometric structure of Euclidean spaces, expressed through the Euclidean distance

$$\|x - y\| \stackrel{\text{def}}{=} \sqrt{\sum_i |x_i - y_i|^2}$$

is artificial on most data and model spaces. This holds even in many cases when the data or models are real vectors. To study the geometry of $\mathcal{X}$ and Θ, it is essential to abandon the realm of Euclidean geometry in favor of a new, more flexible class of geometries. The immediate generalization of Euclidean spaces, Banach and Hilbert spaces, are still vector spaces, and by the arguments above, are not a good model for $\mathcal{X}$ and Θ. Furthermore, the geometries of Banach and Hilbert spaces are quite restricted, as is evident from the undesirable linear scaling of the distance

$$d_E(cx, cy) = |c|\, d_E(x, y) \qquad \forall c \in \mathbb{R}.$$

Despite the fact that most data and model spaces are not Euclidean, they share two important properties: they are smooth and they are locally Euclidean. Manifolds are the natural generalization of Euclidean spaces to locally Euclidean spaces and differentiable manifolds are their smooth counterparts. Riemannian geometry is a mathematical theory of geometries on such smooth, locally Euclidean spaces. In this framework, the geometry of a space is specified by a local inner product $g_x(\cdot,\cdot), x \in \mathcal{X}$ between tangent vectors called the Riemannian metric. This inner product translates into familiar geometric quantities such as distance, curvature and angles. Using the Riemannian geometric approach to study the geometries of $\mathcal{X}$ and Θ allows us to draw upon the vast mathematical literature in this topic. Furthermore, it is an adequate framework as it encompasses most commonly used geometries in machine learning. For example, Euclidean geometry on $\mathcal{X} = \mathbb{R}^n$ is achieved by setting the local metric $g_x(u,v), x \in \mathcal{X}$ to be the Euclidean inner product

$$\delta_x(u,v) \stackrel{\text{def}}{=} \langle u, v \rangle \stackrel{\text{def}}{=} \sum_{i=1}^{n} u_i v_i.$$

The information geometry on a space of statistical models $\Theta \subset \mathbb{R}^n$ is achieved by setting the metric $g_\theta(u,v), \theta \in \Theta$ to be the Fisher information $\mathcal{J}_\theta(u,v)$

$$\mathcal{J}_\theta(u,v) \stackrel{\text{def}}{=} \sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_j \int p(x\,;\theta)\,\frac{\partial}{\partial\theta_i}\log p(x\,;\theta)\,\frac{\partial}{\partial\theta_j}\log p(x\,;\theta)\,\mathrm{d}x \tag{1}$$

where the above integral is replaced with a sum if $\mathcal{X}$ is discrete.

This thesis is concerned with applying tools from Riemannian geometry to study the relationship between statistical learning algorithms and different geometries of $\mathcal{X}$ and Θ. As a result, we gain considerable insight into current learning algorithms and we are able to design powerful new techniques that often outperform the current state-of-the-art.

We start the thesis with Section 2 that contains background from Riemannian geometry that is relevant to most subsequent sections. Additional background that is relevant to a specific section will appear in that section alone. The treatment in Section 2 is short and often not rigorous. For a more complete description most textbooks in Riemannian geometry will suffice. Section 3 gives an overview of relevant research in the interface of machine learning and statistics, and Riemannian geometry and Section 4 applies the mathematical theory of Section 2 to spaces of probability models.

In Section 5 we study the geometry of the space Θ of conditional models underlying the algorithms logistic regression and AdaBoost. We prove the surprising result that both algorithms solve the same primal optimization problem with the only difference being that AdaBoost lacks normalization constraints, hence resulting in non-normalized models. Furthermore, we show that both algorithms implicitly minimize the conditional I-divergence

$$D(p,q) = \sum_{i=1}^{n}\sum_{y}\left(p(y|x_i)\log\frac{p(y|x_i)}{q(y|x_i)} - p(y|x_i) + q(y|x_i)\right)$$

to a uniform model q. Despite the fact that the I-divergence is not a metric distance function, it is related to a distance under a specific geometry on Θ. This geometry is the product Fisher information geometry whose study is pursued in Section 6.

By generalizing the theorems of Čencov and Campbell, Section 6 shows that the only geometry on the space of conditional distributions consistent with a basic set of axioms is the product Fisher information. The results of Sections 5 and 6, provide an axiomatic characterization of the geometries underlying logistic regression and AdaBoost. Apart from providing a substantial new understanding of logistic regression and AdaBoost, this analysis provides, for the first time a theory of information geometry for conditional models.

The axiomatic framework mentioned above provides a natural geometry on the space of distributions Θ. It is less clear what should be an appropriate geometry for the data space $\mathcal{X}$. A common pitfall shared by many classifiers is to assume that $\mathcal{X}$ should be endowed with a Euclidean geometry. Many algorithms, such as radial basis machines and Euclidean k-nearest neighbor, make this assumption explicit. On the other hand, Euclidean geometry is implicitly assumed in linear classifiers such as logistic regression, linear support vector machines, boosting and the perceptron. Careful selection of an appropriate geometry for $\mathcal{X}$ and designing classifiers that respect it should produce better results than the naive Euclidean approach.

In Section 7 we propose the following principle to obtain a natural geometry on the data space. By assuming the data is generated by statistical models in the space $\mathcal{M}$, we embed data points $x \in \mathcal{X}$ into $\mathcal{M}$ and thus obtain a natural geometry on the data space – namely the Fisher geometry on $\mathcal{M}$. For example, consider the data space $\mathcal{X}$ of text documents in normalized term frequency (tf) representation embedded in the space of all multinomial models $\mathcal{M}$. Assuming the documents are generated from multinomial distributions we obtain the maximum likelihood embedding $\hat{\theta}_{MLE} : \mathcal{X} \to \mathcal{M}$ which is equivalent to the inclusion map

$$\iota : \mathcal{X} \hookrightarrow \mathcal{M}, \qquad \iota(x) = x.$$

Turning to Čencov's theorem and selecting the Fisher geometry on $\mathcal{M}$ we obtain a natural geometry on the closure of the data space $\overline{\mathcal{X}} = \mathcal{M}$.

The embedding principle leads to a general framework for adapting existing algorithms to the Fisher geometry. In Section 9 we generalize the notion of linear classifiers to non-Euclidean geometries and derive in detail the multinomial analogue of logistic regression. Similarly, in Section 8 we generalize radial basis kernel to non-Euclidean geometries by approximating the solution of the geometric diffusion equation. In both cases, the resulting non-Euclidean generalizations outperform its Euclidean counterpart, as measured by classification accuracy in several text classification tasks.

The Fisher geometry is a natural choice if the only known information is the statistical family that generates the data. In the presence of actual data $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ it may be possible to induce a geometry that is better suited for it. In Section 10 we formulate a learning principle for

the geometry of $\mathcal{X}$ that is based on maximizing the inverse volume of the given training set. When applied to the space of text documents in tf representation, the learned geometry is similar to, but outperforms the popular tf-idf geometry.

We conclude with a discussion in Section 11. The first two appendices contain technical information relevant to Sections 5 and 10. Appendix D contains a summary of the major contributions included in this thesis, along with relevant publications.

2 Relevant Concepts from Riemannian Geometry

In this section we describe concepts from Riemannian geometry that are relevant to most of the thesis, with the possible exception of Section 5. Other concepts from Riemannian geometry that are useful only for a specific section will be introduced later in the thesis as needed. For more details refer to any textbook discussing Riemannian geometry. Milnor (1963) and Spivak (1975) are particularly well known classics and Lee (2002) is a well-written contemporary textbook.

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.

2.1 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 $\mathbb{R}^m$. 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[1] (Lee, 2000). Unless otherwise noted, for the remainder of this section we assume that all manifolds are of dimension n.

An n-dimensional topological manifold with a boundary is defined similarly to an n-dimensional topological manifold, except that each point has a neighborhood that is homeomorphic to an open subset of the upper half plane

$$\mathbb{H}^n \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : x_n \geq 0\}.$$

[1]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.

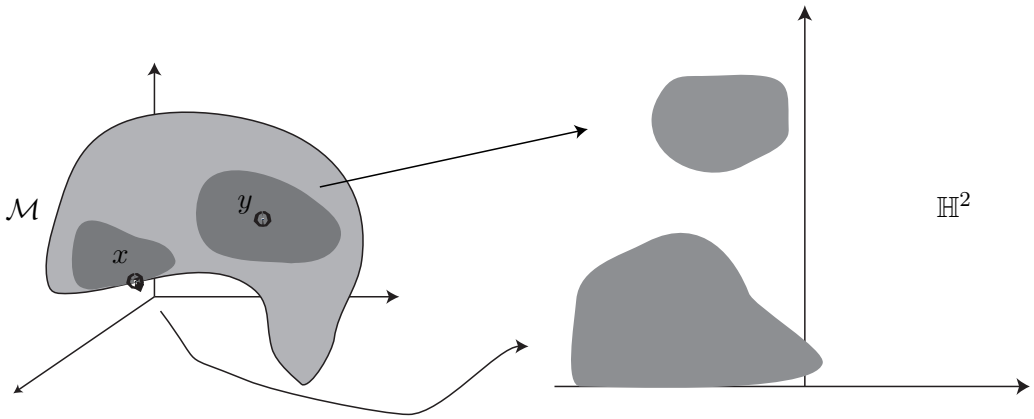

Figure 1: A 2-dimensional manifold with a boundary. The boundary $\partial\mathcal{M}$ is marked by a black contour. For example, x is a boundary point $x \in \partial\mathcal{M}$ while $y \in \operatorname{Int}\mathcal{M}$ is an interior point.

It is possible to show that in this case some points $x \in \mathcal{M}$ have neighborhoods homeomorphic to $U \subset \mathbb{H}^n$ such that $\forall y \in U, y_n > 0$ while other points are homeomorphic to a subset $U \subset \mathbb{H}^n$ that intersects the line $y_n = 0$. These two sets of points are disjoint and are called the interior and boundary of the manifold $\mathcal{M}$ and are denoted by $\operatorname{Int}\mathcal{M}$ and $\partial\mathcal{M}$ respectively (Lee, 2000).

Figure 1 illustrates the concepts associated with a manifold with a boundary. Note that a manifold is a manifold with a boundary but the converse does not hold in general. However, if $\mathcal{M}$ is an n-dimensional manifold with a boundary then $\operatorname{Int}\mathcal{M}$ is an n dimensional manifold and $\partial\mathcal{M}$ is an $n-1$ dimensional manifold. The above definition of boundary and interior of a manifold may differ from the topological notions of boundary and interior, associated with an ambient topological space. When in doubt, we will specify whether we refer to the manifold or topological interior and boundary. We return to manifolds with boundary at the end of this Section.

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 2-dimensional manifold are illustrated in Figure 2.

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

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

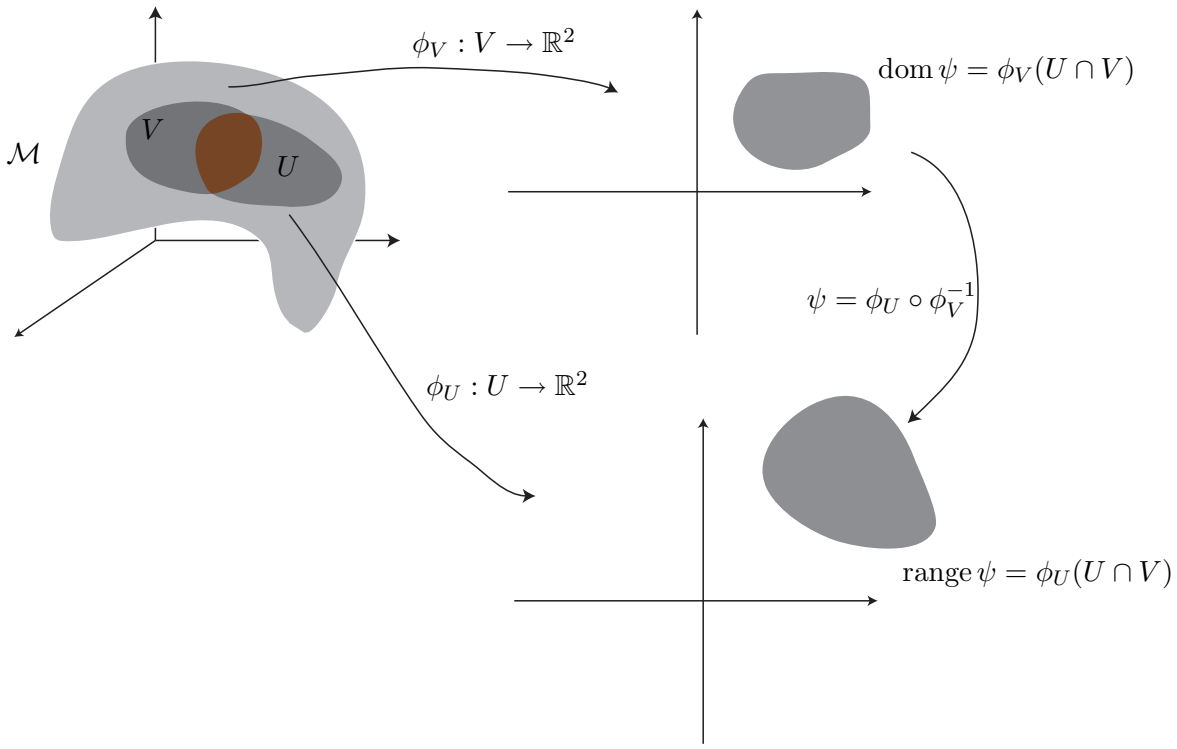

Figure 2: Two neighborhoods U, V in a 2-dimensional manifold $\mathcal{M}$, the coordinate charts ϕ_U, ϕ_V and the transition function ψ between them.

its positive orthant

$$\mathbb{S}_+^n \stackrel{\text{def}}{=} \left\{ x \in \mathbb{R}^{n+1} : \sum_{i=1}^{n} x_i^2 = 1,\ \forall i\ x_i > 0 \right\}, \tag{3}$$

and the n-simplex

$$\mathbb{P}_n \stackrel{\text{def}}{=} \left\{ x \in \mathbb{R}^{n+1} : \sum_{i=1}^{n} x_i = 1,\ \forall i\ x_i > 0 \right\}. \tag{4}$$

We will keep these examples in mind as they will keep appearing throughout the thesis.

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

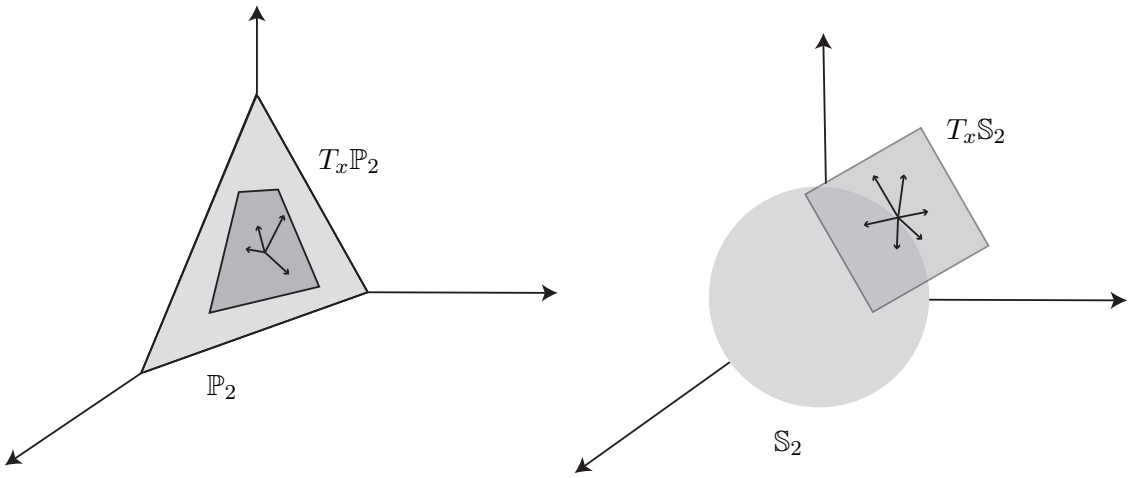

Figure 3: Tangent spaces of the 2-simplex $T_x\mathbb{P}_2$ and the 2-sphere $T_x\mathbb{S}_2$.

2.2 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 $v \in T_x\mathcal{M}$, 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 (Spivak, 1975; Lee, 2002). Intuitively, tangent spaces and tangent vectors 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 Figure 3 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 $\mathbb{S}^n$ defined in (2),(4) 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 Figure 3)

$$T_x\mathbb{P}_n = \left\{ v \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} v_i = 0 \right\} \qquad T_x\mathbb{S}_n = \left\{ v \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} v_i x_i = 0 \right\}. \tag{5}$$

A C^∞ vector field[2] 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 $\mathfrak{X}(\mathcal{M})$ 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 $X \in \mathfrak{X}(\mathcal{M})$ 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.

2.3 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, bi-linear and positive definite

$$\begin{aligned} g_x(u, v) &= g_x(v, u) \\ g_x\left(\sum_{i=1}^n u_i, \sum_{i=1}^n v_i\right) &= \sum_{i=1}^n \sum_{j=1}^n g_x(u_i, v_j) \\ g_x(u, u) &\geq 0 \\ g_x(u, u) &= 0 \Leftrightarrow u = 0 \end{aligned}$$

and is also C^∞ differentiable in x. By the bi-linearity 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 describes 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

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

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

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

[2]The precise definition of a C^∞ vector field requires the definition of the tangent bundle. We do not give this definition since it is somewhat technical and does not contribute much to our discussion.

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. If the manifold in question is clear from the context, we will remove the subscript and use d for the geodesic distance.

A manifold is said to be geodesically complete if any geodesic curve $c(t)$, $t \in [a, b]$, can be extended to be defined for all $t \in \mathbb{R}$. It can be shown, that the following are equivalent

- $(\mathcal{M}, g)$ is geodesically complete
- d_g is a complete metric on $\mathcal{M}$
- closed and bounded subsets of $\mathcal{M}$ are compact.

In particular, compact manifolds are geodesically complete. The Hopf-Rinow theorem asserts that if $\mathcal{M}$ is geodesically complete, then any two points can be joined by a geodesic.

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[4]

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*

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

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. 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 Figure 4, 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*

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

[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.

[4]The push-forward and pull-back maps may be defined more generally using category theory as covariant and contravariant functors (Lee, 2000).

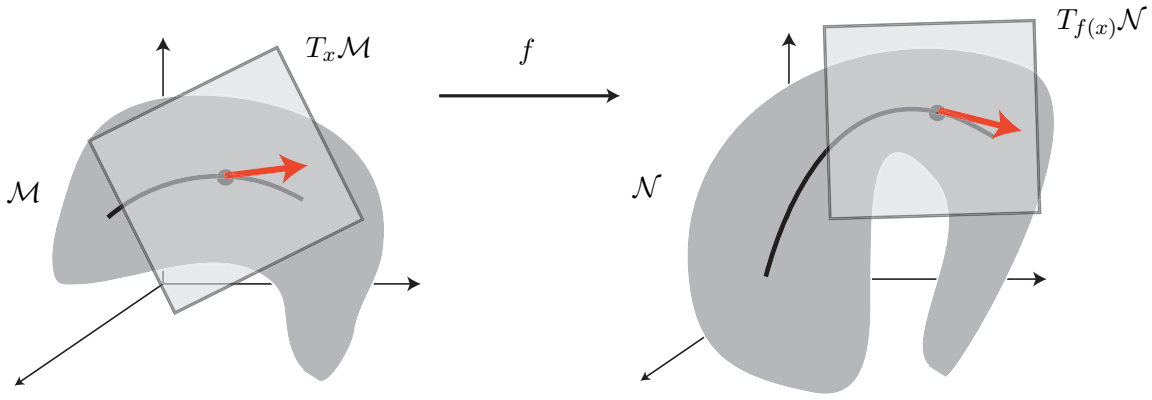

Figure 4: The map $f : \mathcal{M} \to \mathcal{N}$ defines a push forward map $f_* : T_x\mathcal{M} \to T_{f(x)}\mathcal{N}$ that transforms velocity vectors of curves to velocity vectors of the transformed curves.

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*

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

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.

A smooth and Riemannian structure may be defined over a topological manifold with a boundary as well. The definition is a relatively straightforward extension using the notion of differentiability of maps between non-open sets in $\mathbb{R}^n$ (Lee, 2002). Manifolds with a boundary are important for integration and boundary value problems; our use of them will be restricted to Section 8.

In the following section we discuss some previous work and then proceed to examine manifolds of probability distributions and their Fisher geometry.

3 Previous Work

Connections between statistics and Riemannian geometry have been discovered and studied for over fifty years. We give below a roughly chronological overview of this line of research. The overview presented here is not exhaustive. It is intended to review some of the influential research that is closely related to this thesis. For more information on this research area consult the monographs (Murray & Rice, 1993; Kass & Voss, 1997; Amari & Nagaoka, 2000). Additional previous work that is related to a specific section of this thesis will be discussed inside that section.

Rao (1945) was the first to point out that a statistical family can be considered as a Riemannian manifold with the metric determined by the Fisher information quantity. Efron (1975) found a relationship between the geometric curvature of a statistical model and Fisher and Rao's theory of second order efficiency. In this sense, a model with small curvature enjoys nice asymptotic properties and a model with high curvature implies a break-down of these properties. Since Efron's result marks a historic breakthrough of geometry in statistics we describe it below in some detail.

Recall that by Cramér-Rao lower bound, the variance of unbiased estimators is bounded from below by the inverse Fisher information. More precisely, in matrix form we have that $\forall\theta$, $V(\theta) - G^{-1}(\theta)$ is positive semi-definite, where V is the covariance matrix of the estimator and G is the Fisher information matrix[5]. Estimators that achieve this bound asymptotically, for all θ, are called asymptotically efficient, or first order efficient. The prime example for such estimators, assuming some regularity conditions, is the maximum likelihood estimator. The most common example of statistical families for which the regularity conditions hold is the exponential family. Furthermore, in this case, the MLE is a sufficient statistic. These nice properties of efficiency and sufficiency of the MLE do not hold in general for non-exponential families.

The term second order efficiency was coined by Rao, and refers to a subset of consistent and first order efficient estimators $\hat{\theta}(x_1, \ldots, x_n)$ that attain equality in the following general inequality

$$\lim_{n\to\infty} n(i(\theta^{\text{true}}) - i(\hat{\theta}(x_1, \ldots, x_n))) \geq i(\theta^{\text{true}})\gamma^2(\theta^{\text{true}}) \tag{6}$$

where i is the one-dimensional Fisher information and γ some function that depends on θ. Here $x_1, \ldots, x_n$ are sampled iid from $p(x\,;\theta^{\text{true}})$ and $\hat{\theta}(x_1, \ldots, x_n)$ is an estimate of θ^{true} generated by the estimator $\hat{\theta}$. The left hand side of (6) may be interpreted as the asymptotic rate of the loss of information incurred by using the estimated parameter rather than the true parameter. It turns out that the only consistent asymptotically efficient estimator that achieves equality in (6) is the MLE, thereby giving it a preferred place among the class of first order efficient estimators.

The significance of Efron's result is that he identified the function γ in (6) as the Riemannian curvature of the statistical manifold with respect to the exponential connection. Under this connection, exponential families are flat and their curvature is 0. For non-exponential families the curvature γ may be interpreted as measuring the breakdown of asymptotic properties which is surprisingly similar to the interpretation of curvature as measuring the deviation from flatness, expressing in our case an exponential family. Furthermore, Efron showed that the variance of the MLE in non-exponential families exceeds the Cramér-Rao lower bound in approximate proportion to $\gamma^2(\theta)$.

Dawid (1975) points out that Efron's notion of curvature is based on a connection that is not the natural one with respect to the Fisher geometry. A similar notion of curvature may be defined for other connections, and in particular for the Riemannian connection that is compatible with the

[5]The matrix form of the Cramér-Rao lower bound may be written as $V(\theta) - G^{-1}(\theta) \succeq 0$ with equality $V(\theta) = G^{-1}(\theta)$ if the bound is attained.

Fisher information metric. Dawid's comment about the possible statistical significance of curvature with respect to the metric connection remains a largely open question, although some results were obtained by Brody and Houghston (1998). Čencov (1982) introduced a family of connections, later parameterized by α, that include as special cases the exponential connection and the metric connection. Using Amari's parametrization of this family, $\alpha = 1$ corresponds to the exponential connection, $\alpha = 0$ corresponds to the metric connection and the $\alpha = -1$ corresponds to the mixture connection, under which mixture families enjoy 0 curvature (Amari & Nagaoka, 2000).

Čencov (1982) proved that the Fisher information metric is the only Riemannian metric that is preserved under basic probabilistic transformations. These transformations, called congruent embeddings by a Markov morphism, represent transformations of the event space that is equivalent to extracting a sufficient statistic. Later on, Campbell (1986) extended Čencov's result to non-normalized positive models, thus axiomatically characterizing a set of metrics on the positive cone $\mathbb{R}^n_+$.

In his short note Dawid (1977) extended these ideas to infinite dimensional manifolds representing non-parametric sets of densities. More rigorous studies include (Lafferty, 1988) that models the manifold of densities on a Hilbert space and (Pistone & Sempi, 1995) that models the same manifold on non-reflexive Banach spaces called Orlicz spaces. This latter approach, that does not admit a Riemannian structure, was further extended by Pistone and his collaborators and by Grasselli (2001).

Additional research by Barndorff-Nielsen and others considered the connection between geometry and statistics from a different angle. Below is a brief description of some these results. In (Barndorff-Nilsen, 1986) the expected Fisher information is replaced with the observed Fisher information to provide an alternative geometry for a family of statistical models. The observed geometry metric is useful in obtaining approximate expansion of the distribution of the MLE, conditioned on an ancillary statistic. This result continues previous research by Efron and Amari that provides a geometric interpretation for various terms appearing in the Edgeworth expansion of the distribution of the MLE for curved exponential models. Barndorff-Nilsen and Blæsild (1983) studied the relation between certain partitions of an exponential family of models and geometric constructs. The partitions, termed affine dual foliations, refers to a geometric variant of the standard division of $\mathbb{R}^n$ into copies of $\mathbb{R}^{n-1}$. Based on differential geometry, Barndorff-Nilsen and Blæsild (1993) define a set of models called orthogeodesic models that enjoy nice higher-order asymptotic properties. Orthogeodesic models include exponential models with dual affine foliations as well as transformation models and provides an abstract unifying framework for such models.

The geodesic distance under the Fisher metric has been examined in various statistical studies. It is used in statistical testing and estimation as an alternative to Kullback Leibler or Jeffreys divergence (Atkinson & Mitchell, 1981). It is also essentially equivalent to the popular Hellinger distance (Beran, 1977) that plays a strong role in the field of statistical robustness (Tamura & Boos, 1986; Lindsay, 1994; Cutler & Cordero-Brana, 1996).

In a somewhat different line of research, Csiszár (1975) studied the geometry of probability distributions through the notion of I-divergence. In a later paper Csiszár (1991) showed that I-divergence estimation, along with least squares, enjoy nice axiomatic frameworks. While not a distance measure, the I-divergence bears close connection to the geodesic distance under the Fisher information metric (Kullback, 1968). This fact, together with the prevalence of the I-divergence and its special case the Kullback-Leibler divergence, brings these research directions together under the umbrella of information geometry. Amari and Nagaoka (2000) contains some further details on the connection between I-divergence and Fisher geometry.

Information geometry arrived somewhat later in the machine learning literature. Most of the studies in this context were done by Amari's group. Amari examines several geometric learning algorithms for neural networks (Amari, 1995) and shows how to adapt the gradient descent algorithm to information geometry in the context of neural networks (Amari, 1998) and independent component analysis (ICA) (Amari, 1999). Ikeda et al. (2004) interprets several learning algorithms in graphical models such as belief propagation using information geometry and introduces new variations on them. Further information on Amari's effort in different applications including information theory and quantum estimation theory may be obtained from (Amari & Nagaoka, 2000). Saul and Jordan (1997) interpolates between different models based on differential geometric principles and Jaakkola and Haussler (1998) use the Fisher information to enhance a discriminative classifier with generative qualities. Gous (1998) and Hall and Hofmann (2000) use information geometry to represent text documents by affine subfamilies of multinomial models.

In the next section we apply the mathematical framework developed in Section 2 to manifolds of probability models.

4 Geometry of Spaces of Probability Models

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

$$\begin{aligned}\mathcal{J}_\theta(u,v) &\stackrel{\text{def}}{=} \sum_{i,j} u_i v_j \int p(x\,;\theta) \frac{\partial}{\partial\theta_i} \log p(x\,;\theta) \frac{\partial}{\partial\theta_j} \log p(x\,;\theta) \mathrm{d}x \\ &= \sum_{i,j} u_i v_j \; E\left\{ \frac{\partial \log p(x\,;\theta)}{\partial\theta_i} \frac{\partial \log p(x\,;\theta)}{\partial\theta_j} \right\}. \end{aligned} \tag{7}$$

If $\mathcal{X}$ is discrete the above integral is replaced with a sum. An equivalent form of (7) for normalized distributions that is sometimes easier to compute is

$$\begin{aligned}
\mathcal{J}_\theta(u,v) &= \mathcal{J}_\theta(u,v) - \sum_{ij} u_i v_j \frac{\partial^2}{\partial\theta_i \partial\theta_j} \int p(x\,;\theta)\mathrm{d}x \\
&= \sum_{ij} u_i v_j \int p(x\,;\theta) \left(\left(\frac{1}{p(x\,|\,\theta)} \frac{\partial p(x\,;\theta)}{\partial \theta_j} \right) \left(\frac{1}{p(x\,|\,\theta)} \frac{\partial p(x\,;\theta)}{\partial \theta_i} \right) - \frac{1}{p(x\,;\theta)} \frac{\partial^2}{\partial\theta_i \partial\theta_j} p(x\,;\theta) \right) \mathrm{d}x \\
&= -\sum_{ij} u_i v_j \int p(x\,;\theta) \frac{\partial}{\partial\theta_j} \frac{1}{p(x\,|\,\theta)} \frac{\partial p(x\,;\theta)}{\partial\theta_i} \,\mathrm{d}x \\
&= -\sum_{ij} u_i v_j \int p(x\,;\theta) \frac{\partial^2}{\partial\theta_j \partial\theta_i} \log p(x\,;\theta)\,\mathrm{d}x \\
&= \sum_{ij} u_i v_j E\left\{ -\frac{\partial^2}{\partial\theta_j \partial\theta_i} \log p(x\,;\theta) \right\} \qquad (8)
\end{aligned}$$

assuming that we can change the order of the integration and differentiation operators.

In the remainder of this section we examine in detail a few important Fisher geometries. The Fisher geometries of finite dimensional non-parametric space, finite dimensional conditional non-parametric space and spherical normal space are studied next.

4.1 Geometry of Non-Parametric Spaces

In the finite non-parametric setting, the event space $\mathcal{X}$ is a finite set with $|\mathcal{X}| = n$ and $\Theta = \mathbb{P}_{n-1}$, defined in (4), which represents the manifold of all positive probability models over $\mathcal{X}$. 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 (Lee, 2002). Note that the above space Θ is precisely the parametric space of the multinomial family. Hence, the results of this section may be interpreted with respect to the space of all positive distributions on a finite event space, or with respect to the parametric space of the multinomial distribution.

The finiteness of $\mathcal{X}$ 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 vector space called the model space. Such manifolds are called Frechet, Banach or Hilbert manifolds (depending on the model space) and are the topic of a branch of geometry called global analysis (Lang, 1999). Dawid (1977) remarked that an infinite dimensional non-parametric space may be endowed with multinomial geometry leading to spherical geometry on a Hilbert manifold. More rigorous modeling attempts were made by Lafferty (1988) that models the manifold of densities on a Hilbert space and by Pistone and Sempi (1995) that model it on a non-reflexive Banach space. See also the brief discussion on infinite dimensional manifolds representing densities by Amari and Nagaoka (2000) pp. 44-45.

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 (5), this results in the following representation of

$v \in T_\theta \mathbb{P}_{n-1}$

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

Using this representation, the loglikelihood and its derivatives are

$$\begin{aligned} \log p(x\,;\theta) &= \sum_{i=1}^n x_i \log \theta_i \\ \frac{\partial \log p(x\,;\theta)}{\partial \theta_i} &= \frac{x_i}{\theta_i} \\ \frac{\partial^2 \log p(x\,;\theta)}{\partial \theta_i \partial \theta_j} &= -\frac{x_i}{\theta_i^2}\,\delta_{ij} \end{aligned}$$

and using equation (8) the Fisher information metric on $\mathbb{P}_{n-1}$ becomes

$$\mathcal{J}_\theta(u,v) = -\sum_{i=1}^n \sum_{j=1}^n u_i v_j E\left[\frac{\partial^2 \log p(x\,|\,\theta)}{\partial \theta_i \partial \theta_j}\right] = -\sum_{i=1}^n u_i v_i E\left\{-x_i/\theta_i^2\right\} = \sum_{i=1}^n \frac{u_i v_i}{\theta_i}$$

since $Ex_i = \theta_i$. Note that the Fisher metric emphasizes coordinates that correspond to low probabilities. The fact that the metric $\mathcal{J}_\theta(u,v) \to \infty$ when $\theta_i \to 0$ is not problematic since length of curves that involves integrals over g converge.

While geodesic distances are difficult to compute in general, in the present case we can easily compute the geodesics by observing that the standard Euclidean metric on the surface of the positive n-sphere is the pull-back of the Fisher information metric on the simplex. More precisely, the transformation $F(\theta_1, \ldots, \theta_{n+1}) = (2\sqrt{\theta_1}, \ldots, 2\sqrt{\theta_{n+1}})$ is a diffeomorphism of the n-simplex $\mathbb{P}_n$ onto the positive portion of the n-sphere of radius 2

$$\tilde{\mathbb{S}}_+^n = \left\{\theta \in \mathbb{R}^{n+1} \;:\; \sum_{i=1}^{n+1} \theta_i^2 = 4,\ \theta_i > 0\right\}.$$

The inverse transformation is

$$F^{-1} : \tilde{\mathbb{S}}_+^n \to \mathbb{P}_n, \qquad F^{-1}(\theta) = \left(\frac{\theta_1^2}{4}, \ldots, \frac{\theta_{n+1}^2}{4}\right)$$

and its push-forward is

$$F_*^{-1}(u) = \left(\frac{u_1}{2}, \ldots, \frac{u_{n+1}}{2}\right).$$

The metric on $\tilde{\mathbb{S}}_+^n$ obtained by pulling back the Fisher information on $\mathbb{P}_n$ through F^{-1} is

$$\begin{aligned} h_\theta(u,v) &= \mathcal{J}_{\theta^2/4}\left(F_*^{-1}\sum_{k=1}^{n+1} u_k e_k, F_*^{-1}\sum_{l=1}^{n+1} v_l e_l\right) = \sum_{k=1}^{n+1}\sum_{l=1}^{n+1} u_k v_l\, g_{\theta^2/4}(F_*^{-1}e_k, F_*^{-1}e_l) \\ &= \sum_{k=1}^{n+1}\sum_{l=1}^{n+1} u_k v_l \sum_i \frac{4}{\theta_i^2}\,(F_*^{-1}e_k)_i\,(F_*^{-1}e_l)_i = \sum_{k=1}^{n+1}\sum_{l=1}^{n+1} u_k v_l \sum_i \frac{4}{\theta_i^2}\,\frac{\theta_k\delta_{ki}}{2}\,\frac{\theta_l\delta_{li}}{2} = \sum_{i=1}^{n+1} u_i v_i \\ &= \delta_\theta(u,v) \end{aligned}$$

the Euclidean metric on $\tilde{\mathbb{S}}_+^n$ inherited from the embedding Euclidean space $\mathbb{R}^{n+1}$.

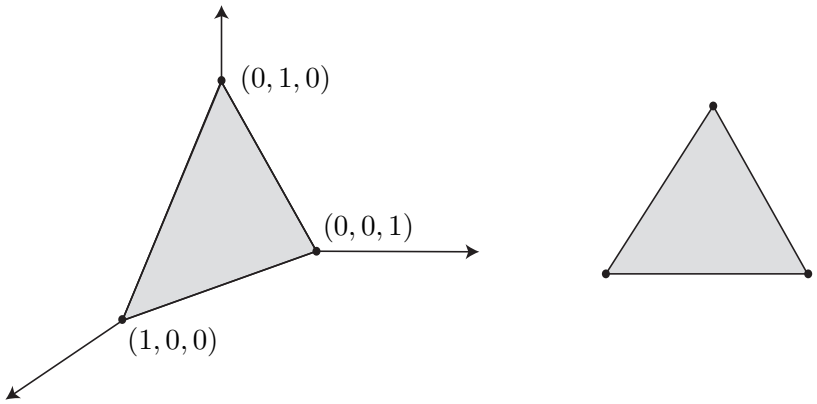

Figure 5: The 2-simplex $\mathbb{P}_2$ may be visualized as a surface in $\mathbb{R}^3$ (left) or as a triangle in $\mathbb{R}^2$ (right).

Since the transformation $F : (\mathbb{P}_n, \mathcal{J}) \to (\tilde{\mathbb{S}}^n_+, \delta)$ is an isometry, the geodesic distance $d_{\mathcal{J}}(\theta, \theta')$ on $\mathbb{P}_n$ may be computed as the shortest curve on $\tilde{\mathbb{S}}^n_+$ connecting $F(\theta)$ and $F(\theta')$. These shortest curves are portions of great circles – the intersection of a two dimensional subspace and $\tilde{\mathbb{S}}^n_+$ whose lengths are

$$d_{\mathcal{J}}(\theta, \theta') = d_\delta(F(\theta), F(\theta')) = 2 \arccos \left(\sum_{i=1}^{n+1} \sqrt{\theta_i \, \theta'_i} \right) . \tag{9}$$

We illustrate these geodesic distances in Figures 5-6. Figure 5 shows how to picture $\mathbb{P}_2$ as a triangle in $\mathbb{R}^2$ and Figure 6 shows the equal distant contours for both Euclidean and Fisher geometries. Will often ignore the factor of 2 in (9) to obtain a more compact notation for the geodesic distance.

The geodesic distances $d_{\mathcal{J}}(\theta, \theta')$ under the Fisher geometry and the Kullback-Leibler divergence $D(\theta, \theta')$ agree up to second order as $\theta \to \theta'$ (Kullback, 1968). Similarly, the Hellinger distance (Beran, 1977)

$$d_H(\theta, \theta') = \sqrt{\sum_i \left(\sqrt{\theta_i} - \sqrt{\theta'_i} \right)^2} \tag{10}$$

is related to $d_{\mathcal{J}}(\theta, \theta')$ by

$$d_H(\theta, \theta') = 2 \sin \left(d_{\mathcal{J}}(\theta, \theta')/4 \right) \tag{11}$$

and thus also agrees with the distance up to second order as $\theta' \to \theta$.

4.2 Geometry of Non-Parametric Conditional Spaces

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

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

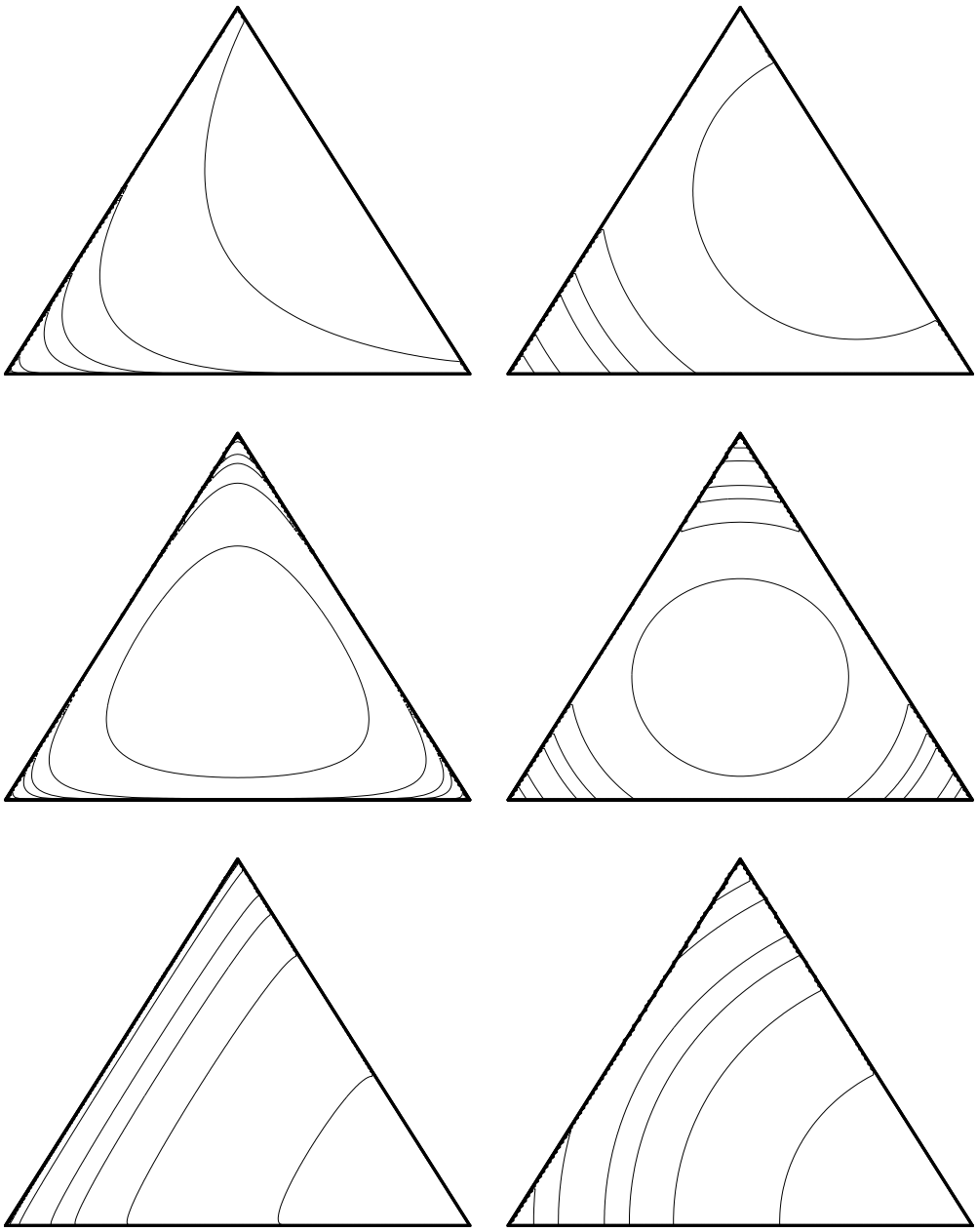

Figure 6: Equal distance contours on $\mathbb{P}_2$ from the upper right edge (top row), the center (center row), and lower right corner (bottom row). The distances are computed using the Fisher information metric (left column) or the Euclidean metric (right column).

For our purposes, it will be sometimes more convenient to work with the more general case of positive non-normalized 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 non-normalized one, we can consider $\mathbb{P}_{m-1}^k$ to be a subset of $\mathbb{R}_+^{k\times m}$. Results obtained for non-normalized models apply then to normalized models as a special case. In addition, some of the notation and formulation is simplified by working with non-normalized models.

In the interest of simplicity, we will often use matrix notation instead of the standard probabilistic notation. A conditional model (either normalized or non-normalized) 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

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

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 sums of the combination coefficients over each row are 0, e.g.

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

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 $u, v \in T_M\mathbb{R}_+^{k\times m}$ or $u, v \in T_M\mathbb{P}_{m-1}^k$ the product Fisher information

$$\begin{aligned}\mathcal{J}_M^k(u_1 \oplus \cdots \oplus u_k, v_1 \oplus \cdots \oplus v_k) &\stackrel{\mathrm{def}}{=} (\mathcal{J} \otimes \cdots \otimes \mathcal{J})_M(u_1 \oplus \cdots \oplus u_k, v_1 \oplus \cdots \oplus v_k)\\ &\stackrel{\mathrm{def}}{=} \sum_{i=1}^k \mathcal{J}_{[M]_i}(u_i, v_i),\end{aligned}$$

where $[A]_i$ is the i-row of the matrix A, $\otimes$ is the tensor product and $\oplus$ is the direct sum decomposition of vectors, reduces to

$$\mathcal{J}_M^k(u, v) = \sum_{i=1}^k \sum_{j=1}^m \frac{u_{ij}v_{ij}}{M_{ij}}. \tag{12}$$

A different way of expressing (12) is by specifying the values of the metric on pairs of basis elements

$$g_M(\partial_{ab}, \partial_{cd}) = \delta_{ac}\delta_{bd}\frac{1}{M_{ab}} \tag{13}$$

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

4.3 Geometry of Spherical Normal Spaces

Given a restricted parametric family $\Theta \subset \mathbb{P}_n$ the Fisher information metric on Θ agrees with the induced metric from the Fisher information metric on $\mathbb{P}_n$. When $\mathcal{X}$ is infinite and Θ is a finite dimensional parametric family, we can still define the Fisher information metric $\mathcal{J}$ on Θ, however without a reference to an embedding non-parametric space. We use this approach to consider the Fisher geometry of the spherical normal distributions on $\mathcal{X} = \mathbb{R}^{n-1}$

$$\{\mathcal{N}(\mu, \sigma I) : \mu \in \mathbb{R}^{n-1}, \sigma \in \mathbb{R}_+\}$$

parameterized by the upper half plane $\Theta = \mathbb{H}^n \cong \mathbb{R}^{n-1} \times \mathbb{R}_+$.

To compute the Fisher information metric for this family, it is convenient to use the expression given by equation (8). Then simple calculations yield, for $1 \leq i, j \leq n-1$

$$\begin{aligned}
[G(\theta)]_{ij} &= -\int_{\mathbb{R}^{n-1}} \frac{\partial^2}{\partial \mu_i \partial \mu_j} \left(-\sum_{k=1}^{n-1} \frac{(x_k - \mu_k)^2}{2\sigma^2} \right) p(x \,|\, \theta)\, dx = \frac{1}{\sigma^2}\, \delta_{ij} \\
[G(\theta)]_{ni} &= -\int_{\mathbb{R}^{n-1}} \frac{\partial^2}{\partial \sigma \partial \mu_i} \left(-\sum_{k=1}^{n-1} \frac{(x_k - \mu_k)^2}{2\sigma^2} \right) p(x \,|\, \theta)\, dx = \frac{2}{\sigma^3} \int_{\mathbb{R}^{n-1}} (x_i - \mu_i)\, p(x \,|\, \theta)\, dx = 0 \\
[G(\theta)]_{nn} &= -\int_{\mathbb{R}^{n-1}} \frac{\partial^2}{\partial \sigma \partial \sigma} \left(-\sum_{k=1}^{n-1} \frac{(x_k - \mu_k)^2}{2\sigma^2} - (n-1) \log \sigma \right) p(x \,|\, \theta)\, dx \\
&= \frac{3}{\sigma^4} \int_{\mathbb{R}^{n-1}} \sum_{k=1}^{n-1} (x_k - \mu_k)^2\, p(x \,|\, \theta)\, dx - \frac{n-1}{\sigma^2} = \frac{2(n-1)}{\sigma^2}.
\end{aligned}$$

Letting θ' be new coordinates defined by $\theta'_i = \mu_i$ for $1 \leq i \leq n-1$ and $\theta'_n = \sqrt{2(n-1)}\, \sigma$, we see that the Gram matrix is given by

$$[G(\theta')]_{ij} = \frac{1}{\sigma^2}\, \delta_{ij} \tag{14}$$

and the Fisher information metric gives the spherical normal manifold a hyperbolic geometry[6]. It is shown by Kass and Voss (1997) that any location-scale family of densities

$$p(x\,; \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma} \right) \quad (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+ \quad f : \mathbb{R} \to \mathbb{R}$$

have a similar hyperbolic geometry. The geodesic curves in the two dimensional hyperbolic space are circles whose centers lie on the line $x_2 = 0$ or vertical lines (considered as circles whose centers lie on the line $x_2 = 0$ with infinite radius) (Lee, 1997). An illustration of these curves appear in Figure 7. To compute the geodesic distance on $\mathbb{H}^n$ we transform points in $\mathbb{H}^n$ to an isometric manifold known as Poincaré's ball. We first define the sphere inversion of x with respect to a sphere S with center a and radius r as

$$I_S(x) = \frac{r^2}{\|x - a\|^2}(x - a) + a.$$

[6]The manifold $\mathbb{H}^n$ with with hyperbolic geometry is often referred to as Poincaré's upper half plane and is a space of constant negative curvature.

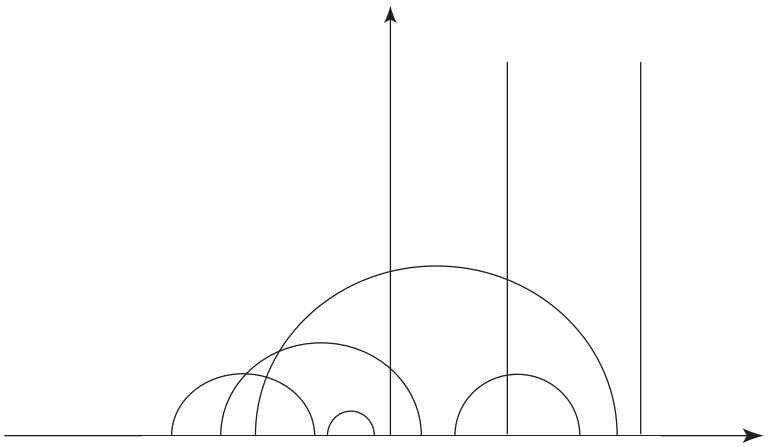

Figure 7: Geodesic curves in $\mathbb{H}^2$ with the hyperbolic metric are circles whose centers lie on the line $x_2 = 0$ or vertical lines.

The Cayley transform is the sphere inversion with respect to a sphere with center $(0, \ldots, 0, -1)$ and radius $\sqrt{2}$. We denote by η the inverse of the Cayley's transform that maps the hyperbolic half plane to Poincaré's ball.

$$\eta(x) = -I_{S'}(-x) \qquad x \in \mathbb{H}^n$$

where S' is a sphere with center at $(0, \ldots, 0, 1)$ and radius $\sqrt{2}$. The geodesic distance in $\mathbb{H}^n$ is then given by

$$d(x, y) = \operatorname{acosh}\left(1 + 2\frac{\|\eta(x) - \eta(y)\|^2}{(1 - \|\eta(x)\|^2)(1 - \|\eta(y)\|^2)}\right) \qquad x, y \in \mathbb{H}^n. \tag{15}$$

For more details see (Bridson & Haefliger, 1999) pages 86–90.

The following sections describe the main contributions of the thesis. The next two sections deal with the geometry of the model space Θ in the context of estimation of conditional models. The later sections study the geometry of the data space $\mathcal{X}$.

5 Geometry of Conditional Exponential Models and AdaBoost

Several recent papers in statistics and machine learning have been devoted to the relationship between boosting and more standard statistical procedures such as logistic regression. In spite of this activity, an easy-to-understand and clean connection between these different techniques has not emerged. Friedman et al. (2000) note the similarity between boosting and stepwise logistic regression procedures, and suggest a least-squares alternative, but view the loss functions of the two

problems as different, leaving the precise relationship between boosting and maximum likelihood unresolved. Kivinen and Warmuth (1999) note that boosting is a form of "entropy projection," and Lafferty (1999) suggests the use of Bregman distances to approximate the exponential loss. Mason et al. (1999) consider boosting algorithms as functional gradient descent and Duffy and Helmbold (2000) study various loss functions with respect to the PAC boosting property. More recently, Collins et al. (2002) show how different Bregman distances precisely account for boosting and logistic regression, and use this framework to give the first convergence proof of AdaBoost. However, in this work the two methods are viewed as minimizing different loss functions. Moreover, the optimization problems are formulated in terms of a reference distribution consisting of the zero vector, rather than the empirical distribution of the data, making the interpretation of this use of Bregman distances problematic from a statistical point of view.

In this section we present a very basic connection between boosting and maximum likelihood for exponential models through a simple convex optimization problem. In this setting, it is seen that the only difference between AdaBoost and maximum likelihood for exponential models, in particular logistic regression, is that the latter requires the model to be normalized to form a probability distribution. The two methods minimize the same I-divergence objective function subject to the same feature constraints. Using information geometry, we show that projecting the exponential loss model onto the simplex of conditional probability distributions gives precisely the maximum likelihood exponential model with the specified sufficient statistics. In many cases of practical interest, the resulting models will be identical; in particular, as the number of features increases to fit the training data the two methods will give the same classifiers. We note that throughout the thesis we view boosting as a procedure for minimizing the exponential loss, using either parallel or sequential update algorithms as presented by Collins et al. (2002), rather than as a forward stepwise procedure as presented by Friedman et al. (2000) and Freund and Schapire (1996).

Given the recent interest in these techniques, it is striking that this connection has gone unobserved until now. However in general, there may be many ways of writing the constraints for a convex optimization problem, and many different settings of the Lagrange multipliers that represent identical solutions. The key to the connection we present here lies in the use of a particular non-standard presentation of the constraints. When viewed in this way, there is no need for special-purpose Bregman divergence as in (Collins et al., 2002) to give a unified account of boosting and maximum likelihood, and we only make use of the standard I-divergence. But our analysis gives more than a formal framework for understanding old algorithms; it also leads to new algorithms for regularizing AdaBoost, which is required when the training data is noisy. In particular, we derive a regularization procedure for AdaBoost that directly corresponds to penalized maximum likelihood using a Gaussian prior. Experiments on UCI data support our theoretical analysis, demonstrate the effectiveness of the new regularization method, and give further insight into the relationship between boosting and maximum likelihood exponential models.

The next section describes an axiomatic characterization of metrics over the space of conditional models and the relationship between the characterized metric and the I-divergence. In this sense this section and the next one should be viewed as one unit as they provide an axiomatic character-

ization of the geometry underlying conditional exponential models such as logistic regression and AdaBoost.

5.1 Definitions

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets of sizes k and m and $\mathbb{P}^k_{m-1}, \mathbb{R}^{k\times m}_+$ be the sets of normalized and non-normalized conditional models as defined in Section 4. Their closures $\overline{\mathbb{P}^k_{m-1}}, \overline{\mathbb{R}^{k\times m}_+}$ represent the sets of non-negative conditional models. Let

$$f = (f_1, \ldots, f_l), \qquad f_j : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

be a sequence functions, which we will refer to as a feature vector. These functions correspond to the weak learners in boosting, and to the sufficient statistics in an exponential model. The empirical distribution associated with a training set $\{(x_i, y_i)\}_{i=1}^n$ is $\tilde{p}(x, y) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i,x}\,\delta_{y_i,y}$. Based on $\tilde{p}(x, y)$ we define marginal and conditional distributions $\tilde{p}(x), \tilde{p}(y|x)$ as usual. We assume that $\tilde{p}(x) > 0$ and that for all x there is a unique $y \in \mathcal{Y}$, denoted by $\tilde{y}(x)$, for which $\tilde{p}(y \,|\, x) > 0$. This assumption, referred to as the consistent data assumption, is made to obtain notation that corresponds to the conventions used to present boosting algorithms; it is not essential to the correspondence between AdaBoost and conditional exponential models presented here. We will use the notation $f(x, y)$ to represent the real vector $(f_1(x, y), \ldots, f_l(x, y))$ and $\langle \cdot, \cdot \rangle$ to be the usual scalar or dot product between real vectors.

The conditional exponential model $q(y \,|\, x\,; \theta)$ associated with the feature vector f is defined by

$$q(y \,|\, x\,; \theta) = \frac{e^{\langle \theta, f(x,y)\rangle}}{\sum_y e^{\langle \theta, f(x,y)\rangle}} \qquad \theta \in \mathbb{R}^l \tag{16}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard scalar product between two vectors. The maximum likelihood estimation problem is to determine a parameter vector θ that maximize the conditional log-likelihood

$$\ell(\theta) \stackrel{\text{def}}{=} \sum_{x,y} \tilde{p}(x, y) \log q(y \,|\, x\,; \theta).$$

The objective function to be minimized in the multi-label boosting algorithm AdaBoost.M2 (Collins et al., 2002) is the exponential loss given by

$$\mathcal{E}(\theta) \stackrel{\text{def}}{=} \sum_{i=1}^n \sum_{y \neq y_i} e^{\,\langle \theta, f(x_i,y) - f(x_i,y_i)\rangle}. \tag{17}$$

In the binary case $\mathcal{Y} = \{-1, +1\}$ and taking $f_j(x, y) = \frac{1}{2} y f_j(x)$ the exponential loss becomes

$$\mathcal{E}(\theta) = \sum_{i=1}^n e^{-y_i \langle \theta, f(x_i)\rangle}, \tag{18}$$

the conditional exponential model becomes the logistic model

$$q(y \,|\, x\,; \theta) \;=\; \frac{1}{1 + e^{-y\langle \theta, f(x)\rangle}}, \tag{19}$$

for which the maximum likelihood problem becomes equivalent to minimizing the logistic loss function

$$-\ell(\theta) = \sum_{i=1}^{n} \log\left(1 + e^{-y_i \langle \theta, f(x_i) \rangle}\right). \tag{20}$$

As has been often noted, the log-loss (20) and the exponential loss (18) are qualitatively different. The exponential loss (18) grows exponentially with increasing negative margin $-y\,\langle \theta, f(x) \rangle$, while the log-loss grows linearly.

5.2 Correspondence Between AdaBoost and Maximum Likelihood

We define the conditional I-divergence with respect to a distribution r over $\mathcal{X}$ as

$$D_r(p, q) \stackrel{\text{def}}{=} \sum_x r(x) \sum_y \left(p(y \,|\, x) \log \frac{p(y \,|\, x)}{q(y \,|\, x)} - p(y \,|\, x) + q(y \,|\, x)\right). \tag{21}$$

It is a non-negative measure of discrepancy between two conditional models $p, q \in \overline{\mathbb{R}_+^{k \times m}}$. If $q \notin \mathbb{R}_+^{k \times m}$, $D_r(p, q)$ may be ∞. The I divergence is not a formal distance function as it does not satisfy symmetry and the triangle inequality. In this section we will always take $r(x) = \tilde{p}(x)$ and hence we omit it from the notation and write $D(p, q) = D_{\tilde{p}}(p, q)$. For normalized conditional models, the I-divergence $D(p, q)$ is equal to the Kullback-Leibler divergence (Kullback, 1968). The formula presented here (21) is a straightforward adaptation of the non-conditional form of the I-divergence studied by Csiszár (1991). The I-divergence comes up in many applications of statistics and machine learning. See (Kullback, 1968; O'Sullivan, 1998; Amari & Nagaoka, 2000) for many examples of such connections.

We define the feasible set $\mathcal{F}(\tilde{p}, f)$ of conditional models associated with $f = (f_1, \ldots, f_l)$ and an empirical distribution $\tilde{p}(x, y)$ as

$$\mathcal{F}(\tilde{p}, f) \stackrel{\text{def}}{=} \left\{ p \in \overline{\mathbb{R}_+^{k \times m}} : \sum_x \tilde{p}(x) \sum_y p(y \,|\, x)\ (f_j(x, y) - E_{\tilde{p}}[f_j \,|\, x]) = 0,\ j = 1, \ldots, l \right\}. \tag{22}$$

Note that this set is non-empty since $\tilde{p} \in \mathcal{F}(\tilde{p}, f)$ and that under the consistent data assumption $E_{\tilde{p}}[f \,|\, x] = f(x, \tilde{y}(x))$. The feasible set represents conditional models that agree with $\tilde{p}$ on the conditional expectation of the features.

Consider now the following two convex optimization problems, labeled P_1 and P_2.

$$\begin{array}{lll} (P_1) & \textit{minimize} & D(p, q_0) \\ & \textit{subject to} & p \in \mathcal{F}(\tilde{p}, f) \end{array} \qquad\qquad \begin{array}{lll} (P_2) & \textit{minimize} & D(p, q_0) \\ & \textit{subject to} & p \in \mathcal{F}(\tilde{p}, f) \\ & & p \in \overline{\mathbb{P}_{m-1}^k}. \end{array}$$

Thus, problem P_2 differs from P_1 only in that the solution is required to be normalized. As we will show, the dual problem P_1^* corresponds to AdaBoost, and the dual problem P_2^* corresponds to maximum likelihood for exponential models.

This presentation of the constraints is the key to making the correspondence between AdaBoost and maximum likelihood. The constraint $\sum_x \tilde{p}(x) \sum_y p(y \mid x) f(x,y) = E_{\tilde{p}}[f]$, which is the usual presentation of the constraints for maximum likelihood (as dual to maximum entropy), doesn't make sense for non-normalized models, since the two sides of the equation may not be on the same scale. Note further that attempting to re-scale by dividing by the mass of p to get

$$\sum_x \tilde{p}(x) \frac{\sum_y p(y \mid x) f(x,y)}{\sum_y p(y \mid x)} = E_{\tilde{p}}[f]$$

would yield nonlinear constraints.

Before we continue, we recall the dual formulation from convex optimization. For mode details refer for example to Section 5 of (Boyd & Vandenberghe, 2004). Given a convex optimization problem

$$\min_{x \in \mathbb{R}^n} f_0(x) \quad \text{subject to} \quad h_i(x) = 0 \quad i = 1, \ldots, r \tag{23}$$

the Lagrangian is defined as

$$\mathcal{L}(x, \theta) \stackrel{\text{def}}{=} f_0(x) - \sum_{i=1}^{r} \theta_i h_i(x). \tag{24}$$

The vector $\theta \in \mathbb{R}^r$ is called the dual variable or the Lagrange multiplier vector. The Lagrange dual function is defined as $h(\theta) = \inf_x \mathcal{L}(x, \theta)$ and the dual problem of the original problem (23), is $\max_\theta h(\theta)$. The dual problem and the original problem, called the primal problem, are equivalent to each other and typically, the easier of the two problems is solved. Both problems are useful, however, as they provide alternative views of the optimization problem.

5.2.1 The Dual Problem (P_1^*)

Applying the above definitions to (P_1), and noting that the term $q(y \mid x)$ in (21) does not play a role in the minimization problem, the Lagrangian is

$$\begin{aligned}
\mathcal{L}_1(p, \theta) &= \sum_x \tilde{p}(x) \sum_y p(y \mid x) \left(\log \frac{p(y \mid x)}{q_0(y \mid x)} - 1 \right) - \sum_i \theta_i \sum_x \tilde{p}(x) \sum_y p(y \mid x)(f_i(x,y) - E_{\tilde{p}}[f_i \mid x]) \\
&= \sum_x \tilde{p}(x) \sum_y p(y \mid x) \left(\log \frac{p(y \mid x)}{q_0(y \mid x)} - 1 - \langle \theta, f(x,y) - E_{\tilde{p}}[f \mid x] \rangle \right).
\end{aligned}$$

The first step is to minimize the Lagrangian with respect to p, which will allow us to express the dual function. Equating the partial derivatives $\frac{\partial \mathcal{L}_1(p,\theta)}{\partial p(y \mid x)}$ to 0 gives

$$\begin{aligned}
0 &= \tilde{p}(x) \left(\log \frac{p(y \mid x)}{q_0(y \mid x)} - 1 - \langle \theta, f(x,y) - E_{\tilde{p}}[f \mid x] \rangle + p(y \mid x) \frac{1}{p(y \mid x)} \right) \\
&= \tilde{p}(x) \left(\log \frac{p(y \mid x)}{q_0(y \mid x)} - \langle \theta, f(x,y) - E_{\tilde{p}}[f \mid x] \rangle \right)
\end{aligned}$$

and we deduce that $z(y \,|\, x\,;\theta) \stackrel{\text{def}}{=} \arg\min_p \mathcal{L}_1(p,\theta)$, is

$$z(y \,|\, x\,;\theta) = q_0(y \,|\, x)\, \exp\left(\sum_j \theta_j \left(f_j(x,y) - E_{\tilde{p}}[f_j \,|\, x]\right)\right).$$

Thus, the dual function $h_1(\theta) = \mathcal{L}_1(z(y \,|\, x\,;\theta),\theta)$ is given by

$$\begin{aligned} h_1(\theta) &= \sum_x \tilde{p}(x) \sum_y q_0(y \,|\, x) e^{\left\langle \theta, f(x,y) - E_{\tilde{p}}[f \,|\, x]\right\rangle} \left(\langle \theta, f(x,y) - E_{\tilde{p}}[f \,|\, x]\rangle - 1 - \langle \theta, f(x,y) - E_{\tilde{p}}[f \,|\, x]\rangle\right) \\ &= -\sum_x \tilde{p}(x) \sum_y q_0(y \,|\, x) e^{\left\langle \theta, f(x,y) - E_{\tilde{p}}[f \,|\, x]\right\rangle}. \end{aligned}$$

The dual problem (P_1^*) is to determine

$$\begin{aligned} \theta^\star &= \arg\max_\theta h_1(\theta) = \arg\min_\theta \sum_x \tilde{p}(x) \sum_y q_0(y \,|\, x) e^{\left\langle \theta, f(x,y) - E_{\tilde{p}}[f \,|\, x]\right\rangle} \\ &= \arg\min_\theta \sum_x \tilde{p}(x) \sum_y q_0(y \,|\, x) e^{\langle \theta, f(x,y) - f(x,\tilde{y}(x)\rangle} \\ &= \arg\min_\theta \sum_x \tilde{p}(x) \sum_{y \neq \tilde{y}(x)} q_0(y \,|\, x) e^{\langle \theta, f(x,y) - f(x,\tilde{y}(x)\rangle}. \end{aligned} \tag{25}$$

5.2.2 The Dual Problem (P_2^*)

To derive the dual for P_2, we add additional Lagrange multipliers μ_x for the constraints $\sum_y p(y \,|\, x) = 1$ and note that if the normalization constraints are satisfied then the other constraints take the form

$$\sum_x \tilde{p}(x) \sum_y p(y \,|\, x) f_j(x,y) = \sum_{x,y} \tilde{p}(x,y) f_j(x,y).$$

The Lagrangian becomes

$$\begin{aligned} \mathcal{L}_2(p,\theta,\mu) =& D(p,q_0) - \sum_j \theta_j \left(\sum_x \tilde{p}(x) \sum_y p(y \,|\, x) f_j(x,y) - \sum_{x,y} \tilde{p}(x,y) f_j(x,y)\right) \\ & - \sum_x \mu_x \left(1 - \sum_y p(y \,|\, x)\right). \end{aligned}$$

Setting the partial derivatives $\frac{\partial \mathcal{L}_2(p,\theta)}{\partial p(y|x)}$ to 0 and noting that in the normalized case we can ignore the last two terms in the I-divergence, we get

$$0 = \tilde{p}(x) \left(\log \frac{p(y|x)}{q_0(y|x)} + 1 - \langle \theta, f(x,y)\rangle\right) + \mu_x$$

from which the minimizer $z(y \,|\, x\,;\theta) \stackrel{\text{def}}{=} \arg\min_p \mathcal{L}_2(p,\theta)$ is seen to be

$$z(y \,|\, x\,;\theta) = q_0(y \,|\, x) e^{\langle \theta, f(x,y)\rangle - 1 - \mu_x/\tilde{p}(x)}.$$

Substituting $z(y \,|\, x\,; \theta)$ in $\mathcal{L}_2$ we obtain the dual function $h_2(\theta, \mu)$. Maximizing the dual function with respect to μ results in a choice of μ_x that ensure the normalization of z

$$z(y \,|\, x\,; \theta) = \frac{1}{Z_x} q_0(y \,|\, x) e^{\langle \theta, f(x,y) \rangle}$$

and maximizing h_2 with respect to θ we get the following dual problem

$$\begin{aligned}
\theta^* &= \arg\max_\theta \sum_x \tilde{p}(x) \sum_y \frac{1}{Z_x(\theta)} q_0(y \,|\, x) e^{\langle \theta, f(x,y) \rangle} (\langle \theta, f(x,y) \rangle - \log Z_x(\theta)) \\
&- \sum_x \tilde{p}(x) \sum_y \langle \theta, f(x,y) \rangle \frac{1}{Z_x(\theta)} q_0(y \,|\, x) e^{\langle \theta, f(x,y) \rangle} + \sum_{xy} \tilde{p}(x,y) \langle \theta, f(x,y) \rangle - 0 \sum_x \mu_x \\
&= \arg\max_\theta - \sum_x \tilde{p}(x) \log Z_x(\theta) \sum_y \frac{1}{Z_x(\theta)} q_0(y \,|\, x) e^{\langle \theta, f(x,y) \rangle} + \sum_{xy} \tilde{p}(x,y) \langle \theta, f(x,y) \rangle \\
&= \arg\max_\theta \sum_{x,y} \tilde{p}(x,y) \langle \theta, f(x,y) \rangle - \sum_{x,y} \tilde{p}(x,y) \log Z_x(\theta) \\
&= \arg\max_\theta \sum_{x,y} \tilde{p}(x,y) \log \frac{1}{Z_x(\theta)} q_0(y \,|\, x) e^{\langle \theta, f(x,y) \rangle} \\
&= \arg\max_\theta \sum_x \tilde{p}(x) \log \frac{1}{Z_x(\theta)} q_0(\tilde{y}(x) \,|\, x) e^{\langle \theta, f(x, \tilde{y}(x)) \rangle}. \qquad (26)
\end{aligned}$$

5.2.3 Special cases

It is now straightforward to derive various boosting and conditional exponential models problems as special cases of the dual problems (25) and (26).

Case 1: AdaBoost.M2. The dual problem (P_1^*) with $q_0(y \,|\, x) = 1$ is the optimization problem of AdaBoost.M2

$$\theta^\star = \arg\min_\theta \sum_x \tilde{p}(x) \sum_{y \neq y_i} e^{\langle \theta, f(x_i, y) - f(x_i, y_i) \rangle} = \arg\min_\theta \mathcal{E}(\theta).$$

Case 2: Binary AdaBoost. The dual problem (P_1^*) with $q_0(y \,|\, x) = 1$, $\mathcal{Y} = \{-1, 1\}$ and $f_j(x,y) = \frac{1}{2} y\, f_j(x)$ is the optimization problem of binary AdaBoost

$$\theta^\star = \arg\min_\theta \sum_x \tilde{p}(x) e^{-y_i \langle \theta, f(x_i) \rangle}.$$

Case 3: Maximum Likelihood for Exponential Models. The dual problem (P_2^*) with $q_0(y \,|\, x) = 1$ is is maximum (conditional) likelihood for a conditional exponential model with sufficient statistics $f_j(x,y)$

$$\theta^\star = \arg\max_\theta \sum_x \tilde{p}(x) \log \frac{1}{Z_x} e^{\langle \theta, f(x, \tilde{y}(x)) \rangle} = \arg\max_\theta \ell(\theta).$$

Case 4: Logistic Regression. The dual problem (P_2^*) with $q_0(y\,|\,x) = 1$, $\mathcal{Y} = \{-1, 1\}$ and $f_j(x, y) = \frac{1}{2} y\, f_j(x)$ is maximum (conditional) likelihood for binary logistic regression.

$$\theta^\star = \arg\max_\theta \sum_x \tilde{p}(x) \frac{1}{1 + e^{-\tilde{y}(x)\langle \theta, f(x)\rangle}}.$$

We note that it is not necessary to scale the features by a constant factor here, as in (Friedman et al., 2000); the correspondence between logistic regression and boosting is direct.

Case 5: Exponential Models with Carrier Density. Taking $q_0(y|x) \neq 1$ to be a non-parametric density estimator in (P_2^*) results in maximum likelihood for exponential models with a carrier density q_0. Such semi-parametric models have been proposed by Efron and Tibshirani (1996) for integrating between parametric and nonparametric statistical modeling and by Della-Pietra et al. (1992) and Rosenfeld (1996) for integrating exponential models and n-gram estimators in language modeling.

Making the Lagrangian duality argument rigorous requires care, because of the possibility that the solution may lie on the (topological) boundary of $\overline{\mathbb{R}_+^{k\times m}}$ or $\overline{\mathbb{P}_{m-1}^k}$.

Let $\mathcal{Q}_1$ and $\mathcal{Q}_2$ be

$$\mathcal{Q}_1(q_0, f) = \left\{ q \in \overline{\mathbb{R}_+^{k\times m}} \;|\; q(y\,|\,x) \;=\; q_0(y\,|\,x)\, e^{\langle \theta, f(x,y) - f(x,\tilde{y}(x))\rangle},\, \theta \in \mathbb{R}^l \right\}$$
$$\mathcal{Q}_2(q_0, f) = \left\{ q \in \overline{\mathbb{P}_{m-1}^k} \;|\; q(y\,|\,x) \;\propto\; q_0(y\,|\,x)\, e^{\langle \theta, f(x,y)\rangle},\, \theta \in \mathbb{R}^l \right\}$$

and the boosting solution $q_{boost}^\star$ and maximum likelihood solution $q_{ml}^\star$ be

$$q_{boost}^\star = \arg\min_{q \in \overline{\mathcal{Q}_1}} \sum_x \tilde{p}(x) \sum_y q(y\,|\,x)$$
$$q_{ml}^\star = \arg\max_{q \in \overline{\mathcal{Q}_2}} \sum_x \tilde{p}(x) \log q(\tilde{y}(x)\,|\,x).$$

The following proposition corresponds to Proposition 4 of (Della Pietra et al., 1997) for the usual Kullback-Leibler divergence; the proof for the I-divergence carries over with only minor changes. In (Della Pietra et al., 2001) the duality theorem is proved for a class of Bregman divergences, including the extended Kullback-Leibler divergence as a special case. Note that we do not require divergences such as $D(0, q)$ as in (Collins et al., 2002), but rather $D(\tilde{p}, q)$, which is more natural and interpretable from a statistical point-of-view.

Proposition 1. *Suppose that $D(\tilde{p}, q_0) < \infty$. Then $q_{boost}^\star$ and $q_{ml}^\star$ exist, are unique, and satisfy*

$$q_{boost}^\star \;=\; \arg\min_{p \in \mathcal{F}} D(p, q_0) \;=\; \arg\min_{q \in \overline{\mathcal{Q}_1}} D(\tilde{p}, q) \tag{27}$$
$$q_{ml}^\star \;=\; \arg\min_{p \in \mathcal{F} \cap \mathbb{P}_{m-1}^k} D(p, q_0) \;=\; \arg\min_{q \in \overline{\mathcal{Q}_2}} D(\tilde{p}, q) \tag{28}$$

Moreover, $q_{ml}^\star$ is computed in terms of $q_{boost}^\star$ as $q_{ml}^\star = \arg\min_{p \in \mathcal{F} \cap \mathbb{P}_{m-1}^k} D(p, q_{boost}^\star)$.

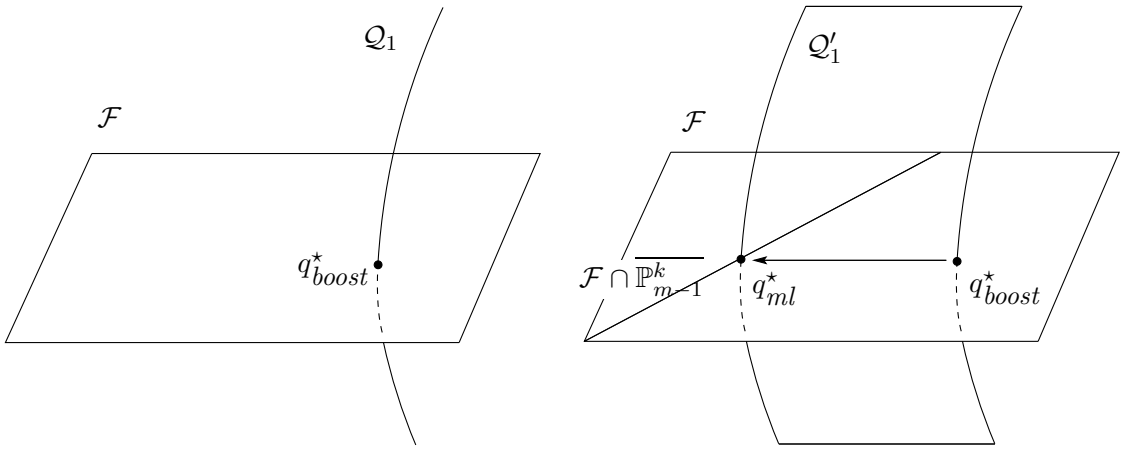

Figure 8: Geometric view of duality. Minimizing the exponential loss finds the member of $\mathcal{Q}_1$ that intersects the feasible set of measures satisfying the moment constraints (left). When we impose the additional constraint that each conditional distribution $q_\theta(y \mid x)$ must be normalized, we introduce a Lagrange multiplier for each training example x, giving a higher-dimensional family $\mathcal{Q}'_1$. By the duality theorem, projecting the exponential loss solution onto the intersection of the feasible set with the simplex of conditional probabilities, $\mathcal{F} \cap \overline{\mathbb{P}^k_{m-1}}$, we obtain the maximum likelihood solution. In many practical cases this projection is obtained by simply normalizing by a constant, resulting in an identical model.

This result has a simple geometric interpretation. The non-normalized exponential family $\mathcal{Q}_1$ intersects the feasible set of measures $\mathcal{F}$ satisfying the constraints (22) at a single point. The algorithms presented in (Collins et al., 2002) determine this point, which is the exponential loss solution $q^\star_{boost} = \arg\min_{q \in \overline{\mathcal{Q}_1}} D(\tilde{p}, q)$ (see Figure 8, left). On the other hand, maximum conditional likelihood estimation for an exponential model with the same features is equivalent to the problem $q^\star_{ml} = \arg\min_{q \in \overline{\mathcal{Q}'_1}} D(\tilde{p}, q)$ where $\mathcal{Q}'_1$ is the exponential family with additional Lagrange multipliers, one for each normalization constraint. The feasible set for this problem is $\mathcal{F} \cap \overline{\mathbb{P}^k_{m-1}}$. Since $\mathcal{F} \cap \overline{\mathbb{P}^k_{m-1}} \subset \mathcal{F}$, by the Pythagorean equality we have that $q^\star_{ml} = \arg\min_{p \in \mathcal{F} \cap \overline{\mathbb{P}^k_{m-1}}} D(p, q^\star_{boost})$ (see Figure 8, right).

5.3 Regularization

Minimizing the exponential loss or the log-loss on real data often fails to produce finite parameters. Specifically, this happens when for some feature f_j

$$\begin{aligned} & f_j(x,y) - f_j(x,\tilde{y}(x)) \geq 0 \text{ for all } y \text{ and } x \text{ with } \tilde{p}(x) > 0 \\ \text{or} \quad & f_j(x,y) - f_j(x,\tilde{y}(x)) \leq 0 \text{ for all } y \text{ and } x \text{ with } \tilde{p}(x) > 0 \end{aligned} \tag{29}$$

This is especially harmful since often the features for which (29) holds are the most important for the purpose of discrimination. The parallel update in (Collins et al., 2002) breaks down in such cases, resulting in parameters going to ∞ or $-\infty$. On the other hand, iterative scaling algorithms work in principle for such features. In practice however, either the parameters θ need

to be artificially capped or the features need to be thrown out altogether, resulting in a partial and less discriminating set of features. Of course, even when (29) does not hold, models trained by maximizing likelihood or minimizing exponential loss can overfit the training data. The standard regularization technique in the case of maximum likelihood employs parameter priors in a Bayesian framework.

In terms of convex duality, a parameter prior for the dual problem corresponds to a "potential" on the constraint values in the primal problem. The case of a Gaussian prior on θ, for example, corresponds to a quadratic potential on the constraint values in the primal problem. Using this correspondence, the connection between boosting and maximum likelihood presented in the previous section indicates how to regularize AdaBoost using Bayesian MAP estimation for non-normalized models, as explained below.

We now consider primal problems over (p, c) where $p \in \mathbb{R}_+^{k \times m}$ and $c \in \mathbb{R}^m$ is a parameter vector that relaxes the original constraints. The feasible set $\mathcal{F}(\tilde{p}, f, c) \subset \mathbb{R}_+^{k \times m}$ allows a violation of the expectation constraints, represented by the vector c

$$\mathcal{F}(\tilde{p}, f, c) \;=\; \left\{ p \in \mathbb{R}_+^{k \times m} \;\mid\; \sum_x \tilde{p}(x) \sum_y p(y \,|\, x)\; (f_j(x, y) - E_{\tilde{p}}[f_j \,|\, x]) = c_j \right\}. \tag{30}$$

A regularized problem for non-normalized models is defined by

$$\begin{array}{lll} (P_{1,\mathrm{reg}}) & \textit{minimize} & D(p, q_0) + U(c) \\ & \textit{subject to} & p \in \mathcal{F}(\tilde{p}, f, c) \end{array}$$

where $U : \mathbb{R}^l \to \mathbb{R}$ is a convex function whose minimum is at $0 \in \mathbb{R}^l$. Intuitively $(P_{1,\mathrm{reg}})$ allows some trade-off between achieving low I divergence to q_0 and some constraint violation, with the exact form of the trade-off represented by the function U. Note that it is possible to choose U in a way that considers some feature constraints more important than others. This may be useful when the values of the features $(f_1, \ldots, f_l)$ are known to be corrupted by noise, where the noise intensity varies among the features.

The dual function of the regularized problem $(P_{1,\mathrm{reg}})$, as derived in Appendix A.3, is

$$h_{1,\mathrm{reg}}(\theta) = h_1(\theta) + U^*(\theta)$$

where $U^*(\theta)$ is the convex conjugate of U. For a quadratic penalty $U(c) = \sum_j \frac{1}{2}\sigma_j^2\, c_j^2$, we have $U^*(\theta) = -\sum_j \frac{1}{2}\sigma_j^{-2}\, \theta_j^2$ and the dual function becomes

$$h_{1,\mathrm{reg}}(\theta) \;=\; -\sum_x \tilde{p}(x) \sum_y q_0(y \,|\, x)\, e^{\langle \theta, f(x,y) - f(x, \tilde{y}(x)) \rangle} - \sum_j \frac{\theta_j^2}{2\sigma_j^2}. \tag{31}$$

A sequential update rule for (31) incurs the small additional cost of solving a nonlinear equation by Newton's method every iteration. See Appendix A.3 for more details. Chen and Rosenfeld (2000) contains a similar regularization for maximum likelihood for exponential models in the context of statistical language modeling.

	Unregularized			*Regularized*		
Data	$\ell_{train}(q_1)$	$\ell_{test}(q_1)$	$\epsilon_{test}(q_1)$	$\ell_{train}(q_2)$	$\ell_{test}(q_2)$	$\epsilon_{test}(q_2)$
Promoters	-0.29	-0.60	0.28	-0.32	-0.50	0.26
Iris	-0.29	-1.16	0.21	-0.10	-0.20	0.09
Sonar	-0.22	-0.58	0.25	-0.26	-0.48	0.19
Glass	-0.82	-0.90	0.36	-0.84	-0.90	0.36
Ionosphere	-0.18	-0.36	0.13	-0.21	-0.28	0.10
Hepatitis	-0.28	-0.42	0.19	-0.28	-0.39	0.19
Breast Cancer Wisconsin	-0.12	-0.14	0.04	-0.12	-0.14	0.04
Pima-Indians	-0.48	-0.53	0.26	-0.48	-0.52	0.25

Table 1: Comparison of unregularized to regularized boosting. For both the regularized and unregularized cases, the first column shows training log-likelihood, the second column shows test log-likelihood, and the third column shows test error rate. Regularization reduces error rate in some cases while it consistently improves the test set log-likelihood measure on all datasets. All entries were averaged using 10-fold cross validation.

5.4 Experiments

We performed experiments on some of the UCI datasets (Blake & Merz, 1998) in order to investigate the relationship between boosting and maximum likelihood empirically. The weak learner was `FindAttrTest` as described in (Freund & Schapire, 1996), and the training set consisted of a randomly chosen 90% of the data. Table 1 shows experiments with regularized boosting. Two boosting models are compared. The first model q_1 was trained for 10 features generated by `FindAttrTest`, excluding features satisfying condition (29). Training was carried out using the parallel update method described in (Collins et al., 2002). The second model, q_2, was trained using the exponential loss with quadratic regularization. The performance was measured using the conditional log-likelihood of the (normalized) models over the training and test set, denoted ℓ_{train} and ℓ_{test}, as well as using the test error rate ϵ_{test}. The table entries were averaged by 10-fold cross validation.

For the weak learner `FindAttrTest`, only the Iris dataset produced features that satisfy (29). On average, 4 out of the 10 features were removed. As the flexibility of the weak learner is increased, (29) is expected to hold more often. On this dataset regularization improves both the test set log-likelihood and error rate considerably. In datasets where q_1 shows significant over-fitting, regularization improves both the log-likelihood measure and the error rate. In cases of little over-fitting (according to the log-likelihood measure), regularization only improves the test set log-likelihood at the expense of the training set log-likelihood, however without affecting much the test set error.

Next we performed a set of experiments to test how much $q^\star_{boost}$ differs from $q^\star_{ml}$, where the boosting model is normalized (after training) to form a conditional probability distribution. For different experiments, `FindAttrTest` generated a different number of features (10–100), and the training set was selected randomly. The plots in Figure 9 show for different datasets the relationship between $\ell_{train}(q^\star_{ml})$ and $\ell_{train}(q^\star_{boost})$ as well as between $\ell_{train}(q^\star_{ml})$ and $D_{train}(q^\star_{ml}, q^\star_{boost})$. The trend

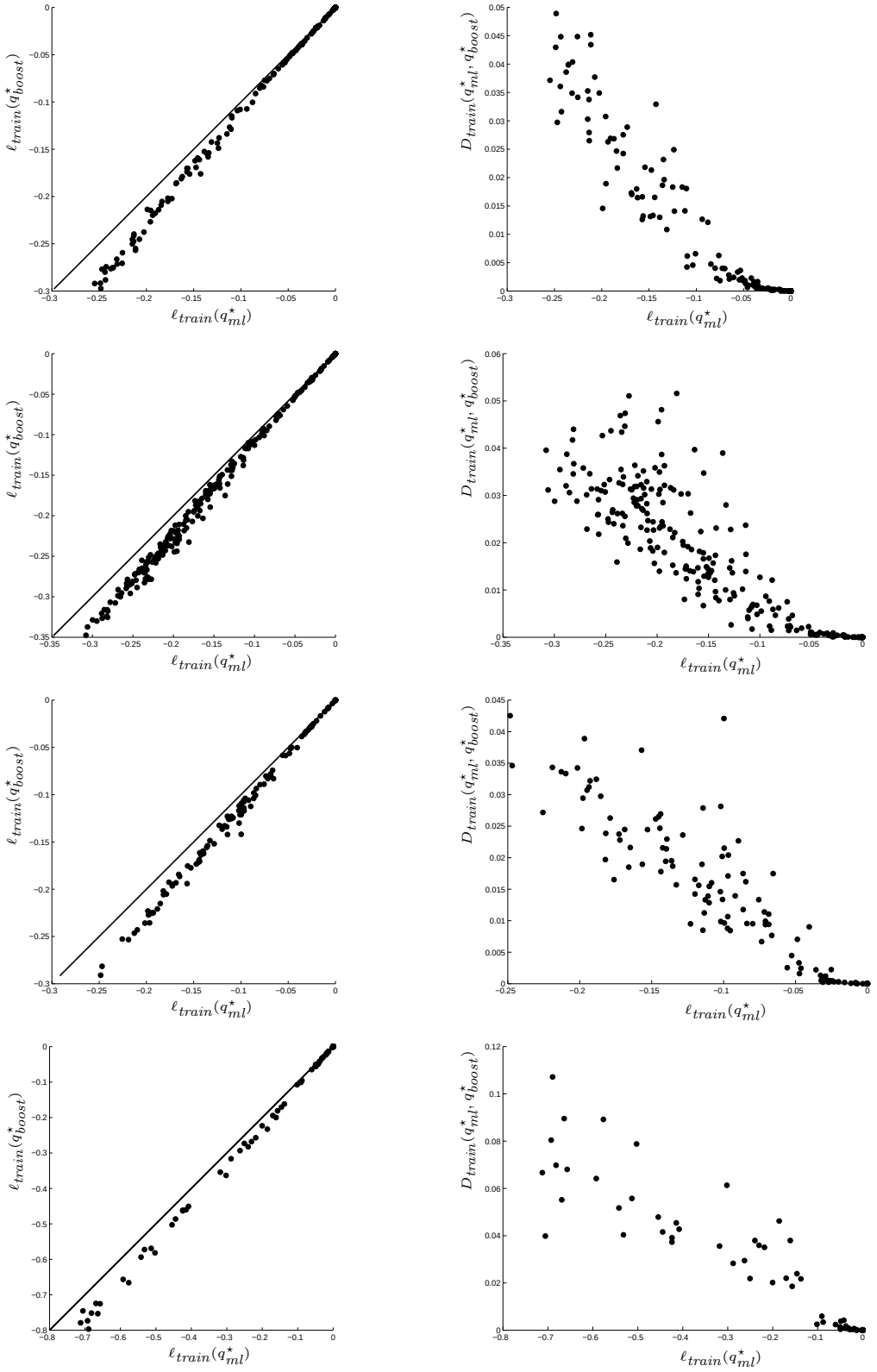

Figure 9: Comparison of AdaBoost and maximum likelihood on four UCI datasets: Hepatitis (top row), Promoters (second row), Sonar (third row) and Glass (bottom row). The left column compares $\ell_{train}(q^\star_{ml})$ to $\ell_{train}(q^\star_{boost})$, and the right column compares $\ell_{train}(q^\star_{ml})$ to $D_{train}(q^\star_{ml}, q^\star_{boost})$.

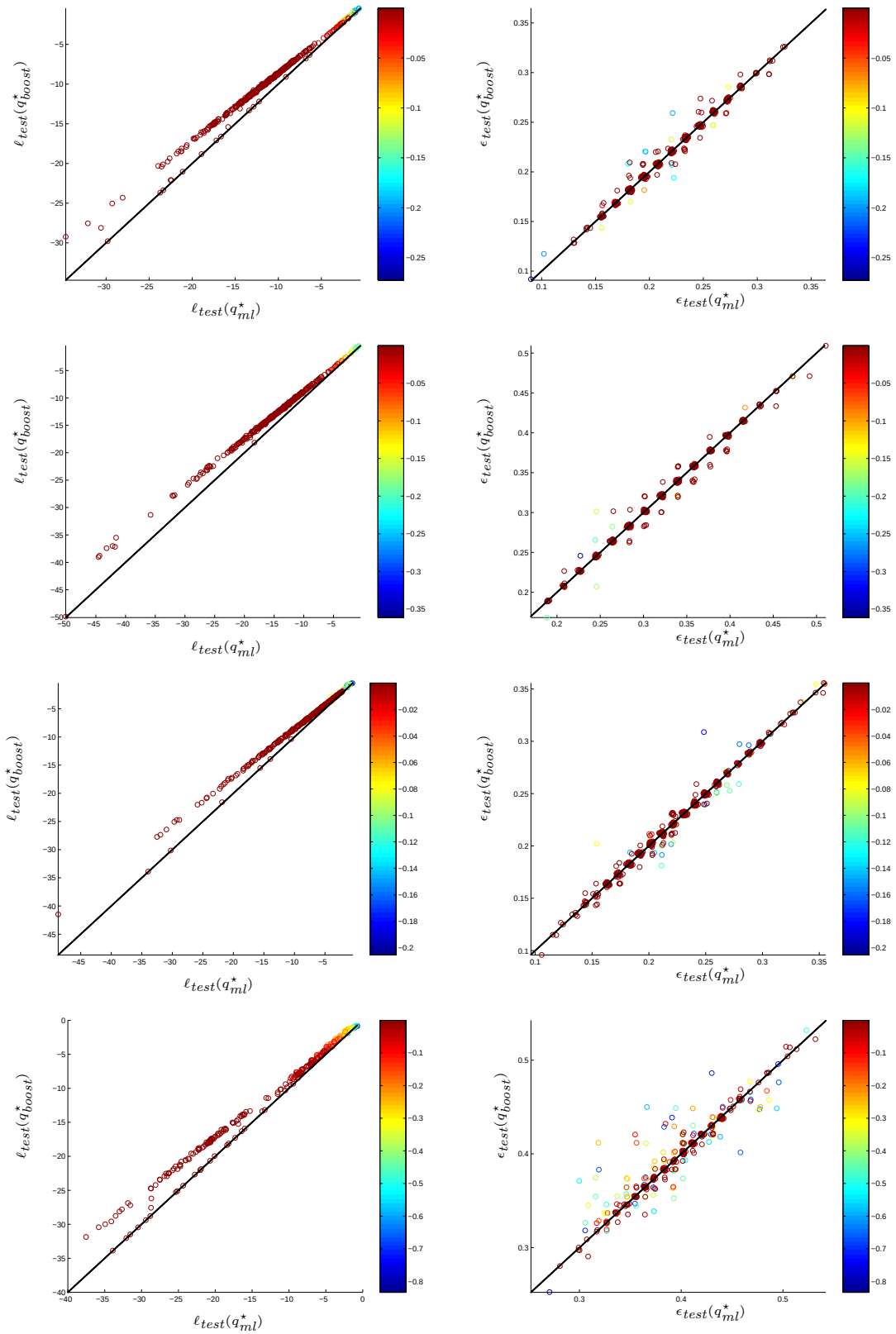

Figure 10: Comparison of AdaBoost and maximum likelihood on the same UCI datasets as in the previous figure. The left column compares the test likelihoods, $\ell_{test}(q^\star_{ml})$ to $\ell_{test}(q^\star_{boost})$, and the right column compares test error rates, $\epsilon_{test}(q^\star_{ml})$ to $\epsilon_{test}(q^\star_{boost})$. In each plot, the color represents the *training* likelihood $\ell_{train}(q^\star_{ml})$; red corresponds to fitting the training data well.

is the same in each data set: as the number of features increases so that the training data is more closely fit ($\ell_{train}(q_{ml}) \to 0$), the boosting and maximum likelihood models become more similar, as measured by the I-divergence.

The plots in Figure 10 show the relationship between the test set log-likelihoods, $\ell_{test}(q^\star_{ml})$ to $\ell_{test}(q^\star_{boost})$, together with the test set error rates $\epsilon_{test}(q^\star_{ml})$ and $\epsilon_{test}(q^\star_{boost})$. In these figures the testing set was chosen to be 50% of the total data. The color represents the *training* data log-likelihood, $\ell_{train}(q^\star_{ml})$, with the color red corresponding to high likelihood. In order to indicate the number of points at each error rate, each circle was shifted by a small random value to avoid points falling on top of each other.

While the plots in Figure 9 indicate that $\ell_{train}(q^\star_{ml}) > \ell_{train}(q^\star_{boost})$, as expected, on the test data the linear trend is reversed, so that $\ell_{test}(q^\star_{ml}) < \ell_{test}(q^\star_{boost})$. This is a result of the fact that for the above data-sets and features, little over-fitting occurs and the more aggressive exponential loss criterion is superior to the more relaxed log-loss criterion. However, as $\ell(q^\star_{ml}) \longrightarrow 0$, the two models come to agree. Appendix A.4 shows that for any exponential model $q_\theta \in \mathcal{Q}_2$,

$$D_{train}(q^\star_{ml}, q_\theta) \;\; = \;\; \ell(q^\star_{ml}) - \ell(q_\theta). \tag{32}$$

By taking $q_\theta = q^\star_{boost}$ it is seen that as the difference between $\ell(q^\star_{ml})$ and $\ell(q^\star_{boost})$ gets smaller, the divergence between the two models also gets smaller.

The results are consistent with the theoretical analysis. As the number of features is increased so that the training data is fit more closely, the model matches the empirical distribution $\tilde{p}$ and the normalizing term $Z(x)$ becomes a constant. In this case, normalizing the boosting model $q^\star_{boost}$ does not violate the constraints, and results in the maximum likelihood model.

In Appendix A.1,A.2 we derive update rules for exponential loss minimization. These update rules are derived by minimizing an auxiliary function that bounds from above the reduction in loss. See (Collins et al., 2002) for the definition of an auxiliary function and proofs that these functions are indeed auxiliary functions. The derived update rules are similar to the ones derived by Collins et al. (2002), except that we do not assume that $M = \max_{i,y} \sum_j |f_j(x,y) - f_j(x,\tilde{y})| < 1$. In Appendix A.3 the regularized formulation is shown in detail and a sequential update rule is derived. Appendix A.4 contains a proof for (32).

The next section derives an axiomatic characterization of the geometry of conditional models. It then builds on the results of this section to given an axiomatic characterization of the geometry underlying conditional exponential models and AdaBoost.

6 Axiomatic Geometry for Conditional Models

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 (Lemma 11.3, (Čencov, 1982)). 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 on, Campbell extended Čencov's result to include non-normalized positive models (Campbell, 1986).

The theorems of Čencov and Campbell are particularly interesting since the Fisher information is pervasive in statistics and machine learning. It is the asymptotic variance of the maximum likelihood estimators under some regularity conditions. Cramer 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 non-informative prior. It is tightly connected to the Kullback-Leibler divergence which 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 non-conditional case, are not supported by theoretical considerations.

In this section 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 non-normalized 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 connection between AdaBoost and constrained minimization of I-divergence described in Section 5.

Throughout the section 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.

In the characterization theorem we will make use of the fact that $\mathbb{P}^k_{m-1} \cap \mathbb{Q}^{k\times m}$ and $\mathbb{R}^{k\times m}_+ \cap \mathbb{Q}^{k\times m} = \mathbb{Q}^{k\times m}_+$ are dense in $\mathbb{P}^k_{m-1}$ and $\mathbb{R}^{k\times m}_+$ respectively. 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 non-rational models follows from its continuous extension to $\mathbb{R}^{k\times m}_+$.

In Section 6.1 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 6.2.

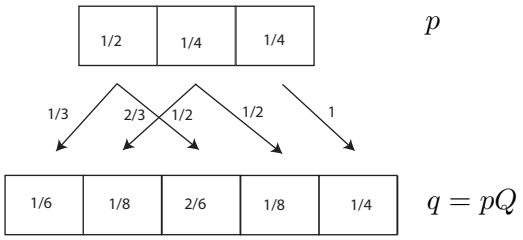

Figure 11: Congruent embedding by a Markov morphism of $p = (1/2, 1/4, 1/4)$.

6.1 Congruent Embeddings by Markov Morphisms of Conditional Models

The characterization result of Section 6.2 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 (1982) and Campbell (1986) 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*

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

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

$$\begin{pmatrix} 1/3 & 0 & 2/3 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{33}$$

Obviously, the columns of any $\mathcal{A}$-stochastic matrix have precisely one non-zero 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 (33) 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 (Figure 3)

The transformation $q \mapsto qQ$ is injective and therefore invertible. For example, the inverse transformation to Q in (33) 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 non-conditional models. The straightforward generalization to conditional models involves performing a similar transformation on the response space $\mathcal{Y}$ for every non-conditional model $p(\cdot|x_i)$ followed by transforming the explanatory space $\mathcal{X}$. It is formalized in the definitions below and illustrated in Figure 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*

$$[M \otimes Q]_{ij} = \sum_{s=1}^{m} M_{is} Q^{(i)}_{sj} = [MQ^{(i)}]_{ij}. \tag{34}$$

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

Definition 6. *Let $\mathcal{B}$ be a k sized partition of $\{1,\dots,l\}$ and $\{\mathcal{A}^{(i)}\}_{i=1}^{k}$ be a set of m sized partitions of $\{1,\dots,n\}$. 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*

$$f : \mathbb{R}_+^{k\times m} \to \mathbb{R}_+^{l\times n} \quad f(M) = R^{\top}(M \otimes Q) \tag{35}$$

is termed a congruent embeddings 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 equation (35) is

$$[f(M)]_{ij} = \sum_{s=1}^{k}\sum_{t=1}^{m} R_{si} Q^{(s)}_{tj} M_{st} \tag{36}$$

with the above sum containing precisely one non-zero term since every column of $Q^{(s)}$ and R contains only one non-zero 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

$$f_*(\partial_{ab}) = \sum_{i=1}^{l}\sum_{j=1}^{n} R_{ai} Q^{(a)}_{bj} \partial'_{ij} \tag{37}$$

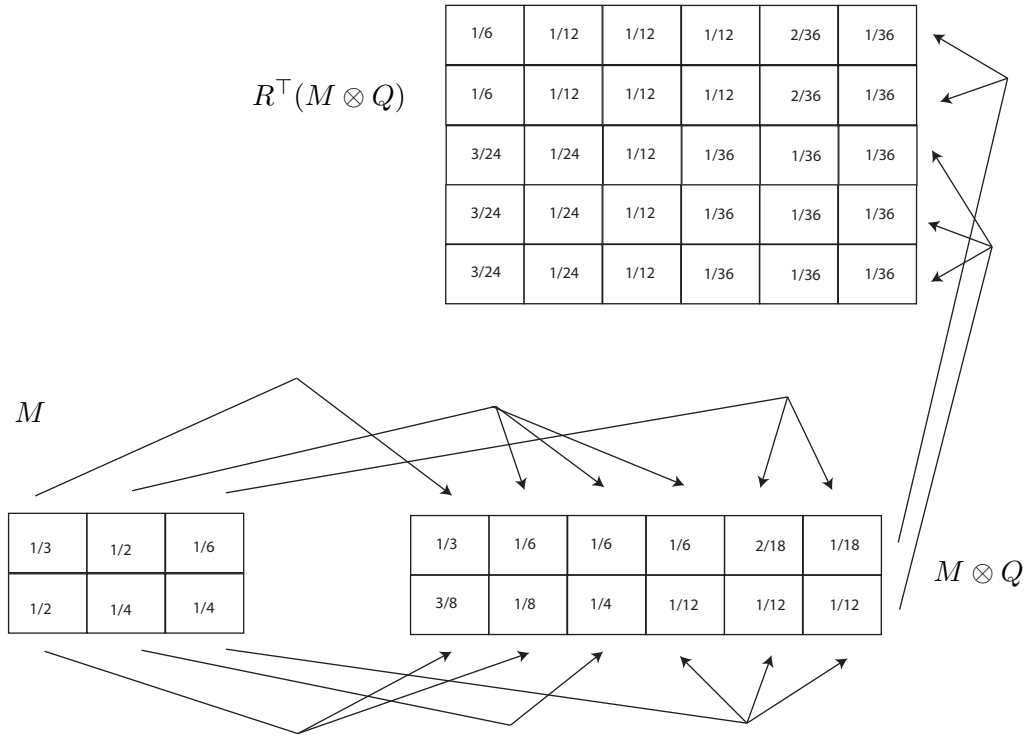

Figure 12: Congruent embedding by a Markov morphism of $\mathbb{R}_+^{2\times 3}$ into $\mathbb{R}_+^{5\times 6}$.

where $\{\partial_{ab}\}_{a,b}$ and $\{\partial'_{ij}\}_{ij}$ are the bases of $T_M\mathbb{R}_+^{k\times m}$ and $\partial'_{ij} \in T_{f(M)}\mathbb{R}_+^{l\times n}$ respectively. Using definition 2 and equation (37), the pull-back of a metric g on $\mathbb{R}_+^{l\times n}$ through $f \in \mathfrak{F}_{k,m}^{l,n}$ is

$$(f^*g)_M(\partial_{ab}, \partial_{cd}) = g_{f(M)}(f_*\partial_{ab}, f_*\partial_{cd}) = \sum_{i=1}^{l}\sum_{j=1}^{n}\sum_{s=1}^{l}\sum_{t=1}^{n} R_{ai}R_{cs}Q_{bj}^{(a)}Q_{dt}^{(c)}g_{f(M)}(\partial'_{ij}, \partial'_{st}). \tag{38}$$

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 non-zero elements and if all its positive entries are identical.*

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

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

We proceed in the next section to state and prove the characterization theorem.

6.2 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 $f \in \mathfrak{F}^{l,n}_{k,m}$ to be isometries. The main results of the section are Theorems 1 and 1 below followed by Corollary 1. The proof of Theorem 1 bears some similarity to the proof of Campbell's theorem (Campbell, 1986) which in turn is related to the proof technique used in Khinchin's characterization of the entropy (Khinchin, 1957). Throughout the section we avoid Čencov's style of using category theory and use only standard techniques in differential geometry.

6.2.1 Three Useful Transformation

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 on.

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 2. *Maps in $\mathfrak{F}^{l,n}_{k,m}$ are norm preserving:*

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

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

$$|M| = |M \otimes Q| = |(M \otimes Q)^\top| = |(M \otimes Q)^\top R| = |R^\top (M \otimes Q)| = |f(M)|.$$

□

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

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

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

$$\mathfrak{h}^\Pi_{\sigma *}(\partial_{ab}) = \partial'_{\sigma(a)\pi^{(a)}(b)} \tag{39}$$

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

$$g_M(\partial_{ab}, \partial_{cd}) = g_{\mathfrak{h}^{\Pi}_{\sigma}(M)}(\partial_{\sigma(a)\pi^{(a)}(b)}, \partial_{\sigma(c)\pi^{(c)}(d)}) \tag{40}$$

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

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

Proposition 3. *Given $\partial_{a_1b_1}, \partial_{a_2b_2}, \partial_{c_1d_1}, \partial_{c_2d_2}$ with $a_1 \neq c_1$ and $a_2 \neq c_2$ there exists σ, Π such that*

$$\mathfrak{h}^{\Pi}_{\sigma *}(\partial_{a_1b_1}) = \partial_{a_2b_2} \qquad \mathfrak{h}^{\Pi}_{\sigma *}(\partial_{c_1d_1}) = \partial_{c_2d_2}. \tag{41}$$

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

The second transformation $\mathfrak{r}_{zw} \in \mathfrak{F}^{kz,mw}_{k,m}$, 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 m non-zero entries of value $1/m$ and each row of R has precisely z non-zero 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 equation (37) 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

$$(\mathfrak{r}^{*}_{zw}g)_M(\partial_{ab}, \partial_{cd}) = \frac{1}{(zw)^2}\sum_{i=1}^{z}\sum_{j=1}^{w}\sum_{s=1}^{z}\sum_{t=1}^{w} g_{\mathfrak{r}_{zw}(M)}(\partial'_{\pi_1(i),\sigma_1(j)}, \partial'_{\pi_2(s),\sigma_2(t)}), \tag{42}$$

again, for some permutations $\pi_1, \pi_2, \sigma_1, \sigma_2$.

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 natural numbers. Given a rational model M, the third mapping

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

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 $|\tilde{M}_i|$ non-zero 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 $\tilde{M}_{ij}\prod_{s\neq i}|\tilde{M}_s|$ non-zero 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 the rational models M into a constant matrix.

Proposition 4. *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

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

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

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

for some r that depends on i, j. □

A straightforward calculation using equation (37) and the definition of $\mathfrak{y}_{M*}$ above shows that the push-forward of $\mathfrak{y}_M$ is

$$\mathfrak{y}_{M*}(\partial_{ab}) = \frac{\sum_{i=1}^{|\tilde{M}_a|}\ \sum_{j=1}^{\tilde{M}_{ab}\prod_{l\neq a}|\tilde{M}_l|}\partial'_{\pi(i)\sigma(j)}}{\tilde{M}_{ab}\prod_i|\tilde{M}_i|}. \tag{43}$$

for some permutations π, σ that depend on M, s, t. Substituting equation (43) in equation (38) gives the pull-back

$$(\mathfrak{y}_M^* g)_M(\partial_{ab}, \partial_{cd}) = \frac{\sum_i\sum_s\sum_j\sum_t g_{\mathfrak{y}_M(M)}(\partial_{\pi_1(i)\sigma_1(j)}, \partial_{\pi_2(s)\sigma_2(t)})}{\tilde{M}_{ab}\tilde{M}_{cd}\prod_s|\tilde{M}_s|^2} \tag{44}$$

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

6.2.2 The Characterization Theorem

Theorems 1 and 2 below are the main result of Section 6.

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*

$$g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = \\ A(|M|) + \delta_{ac}\left(\frac{|M|}{|M_a|}B(|M|) + \delta_{bd}\frac{|M|}{M_{ab}}C(|M|)\right) \tag{45}$$

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

Proof. The proof below 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 non-rational 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 3 and equation (40) to pull back through a corresponding $\mathfrak{h}_\sigma^\Pi$ to obtain

$$g_U^{(k,m)}(\partial_{a_1b_1}, \partial_{c_1d_1}) = g_{\mathfrak{h}_\sigma^\Pi(U)}^{(k,m)}(\partial_{a_2b_2}, \partial_{c_2d_2}) = g_U^{(k,m)}(\partial_{a_2b_2}, \partial_{c_2d_2}). \tag{46}$$

Since (46) 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 Figure 13, 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 equation (38), 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 $\mathfrak{r}_{zw}$ (42) we have a sum of z^2w^2 metrics between vectors of disjoint rows

$$\hat{A}(k, m, |U|) = g_U^{(k,m)}(\partial_{ab}, \partial_{cd}) = \frac{(zw)^2}{(zw)^2}\hat{A}(kz, mw, |\mathfrak{r}_{zw}(U)|) = \hat{A}(kz, mw, |U|) \tag{47}$$

since $\mathfrak{r}_{zw}(U)$ is a constant matrix with the same norm as U. Equation (47) 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

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

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 equation (44) we have

$$g_M^{(k,m)}(\partial_{ab}, \partial_{cd}) = \frac{\tilde{M}_{ab}\tilde{M}_{cd}\prod_s |\tilde{M}_s|^2}{\tilde{M}_{ab}\tilde{M}_{cd}\prod_s |\tilde{M}_s|^2}A(|\mathfrak{y}_M(M)|) = A(|M|). \tag{48}$$

Again, we made use of the fact that in the pull-back equation (44) all the terms in the sum are metrics between vectors of different rows.

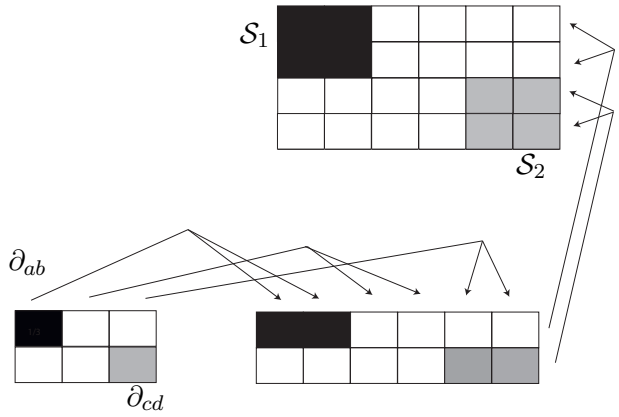

Figure 13: Pushing forward $\partial_{ab}, \partial_{cd}$ for $a \neq c$ through any $f \in \mathfrak{F}_{k,m}^{l,n}$ results in two sets of basis vectors $\mathcal{S}_1$ (black) and $\mathcal{S}_2$ (gray) for which every pair of vectors $\{(v,u) : v \in \mathcal{S}_1, u \in \mathcal{S}_2\}$ are in disjoint rows.

Finally, since $\mathbb{Q}_+^{k\times m}$ is dense in $\mathbb{R}_+^{k\times m}$ and $g_M^{(k,m)}$ is continuous in M, equation (48) 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 $\partial_{a_1,b_1}, \partial_{c_1,d_1}$ 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

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

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 Figure 14. 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.

Pulling back through $\mathfrak{r}_{zw}$ according to equation (42) and the above observation we obtain

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

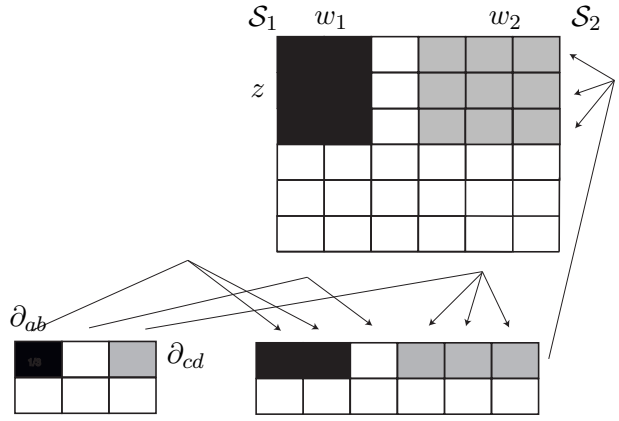

Figure 14: Let f_* push forward ∂_{ab} to a set of vectors $\mathcal{S}_1$ (black) organized in z rows and w_1 columns and $\partial_{ab}, b \neq d$ to a set of vectors (gray) $\mathcal{S}_2$ organized in z rows and w_2 columns. Then counting the pairs of vectors $\mathcal{S}_1 \times \mathcal{S}_2$ we obtain 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.

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

$$g_U^{(k,m)}(\partial_{ab}, \partial_{ad}) = A(|U|) + kB(|U|). \tag{49}$$

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 Figure 14, 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 (44) the pull-back through $\mathfrak{y}_M$ is

$$\begin{aligned}
g_M^{(k,m)}(\partial_{ab}, \partial_{ad}) =& \frac{|\tilde{M}_a|\tilde{M}_{ab}\prod_{l\neq a}|\tilde{M}_l|(|\tilde{M}_a|-1)\tilde{M}_{ad}\prod_{l\neq a}|\tilde{M}_l|}{\tilde{M}_{ab}\tilde{M}_{ad}\prod_i|\tilde{M}_i|^2}A(|M|) \\
&+ \frac{|\tilde{M}_a|\tilde{M}_{ab}\tilde{M}_{ad}\prod_{l\neq a}|\tilde{M}_l|^2}{\tilde{M}_{ab}\tilde{M}_{ad}\prod_i|\tilde{M}_i|^2}\Big(A(|M|) + B(|M|)\sum_i|\tilde{M}_i|\Big) \\
=& \frac{|\tilde{M}_a|-1}{|\tilde{M}_a|}A(|M|) + \frac{1}{|\tilde{M}_a|}\Big(A(|M|) + B(|M|)\sum_i|\tilde{M}_i|\Big) \\
=& A(|M|) + \frac{|\tilde{M}|}{|\tilde{M}_a|}B(|M|) = A(|M|) + \frac{|M|}{|M_a|}B(|M|).
\end{aligned} \tag{50}$$

The first term in the sums above 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 $g^{(k,m)}$ equation (50) 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 $\mathfrak{h}^{\Pi}_{\sigma}$ 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 the 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 (38) into the three cases under considerations.

Pulling back through $\mathfrak{r}_{zw}$ (42) we obtain

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

which after rearrangement and dividing by km gives

$$\frac{\hat{C}(k,m,|U|)-A(|U|)-kB(|U|)}{km} = \frac{\hat{C}(kz,mw,|U|)-A(|U|)-kzB(|U|)}{kzmw}. \tag{51}$$

It follows that the left side of (51) equals a function $C(|U|)$ for some $C\in C^\infty(\mathbb{R}_+,\mathbb{R})$ 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 $\mathfrak{y}_M$ (44) 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

$$\begin{aligned}g_M^{(k,m)}(\partial_{ab},\partial_{ab}) =&\frac{|\tilde{M}_a|-1}{|\tilde{M}_a|}A(|M|) + \Big(\frac{1}{|\tilde{M}_a|}-\frac{1}{\tilde{M}_{ab}\prod_i|\tilde{M}_i|}\Big)\Big(A(|M|)+B(|M|)\sum_i|\tilde{M}_i|\Big) \qquad (52)\\ &+\frac{A(|M|)+B(|M|)\sum_i|\tilde{M}_i|+C(|M|)\prod_j|\tilde{M}_j|\sum_i|\tilde{M}_i|}{\tilde{M}_{ab}\prod_i|\tilde{M}_i|}\\ =&A(|M|)+\frac{|\tilde{M}|}{|\tilde{M}_a|}B(|M|)+\frac{|\tilde{M}|}{\tilde{M}_{ab}}C(|M|) = A(|M|)+\frac{|M|}{|M_a|}B(|M|)+\frac{|M|}{M_{ab}}C(|M|).\end{aligned}$$

Since the positive rational matrices are dense in $\mathbb{R}_+^{k\times m}$ and the metric $g_M^{(k,m)}$ is continuous in M, equation (52) 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*

$$g_M^{(k,m)}(\partial_{ab},\partial_{cd}) = A(|M|) + \delta_{ac}\left(\frac{|M|}{|M_a|}B(|M|) + \delta_{bd}\frac{|M|}{M_{ab}}C(|M|)\right) \tag{53}$$

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

Proof. To prove the theorem we need to show that

$$\forall M \in \mathbb{R}_+^{k\times m}, \quad \forall f \in \mathfrak{F}_{k,m}^{l,n}, \quad \forall u, v \in T_M\mathbb{R}_+^{k\times m}, \qquad g_M^{(k,m)}(u,v) = g_{f(M)}^{(l,n)}(f_*u, f_*v). \tag{54}$$

Considering arbitrary $M \in \mathbb{R}_+^{k\times m}$ and $f \in \mathfrak{F}_{k,m}^{l,n}$ we have by equation (38)

$$g_{f(M)}^{(l,n)}(f_*\partial_{ab}, f_*\partial_{cd}) = \sum_{i=1}^{l}\sum_{j=1}^{n}\sum_{s=1}^{l}\sum_{t=1}^{n} R_{ai}R_{cs}Q_{bj}^{(a)}Q_{dt}^{(c)}g_{f(M)}^{(l,n)}(\partial'_{ij},\partial'_{st}). \tag{55}$$

For $a \neq c$, using the metric form of equation (53), the right hand side of equation (55) 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 equation (55) reduces to

$$\begin{aligned} &A(|f(M)|)\sum_{i=1}^{l}\sum_{j=1}^{n}\sum_{s=1}^{l}\sum_{t=1}^{n} R_{ai}R_{cs}Q_{bj}^{(a)}Q_{dt}^{(c)} + B(|f(M)|)\sum_i \frac{|f(M)|}{|[f(M)]_i|}R_{ai}^2\sum_j\sum_t Q_{bj}^{(a)}Q_{dt}^{(a)} \\ &= A(|M|) + B(|M|)|M|\sum_i \frac{R_{ai}^2}{|[f(M)]_i|}. \end{aligned} \tag{56}$$

Recall from equation (36) that

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

Summing over j we obtain

$$|[f(M)]_i| = \sum_{s=1}^{k} R_{si}\sum_{t=1}^{m} M_{st}\sum_j Q_{tj}^{(s)} = \sum_{s=1}^{k} R_{si}|M_s|. \tag{57}$$

Since every column of R has precisely one non-zero element it follows from (57) that R_{ai} is either 0 or $\frac{|[f(M)]_i|}{|M_a|}$ which turns equation (56) into

$$g_{f(M)}^{(l,n)}(f_*\partial_{ab}, f_*\partial_{ad}) = A(|M|) + B(|M|)|M|\sum_{i:R_{ai}\neq 0}\frac{R_{ai}}{|M_a|} = A(|M|) + B(|M|)\frac{|M|}{|M_a|} = g_M^{(k,m)}(\partial_{ab},\partial_{ad}).$$

Finally, for the case $a = c, b = d$ the right hand side of equation (55) 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 equation (36)

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

there is a unique positive element, $R_{ai}Q_{bj}^{(a)}$ is either $[f(M)]_{ij}/M_{ab}$ or 0. It follows then that equation (55) equals

$$\begin{aligned} g_{f(M)}^{(l,n)}(f_*\partial_{ab}, f_*\partial_{ab}) =& A(|M|) + B(|M|)\frac{|M|}{|M|_a} + C(|M|)|M| \sum_{i:R_{ai}\neq 0}\sum_{j:Q_{bj}^{(a)}\neq 0} \frac{R_{ai}Q_{bj}^{(a)}}{M_{ab}} \\ &= A(|M|) + B(|M|)\frac{|M|}{|M_a|} + C(|M|)\frac{|M|}{M_{ab}} = g_M^{(k,m)}(\partial_{ab}, \partial_{ab}). \end{aligned}$$

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 $\partial_{ab}, \partial_{cd}$ and hence the condition in (54) 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. □

6.2.3 Normalized Conditional Models

A stronger statement can be said in the case of normalized conditional models. In this case, it turns out that the choices of A and B are immaterial and equation (45) 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, equation* (45) *in theorem 1 reduces to the product Fisher information metric up to a multiplicative constant.*

Proof. For $u, v \in T_M\mathbb{P}_{m-1}^k$ 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

$$g_M^{(k,m)}(u, v) = \left(\sum_{ij} u_{ij}\right)\left(\sum_{ij} v_{ij}\right) A(|M|) + \sum_i \left(\sum_j u_{ij}\right)\left(\sum_j v_{ij}\right)\frac{|M|}{|M_i|}B(|M|) \tag{58}$$

$$+ \sum_{ij} u_{ij}v_{ij}\frac{|M|C(|M|)}{M_{ij}} = kC(k)\sum_{ij}\frac{u_{ij}v_{ij}}{M_{ij}} \tag{59}$$

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. □

6.3 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

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

where Z is the normalization factor. A more general form, demonstrated in Section 5 that is appropriate for $2 \leq |\mathcal{Y}| < \infty$ is

$$p(y|x\,;\theta) = \frac{1}{Z} e^{\sum_i \theta_i f_i(x,y)}, \quad x, \theta \in \mathbb{R}^n, y \in \mathcal{Y} \tag{60}$$

where $f_i : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ are arbitrary feature functions. The model (60) is a conditional exponential model and the parameters θ are normally obtained by maximum likelihood estimation for a training set $\{(x_j, y_j)\}_{j=1}^N$

$$\arg\max_\theta \sum_{j=1}^N \sum_i \theta_i f_i(x_j, y_j) - \sum_{j=1}^N \log \sum_{y' \in \mathcal{Y}} e^{\sum_i \theta_i f_i(x_j, y')}. \tag{61}$$

AdaBoost is a linear classifier, usually viewed as an incremental ensemble methods that combines weak learners (Schapire, 2002). The incremental rule that AdaBoost uses to select the weight vector θ is known to greedily minimize the exponential loss

$$\arg\min_\theta \sum_j \sum_{y \neq y_j} e^{\sum_i \theta_i (f_i(x_j, y) - f_i(x_j, y_j))} \tag{62}$$

associated with a non-normalized model

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

By moving to the convex primal problems that correspond to maximum likelihood for logistic regression (61) and minimum exponential loss for AdaBoost (62) a close connection between the two algorithms appear cf. Section 5. Both problems selects a model that minimizes the I-divergence (21)

$$D_r(p, q) = \sum_x r(x) \sum_y \left(p(y|x) \log \frac{p(y|x)}{q(y|x)} - p(y|x) + q(y|x) \right).$$

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 non-normalized conditional models and reduces to the conditional Kullback-Leibler divergence for normalized models. The conditional form above (21) is a generalization of the non-normalized divergence for probability measures studied by Csiszár (Csiszár, 1991).

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 $\epsilon = 0$

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

The first order terms

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

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

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

at $\epsilon = 0$ are $\delta_{y_1 y_2}\delta_{x_1 x_2}\frac{r(x_1)}{p(y_1|x_1)}$. 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 (45) 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 (45) 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 non-negative conditional models.

The above 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.

6.4 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 non-normalized 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 (Campbell, 1986) and corollary 1 reduces to Čencov's theorem (Lemma 11.3 of (Čencov, 1982)).

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 section.

As mentioned in Section 3, the proper framework for considering non-negative models is a manifold with corners (Lee, 2002). The theorem stated here carries over by the continuity of the metric from the manifold of positive models to its closure. Extensions to infinite $\mathcal{X}$ or $\mathcal{Y}$ poses considerable difficulty. For a brief discussion of infinite dimensional manifolds representing densities see (Amari & Nagaoka, 2000) pp. 44-45.

The characterized metric (45) has three additive components. The first one represents a components 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 non-zero 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 non-normalized 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 non-robust. Indeed, in the statistical literature, the maximum likelihood estimator is often replaced by more robust estimators, among them the minimum Hellinger distance estimator (Beran, 1977; Lindsay, 1994). 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.

So far, we examined the geometry of the model space Θ. In the remainder of this thesis we turn to studying machine learning algorithms in the context of geometric assumption on the data space $\mathcal{X}$.

7 Data Geometry Through the Embedding Principle

The standard practice in machine learning is to represent data points as vectors in a Euclidean space, and then process them under the assumptions of Euclidean geometry. Such an embedding is often done regardless of the origin of the data. It is a common practice for inherently real valued data, such as measurements of physical quantities as well as for categorical data such as text documents or boolean data.

Two classes of learning algorithms, which make such assumptions, are radial basis machines and linear classifiers. Radial basis machines are algorithms that are based on the radial basis or Gaussian function

$$K(x, y) = c \exp\left(\frac{1}{\sigma^2}\|x - y\|^2\right) = c \exp\left(\frac{1}{\sigma^2}\sum_i |x_i - y_i|^2\right)$$

which is in turn based on the Euclidean normed distance. Linear classifiers, such as boosting, logistic regression, linear SVM and the perceptron make an implicit Euclidean assumption by choosing the class of linear hyperplane separators. Furthermore, the training phase of many linear classifiers is based on Euclidean arguments such as the margin. Section 9 contains more details on this point.

In accordance with the earlier treatment of the model space geometry, we would like to investigate a more appropriate geometry for $\mathcal{X}$ and its implications on practical algorithms. It is likely that since data come from different sources, the appropriate geometry should be problem dependent.

More specifically, the data should dictate first the topological space and then the geometric structure to endow it with. It seems that selecting a geometry for the data space is much harder than for the model space, where Čencov's theorem offered a natural candidate.

The first step toward selecting a metric for $\mathcal{X}$ is assuming that the data comes from a set of distributions $\{p(x\,;\theta) : \theta \in \Theta\}$. In this case we assume that every data point is associated with a potentially different distribution. Since there is already a natural geometric structure on the model space Θ, we can obtain a geometry for the data points by identifying them with the corresponding probabilistic models and taking the Fisher geometry on Θ.

Under a frequentist interpretation, we can associate each data point with a single model θ. An obvious choice for the mapping $x \mapsto \theta$ is the maximum likelihood estimator $\hat{\theta} : \mathcal{X} \to \Theta$, which enjoys both nice properties and often satisfactory performance. We can then measure geometric quantities such as distance between two data points x, y by their geometric counterparts on Θ with respect to the Fisher geometry

$$d(x, y) = d_{\mathcal{J}}(\hat{\theta}(x), \hat{\theta}(y)). \tag{63}$$

Under a Bayesian interpretation, we can associate a posterior $p(\theta \,|\, x)$ with each data point x. The quantities that correspond to geometric measurements on Θ transform then into random variables. For example, distance between two data points x, y becomes a random variable whose posterior mean is

$$E_{\theta|x} d(x, y) = \iint d_{\mathcal{J}}(\theta, \eta) p(\theta \,|\, x) p(\eta \,|\, x) \mathrm{d}\theta \mathrm{d}\eta. \tag{64}$$

While there is a clearly defined geometry on Θ, the resulting concepts on $\mathcal{X}$ may be different from what is expected by a geometry on $\mathcal{X}$. For example, the functions in (63)-(64) may not be metric distance function (satisfying positivity, symmetry and triangle inequality) on $\mathcal{X}$, as happens for example when $\hat{\theta}$ is not injective. Another possible disparity occurs if $\hat{\theta}$ is not surjective and significant areas of Θ are not represented by the data.

As a result, we will be more interested in cases where $\hat{\theta}$ is injective and its image $\hat{\theta}(\mathcal{X})$ is dense in Θ. Such is the case of text document representation, the main application area of this thesis. A common assumption for text document representation is to disregard the order of the words. This assumption, termed bag of words representation, is almost always used for text classification task. Under this assumption, a document is represented as a vector of word counts $x \in \mathbb{N}^{|V|}$ where V is the set of distinct words commonly known as the dictionary. In order to treat long and short documents on equal footing, it is furthermore common to divide the above representation by the length of the document resulting in a non-negative vector that sum to 1. This is the representation that we assume in this thesis, and from now on we assume that text documents $x \in \mathcal{X}$ are given in this representation.

Assuming that a multinomial model generates the documents, we find a close relationship between $\mathcal{X}$ and Θ. The data space $\mathcal{X}$ is in fact a subset of the non-negative simplex $\overline{\mathbb{P}_n}$ which is precisely the model space Θ (in this case, $n+1$ is the size of the dictionary). More specifically, $\mathcal{X}$ is the subset of $\overline{\mathbb{P}_n} = \Theta$ with all rational coordinates

$$\mathcal{X} = \overline{\mathbb{P}_n} \cap \mathbb{Q}^{n+1}.$$

Identifying $\mathcal{X}$ as a subset of $\overline{\mathbb{P}_n}$, we see that the maximum likelihood mapping $\hat{\theta} : \mathcal{X} \to \Theta$ is the injective inclusion map $\iota : \mathcal{X} \to \Theta, \iota(x) = x$, and

$$\Theta = \overline{\mathbb{P}_n} = \overline{\overline{\mathbb{P}_n} \cap \mathbb{Q}^{n+1}} = \overline{\mathcal{X}}.$$

In other words, the maximum likelihood mapping is an injective embedding of $\mathcal{X}$ onto a dense set in Θ.

Since $\mathcal{X}$ is nowhere dense, it is not suitable for continuous treatment. It should, at the very least be embedded in a complete space in which every Cauchy sequence convergence. Replacing the document space $\mathcal{X}$ by its completion $\overline{\mathbb{P}_n}$, as we propose above, is the smallest embedding that is sufficient for continuous treatment.

Once we assume that the data x_i is sampled from a model $p(x\,;\theta_i^{\text{true}})$ there is some justification in replacing the data by points $\{\theta_i^{\text{true}}\}_i$ on a Riemannian manifold $(\Theta, \mathcal{J})$. Since the models generated the data, in some sense they contain the essence of the data and the Fisher geometry is motivated by Čencov's theorem. The weakest part of the embedding framework, and apparently the most arbitrary, is the choice of the particular embedding. In the next two subsections, we address this concern by examining the properties of different embeddings.

7.1 Statistical Analysis of the Embedding Principle

As mentioned previously, one possible embedding is the maximum likelihood embedding $\hat{\theta}^{\text{mle}} : \mathcal{X} \to \Theta$. The nice asymptotic properties that the MLE enjoys seem to be irrelevant to our purposes since it is employed for a single data point. If we assume that the data parameters are clustered with a finite number of clusters $\theta_1^{\text{true}}, \cdots, \theta_C^{\text{true}}$, then as the total number of examples increases, we obtain increasing numbers of data points per cluster. In this case the MLE may enjoy the asymptotic properties of first order efficiency. If no clustering exists, then the number of parameters grow proportionally to the number of data points. Such is the situation in non-parametric statistics which becomes a more relevant framework than parametric statistics.

Before we continue, we review some definitions from classical decision theory. For more details, see for example (Schervish, 1995). Given an estimator $T : \mathcal{X}^n \to \Theta$, and a loss function $l : \Theta \times \Theta \to \mathbb{R}$, we can define the risk

$$r_T : \Theta \to \mathbb{R} \qquad r_T(\theta) = \int_{\mathcal{X}^n} p(x_1, \ldots, x_n\,;\theta)\; l(T(x_1, \ldots, x_n), \theta)\; \mathrm{d}x_1 \ldots \mathrm{d}x_n. \tag{65}$$

The risk $r_T(\theta)$ measures the average loss of the estimator T assuming that the true parameter value is θ. If there exists another estimator T' that dominates T i.e.

$$\forall\theta \quad r_{T'}(\theta) \leq r_T(\theta) \quad \text{and} \quad \exists\theta \quad r_{T'}(\theta) < r_T(\theta)$$

we say that the estimator T is inadmissable. We will assume from now on that the loss function is the mean squared error function.

Assuming no clustering, the admissability of the MLE estimator is questionable, as was pointed by James and Stein who stunned the statistics community by proving that the James-Stein estimator $\hat{\theta}^{\text{JS}} : \mathbb{R} \to \mathbb{R}$ for $N(\theta, 1)$ given by

$$\hat{\theta}^{\text{JS}}(x_i) = \left(1 - \frac{n-2}{\sum_i x_i^2}\right) x_i$$

dominates[7] the MLE in mean squared error for $n \geq 3$ (Stein, 1955; Efron & Morris, 1977). Here the data $x_1, \ldots, x_n$ is sampled from $N(\mu_1, 1), \ldots, N(\mu_n, 1)$ and the parameter space is $\Theta = \mathbb{R}$. James-Stein estimator and other shrinkage estimators are widely studied in statistics. Such estimators, however, depend on the specific distribution being estimated and are usually difficult to come up with.

A similar result from a different perspective may be obtained by considering the Bayesian framework. In this framework, θ_i are iid random variables with a common, but unknown, distribution $p(\theta|\phi) = \prod_i p(\theta_i|\phi)$. If ϕ is unknown our marginal distribution is a mixture of iid distributions[8]

$$p(\theta) = \int \prod_i p(\theta_i|\phi)p(\phi)\,\mathrm{d}\phi.$$

Introducing the data x_i, we then have the following form for the posterior

$$p(\theta_1, \ldots, \theta_n|x_1, \ldots x_n) \propto \prod_i p(x_i|\theta_i) \int \prod_i p(\theta_i|\phi)p(\phi)\,\mathrm{d}\phi = \int \prod_i p(x_i|\theta_i)p(\theta_i|\phi)p(\phi)\,\mathrm{d}\phi. \tag{66}$$

Recall that in the Bayesian setting, the embedding becomes a random variable. Furthermore, as seen from equation (66) if ϕ is unknown, the embedding of the data $x_1, \ldots, x_n$ cannot be decoupled into independent random embeddings and the posterior of the entire parameter set has to be considered.

In the Bayesian framework, our distributional assumptions dictate the form of the posterior and there is no arbitrariness such as a selection of a specific point estimator. This benefit, as is often the case in Bayesian statistics, comes at the expense of added computational complexity. Given a data set $x_1, \ldots, x_n$, the distance between two data points becomes a random variable. Assuming we want to use a nearest neighbor classifier, a reasonable thing would be to consider the posterior

[7] Interestingly, the James-Stein estimator is inadmissable as well as it is further dominated by another estimator.

[8] the mixture of iid distribution is also motivated by de Finetti's theorem that singles it out as the only possibly distribution if $n \to \infty$ and θ_i are exchangeable, rather than independent

mean of the geodesic distance $E_{\theta|x} d(x_i, x_j)$. If ϕ is unknown, equation (64) should be replaced with the following high dimensional integral

$$E_{\theta|x} d(x_i, x_j) \propto \iint \prod_i p(x_i|\theta_i) p(\theta_i|\phi) p(\phi)\, \mathrm{d}\phi\, d(\theta_i, \theta_j)\, \mathrm{d}\theta.$$

The empirical Bayes framework achieves a compromise between the frequentist and the Bayesian approaches (Robbins, 1955; Casella, 1985). In this approach, a deterministic embedding $\hat{\theta}$ for x_i is obtained by using the entire data set $x_1, \ldots, x_n$ to approximate the regression function $E(\theta_i|x_i)$. For some restricted cases, the empirical Bayes procedure enjoys asymptotic optimality. However, both the analysis and the estimation itself depend heavily on the specific case at hand.

In the next sections we use the embedding principle to improve several classification algorithms. In Section 7 we propose a generalization of the radial basis function, adapted to the Fisher geometry of the embedded data space. In Section 9 we define the class of hyperplane separators and margin quantity to arbitrary data geometries and work out the detailed generalization of logistic regression to multinomial geometry. Finally, Section 10 goes beyond the Fisher geometry and attempts to learn a local metric, adapted to the provided training set. Experimental results in these sections show that the above generalizations of known algorithms to non-Euclidean geometries outperform their Euclidean counterparts.

8 Diffusion Kernels on Statistical Manifolds

The use of Mercer kernels for transforming linear classification and regression schemes into nonlinear methods is a fundamental idea, one that was recognized early in the development of statistical learning algorithms such as the perceptron, splines, and support vector machines (Aizerman et al., 1964; Kimeldorf & Wahba, 1971; Boser et al., 1992). The recent resurgence of activity on kernel methods in the machine learning community has led to the further development of this important technique, demonstrating how kernels can be key components in tools for tackling nonlinear data analysis problems, as well as for integrating data from multiple sources.

Kernel methods can typically be viewed either in terms of an implicit representation of a high dimensional feature space, or in terms of regularization theory and smoothing (Poggio & Girosi, 1990). In either case, most standard Mercer kernels such as the Gaussian or radial basis function kernel require data points to be represented as vectors in Euclidean space. This initial processing of data as real-valued feature vectors, which is often carried out in an *ad hoc* manner, has been called the "dirty laundry" of machine learning Dietterich (2002)—while the initial Euclidean feature representation is often crucial, there is little theoretical guidance on how it should be obtained. For example in text classification, a standard procedure for preparing the document collection for the application of learning algorithms such as support vector machines is to represent each document as a vector of scores, with each dimension corresponding to a term, possibly after scaling by an inverse document frequency weighting that takes into account the distribution of terms in the collection Joachims (2000). While such a representation has proven to be effective, the statistical justification of such a transform of categorical data into Euclidean space is unclear.

Recent work by Kondor and Lafferty (2002) was directly motivated by this need for kernel methods that can be applied to discrete, categorical data, in particular when the data lies on a graph. Kondor and Lafferty (2002) propose the use of discrete diffusion kernels and tools from spectral graph theory for data represented by graphs. In this section, we propose a related construction of kernels based on the heat equation. The key idea in our approach is to begin with a statistical family that is natural for the data being analyzed, and to represent data as points on the statistical manifold associated with the Fisher information metric of this family. We then exploit the geometry of the statistical family; specifically, we consider the heat equation with respect to the Riemannian structure given by the Fisher metric, leading to a Mercer kernel defined on the appropriate function spaces. The result is a family of kernels that generalizes the familiar Gaussian kernel for Euclidean space, and that includes new kernels for discrete data by beginning with statistical families such as the multinomial. Since the kernels are intimately based on the geometry of the Fisher information metric and the heat or diffusion equation on the associated Riemannian manifold, we refer to them here as information diffusion kernels.

One apparent limitation of the discrete diffusion kernels of (Kondor & Lafferty, 2002) is the difficulty of analyzing the associated learning algorithms in the discrete setting. This stems from the fact that general bounds on the spectra of finite or even infinite graphs are difficult to obtain, and research has concentrated on bounds on the first eigenvalues for special families of graphs. In contrast, the kernels we investigate here are over continuous parameter spaces even in the case where the underlying data is discrete, leading to more amenable spectral analysis. We can draw on the considerable body of research in differential geometry that studies the eigenvalues of the geometric Laplacian, and thereby apply some of the machinery that has been developed for analyzing the generalization performance of kernel machines in our setting.

Although the framework proposed is fairly general, in this section we focus on the application of these ideas to text classification, where the natural statistical family is the multinomial. In the simplest case, the words in a document are modeled as independent draws from a fixed multinomial; non-independent draws, corresponding to n-grams or more complicated mixture models are also possible. For n-gram models, the maximum likelihood multinomial model is obtained simply as normalized counts, and smoothed estimates can be used to remove the zeros. This mapping is then used as an embedding of each document into the statistical family, where the geometric framework applies. We remark that the perspective of associating multinomial models with individual documents has recently been explored in information retrieval, with promising results (Ponte & Croft, 1998; Zhai & Lafferty, 2001).

The statistical manifold of the n-dimensional multinomial family comes from an embedding of the multinomial simplex into the n-dimensional sphere which is isometric under the Fisher information metric. Thus, the multinomial family can be viewed as a manifold of constant positive curvature. As discussed below, there are mathematical technicalities due to corners and edges on the boundary of the multinomial simplex, but intuitively, the multinomial family can be viewed in this way as a Riemannian manifold with boundary; we address the technicalities by a "rounding" procedure on the simplex. While the heat kernel for this manifold does not have a closed form, we can

approximate the kernel in a closed form using the leading term in the parametrix expansion, a small time asymptotic expansion for the heat kernel that is of great use in differential geometry. This results in a kernel that can be readily applied to text documents, and that is well motivated mathematically and statistically.

We present detailed experiments for text classification, using both the WebKB and Reuters data sets, which have become standard test collections. Our experimental results indicate that the multinomial information diffusion kernel performs very well empirically. This improvement can in part be attributed to the role of the Fisher information metric, which results in points near the boundary of the simplex being given relatively more importance than in the flat Euclidean metric. Viewed differently, effects similar to those obtained by heuristically designed term weighting schemes such as inverse document frequency are seen to arise automatically from the geometry of the statistical manifold.

The section is organized as follows. In Section 8.1 we define the relevant concepts from Riemannian geometry, that have not been described in Section 2 and then proceed to define the heat kernel for a general manifold, together with its parametrix expansion. In Section 8.3, we derive bounds on covering numbers and Rademacher averages for various learning algorithms that use the new kernels, borrowing results from differential geometry on bounds for the geometric Laplacian. Section 8.4 describes the results of applying the multinomial diffusion kernels to text classification, and we conclude with a discussion of our results in Section 8.6.

8.1 Riemannian Geometry and the Heat Kernel

We begin by briefly reviewing some relevant concepts from Riemannian geometry that will be used in the construction of information diffusion kernels. These concepts complement the ones defined in Section 2. Using these concepts, the heat kernel is defined, and its basic properties are presented. An excellent introductory account of this topic is given by Rosenberg (1997), and an authoritative reference for spectral methods in Riemannian geometry is (Schoen & Yau, 1994).

The construction of our kernels is based on the Laplacian[9]. One way to describe the appropriate generalization of the Euclidean Laplacian to arbitrary Riemannian manifold is through the notions of gradient and divergence. The gradient of a function is defined as the vector field that satisfies

$$\text{grad} : C^\infty(\mathcal{M}, \mathbb{R}) \to \mathfrak{X}(\mathcal{M}) \qquad g_p(\text{grad}\, f|_p, X_p) \;=\; X_p(f) \tag{67}$$

for every vector field $X \in \mathfrak{X}(\mathcal{M})$ and every point $p \in \mathcal{M}$. In local coordinates, the gradient is given by

$$(\text{grad}\, f|_p)_i = \sum_j [G^{-1}(p)]_{ij} \frac{\partial f(p)}{\partial x_j} \tag{68}$$

where $G(p)$ is the gram matrix associated with the metric g (see Section 2).

[9] As described by Nelson (1968), "The Laplace operator in its various manifestations is the most beautiful and central object in all of mathematics. Probability theory, mathematical physics, Fourier analysis, partial differential equations, the theory of Lie groups, and differential geometry all revolve around this sun, and its light even penetrates such obscure regions as number theory and algebraic geometry."

The divergence operator

$$\operatorname{div} : \mathfrak{X}(\mathcal{M}) \to C^\infty(\mathcal{M}, \mathbb{R})$$

is defined to be the adjoint of the gradient, allowing "integration by parts" on manifolds with special structure. In local coordinates, the divergence is the function

$$\operatorname{div} X\,(p) = \frac{1}{\sqrt{\det g(p)}} \sum_i \frac{\partial}{\partial x_i} \left(\sqrt{\det g(p)}\, (X_p)_i \right) \tag{69}$$

where $\det g(p)$ is the determinant[10] of the Gram matrix $G(p)$.

Finally, the Laplace-Beltrami operator or the Laplacian is defined by[11]

$$\Delta : C^\infty(\mathcal{M}, \mathbb{R}) \to C^\infty(\mathcal{M}, \mathbb{R}) \qquad \Delta = \operatorname{div} \circ \operatorname{grad} \tag{70}$$

which in local coordinates is given by

$$\Delta f\,(p) = \frac{1}{\sqrt{\det g(p)}} \sum_{ij} \frac{\partial}{\partial x_i} \left([G(p)^{-1}]_{ij} \sqrt{\det g(p)}\, \frac{\partial f}{\partial x_j} \right). \tag{71}$$

These definitions preserve the familiar intuitive interpretation of the usual operators in Euclidean geometry; in particular, the gradient $\operatorname{grad} f$ points in the direction of steepest ascent of f and the divergence $\operatorname{div} X$ measures outflow minus inflow of liquid or heat flowing according to the vector field X.

8.1.1 The Heat Kernel

The Laplacian is used to model how heat will diffuse throughout a geometric manifold; the flow $f(x,t)$, at point x and time t, is governed by the following second order partial differential equation with initial conditions

$$\frac{\partial f}{\partial t} - \Delta f = 0 \tag{72}$$

$$f(x,0) = f_0(x). \tag{73}$$

The value $f(x,t)$ describes the heat at location x and time t, beginning from an initial distribution of heat given by $f_0(x)$ at time zero. The heat or diffusion kernel $K_t(x,y)$ is the solution to the heat equation $f(x,t)$ with initial condition given by Dirac's delta function δ_y. As a consequence of the linearity of the heat equation, the heat kernel can be used to generate the solution to the heat equation with arbitrary initial conditions, according to

$$f(x,t) = \int_{\mathcal{M}} K_t(x,y)\, f(y)\, \mathrm{d}y. \tag{74}$$

[10] Most definition of the divergence, require the manifold to be oriented. We ignore this issue because it is not important to what follows and we will always work with orientable manifolds.

[11] There is no general agreement about the sign convention for the Laplacian. Many authors define the Laplacian as the negative of the present definition.

When $\mathcal{M} = \mathbb{R}$ with the Euclidean metric, the heat kernel is the familiar Gaussian kernel, so that the solution to the heat equation is expressed as

$$f(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{\mathbb{R}} e^{-\frac{(x-y)^2}{4t}} f(y)\,\mathrm{d}y \tag{75}$$

and it is seen that as $t \to \infty$, the heat diffuses out "to infinity" so that $f(x,t) \to 0$.

When $\mathcal{M}$ is compact the Laplacian has discrete eigenvalues $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \cdots$ with corresponding eigenfunctions ϕ_i satisfying $\Delta\phi_i = -\lambda_i\phi_i$. When the manifold has a boundary, appropriate boundary conditions must be imposed in order for Δ to be self-adjoint. Dirichlet boundary conditions set $\phi_i|_{\partial\mathcal{M}} = 0$ and Neumann boundary conditions require $\frac{\partial\phi_i}{\partial\nu}\Big|_{\partial\mathcal{M}} = 0$ where ν is the outer normal direction. The following theorem summarizes the basic properties for the kernel of the heat equation on $\mathcal{M}$; we refer to (Schoen & Yau, 1994) for a proof.

Theorem 3. *Let $\mathcal{M}$ be a complete Riemannian manifold. Then there exists a function $K \in C^\infty(\mathbb{R}_+ \times M \times M)$, called the heat kernel, which satisfies the following properties for all $x, y \in M$, with $K_t(\cdot,\cdot) = K(t,\cdot,\cdot)$*

1. $K_t(x,y) = K_t(y,x)$
2. $\lim_{t\to 0} K_t(x,y) = \delta_x(y)$
3. $\left(\Delta - \frac{\partial}{\partial t}\right) K_t(x,y) = 0$
4. $K_t(x,y) = \int_{\mathcal{M}} K_{t-s}(x,z)K_s(z,y)\,dz$ *for any* $s > 0$.

If in addition M is compact, then K_t can be expressed in terms of the eigenvalues and eigenfunctions of the Laplacian as $K_t(x,y) = \sum_{i=0}^{\infty} e^{-\lambda_i t}\phi_i(x)\,\phi_i(y)$.

Properties 2 and 3 imply that $K_t(x,y)$ solves the heat equation in x, starting from y. It follows that $e^{t\Delta}f(x) = f(x,t) = \int_M K_t(x,y)\,f(y)\,\mathrm{d}y$ solves the heat equation with initial conditions $f(x,0) = f(x)$, since

$$\begin{aligned}
\frac{\partial f(x,t)}{\partial t} &= \int_M \frac{\partial K_t(x,y)}{\partial t} f(y)\,\mathrm{d}y && (76)\\
&= \int_M \Delta K_t(x,y)\,f(y)\,\mathrm{d}y && (77)\\
&= \Delta \int_M K_t(x,y)\,f(y)\,\mathrm{d}y && (78)\\
&= \Delta f(x) && (79)
\end{aligned}$$

and $\lim_{t\to 0} f(x,t) = \int_M \lim_{t\to 0} K_t(x,y)\,dy = f(x)$. Property 4 implies that $e^{t\Delta}e^{s\Delta} = e^{(t+s)\Delta}$, which has the physically intuitive interpretation that heat diffusion for time t is the composition of heat diffusion up to time s with heat diffusion for an additional time $t - s$. Since $e^{t\Delta}$ is a positive operator,

$$\int_M \int_M K_t(x,y)f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y = \left\langle f, e^{t\Delta}f\right\rangle \geq 0 \tag{80}$$

and $K_t(x,y)$ is positive-definite. In the compact case, positive-definiteness follows directly from the expansion $K_t(x,y) = \sum_{i=0}^{\infty} e^{-\lambda_i t}\phi_i(x)\,\phi_i(y)$, which shows that the eigenvalues of K_t as an integral operator are $e^{-\lambda_i t}$. Together, these properties show that K_t defines a Mercer kernel.

The heat kernel $K_t(x, y)$ is a natural candidate for measuring the similarity between points between $x, y \in \mathcal{M}$, while respecting the geometry encoded in the metric g. Furthermore it is, unlike the geodesic distance, a Mercer kernel – a fact that enables its use in statistical kernel machines. When this kernel is used for classification, as in our text classification experiments presented in Section 8.4, the discriminant function $y_t(x) = \sum_i \alpha_i y_i K_t(x, x_i)$ can be interpreted as the solution to the heat equation with initial temperature $y_0(x_i) = \alpha_i\, y_i$ on labeled data points x_i, and $y_0(x) = 0$ elsewhere.

8.1.2 The parametrix expansion

For most geometries, there is no closed form solution for the heat kernel. However, the short time behavior of the solutions can be studied using an asymptotic expansion, called the *parametrix expansion*. In fact, the existence of the heat kernel, as asserted in the above theorem, is most directly proven by first showing the existence of the parametrix expansion. Although it is local, the parametrix expansion contains a wealth of geometric information, and indeed much of modern differential geometry, notably index theory, is based upon this expansion and its generalizations. In Section 8.4 we will employ the first-order parametrix expansion for text classification.

Recall that the heat kernel on n-dimensional Euclidean space is given by

$$K_t^{\text{Euclid}}(x, y) \;=\; (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{\|x - y\|^2}{4t}\right) \tag{81}$$

where $\|x-y\|^2 = \sum_{i=1}^n |x_i - y_i|^2$ is the squared Euclidean distance between x and y. The parametrix expansion approximates the heat kernel locally as a correction to this Euclidean heat kernel. It is given by

$$P_t^{(m)}(x, y) = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{d^2(x, y)}{4t}\right) (\psi_0(x, y) + \psi_1(x, y)t + \cdots + \psi_m(x, y)t^m) \tag{82}$$

where d is the geodesic distance and ψ_k are recursively obtained by solving the heat equation approximately to order t^m, for small diffusion time t. Denoting $K_t^{(m)}(x, y) = P_t^{(m)}(x, y)$ we thus obtain an approximation for the heat kernel, that converges as $t \to 0$ and $x \to y$. For further details refer to (Schoen & Yau, 1994; Rosenberg, 1997).

While the parametrix $K_t^{(m)}$ is not in general positive-definite, and therefore does not define a Mercer kernel, it is positive-definite for t sufficiently small. In particular, define $f(t) = \min \operatorname{spec}\left(K_t^{(m)}\right)$, where $\min \operatorname{spec}$ denotes the smallest eigenvalue. Then f is a continuous function with $f(0) = 1$ since $K_0^{(m)} = I$. Thus, there is some time interval $[0, \epsilon)$ for which $K_t^{(m)}$ is positive-definite in case $t \in [0, \epsilon)$.

The following two basic examples illustrate the geometry of the Fisher information metric and the associated diffusion kernel it induces on a statistical manifold. Under the Fisher information metric, the spherical normal family corresponds to a manifold of constant negative curvature, and the multinomial corresponds to a manifold of constant positive curvature. The multinomial will be the most important example that we develop, and we will report extensive experiments with the resulting kernels in Section 8.4.

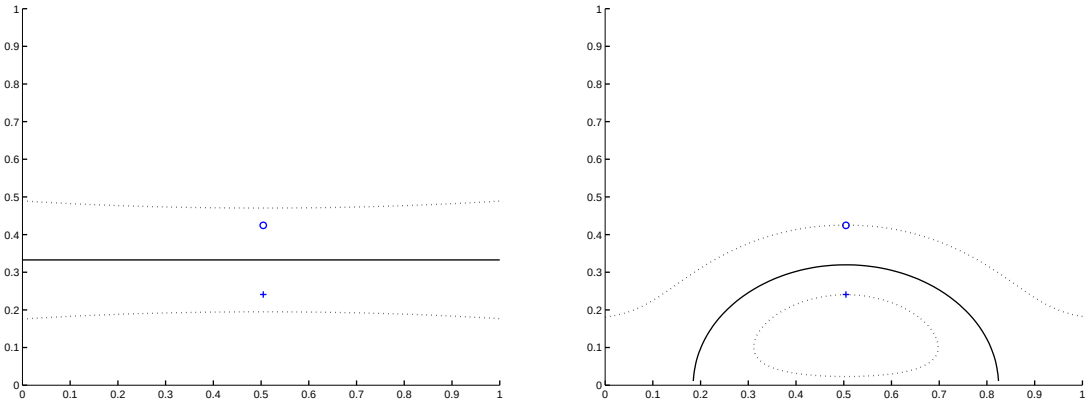

Figure 15: Example decision boundaries for a kernel-based classifier using information diffusion kernels for spherical normal geometry with $d = 2$ (right), which has constant negative curvature, compared with the standard Gaussian kernel for flat Euclidean space (left). Two data points are used, simply to contrast the underlying geometries. The curved decision boundary for the diffusion kernel can be interpreted statistically by noting that as the variance decreases the mean is known with increasing certainty.

The heat kernel on the hyperbolic space $\mathbb{H}^n$ has the following closed form (Grigor'yan & Noguchi, 1998). For odd $n = 2m + 1$ it is given by

$$K_t(x, x') = \frac{(-1)^m}{2^m \pi^m} \frac{1}{\sqrt{4\pi t}} \left(\frac{1}{\sinh r} \frac{\partial}{\partial r} \right)^m \exp\left(-m^2 t - \frac{r^2}{4t} \right) \tag{83}$$

and for even $n = 2m + 2$ it is given by

$$K_t(x, x') = \frac{(-1)^m}{2^m \pi^m} \frac{\sqrt{2}}{\sqrt{4\pi t^3}} \left(\frac{1}{\sinh r} \frac{\partial}{\partial r} \right)^m \int_r^\infty \frac{s \exp\left(-\frac{(2m+1)^2 t}{4} - \frac{s^2}{4t} \right)}{\sqrt{\cosh s - \cosh r}} \, \mathrm{d}s \tag{84}$$

where $r = d(x, x')$ is the geodesic distance between the two points in $\mathbb{H}^n$ given by equation (15). If only the mean $\theta = \mu$ is unspecified, then the associated kernel is the standard RBF or Gaussian kernel. The hyperbolic geometry is illustrated in Figure 15 where decision boundaries of SVM with the diffusion kernel are plotted for both Euclidean and hyperbolic geometry.

Unlike the explicit expression for the Gaussian geometry discussed above, there is no explicit form for the heat kernel on the sphere, nor on the positive orthant of the sphere. We will therefore resort to the parametrix expansion to derive an approximate heat kernel for the multinomial geometry.

For the n-sphere it can be shown (Berger et al., 1971) that the function ψ_0 of in the parametrix expansion, which is the leading order correction of the Gaussian kernel under the Fisher information metric, is given by

$$\begin{aligned} \psi_0(r) &= \left(\frac{\sqrt{\det g}}{r^{n-1}} \right)^{-\frac{1}{2}} = \left(\frac{\sin r}{r} \right)^{-\frac{(n-1)}{2}} \\ &= 1 + \frac{(n-1)}{12} r^2 + \frac{(n-1)(5n-1)}{1440} r^4 + O(r^6). \end{aligned}$$

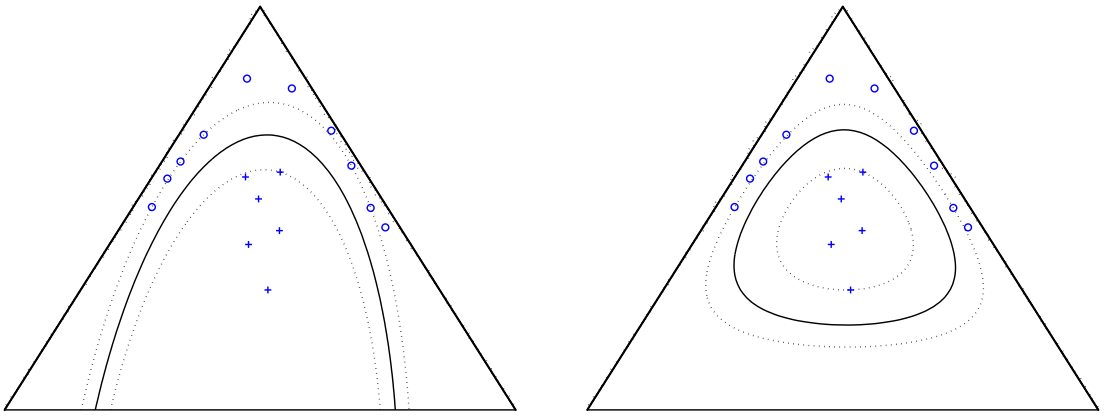

Figure 16: Example decision boundaries using support vector machines with information diffusion kernels for trinomial geometry on the 2-simplex (top right) compared with the standard Gaussian kernel (left).

In our experiments we approximate the diffusion kernel using $\psi_0 \equiv 1$ and obtain

$$K(\theta, \theta') = (4\pi t)^{-\frac{n}{2}} e^{-\frac{1}{t}\arccos^2(\sum_i \sqrt{\theta_i \theta_i'})}. \tag{85}$$

In Figure 16 the kernel (85) is compared with the standard Euclidean space Gaussian kernel for the case of the trinomial model, $d = 2$, using an SVM classifier.

8.2 Rounding the Simplex

The case of multinomial geometry poses some technical complications for the analysis of diffusion kernels, due to the fact that the open simplex is not complete, and moreover, its closure is not a differentiable manifold with boundary. Thus, it is technically not possible to apply several results from differential geometry, such as bounds on the spectrum of the Laplacian, as adopted in Section 8.3. We now briefly describe a technical "patch" that allows us to derive all of the needed analytical results, without sacrificing in practice any of the methodology that has been derived so far. The idea is to "round the corners" of $\overline{\mathbb{P}_n}$ to obtain a compact manifold with boundary, and that closely approximates the original simplex $\mathbb{P}_n$.

For $\epsilon > 0$, let $B_\epsilon(x) = \{y \,|\, \|x - y\| < \epsilon\}$ be the open Euclidean ball of radius ϵ centered at x and $C_\epsilon(\mathbb{P}_n)$ be

$$C_\epsilon(\mathbb{P}_n) = \left\{x \in \overline{\mathbb{P}_n} \;:\; B_\epsilon(x) \subset \mathbb{P}_n\right\} \tag{86}$$

The ϵ-rounded simplex is then defined as the closure of

$$\mathbb{P}_n^\epsilon = \bigcup_{x \in C_\epsilon(\mathbb{P}_n)} B_\epsilon(x). \tag{87}$$

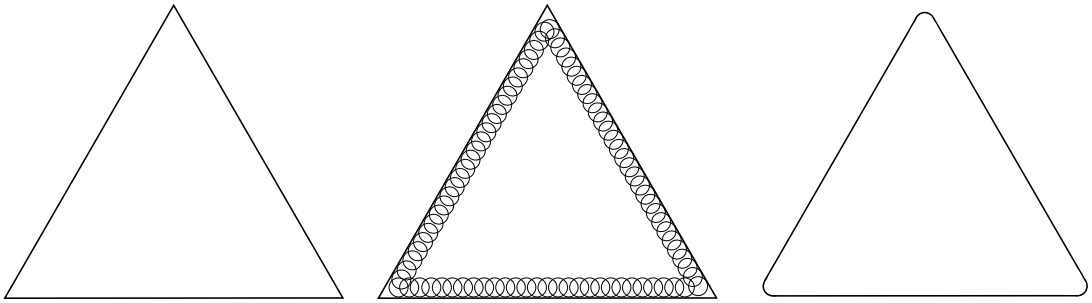

Figure 17: Rounding the simplex. Since the closed simplex is not a manifold with boundary, we carry out a "rounding" procedure to remove edges and corners. The ϵ-rounded simplex is the closure of the union of all ϵ-balls lying within the open simplex.

The above rounding procedure that yields $\overline{\mathbb{P}_2^\epsilon}$ is suggested by Figure 17. Note that in general the ϵ-rounded simplex $\overline{\mathbb{P}_n^\epsilon}$ will contain points with a single, but not more than one component having zero probability and it forms a compact manifold with boundary whose image under the isometry F described above is a compact submanifold with boundary of the n-sphere. Since, by choosing ϵ small enough we can approximate $\overline{\mathbb{P}_n}$ arbitrarily well (both in the Euclidean and geodesic distances), no harm is done by assuming that we are dealing with a rounded compact manifold with a boundary.

8.3 Spectral Bounds on Covering Numbers and Rademacher Averages

We now turn to establishing bounds on the generalization performance of kernel machines that use information diffusion kernels. We begin by adopting the approach of Guo et al. (2002), estimating covering numbers by making use of bounds on the spectrum of the Laplacian on a Riemannian manifold, rather than on VC dimension techniques; these bounds in turn yield bounds on the expected risk of the learning algorithms. Our calculations give an indication of how the underlying geometry influences the entropy numbers, which are inverse to the covering numbers. We then show how bounds on Rademacher averages may be obtained by plugging in the spectral bounds from differential geometry. The primary conclusion that is drawn from these analyses is that from the point of view of generalization error bounds, information diffusion kernels behave essentially the same as the standard Gaussian kernel.

8.3.1 Covering Numbers

We begin by recalling the main result of Guo et al. (2002), modifying their notation slightly to conform with ours. Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact subset of d-dimensional Euclidean space, and suppose that $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a Mercer kernel. Denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ the eigenvalues of K, i.e., of the mapping $f \mapsto \int_{\mathcal{X}} K(\cdot, y)\, f(y)\, dy$, and let $\psi_j(\cdot)$ denote the corresponding eigenfunctions. We assume that $C_K \stackrel{\text{def}}{=} \sup_j \|\psi_j\|_\infty < \infty$.

Given m points $x_i \in \mathcal{X}$, the kernel hypothesis class for $\boldsymbol{x} = \{x_i\}$ with weight vector bounded by R is defined as the collection of functions on $\boldsymbol{x}$ given by

$$\mathcal{F}_R(\boldsymbol{x}) = \{f : f(x_i) = \langle w, \Phi(x_i)\rangle \text{ for some } \|w\| \leq R\} \tag{88}$$

where $\Phi(\cdot)$ is the mapping from M to feature space defined by the Mercer kernel, and $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the corresponding Hilbert space inner product and norm. It is of interest to obtain uniform bounds on the covering numbers $\mathcal{N}(\epsilon, \mathcal{F}_R(\boldsymbol{x}))$, defined as the size of the smallest ϵ-cover of $\mathcal{F}_R(\boldsymbol{x})$ in the metric induced by the norm $\|f\|_{\infty,\boldsymbol{x}} = \max_{i=1,\ldots,m} |f(x_i)|$. The following is the main result of (Guo et al., 2002).

Theorem 4. *Given an integer $n \in \mathbb{N}$, let j_n^* denote the smallest integer j for which*

$$\lambda_{j+1} < \left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right)^{\frac{1}{j}} \tag{89}$$

and define

$$\epsilon_n^* = 6C_K R \sqrt{j_n^* \left(\frac{\lambda_1 \cdots \lambda_{j_n^*}}{n^2}\right)^{\frac{1}{j_n^*}} + \sum_{i=j_n^*}^{\infty} \lambda_i}\,. \tag{90}$$

Then $\sup_{\{x_i\}\in M^m} \mathcal{N}(\epsilon_n^*, \mathcal{F}_R(\boldsymbol{x})) \leq n$.

To apply this result, we will obtain bounds on the indices j_n^* using spectral theory in Riemannian geometry. The following bounds on the eigenvalues of the Laplacian are due to (Li & Yau, 1980).

Theorem 5. *Let M be a compact Riemannian manifold of dimension d with non-negative Ricci curvature, and let $0 < \mu_1 \leq \mu_2 \leq \cdots$ denote the eigenvalues of the Laplacian with Dirichlet boundary conditions. Then*

$$c_1(d)\left(\frac{j}{V}\right)^{\frac{2}{d}} \;\leq\; \mu_j \;\leq\; c_2(d)\left(\frac{j+1}{V}\right)^{\frac{2}{d}} \tag{91}$$

where V is the volume of M and c_1 and c_2 are constants depending only on the dimension.

Note that the manifold of the multinomial model satisfies the conditions of this theorem. Using these results we can establish the following bounds on covering numbers for information diffusion kernels. We assume Dirichlet boundary conditions; a similar result can be proven for Neumann boundary conditions. We include the constant $V = \mathrm{vol}(M)$ and diffusion coefficient t in order to indicate how the bounds depend on the geometry.

Theorem 6. *Let M be a compact Riemannian manifold, with volume V, satisfying the conditions of Theorem 5. Then the covering numbers for the Dirichlet heat kernel K_t on M satisfy*

$$\log \mathcal{N}(\epsilon, \mathcal{F}_R(\boldsymbol{x})) = O\left(\left(\frac{V}{t^{\frac{d}{2}}}\right) \log^{\frac{d+2}{2}}\left(\frac{1}{\epsilon}\right)\right) \tag{92}$$

Proof. By the lower bound in Theorem 5, the Dirichlet eigenvalues of the heat kernel $K_t(x,y)$, which are given by $\lambda_j = e^{-t\mu_j}$, satisfy $\log \lambda_j \leq -tc_1(d)\left(\frac{j}{V}\right)^{\frac{2}{d}}$. Thus,

$$-\frac{1}{j}\log\left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right) \;\geq\; \frac{tc_1}{j}\sum_{i=1}^{j}\left(\frac{i}{V}\right)^{\frac{2}{d}} + \frac{2}{j}\log n \;\geq\; tc_1\frac{d}{d+2}\left(\frac{j}{V}\right)^{\frac{2}{d}} + \frac{2}{j}\log n \tag{93}$$

where the second inequality comes from $\sum_{i=1}^{j} i^p \geq \int_0^j x^p\,dx = \frac{j^{p+1}}{p+1}$. Now using the upper bound of Theorem 5, the inequality $j_n^* \leq j$ will hold if

$$tc_2\left(\frac{j+2}{V}\right)^{\frac{2}{d}} \;\geq\; -\log\lambda_{j+1} \;\geq\; tc_1\frac{d}{d+2}\left(\frac{j}{V}\right)^{\frac{2}{d}} + \frac{2}{j}\log n \tag{94}$$

or equivalently

$$\frac{tc_2}{V^{\frac{2}{d}}}\left(j(j+2)^{\frac{2}{d}} - \frac{c_1}{c_2}\frac{d}{d+2}j^{\frac{d+2}{d}}\right) \;\geq\; 2\log n \tag{95}$$

The above inequality will hold in case

$$j \geq \left\lceil \left(\frac{2V^{\frac{2}{d}}}{t(c_2 - c_1\frac{d}{d+2})}\log n\right)^{\frac{d}{d+2}} \right\rceil \;\geq\; \left\lceil \left(\frac{V^{\frac{2}{d}}(d+2)}{tc_1}\log n\right)^{\frac{d}{d+2}} \right\rceil \tag{96}$$

since we may assume that $c_2 \geq c_1$; thus, $j_n^* \leq \left\lceil \overline{c}_1\left(\frac{V^{\frac{2}{d}}}{t}\log n\right)^{\frac{d}{d+2}} \right\rceil$ for a new constant $\overline{c}_1(d)$. Plugging this bound on j_n^* into the expression for ϵ_n^* in Theorem 2 and using

$$\sum_{i=j_n^*}^{\infty} e^{-i^{\frac{2}{d}}} = O\left(e^{-j_n^{*\frac{2}{d}}}\right) \tag{97}$$

we have after some algebra that

$$\log\left(\frac{1}{\epsilon_n}\right) = \Omega\left(\left(\frac{t}{V^{\frac{2}{d}}}\right)^{\frac{d}{d+2}}\log^{\frac{2}{d+2}} n\right) \tag{98}$$

Inverting the above expression in $\log n$ gives equation (92). □

We note that Theorem 4 of (Guo et al., 2002) can be used to show that this bound does not, in fact, depend on m and $\boldsymbol{x}$. Thus, for fixed t the covering numbers scale as $\log\mathcal{N}(\epsilon,\mathcal{F}) = O\left(\log^{\frac{d+2}{2}}\left(\frac{1}{\epsilon}\right)\right)$, and for fixed ϵ they scale as $\log\mathcal{N}(\epsilon,\mathcal{F}) = O\left(t^{-\frac{d}{2}}\right)$ in the diffusion time t.

8.3.2 Rademacher Averages

We now describe a different family of generalization error bounds that can be derived using the machinery of Rademacher averages Bartlett and Mendelson (2002); Bartlett et al. (2003). The bounds fall out directly from the work of (Mendelson, 2003) on computing local averages for kernel-based function classes, after plugging in the eigenvalue bounds of Theorem 3.

As seen above, covering number bounds are related to a complexity term of the form

$$C(n) = \sqrt{j_n^* \left(\frac{\lambda_1 \cdots \lambda_{j_n^*}}{n^2} \right)^{\frac{1}{j_n^*}} + \sum_{i=j_n^*}^{\infty} \lambda_i} \tag{99}$$

In the case of Rademacher complexities, risk bounds are instead controlled by a similar, yet simpler expression of the form

$$C(r) = \sqrt{j_r^* \, r + \sum_{i=j_r^*}^{\infty} \lambda_i} \tag{100}$$

where now j_r^* is the smallest integer j for which $\lambda_j < r$ Mendelson (2003), with r acting as a parameter bounding the error of the family of functions. To place this into some context, we quote the following results from (Bartlett et al., 2003) and (Mendelson, 2003), which apply to a family of loss functions that includes the quadratic loss; we refer to (Bartlett et al., 2003) for details on the technical conditions.

Let $(X_1, Y_1), (X_2, Y_2) \ldots, (X_n, Y_n)$ be an independent sample from an unknown distribution P on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} \subset \mathbb{R}$. For a given loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, and a family $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathcal{Y}$, the objective is to minimize the expected loss $E[\ell(f(X), Y)]$. Let $E\ell_{f^*} = \inf_{f \in \mathcal{F}} E\ell_f$, where $\ell_f(X, Y) = \ell(f(X), Y)$, and let $\hat{f}$ be any member of $\mathcal{F}$ for which $E_n \ell_{\hat{f}} = \inf_{f \in \mathcal{F}} E_n \ell_f$ where E_n denotes the empirical expectation. The *Rademacher average* of a family of functions $\mathfrak{G} = \{g : \mathcal{X} \to \mathbb{R}\}$ is defined as the expectation $ER_n\mathfrak{G} = E\left[\sup_{g \in \mathfrak{G}} R_n g\right]$ with $R_n g = \frac{1}{n} \sum_{i=1}^n \sigma_i \, g(X_i)$, where $\sigma_1, \ldots, \sigma_n$ are independent *Rademacher* random variables; that is, $p(\sigma_i = 1) = p(\sigma_i = -1) = \frac{1}{2}$.

Theorem 7. *Let $\mathcal{F}$ be a convex class of functions and define ψ by*

$$\psi(r) = a \, ER_n \left\{ f \in \mathcal{F} : E(f - f^*)^2 \leq r \right\} + \frac{b \, x}{n} \tag{101}$$

where a and b are constants that depend on the loss function ℓ. Then when $r \geq \psi(r)$,

$$E\left(\ell_{\hat{f}} - \ell_{f^*} \right) \leq c \, r + \frac{d \, x}{n} \tag{102}$$

with probability at least $1 - e^{-x}$, where c and d are additional constants.

Moreover, suppose that K is a Mercer kernel and $\mathcal{F} = \{f \in \mathcal{H}_K : \|f\|_K \leq 1\}$ is the unit ball in the reproducing kernel Hilbert space associated with K. Then

$$\psi(r) \leq a \sqrt{\frac{2}{n} \sum_{j=1}^{\infty} \min\{r, \lambda_j\}} + \frac{bx}{n} \tag{103}$$

Thus, to bound the excess risk for kernel machines in this framework it suffices to bound the term

$$\widetilde{\psi}(r) = \sqrt{\sum_{j=1}^{\infty} \min\{r, \lambda_j\}} \tag{104}$$

involving the spectrum. Given bounds on the eigenvalues, this is typically easy to do.

Theorem 8. *Let M be a compact Riemannian manifold, satisfying the conditions of Theorem 5. Then the Rademacher term $\widetilde{\psi}$ for the Dirichlet heat kernel K_t on M satisfies*

$$\widetilde{\psi}(r) \leq C\sqrt{\left(\frac{r}{t^{\frac{d}{2}}}\right)\log^{\frac{d}{2}}\left(\frac{1}{r}\right)} \tag{105}$$

for some constant C depending on the geometry of M.

Proof. We have that

$$\begin{aligned}
\widetilde{\psi}^2(r) &= \sum_{j=1}^{\infty}\min\{r,\lambda_j\} && (106)\\
&= j_r^*\, r + \sum_{j=j_r^*}^{\infty} e^{-t\mu_j} && (107)\\
&\leq j_r^*\, r + \sum_{j=j_r^*}^{\infty} e^{-tc_1 j^{\frac{2}{d}}} && (108)\\
&\leq j_r^*\, r + Ce^{-tc_1 {j_r^*}^{\frac{2}{d}}} && (109)
\end{aligned}$$

for some constant C, where the first inequality follows from the lower bound in Theorem 5. But $j_r^* \leq j$ in case $\log\lambda_{j+1} > r$, or, again from Theorem 5, if

$$t\, c_2(j+1)^{\frac{2}{d}} \leq -\log\lambda_j < \log\frac{1}{r} \tag{110}$$

or equivalently,

$$j_r^* \leq \frac{C'}{t^{\frac{d}{2}}}\log^{\frac{d}{2}}\left(\frac{1}{r}\right) \tag{111}$$

It follows that

$$\widetilde{\psi}^2(r) \leq C''\left(\frac{r}{t^{\frac{d}{2}}}\right)\log^{\frac{d}{2}}\left(\frac{1}{r}\right) \tag{112}$$

for some new constant C''. □

From this bound, it can be shown that, with high probability,

$$E\left(\ell_{\hat{f}} - \ell_{f^*}\right) = O\left(\frac{\log^{\frac{d}{2}} n}{n}\right) \tag{113}$$

which is the behavior expected of the Gaussian kernel for Euclidean space.

Thus, for both covering numbers and Rademacher averages, the resulting bounds are essentially the same as those that would be obtained for the Gaussian kernel on the flat d-dimensional torus, which is the standard way of "compactifying" Euclidean space to get a Laplacian having only discrete spectrum; the results of (Guo et al., 2002) are formulated for the case $d = 1$, corresponding to the circle. While the bounds for information diffusion kernels were derived for the case of positive curvature, which apply to the special case of the multinomial, similar bounds for general manifolds with curvature bounded below by a negative constant should also be attainable.

8.4 Experimental Results for Text Classification

In this section we present the application of multinomial diffusion kernels to the problem of text classification. Text processing can be subject to some of the "dirty laundry" referred to in the introduction—documents are cast as Euclidean space vectors with special weighting schemes that have been empirically honed through applications in information retrieval, rather than inspired from first principles. However for text, the use of multinomial geometry is natural and well motivated; our experimental results offer some insight into how useful this geometry may be for classification.

We consider several embeddings $\hat{\theta} : \mathcal{X} \to \Theta$ of documents in bag of words representation into the probability simplex. The term frequency (tf) representation uses normalized counts; the corresponding embedding is the maximum likelihood estimator for the multinomial distribution

$$\widehat{\theta}_{\text{tf}}(x) = \left(\frac{x_1}{\sum_i x_i}, \ldots, \frac{x_{n+1}}{\sum_i x_i} \right). \tag{114}$$

Another common representation is based on term frequency, inverse document frequency (tf-idf). This representation uses the distribution of terms across documents to discount common terms; the document frequency df_v of term v is defined as the number of documents in which term v appears. Although many variants have been proposed, one of the simplest and most commonly used embeddings is

$$\widehat{\theta}_{\text{tf-idf}}(x) = \left(\frac{x_1 \log(D/df_1)}{\sum_i x_i \log(D/df_i)}, \ldots, \frac{x_{n+1} \log(D/df_{n+1})}{\sum_i x_i \log(D/df_i)} \right) \tag{115}$$

where D is the number of documents in the corpus.

In text classification applications the tf and tf-idf representations are typically normalized to unit length in the L_2 norm rather than the L_1 norm, as above Joachims (2000). For example, the tf representation with L_2 normalization is given by

$$x \mapsto \left(\frac{x_1}{\sum_i x_i^2}, \ldots, \frac{x_{n+1}}{\sum_i x_i^2} \right) \tag{116}$$

and similarly for tf-idf. When used in support vector machines with linear or Gaussian kernels, L_2-normalized tf and tf-idf achieve higher accuracies than their L_1-normalized counterparts. However, for the diffusion kernels, L_1 normalization is necessary to obtain an embedding into the simplex. These different embeddings or feature representations are compared in the experimental results reported below.

The three kernels that we compare are the linear kernel

$$K^{\text{Lin}}(\theta, \theta') \sum_{v=1}^{n+1} \theta_v \, \theta'_v, \tag{117}$$

the Gaussian kernel

$$K_\sigma^{\text{Gauss}}(\theta', \theta') \;=\; (2\pi\sigma)^{-\frac{n+1}{2}} \exp\left(-\frac{\sum_{i=1}^{n+1} |\theta_i - \theta'_i|^2}{2\sigma^2} \right) \tag{118}$$

and the multinomial diffusion kernel approximation

$$K_t^{\text{Mult}}(\theta, \theta') = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{1}{t} \arccos^2\left(\sum_{i=1}^{n+1} \sqrt{\theta_i \theta_i'}\right)\right). \tag{119}$$

In our experiments, the multinomial diffusion kernel using the tf embedding was compared to the linear or Gaussian kernel with tf and tf-idf embeddings using a support vector machine classifier on the WebKB and Reuters-21578 collections, which are standard data sets for text classification.

The WebKB dataset contains web pages found on the sites of four universities Craven et al. (1998). The pages were classified according to whether they were student, faculty, course, project or staff pages; these categories contain 1641, 1124, 929, 504 and 137 instances, respectively. Since only the `student`, `faculty`, `course` and `project` classes contain more than 500 documents each, we restricted our attention to these classes. The Reuters-21578 dataset is a collection of newswire articles classified according to news topic Lewis and Ringuette (1994). Although there are more than 135 topics, most of the topics have fewer than 100 documents; for this reason, we restricted our attention to the following five most frequent classes: `earn`, `acq`, `moneyFx`, `grain` and `crude`, of sizes 3964, 2369, 717, 582 and 578 documents, respectively.

For both the WebKB and Reuters collections we created two types of binary classification tasks. In the first task we designate a specific class, label each document in the class as a "positive" example, and label each document on any of the other topics as a "negative" example. In the second task we designate a class as the positive class, and choose the negative class to be the most frequent remaining class (`student` for WebKB and `earn` for Reuters). In both cases, the size of the training set is varied while keeping the proportion of positive and negative documents constant in both the training and test set.

Figure 18 shows the test set error rate for the WebKB data, for a representative instance of the one-versus-all classification task; the designated class was `course`. The results for the other choices of positive class were qualitatively very similar; all of the results are summarized in Table 2. Similarly, Figure 20 shows the test set error rates for two of the one-versus-all experiments on the Reuters data, where the designated classes were chosen to be `acq` and `moneyFx`. All of the results for Reuters one-versus-all tasks are shown in Table 4.

Figure 19 and Figure 21 show representative results for the second type of classification task, where the goal is to discriminate between two specific classes. In the case of the WebKB data the results are shown for `course` vs. `student`. In the case of the Reuters data the results are shown for `moneyFx` vs. `earn` and `grain` vs. `earn`. Again, the results for the other classes are qualitatively similar; the numerical results are summarized in Tables 3 and 5.

In these figures, the leftmost plots show the performance of tf features while the rightmost plots show the performance of tf-idf features. As mentioned above, in the case of the diffusion kernel we use L_1 normalization to give a valid embedding into the probability simplex, while for the linear

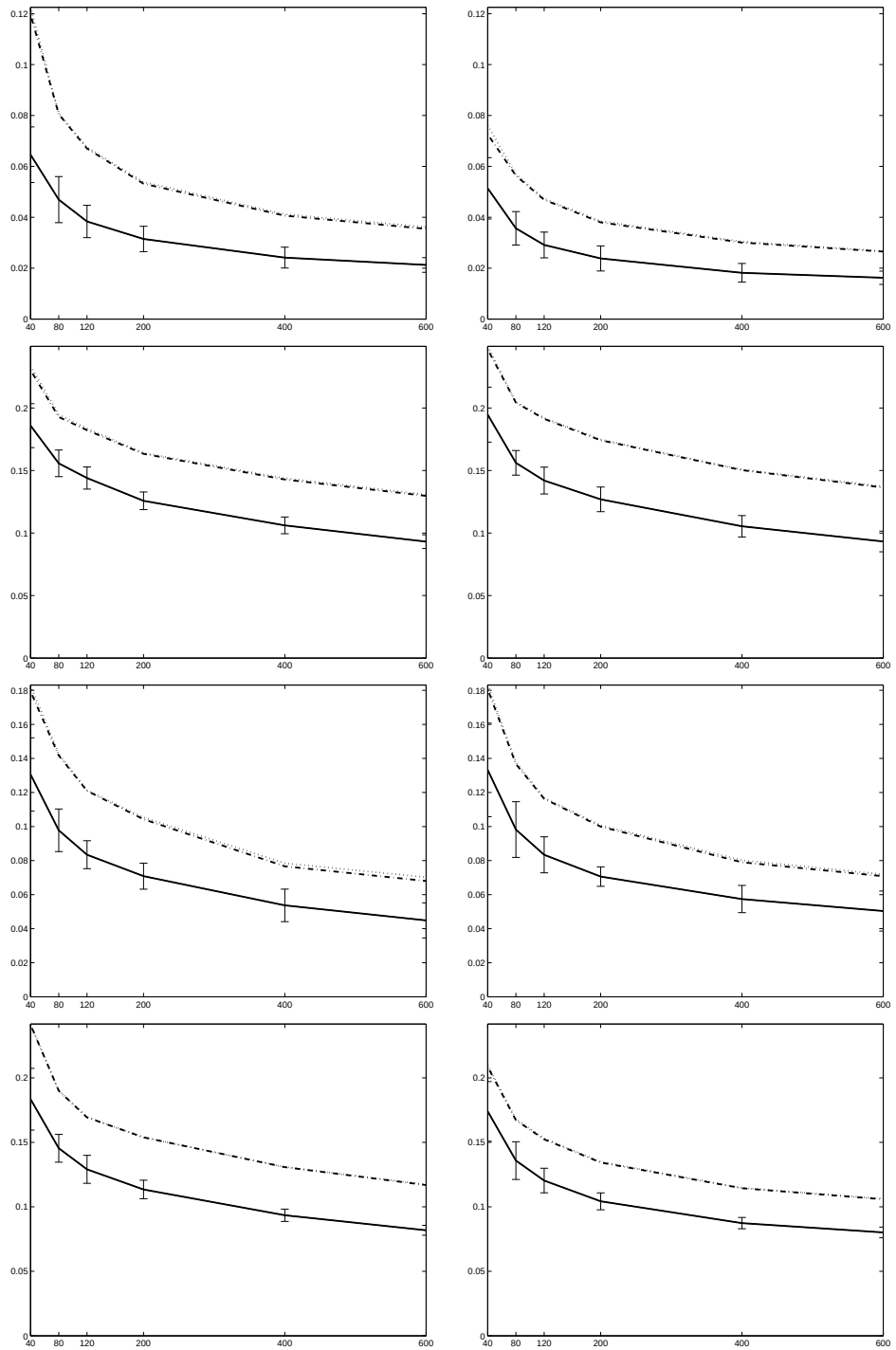

Figure 18: Experimental results on the WebKB corpus, using SVMs for linear (dotted) and Gaussian (dash-dotted) kernels, compared with the diffusion kernel for the multinomial (solid). Classification error for the task of labeling `course` (top row), `faculty` (second row), `project` (third row), and `student`(bottom row) is shown in these plots, as a function of training set size. The left plot uses tf representation and the right plot uses tf-idf representation. The curves shown are the error rates averaged over 20-fold cross validation. The results for the other "1 vs. all" labeling tasks are qualitatively similar, and therefore not shown.

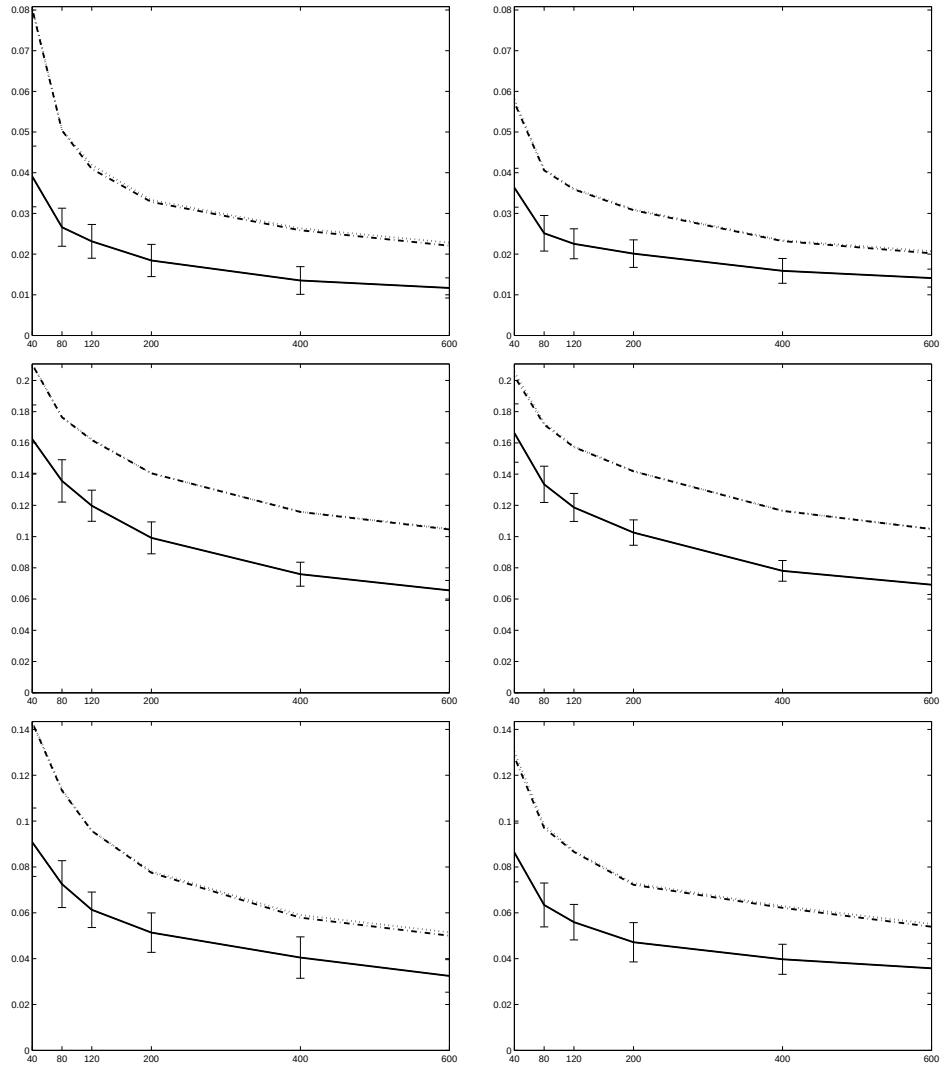

Figure 19: Results on the WebKB corpus, using SVMs for linear (dotted) and Gaussian (dash-dotted) kernels, compared with the diffusion kernel (solid). The tasks are `course` vs. `student` (top row), `faculty` vs. `student` (top row) and `project` vs. `student` (top row). The left plot uses tf representation and the right plot uses tf-idf representation. Results for other label pairs are qualitatively similar. The curves shown are the error rates averaged over 20-fold cross validation.

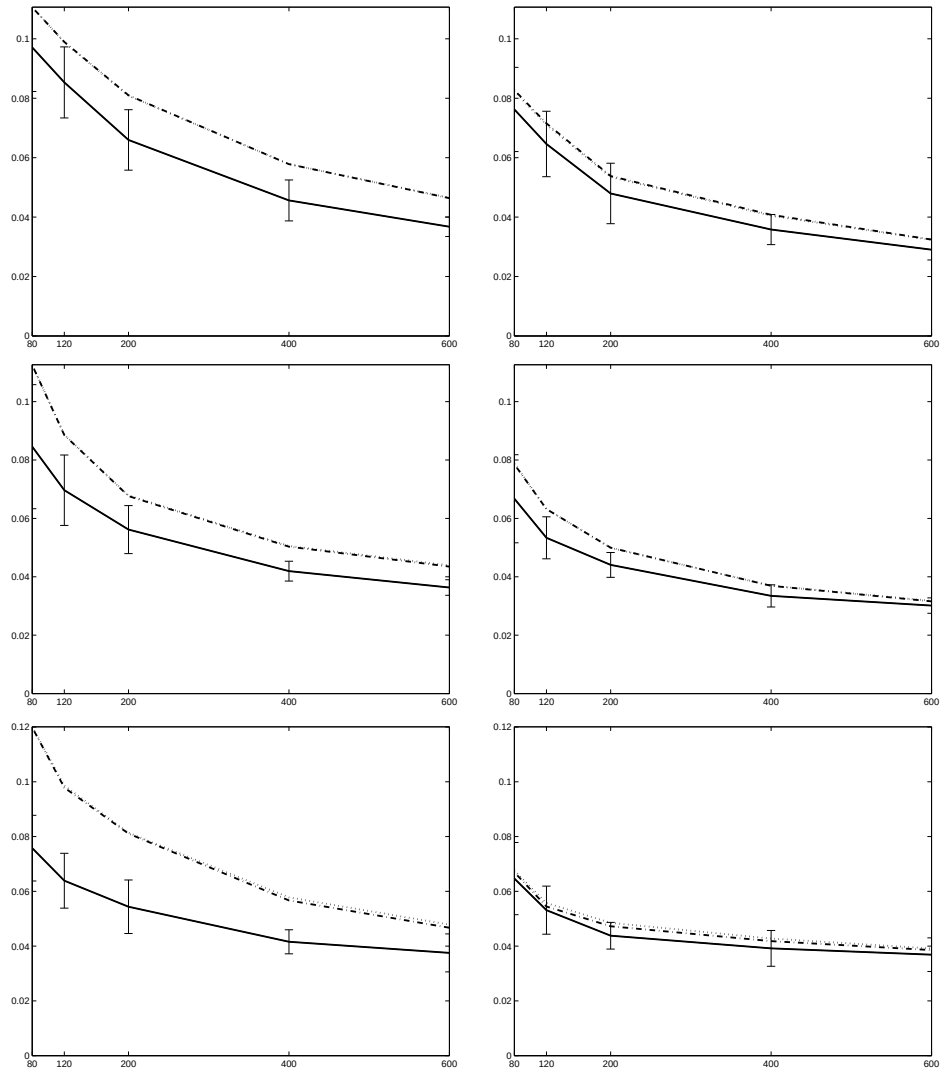

Figure 20: Experimental results on the Reuters corpus, using SVMs for linear (dotted) and Gaussian (dash-dotted) kernels, compared with the diffusion kernel (solid). The tasks are classifying `earn` (top row), `acq` (second row), `moneyFx` (bottom row) vs. the rest. Plots for the other classes are qualitatively similar. The left column uses tf representation and the right column uses tf-idf. The curves shown are the error rates averaged over 20-fold cross validation.

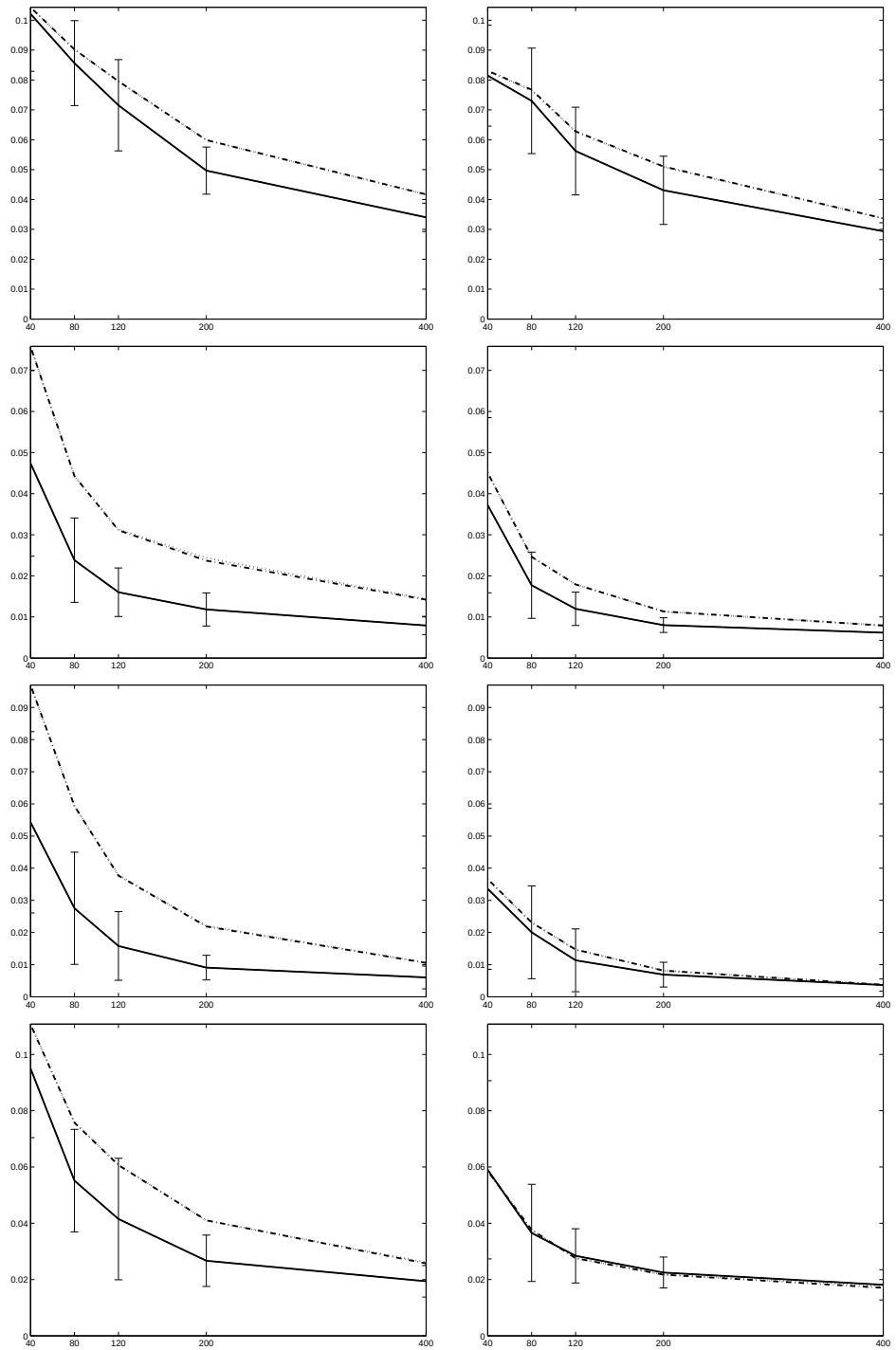

Figure 21: Experimental results on the Reuters corpus, using SVMs for linear (dotted) and Gaussian (dash-dotted) kernels, compared with the diffusion (solid). The tasks are `acq` vs. `earn` (top row), `moneyFx` vs. `earn` (top row), `grain` vs. `earn` (top row), `crude` vs. `earn` (top row). The left column uses tf representation and the right column uses tf-idf representation. The left column uses tf representation and the right column uses tf-idf. The curves shown are the error rates averaged over 20-fold cross validation.

		tf Representation			tf-idf Representation		
Task	L	Linear	Gaussian	Diffusion	Linear	Gaussian	Diffusion
course vs. all	40	0.1225	0.1196	**0.0646**	0.0761	0.0726	**0.0514**
	80	0.0809	0.0805	**0.0469**	0.0569	0.0564	**0.0357**
	120	0.0675	0.0670	**0.0383**	0.0473	0.0469	**0.0291**
	200	0.0539	0.0532	**0.0315**	0.0385	0.0380	**0.0238**
	400	0.0412	0.0406	**0.0241**	0.0304	0.0300	**0.0182**
	600	0.0362	0.0355	**0.0213**	0.0267	0.0265	**0.0162**
faculty vs. all	40	0.2336	0.2303	**0.1859**	0.2493	0.2469	**0.1947**
	80	0.1947	0.1928	**0.1558**	0.2048	0.2043	**0.1562**
	120	0.1836	0.1823	**0.1440**	0.1921	0.1913	**0.1420**
	200	0.1641	0.1634	**0.1258**	0.1748	0.1742	**0.1269**
	400	0.1438	0.1428	**0.1061**	0.1508	0.1503	**0.1054**
	600	0.1308	0.1297	**0.0931**	0.1372	0.1364	**0.0933**
project vs. all	40	0.1827	0.1793	**0.1306**	0.1831	0.1805	**0.1333**
	80	0.1426	0.1416	**0.0978**	0.1378	0.1367	**0.0982**
	120	0.1213	0.1209	**0.0834**	0.1169	0.1163	**0.0834**
	200	0.1053	0.1043	**0.0709**	0.1007	0.0999	**0.0706**
	400	0.0785	0.0766	**0.0537**	0.0802	0.0790	**0.0574**
	600	0.0702	0.0680	**0.0449**	0.0719	0.0708	**0.0504**
student vs. all	40	0.2417	0.2411	**0.1834**	0.2100	0.2086	**0.1740**
	80	0.1900	0.1899	**0.1454**	0.1681	0.1672	**0.1358**
	120	0.1696	0.1693	**0.1291**	0.1531	0.1523	**0.1204**
	200	0.1539	0.1539	**0.1134**	0.1349	0.1344	**0.1043**
	400	0.1310	0.1308	**0.0935**	0.1147	0.1144	**0.0874**
	600	0.1173	0.1169	**0.0818**	0.1063	0.1059	**0.0802**

Table 2: Experimental results on the WebKB corpus, using SVMs for linear, Gaussian, and multinomial diffusion kernels. The left columns use tf representation and the right columns use tf-idf representation. The error rates shown are averages obtained using 20-fold cross validation. The best performance for each training set size L is shown in boldface. All differences are statistically significant according to the paired t test at the 0.05 level.

		tf Representation			tf-idf Representation		
Task	L	Linear	Gaussian	Diffusion	Linear	Gaussian	Diffusion
course vs. student	40	0.0808	0.0802	**0.0391**	0.0580	0.0572	**0.0363**
	80	0.0505	0.0504	**0.0266**	0.0409	0.0406	**0.0251**
	120	0.0419	0.0409	**0.0231**	0.0361	0.0359	**0.0225**
	200	0.0333	0.0328	**0.0184**	0.0310	0.0308	**0.0201**
	400	0.0263	0.0259	**0.0135**	0.0234	0.0232	**0.0159**
	600	0.0228	0.0221	**0.0117**	0.0207	0.0202	**0.0141**
faculty vs. student	40	0.2106	0.2102	**0.1624**	0.2053	0.2026	**0.1663**
	80	0.1766	0.1764	**0.1357**	0.1729	0.1718	**0.1335**
	120	0.1624	0.1618	**0.1198**	0.1578	0.1573	**0.1187**
	200	0.1405	0.1405	**0.0992**	0.1420	0.1418	**0.1026**
	400	0.1160	0.1158	**0.0759**	0.1166	0.1165	**0.0781**
	600	0.1050	0.1046	**0.0656**	0.1050	0.1048	**0.0692**
project vs. student	40	0.1434	0.1430	**0.0908**	0.1304	0.1279	**0.0863**
	80	0.1139	0.1133	**0.0725**	0.0982	0.0970	**0.0634**
	120	0.0958	0.0957	**0.0613**	0.0870	0.0866	**0.0559**
	200	0.0781	0.0775	**0.0514**	0.0729	0.0722	**0.0472**
	400	0.0590	0.0579	**0.0405**	0.0629	0.0622	**0.0397**
	600	0.0515	0.0500	**0.0325**	0.0551	0.0539	**0.0358**

Table 3: Experimental results on the WebKB corpus, using SVMs for linear, Gaussian, and multinomial diffusion kernels. The left columns use tf representation and the right columns use tf-idf representation. The error rates shown are averages obtained using 20-fold cross validation. The best performance for each training set size L is shown in boldface. All differences are statistically significant according to the paired t test at the 0.05 level.

		tf Representation			tf-idf Representation		
Task	L	Linear	Gaussian	Diffusion	Linear	Gaussian	Diffusion
earn vs. all	80	0.1107	0.1106	**0.0971**	0.0823	0.0827	**0.0762**
	120	0.0988	0.0990	**0.0853**	0.0710	0.0715	**0.0646**
	200	0.0808	0.0810	**0.0660**	0.0535	0.0538	**0.0480**
	400	0.0578	0.0578	**0.0456**	0.0404	0.0408	**0.0358**
	600	0.0465	0.0464	**0.0367**	0.0323	0.0325	**0.0290**
acq vs. all	80	0.1126	0.1125	**0.0846**	0.0788	0.0785	**0.0667**
	120	0.0886	0.0885	**0.0697**	0.0632	0.0632	**0.0534**
	200	0.0678	0.0676	**0.0562**	0.0499	0.0500	**0.0441**
	400	0.0506	0.0503	**0.0419**	0.0370	0.0369	**0.0335**
	600	0.0439	0.0435	**0.0363**	0.0318	0.0316	**0.0301**
moneyFx vs. all	80	0.1201	0.1198	**0.0758**	0.0676	0.0669	**0.0647***
	120	0.0986	0.0979	**0.0639**	0.0557	0.0545	**0.0531***
	200	0.0814	0.0811	**0.0544**	0.0485	0.0472	**0.0438**
	400	0.0578	0.0567	**0.0416**	0.0427	0.0418	**0.0392**
	600	0.0478	0.0467	**0.0375**	0.0391	0.0385	**0.0369***
grain vs. all	80	0.1443	0.1440	**0.0925**	0.0536	**0.0518***	0.0595
	120	0.1101	0.1097	**0.0717**	0.0476	**0.0467***	0.0494
	200	0.0793	0.0786	**0.0576**	0.0430	**0.0420***	0.0440
	400	0.0590	0.0573	**0.0450**	0.0349	**0.0340***	0.0365
	600	0.0517	0.0497	**0.0401**	0.0290	**0.0284***	0.0306
crude vs. all	80	0.1396	0.1396	**0.0865**	0.0502	**0.0485***	0.0524
	120	0.0961	0.0953	**0.0542**	0.0446	**0.0425***	0.0428
	200	0.0624	0.0613	**0.0414**	0.0388	0.0373	**0.0345***
	400	0.0409	0.0403	**0.0325**	0.0345	0.0337	**0.0297**
	600	0.0379	0.0362	**0.0299**	0.0292	0.0284	**0.0264***

Table 4: Experimental results on the Reuters corpus, using SVMs for linear, Gaussian, and multinomial diffusion kernels. The left columns use tf representation and the right columns use tf-idf representation. The error rates shown are averages obtained using 20-fold cross validation. The best performance for each training set size L is shown in boldface. An asterisk (*) indicates that the difference is not statistically significant according to the paired t test at the 0.05 level.

		tf Representation			tf-idf Representation		
Task	L	Linear	Gaussian	Diffusion	Linear	Gaussian	Diffusion
acq vs. earn	40	0.1043	0.1043	**0.1021***	0.0829	0.0831	**0.0814***
	80	0.0902	0.0902	**0.0856***	0.0764	0.0767	**0.0730***
	120	0.0795	0.0796	**0.0715**	0.0626	0.0628	**0.0562**
	200	0.0599	0.0599	**0.0497**	0.0509	0.0511	**0.0431**
	400	0.0417	0.0417	**0.0340**	0.0336	0.0337	**0.0294**
moneyFx vs. earn	40	0.0759	0.0758	**0.0474**	0.0451	0.0451	**0.0372***
	80	0.0442	0.0443	**0.0238**	0.0246	0.0246	**0.0177**
	120	0.0313	0.0311	**0.0160**	0.0179	0.0179	**0.0120**
	200	0.0244	0.0237	**0.0118**	0.0113	0.0113	**0.0080**
	400	0.0144	0.0142	**0.0079**	0.0080	0.0079	**0.0062**
grain vs. earn	40	0.0969	0.0970	**0.0543**	0.0365	0.0366	**0.0336***
	80	0.0593	0.0594	**0.0275**	0.0231	0.0231	**0.0201***
	120	0.0379	0.0377	**0.0158**	0.0147	0.0147	**0.0114***
	200	0.0221	0.0219	**0.0091**	0.0082	0.0081	**0.0069***
	400	0.0107	0.0105	**0.0060**	0.0037	0.0037	**0.0037***
crude vs. earn	40	0.1108	0.1107	**0.0950**	**0.0583***	0.0586	0.0590
	80	0.0759	0.0757	**0.0552**	0.0376	0.0377	**0.0366***
	120	0.0608	0.0607	**0.0415**	0.0276	**0.0276***	0.0284
	200	0.0410	0.0411	**0.0267**	**0.0218***	0.0218	0.0225
	400	0.0261	0.0257	**0.0194**	0.0176	**0.0171***	0.0181

Table 5: Experimental results on the Reuters corpus, using support vector machines for linear, Gaussian, and multinomial diffusion kernels. The left columns use tf representation and the right columns use tf-idf representation. The error rates shown are averages obtained using 20-fold cross validation. The best performance for each training set size L is shown in boldface. An asterisk (*) indicates that the difference is not statistically significant according to the paired t test at the 0.05 level.

Category	Linear	RBF	Diffusion
earn	0.01159	0.01159	**0.01026**
acq	0.01854	0.01854	**0.01788**
money-fx	0.02418	0.02451	**0.02219**
grain	0.01391	0.01391	**0.01060**
crude	0.01755	0.01656	**0.01490**
trade	0.01722	**0.01656**	0.01689
interest	0.01854	0.01854	**0.01689**
ship	0.01324	0.01324	**0.01225**
wheat	0.00894	0.00794	**0.00629**
corn	0.00794	0.00794	**0.00563**

Table 6: Test set error rates for the Reuters top 10 classes using tf features. The train and test sets were created using the Mod-Apt split.

and Gaussian kernels we use L_2 normalization, which works better empirically than L_1 for these kernels. The curves show the test set error rates averaged over 20 iterations of cross validation as a function of the training set size. The error bars represent one standard deviation. For both the Gaussian and diffusion kernels, we test scale parameters ($\sqrt{2}\sigma$ for the Gaussian kernel and $2t^{1/2}$ for the diffusion kernel) in the set $\{0.5, 1, 2, 3, 4, 5, 7, 10\}$. The results reported are for the best parameter value in that range.

We also performed experiments with the popular Mod-Apte train and test split for the top 10 categories of the Reuters collection. For this split, the training set has about 7000 documents and is highly biased towards negative documents. We report in Table 6 the test set accuracies for the tf representation. For the tf-idf representation, the difference between the different kernels is not statistically significant for this amount of training and test data. The provided train set is more than enough to achieve outstanding performance with all kernels used, and the absence of cross validation data makes the results too noisy for interpretation.

In Table 7 we report the F1 measure rather than accuracy, since this measure is commonly used in text classification. The last column of the table compares the presented results with the published results of Zhang and Oles (2001), with a + indicating the diffusion kernel F1 measure is greater than the result published by Zhang and Oles (2001) for this task.

Our results are consistent with previous experiments in text classification using SVMs, which have observed that the linear and Gaussian kernels result in very similar performance (Joachims et al., 2001). However the multinomial diffusion kernel significantly outperforms the linear and Gaussian kernels for the tf representation, achieving significantly lower error rate than the other kernels. For the tf-idf representation, the diffusion kernel consistently outperforms the other kernels for the WebKb data and usually outperforms the linear and Gaussian kernels for the Reuters data. The Reuters data is a much larger collection than WebKB, and the document frequency statistics,

Category	Linear	RBF	Diffusion	±
earn	0.9781	0.9781	**0.9808**	−
acq	0.9626	0.9626	**0.9660**	+
money-fx	0.8254	0.8245	**0.8320**	+
grain	0.8836	0.8844	**0.9048**	−
crude	0.8615	0.8763	**0.8889**	+
trade	0.7706	0.7797	**0.8050**	+
interest	**0.8263**	**0.8263**	0.8221	+
ship	0.8306	0.8404	**0.8827**	+
wheat	0.8613	0.8613	**0.8844**	−
corn	0.8727	0.8727	**0.9310**	+

Table 7: F1 measure for the Reuters top 10 classes using tf features. The train and test sets were created using the Mod-Apte split. The last column compares the presented results with the published results of (Zhang & Oles, 2001), with a + indicating the diffusion kernel F1 measure is greater than the result published in (Zhang & Oles, 2001) for this task.

which are the basis for the inverse document frequency weighting in the tf-idf representation, are evidently much more effective on this collection. It is notable, however, that the multinomial information diffusion kernel achieves at least as high an accuracy without the use of any heuristic term weighting scheme. These results offer evidence that the use of multinomial geometry is both theoretically motivated and practically effective for document classification.

8.5 Experimental Results for Gaussian Embedding

In this section we report experiments on synthetic data demonstrating the applicability of the heat kernel on the hyperbolic space $\mathbb{H}^n$ that corresponds to the manifold of spherical normal distributions.

Embedding data points in the hyperbolic space is more complicated than the multinomial case. Recall that in the multinomial case, we embedded points by computing the maximum likelihood estimator the the data. A similar method would fail for $\mathbb{H}^n$ embedding since all the data points will be mapped to normal distributions with variance 0, which will result in a degenerate geometry that is equivalent to the Euclidean one (see Section 4.3).

A more realistic embedding can be achieved by sampling from the posterior of a Dirichlet Process Mixture Model (DPMM) (Ferguson, 1973; Blackwell & MacQueen, 1973; Antoniak, 1974). A Dirichlet Process Mixture model, based on the spherical Normal distribution, associates with the data $x_1, \ldots, x_m \in \mathbb{R}^n$ a posterior $p(\theta_1, \ldots, \theta_m | x_1, \ldots, x_m)$ where $\theta_i \in \mathbb{H}^{n+1}$. Instead of going into the description and properties of Dirichlet Process Mixture Model we refer the interested reader to the references above, and to (Escobar & West, 1995; Neal, 2000) for relevant Monte Carlo approximations.

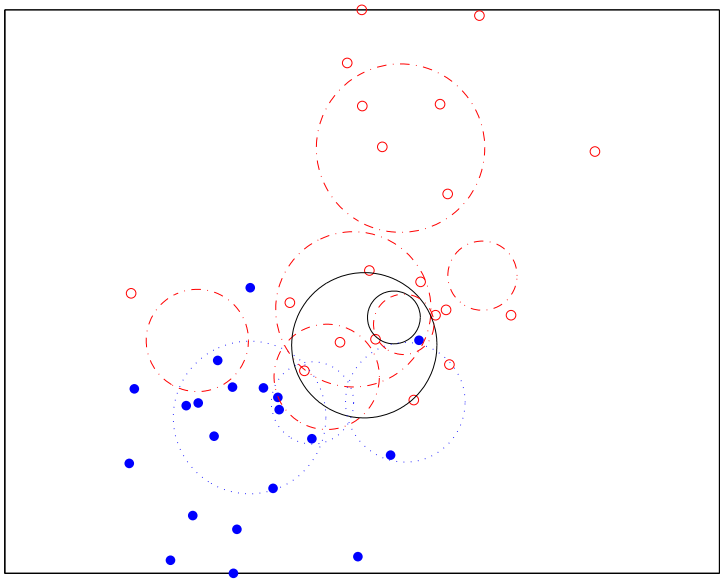

Figure 22: A sample from the posterior of a DPMM based on data from two Gaussians $N((-1,-1)^\top, I)$ (solid blue dots) and $N((1,1)^\top, I)$ (hollow red dots). The sample is illustrated by the circles that represent one standard deviation centered around the mean. Blue dotted circles represent a parameter associated with a point from $N((-1,-1)^\top, I)$, red dashed circles represent a parameter associated with a point from $N((1,1)^\top, I)$ and solid black circles represent parameters associated with points from both Gaussians.

Obtaining T samples from the DPMM posterior $\{\theta_i^{(t)}\}_{i=1,t=1}^{m,T}$ we can measure the similarity between points x_i, x_j as

$$\tilde{K}_t(x_i, x_j) = \frac{1}{T} \sum_{t=1}^{T} K_t \left(\theta_i^{(t)}, \theta_j^{(t)}\right) \tag{120}$$

where K_t is the heat kernel on the hyperbolic space given by equations (83)–(84). The above definition of $\tilde{K}$ has the interpretation of being approximately the mean of the heat kernel – which is now a random variable under the DPMM posterior $p(\theta|x)$. The details of Gibbs sampling from the posterior of a spherical Normal based DPMM is given in Appendix B.

Some intuition may be provided by Figure 22[12]. A sample from the posterior of a DPMM based on data from two Gaussians $N((-1,-1)^\top, I)$ (solid blue dots) and $N((1,1)^\top, I)$ (hollow red dots). The sample is illustrated by the circles that represent one standard deviation centered around the mean. Blue dotted circles represent a parameter associated with a point from $N((-1,-1)^\top, I)$, red dashed circles represent a parameter associated with a point from $N((1,1)^\top, I)$ and solid black circles represent parameters associated with points from both Gaussians.

[12]This Figure is better displayed in color.

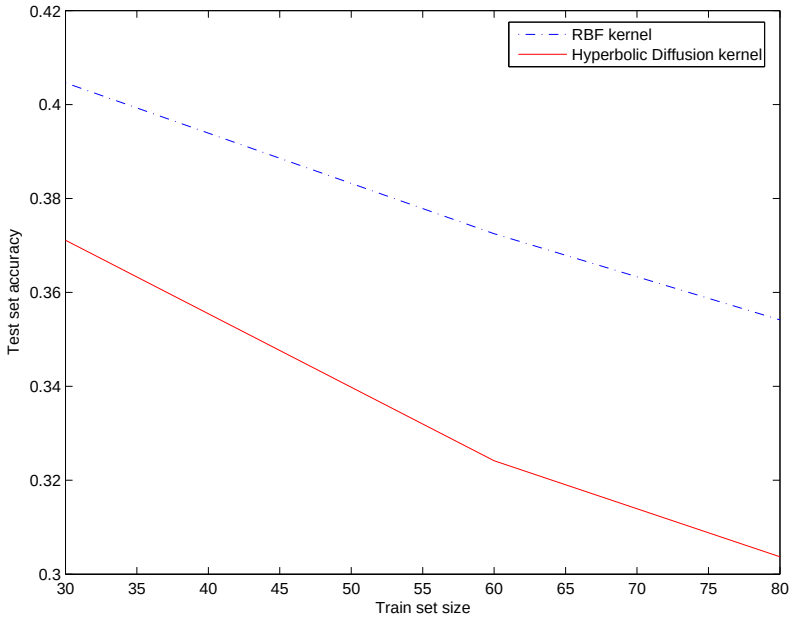

Figure 23: Test set error rate for SVM based on standard RBF kernel and the mean hyperbolic heat kernel $\tilde{K}$ as a function of the train set size, after 20-fold cross validation. The data was generated from two significantly overlapping Gaussians $N((-2,-2)^\top, 5I), N((2,2)^\top, 5I)$.

After generating data from two significantly overlapping Gaussians $N((-2,-2)^\top, 5I), N((2,2)^\top, 5I)$, and sampling 20 samples from the posterior we compared the performance of a standard RBF kernel and $\tilde{K}$. The test set accuracy for SVM, after 20-fold cross validation, as a function of the train set size is displayed in Figure 23. The mean hyperbolic heat kernel outperforms the RBF kernel consistently.

8.6 Discussion

In this section we introduced a family of kernels that is intimately based on the geometry of the Riemannian manifold associated with a statistical family through the Fisher information metric. The metric is canonical in the sense that it is uniquely determined by requirements of invariance (Čencov, 1982), and moreover, the choice of the heat kernel is natural because it effectively encodes a great deal of geometric information about the manifold. While the geometric perspective in statistics has most often led to reformulations of results that can be viewed more traditionally, the kernel methods developed here clearly depend crucially on the geometry of statistical families.

The main application of these ideas has been to develop the multinomial diffusion kernel. Our experimental results indicate that the resulting diffusion kernel is indeed effective for text classification using support vector machine classifiers, and can lead to significant improvements in accuracy compared with the use of linear or Gaussian kernels, which have been the standard for this application. The results of Section 8.4 are notable since accuracies better or comparable to those obtained using heuristic weighting schemes such as tf-idf are achieved directly through the geomet-

ric approach. In part, this can be attributed to the role of the Fisher information metric; because of the square root in the embedding into the sphere, terms that are infrequent in a document are effectively up-weighted, and such terms are typically rare in the document collection overall. The primary degree of freedom in the use of information diffusion kernels lies in the specification of the mapping of data to model parameters. For the multinomial, we have used the maximum likelihood mapping. The use of other model families and mappings remains an interesting direction to explore.

While kernel methods generally are "model free," and do not make distributional assumptions about the data that the learning algorithm is applied to, statistical models offer many advantages, and thus it is attractive to explore methods that combine data models and purely discriminative methods. Our approach combines parametric statistical modeling with non-parametric discriminative learning, guided by geometric considerations. In these aspects it is related to the methods proposed by Jaakkola and Haussler (1998). However, the kernels proposed in the current section differ significantly from the Fisher kernel of Jaakkola and Haussler (1998). In particular, the latter is based on the score $\operatorname{grad}_\theta \log p(X \mid \hat{\theta})$ at a single point $\hat{\theta}$ in parameter space. In the case of an exponential family model it is given by a covariance $K_F(x, x') = \sum_i \left(x_i - E_{\hat{\theta}}[X_i]\right)\left(x'_i - E_{\hat{\theta}}[X_i]\right)$; this covariance is then heuristically exponentiated. In contrast, information diffusion kernels are based on the full geometry of the statistical family, and yet are also invariant under re-parameterizations of the family. In other conceptually related work, Belkin and Niyogi (2003) suggest measuring distances on the data graph to approximate the underlying manifold structure of the data. In this case the underlying geometry is inherited from the embedding Euclidean space rather than the Fisher geometry.

While information diffusion kernels are very general, they will be difficult to compute in many cases – explicit formulas such as equations (83–84) for hyperbolic space are rare. To approximate an information diffusion kernel it may be attractive to use the parametrices and geodesic distance between points, as we have done for the multinomial. In cases where the distance itself is difficult to compute exactly, a compromise may be to approximate the geodesic distance between nearby points in terms of the Kullback-Leibler divergence. In effect, this approximation is already incorporated into the kernels recently proposed by Moreno et al. (2004) for multimedia applications, which have the form $K(\theta, \theta') \propto \exp(-\alpha D(\theta, \theta')) \approx \exp(-2\alpha d^2(\theta, \theta'))$, and so can be viewed in terms of the leading order approximation to the heat kernel. The results of Moreno et al. (2004) are suggestive that diffusion kernels may be attractive not only for multinomial geometry, but also for much more complex statistical families.

9 Hyperplane Margin Classifiers

Linear classifiers are a mainstay of machine learning algorithms, forming the basis for techniques such as the perceptron, logistic regression, boosting, and support vector machines. A linear classifier, parameterized by a vector $w \in \mathbb{R}^n$, classifies examples according to the decision rule $\hat{y}(x) = \operatorname{sign}\left(\sum_i w_i \phi_i(x)\right) = \operatorname{sign}(\langle w, x \rangle) \in \{-1, +1\}$, following the common practice of identifying x with the feature vector $\phi(x)$. The differences between different linear classifiers lie in the criteria and algorithms used for selecting the parameter vector w based on a training set.

Geometrically, the decision surface of a linear classifier is formed by a hyperplane or linear subspace in n-dimensional Euclidean space, $\{x \in \mathbb{R}^n : \langle x, w \rangle = 0\}$ where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. (In both the algebraic and geometric formulations, a bias term is sometimes added; we prefer to absorb the bias into the notation given by the inner product, by setting $x_n = 1$ for all x.) The linearity assumption made by such classifiers can be justified on purely computational grounds; linear classifiers are generally easy to train, and the linear form is simple to analyze and compute.

Modern learning theory emphasizes the tension between fitting the training data well and the more desirable goal of achieving good generalization. A common practice is to choose a model that fits the data closely, but from a restricted class of models. The model class needs to be sufficiently rich to allow the choice of a good hypothesis, yet not so expressive that the selected model is likely to overfit the data. Hyperplane classifiers are attractive for balancing these two goals. Indeed, linear hyperplanes are a rather restricted set of models, but they enjoy many unique properties. For example, given two points $x, y \in \mathbb{R}^n$, the set of points equidistant from x and y is a hyperplane; this lies behind the intuition that a hyperplane is the correct geometric shape for separating sets of points. Similarly, a hyperplane is the best decision boundary to separate two Gaussian distributions of equal covariance. Another distinguishing property is that a hyperplane in $\mathbb{R}^n$ is isometric to $\mathbb{R}^{n-1}$, and can therefore be thought of as a reduced dimension version of the original feature space. Finally, a linear hyperplane is the union of straight lines, which are distance minimizing curves, or geodesics, in Euclidean geometry.

However, a fundamental assumption is implicitly associated with linear classifiers, since they are based crucially on the use of the Euclidean geometry of $\mathbb{R}^n$. If the data or features at hand lack a Euclidean structure, the arguments above for linear classifiers break down; arguably, there is lack of Euclidean geometry for the feature vectors in most applications. This section studies analogues of linear hyperplanes as a means of obtaining simple, yet effective classifiers when the data can be represented in terms of a natural geometric structure that is only locally Euclidean. This is the case for categorical data that is represented in terms of multinomial models, for which the associated geometry is spherical.

Because of the complexity of the notion of linearity in general Riemannian spaces, we focus our attention on the multinomial manifold, which permits a relatively simple analysis. Hyperplanes in multinomial manifold is discussed in Section 9.2. The construction and training of margin based models is discussed in Section 9.3, with an emphasis on spherical logistic regression. A brief examination of linear hyperplanes in general Riemannian manifolds appears in Section 9.4 followed by experimental results for text classification given in Section 9.5. Concluding remarks are made in Section **??**.

9.1 Hyperplanes and Margins on $\mathbb{S}^n$

This section generalizes the notion of linear hyperplanes and margins to the n-sphere $\mathbb{S}^n = \{x \in \mathbb{R}^{n+1} : \sum_i x_i^2 = 1\}$. A similar treatment on the positive n-sphere $\mathbb{S}^n_+$ is more complicated, and is

postponed to the next section. In the remainder of this section we denote points on $\mathbb{P}_n, \mathbb{S}^n$ or $\mathbb{S}^n_+$ as vectors in $\mathbb{R}^{n+1}$ using the standard basis of the embedding space. The notation $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ will be used for the Euclidean inner product and norm.

A hyperplane on $\mathbb{S}^n$ is defined as $H_u = \mathbb{S}^n \cap E_u$ where E_u is an n-dimensional linear subspace of $\mathbb{R}^{n+1}$ associated with the normal vector u. We occasionally need to refer to the unit normal vector (according to the Euclidean norm) and denote it by $\hat{u}$. H_u is an $n-1$ dimensional submanifold of $\mathbb{S}^n$ which is isometric to $\mathbb{S}^{n-1}$ (Bridson & Haefliger, 1999). Using the common notion of the distance of a point from a set $d(x, S) = \inf_{y \in S} d(x, y)$ we make the following definitions.

Definition 8. *Let X be a metric space. A decision boundary is a subset of X that separates X into two connected components. The margin of x with respect to a decision boundary H is $d(x, H) = \inf_{y \in H} d(x, y)$.*

Note that this definition reduces to the common definition of margin for Euclidean geometry and affine hyperplanes.

In contrast to Gous (1998), our submanifolds are intersections of the sphere with linear subspaces, not affine sets. One motivation for the above definition of hyperplane as the correct generalization of a Euclidean hyperplane is that H_u is the set of points equidistant from $x, y \in \mathbb{S}^n$ in the spherical metric. Further motivation is given in Section 9.4.

Before we can obtain a closed form expression for margins on $\mathbb{S}^n$ we need the following definitions.

Definition 9. *Given a point $x \in \mathbb{R}^{n+1}$, we define its reflection with respect to E_u as*

$$r_u(x) = x - 2\langle x, \hat{u} \rangle \hat{u}.$$

Note that if $x \in \mathbb{S}^n$ then $r_u(x) \in \mathbb{S}^n$ as well, since $\|r_u(x)\|^2 = \|x\|^2 - 4\langle x, \hat{u} \rangle^2 + 4\langle x, \hat{u} \rangle^2 = 1$.

Definition 10. *The projection of $x \in \mathbb{S}^n \setminus \{\hat{u}\}$ on H_u is defined to be*

$$p_u(x) = \frac{x - \langle x, \hat{u} \rangle \hat{u}}{\sqrt{1 - \langle x, \hat{u} \rangle^2}}.$$

Note that $p_u(x) \in H_u$, since $\|p_u(x)\| = 1$ and $\langle x - \langle x, \hat{u} \rangle \hat{u}, \hat{u} \rangle = \langle x, \hat{u} \rangle - \langle x, \hat{u} \rangle \|\hat{u}\|^2 = 0$. The term projection is justified by the following proposition.

Proposition 5. *Let $x \in \mathbb{S}^n \setminus (H_u \cup \{\hat{u}\})$. Then*

$$\begin{aligned}
&(a) \quad d(x, q) = d(r_u(x), q) \quad \forall q \in H_u \\
&(b) \quad d(x, p_u(x)) = \arccos\left(\sqrt{1 - \langle x, \hat{u} \rangle^2}\right) \\
&(c) \quad d(x, H_u) = d(x, p_u(x)).
\end{aligned}$$

Proof. Since $q \in H_u$,

$$\begin{aligned}\cos d(r_u(x), q) &= \langle x - 2\langle x, \hat{u}\rangle \hat{u}, q\rangle \\ &= \langle x, q\rangle - 2\langle x, \hat{u}\rangle\langle \hat{u}, q\rangle \\ &= \langle x, q\rangle = \cos d(x, q)\end{aligned}$$

and (a) follows. Assertion (b) follows from

$$\cos d(x, p_u(x)) = \left\langle x, \frac{x - \langle x, \hat{u}\rangle \hat{u}}{\sqrt{1 - \langle x, \hat{u}\rangle^2}} \right\rangle = \frac{1 - \langle x, \hat{u}\rangle^2}{\sqrt{1 - \langle x, \hat{u}\rangle^2}}.$$

Finally, to prove (c) note that by the identity $\cos 2\theta = 2\cos^2\theta - 1$,

$$\begin{aligned}\cos(2d(x, p_u(x))) &= 2\cos^2(d(x, p_u(x))) - 1 \\ &= 1 - 2\langle x, \hat{u}\rangle^2 = \cos(d(x, r_u(x)))\end{aligned}$$

and hence $d(x, p_u(x)) = \frac{1}{2} d(x, r_u(x))$. The distance $d(x, q), q \in H_u$ cannot be any smaller than $d(x, p_u(x))$ since this would result in a path from x to $r_u(x)$ of length shorter than the geodesic $d(x, r_u(x))$. □

Parts (b) and (c) of Proposition 5 provide a closed form expression for the $\mathbb{S}^n$ margin analogous to the Euclidean unsigned margin $|\langle x, \hat{u}\rangle|$. Similarly, the $\mathbb{S}^n$ analogue of the Euclidean signed margin $y\langle \hat{u}, x\rangle$ is

$$y \frac{\langle x, \hat{u}\rangle}{|\langle x, \hat{u}\rangle|} \arccos\left(\sqrt{1 - \langle x, \hat{u}\rangle^2}\right).$$

A plot of the signed margin as a function of $\langle x, \hat{u}\rangle$ and a geometric interpretation of the spherical margin appear in Figure 24.

9.2 Hyperplanes and Margins on $\mathbb{S}^n_+$

A hyperplane on the positive n-sphere $\mathbb{S}^n_+$ is defined as $H_{u+} = E_u \cap \mathbb{S}^n_+$, assuming it is non-empty. This definition leads to a margin concept $d(x, H_{u+})$ different from the $\mathbb{S}^n$ margin $d(x, H_u)$ since

$$\begin{aligned}d(x, H_{u+}) &= \inf_{y \in E_u \cap \mathbb{S}^n_+} d(x, y) \\ &\geq \inf_{y \in E_u \cap \mathbb{S}^n} d(x, y) = d(x, H_u).\end{aligned}$$

The above infimum is attained by the continuity of d and compactness of $E_u \cap \mathbb{S}^n_+$ justifying the notation $q = \arg\min_{y \in E_u \cap \mathbb{S}^n_+} d(x, y)$ as a point realizing the margin distance $d(x, H_{u+})$.

The following theorem will be useful in computing $d(x, H_{u+})$. For a proof see Bridson and Haefliger (1999) page 17.

Theorem 9. (**The Spherical Law of Cosines**)
Consider a spherical triangle with geodesic edges of lengths a, b, c, where γ is the vertex angle opposite to edge c. Then

$$\cos c = \cos a \cos b + \sin a \sin b \cos \gamma.$$

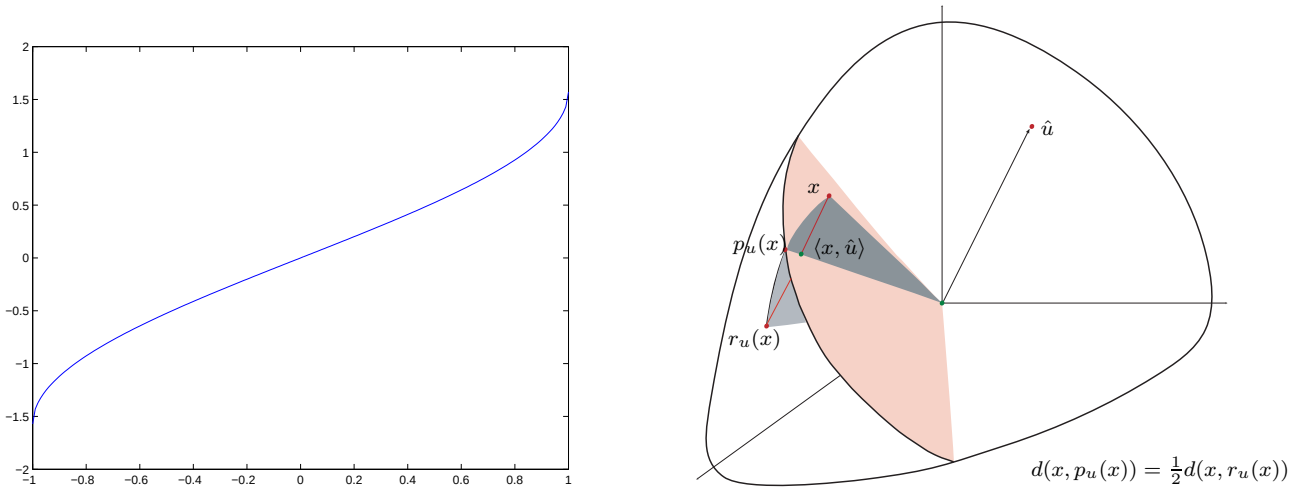

Figure 24: The signed margin $\text{sign}(\langle x, \hat{y} \rangle) d(x, H_u)$ as a function of $\langle x, \hat{u} \rangle$, which lies in the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$ (left) and a geometric interpretation of the spherical margin (right).

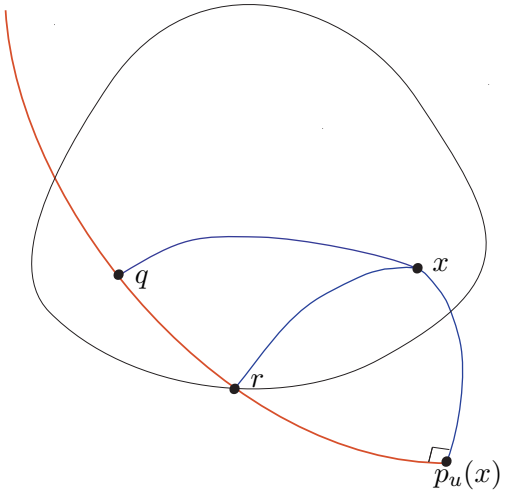

Figure 25: The spherical law of cosines implies $d(r, x) \leq d(q, x)$.

We have the following corollaries of Proposition 5.

Proposition 6. *If $x \in \mathbb{S}^n_+$ and $p_u(x) \in \mathbb{S}^n_+$ then*

$$p_u(x) = \operatorname*{arg\,min}_{y \in \mathbb{S}^n_+ \cap E_u} d(x, y)$$
$$d(x, H_u) = d(x, H_{u+})$$

Proof. This follows immediately from the fact that $p_u(x) = \arg\min_{y \in \mathbb{S}^n \cap E_u} d(x, y)$ and from $\mathbb{S}^n_+ \cap E_u \subset \mathbb{S}^n \cap E_u$. □

Proposition 7. *For $x \in \mathbb{S}^n_+$ and $p_u(x) \notin \mathbb{S}^n_+$ we have*

$$q = \operatorname*{arg\,min}_{y \in \mathbb{S}^n_+ \cap E_u} d(x, y) \in \partial \mathbb{S}^n_+$$

where $\partial \mathbb{S}^n_+$ is the boundary of $\mathbb{S}^n_+$.

Proof. Assume that $q \notin \partial \mathbb{S}^n_+$ and connect q and $p_u(x)$ by a minimal geodesic α. Since $p_u(x) \notin \mathbb{S}^n_+$, the geodesic α intersects the boundary $\partial \mathbb{S}^n_+$ at a point r. Since $q, p_u(x) \in H_u$ and H_u is geodesically convex, $\alpha \subset H_u$. Now, since $p_u(x) = \arg\min_{y \in \alpha} d(y, x)$, the geodesic from x to $p_u(x)$ and α intersect orthogonally (this is an elementary result in Riemannian geometry, e.g. Lee (1997) p. 113). Using the spherical law of cosines, applied to the spherical triangles $(q, x, p_u(x))$ and $(r, x, p_u(x))$ (see Figure 25), we deduce that

$$\begin{aligned} \cos d(x, q) &= \cos d(q, p_u(x)) \cos d(x, p_u(x)) \\ &\leq \cos d(r, p_u(x)) \cos d(x, p_u(x)) \\ &= \cos d(x, r) \end{aligned}$$

Hence r is closer to x than q. This contradicts the definition of q; thus q can not lie in the interior of $\mathbb{S}^n_+$. □

Before we proceed to compute $d(x, H_{u+})$ for $p_u(x) \notin \mathbb{S}^n_+$ we define the following concepts.

Definition 11. *The boundary of $\mathbb{S}^n$ and $\mathbb{S}^n_+$ with respect to $A \subset \{1, \ldots, n+1\}$ is*

$$\partial_A \mathbb{S}^n = \mathbb{S}^n \cap \{x \in \mathbb{R}^{n+1} : \forall i \in A,\, x_i = 0\} \cong \mathbb{S}^{n-|A|}$$
$$\partial_A \mathbb{S}^n_+ = \mathbb{S}^n_+ \cap \{x \in \mathbb{R}^{n+1} : \forall i \in A,\, x_i = 0\} \cong \mathbb{S}^{n-|A|}_+$$

Note that if $A \subset A'$ then $\partial_{A'} \mathbb{S}^n_+ \subset \partial_A \mathbb{S}^n_+$. We use the notation $\langle \cdot, \cdot \rangle_A$ and $\|\cdot\|_A$ to refer to the Euclidean inner product and norm, where the summation is restricted to indices *not* in A.

Definition 12. *Given $x \in \mathbb{S}^n$ we define $x|_A \in \partial_A \mathbb{S}^n$ as*

$$(x|_A)_i = \begin{cases} 0 & i \in A \\ x_i / \|x\|_A & i \notin A\,. \end{cases}$$

We abuse the notation by identifying $x|_A$ also with the corresponding point on $\mathbb{S}^{n-|A|}$ under the isometry $\partial_A \mathbb{S}^n \cong \mathbb{S}^{n-|A|}$ mentioned in Definition 11. Note that if $x \in \mathbb{S}^n_+$ then $x|_A \in \partial_A \mathbb{S}^n_+$. The following proposition computes the $\mathbb{S}^n_+$ margin $d(x, H_{u+})$ given the boundary set of $q = \arg\min_{y \in \mathbb{S}^n_+ \cap E_u} d(y, x)$.

Proposition 8. *Let $\hat{u} \in \mathbb{R}^{n+1}$ be a unit vector, $x \in \mathbb{S}^n_+$ and $q = \arg\min_{y \in \mathbb{S}^n_+ \cap E_u} d(y, x) \in \partial_A \mathbb{S}^n_+$ where A is the (possibly empty) set $A = \{1 \le i \le n+1 : q_i = 0\}$. Then*

$$d(x, H_{u+}) = \arccos\left(\|x\|_A \sqrt{1 - \langle x|_A, \hat{u}|_A \rangle^2} \right)$$

Proof. If $p_u(x) \in \mathbb{S}^n_+$ then the proposition follows from earlier propositions and the fact that when $A = \emptyset$, $\|x\|_A = \|x\| = 1$ and $v|_A = v$. We thus restrict our attention to the case of $A \neq \emptyset$.

For all $I \subset \{1, \ldots, n+1\}$ we have

$$\begin{aligned}
\arg\min_{y \in \partial_I \mathbb{S}^n_+ \cap E_u} d(x, y) &= \arg\max_{y \in \partial_I \mathbb{S}^n_+ \cap E_u} \langle x, y \rangle \\
&= \arg\max_{y \in \partial_I \mathbb{S}^n_+ \cap E_u} \langle x, y \rangle_I \\
&= \arg\max_{y \in \mathbb{S}^{n-|I|}_+ \cap E_{u|I}} \|x\|_I \langle x|_I, y \rangle \\
&= \arg\min_{y \in \mathbb{S}^{n-|I|}_+ \cap E_{u|I}} d(x|_I, y) \,.
\end{aligned}$$

It follows that

$$q|_A = \arg\min_{y \in \mathbb{S}^{n-|A|}_+ \cap E_{u|A}} d(x|_A, y). \tag{121}$$

By Proposition 7 applied to $\mathbb{S}^{n-|A|}$ we have that since $q|_A$ lies in the interior of $\mathbb{S}^{n-|A|}$ then so does

$$p_{u|A}(x|_A) = \frac{x|_A - \langle x|_A, \hat{u}|_A \rangle \hat{u}|_A}{\sqrt{1 - \langle x|_A, \hat{u}|_A \rangle^2}}, \quad x|_A, \hat{u}|_A \in \mathbb{S}^{n-|A|}_+.$$

Using Proposition 5 applied to $\mathbb{S}^{n-|A|}$ we can compute $d(x, H_{u+})$ as

$$\begin{aligned}
d(x, p_{u|A}(x|_A)) &= \arccos\left\langle x, \frac{x|_A - \langle x|_A, \hat{u}|_A \rangle \hat{u}|_A}{\sqrt{1 - \langle x|_A, \hat{u}|_A \rangle^2}} \right\rangle \\
&= \arccos \frac{\|x\|_A - \langle x|_A, \hat{u}|_A \rangle \langle x, \hat{u}|_A \rangle}{\sqrt{1 - \langle x|_A, \hat{u}|_A \rangle^2}} \\
&= \arccos\left(\|x\|_A \sqrt{1 - \langle x|_A, \hat{u}|_A \rangle^2} \right).
\end{aligned}$$

□

In practice the boundary set A of q is not known. In our experiments we set $A = \{i : (p_u(x))_i \le 0\}$; in numerical simulations in low dimensions, the true boundary never lies outside of this set.

9.3 Logistic Regression on the Multinomial Manifold

The logistic regression model $p(y \mid x) = \frac{1}{Z} \exp(y\langle x, w\rangle)$, with $y \in \{-1, 1\}$, assumes Euclidean geometry. It can be re-expressed as

$$\begin{aligned} p(y \mid x\,; u) &\propto \exp\left(y \left\|u\right\| \langle x, \hat{u}\rangle\right) \\ &= \exp\left(y \operatorname{sign}(\langle x, \hat{u}\rangle)\, \theta d(x, H_u)\right) \end{aligned}$$

where d is the Euclidean distance of x from the hyperplane that corresponds to the normal vector $\hat{u}$, and where $\theta = \|u\|$ is a parameter.

The generalization to spherical geometry involves simply changing the margin to reflect the appropriate geometry:

$$\begin{aligned} &p(y|x\,; \hat{u}, \theta) \propto \\ &\exp\left(y \operatorname{sign}(\langle x, \hat{u}\rangle)\ \theta \arccos\left(\left\|x\right\|_A \sqrt{1 - \langle x|_A, \hat{u}|_A\rangle^2}\right)\right). \end{aligned}$$

Denoting $s_x = y \operatorname{sign}(\langle x, \hat{u}\rangle)$, the log-likelihood of the example (x, y) is

$$\begin{aligned} &\ell(\hat{u}, \theta\,; (x, y)) \\ &= -\log\left(1 + e^{-2s_x\theta \arccos\left(\left\|x\right\|_A \sqrt{1 - \langle x|_A, \hat{u}|_A\rangle^2}\right)}\right). \end{aligned}$$

We compute the derivatives of the log-likelihood in several steps, using the chain rule and the notation $z = \langle x|_A, \hat{u}|_A\rangle$. We have

$$\frac{\partial \arccos\left(\left\|x\right\|_A \sqrt{1 - z^2}\right)}{\partial z} = \frac{z \left\|x\right\|_A}{\sqrt{1 - \left\|x\right\|_A^2 (1 - z^2)}\sqrt{1 - z^2}}$$

and hence

$$\frac{\partial \ell(\hat{u}, \theta\,; (x, y))}{\partial z} = \frac{2s_x\theta z \left\|x\right\|_A / (1 + e^{2s_x\theta \arccos\left(\left\|x\right\|_A \sqrt{1 - z^2}\right)})}{\sqrt{1 - \left\|x\right\|_A^2 (1 - z^2)}\sqrt{1 - z^2}}. \tag{122}$$

The log-likelihood derivative with respect to $\hat{u}_i$ is equation (122) times

$$\frac{\partial \langle x|_A, \hat{u}|_A\rangle}{\partial \hat{u}_i} = \begin{cases} 0 & i \in A \\ \frac{(x|_A)_i}{\left\|\hat{u}\right\|_A} - \hat{u}_i \frac{\langle x|_A, \hat{u}|_A\rangle}{\left\|u\right\|_A^2} & i \notin A\,. \end{cases}$$

The log-likelihood derivative with respect to θ is

$$\frac{\partial \ell(\hat{u}, \theta\,; (x, y))}{\partial \theta} = \frac{2s_x \arccos(\left\|x\right\|_A \sqrt{1 - z^2})}{1 + e^{2s_x\theta \arccos(\left\|x\right\|_A \sqrt{1 - z^2})}}.$$

Optimizing the log-likelihood with respect to $\hat{u}$ requires care. Following the gradient $\hat{u}^{(t+1)} = \hat{u}^{(t)} + \alpha \text{grad}\, \ell(\hat{u}^{(t)})$ results in a non-normalized vector. Performing the above gradient descent step followed by normalization has the effect of moving along the sphere in a curve whose tangent vector at $\hat{u}^{(t)}$ is the projection of the gradient onto the tangent space $T_{\hat{u}^{(t)}}\mathbb{S}^n$. This is the technique used in the experiments described in Section 9.5.

Note that the spherical logistic regression model has $n+1$ parameters in contrast to the $n+2$ parameters of Euclidean logistic regression. This is in accordance with the intuition that a hyperplane separating an n-dimensional manifold should have n parameters. The extra parameter in the Euclidean logistic regression is an artifact of the embedding of the n-dimensional multinomial space, on which the data lies, into an $(n+1)$-dimensional Euclidean space.

The derivations and formulations above assume spherical data. If the data lies on the multinomial manifold, the isometry π mentioned earlier has to precede these calculations. The net effect is that x_i is replaced by $\sqrt{x_i}$ in the model equation, and in the log-likelihood and its derivatives.

Synthetic data experiments contrasting Euclidean logistic regression and spherical logistic regression on $\mathbb{S}^n_+$, as described in this section, are shown in Figure 26. The leftmost column shows an example where both models give a similar solution. In general, however, as is the case in the other two columns, the two models yield significantly different decision boundaries.

9.4 Hyperplanes in Riemannian Manifolds

The definition of hyperplanes in general Riemannian manifolds has two essential components. In addition to discriminating between two classes, hyperplanes should be regular in some sense with respect to the geometry. In Euclidean geometry, the two properties of discrimination and regularity coincide, as every affine subspace of dimension $n-1$ separates $\mathbb{R}^n$ into two regions. In general, however, these two properties do not necessarily coincide, and have to be considered separately.

The separation property implies that if N is a hyperplane of M then $M \setminus N$ has two connected components. Note that this property is topological and independent of the metric. The linearity property is generalized through the notion of auto-parallelism explained below. The following definitions and propositions are taken from Spivak (1975), Volume 3. All the connections described below ∇ are the metric connections inherited from the metric g.

Definition 13. *Let $(\mathcal{M}, g)$ be a Riemannian manifold and ∇ the metric connection. A submanifold $N \subset M$ is auto-parallel if parallel translation in M along a curve $C \subset N$ takes vectors tangent to N to vectors tangent to N.*

Proposition 9. *A submanifold $N \subset M$ is auto-parallel if and only if*

$$X, Y \in T_pN \Rightarrow \nabla_X Y \in T_pN.$$

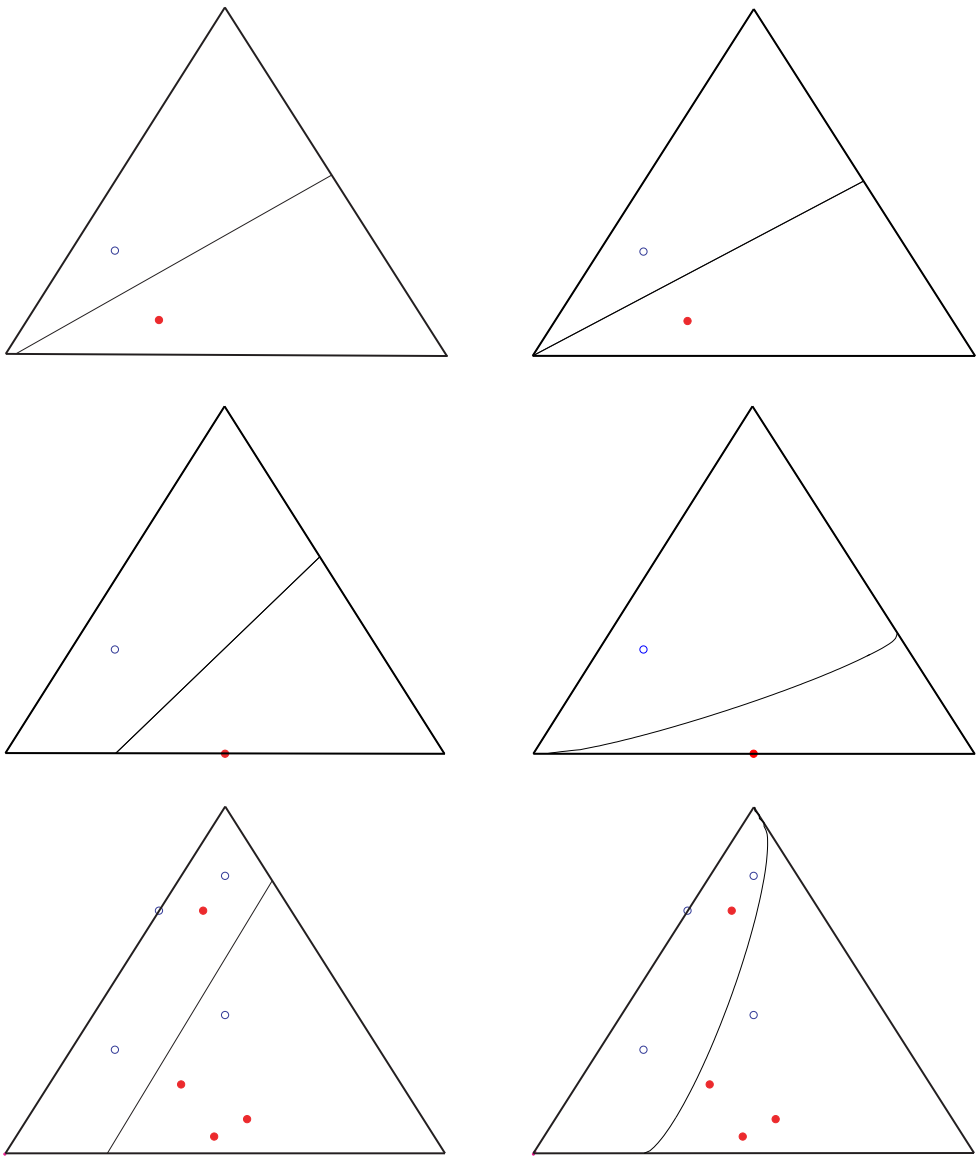

Figure 26: Experiments contrasting Euclidean logistic regression (left column) with multinomial logistic regression (right column) for several toy data sets in $\mathbb{P}^2$.

Definition 14. *A submanifold N of M is totally geodesic at $p \in N$ if every geodesic γ in M with $\gamma(0) = p, \gamma'(0) \in T_pN$ remains in N on some interval $(-\epsilon, \epsilon)$. The submanifold N is said to be totally geodesic if it is totally geodesic at every point.*

As a consequence, we have that N is totally geodesic if and only if every geodesic in N is also a geodesic in M.

Proposition 10. *Let N be a submanifold of (M, ∇). Then*

1. *If N is auto-parallel in M then N is totally geodesic.*
2. *If M is totally geodesic and ∇ is symmetric then M is auto-parallel.*

Since the metric connection is symmetric, the last proposition gives a complete equivalence between auto-parallelism and totally geodesic submanifolds.

We can now define linear hyperplanes on Riemannian manifolds.

Definition 15. *A linear decision boundary N in M is an auto-parallel submanifold of M such that $M \setminus N$ has two connected components.*

Several observations are in order. First note that if M is an n-dimensional manifold, the separability condition requires N to be an $(n-1)$-dimensional submanifold. It is easy to see that every affine subspace of $\mathbb{R}^n$ is totally geodesic and hence auto-parallel. Conversely, since the metric connection is symmetric, every auto-parallel submanifold of Euclidean space that separates it is an affine subspace. As a result, we have that our generalization does indeed reduce to affine subspaces under Euclidean geometry. Similarly, the above definition reduces to spherical hyperplanes $H_u \cap \mathbb{S}^n$ and $H_u \cap \mathbb{S}^n_+$. Another example is the hyperbolic half plane $\mathbb{H}^2$ where the linear decision boundaries are half-circles whose centers lie on the x axis.

Hyperplanes on $\mathbb{S}^n$ have the following additional nice properties. They are the set of equidistant points from $x, y \in \mathbb{S}^n$ (for some x, y), they are isometric to $\mathbb{S}^{n-1}$ and they are parameterized by n parameters. These properties are particular to the sphere and do not hold in general (Bridson & Haefliger, 1999).

9.5 Experiments

A natural embedding of text documents in the multinomial simplex is the L_1 normalized term-frequency (tf) representation (Joachims, 2000)

$$\hat{\theta}(x) = \left(\frac{x_1}{\sum_i x_i}, \ldots, \frac{x_{n+1}}{\sum_i x_i} \right).$$

Using this embedding we compared the performance of spherical logistic regression with Euclidean logistic regression. Since Euclidean logistic regression often performs better with L_2 normalized tf representation, we included these results as well.

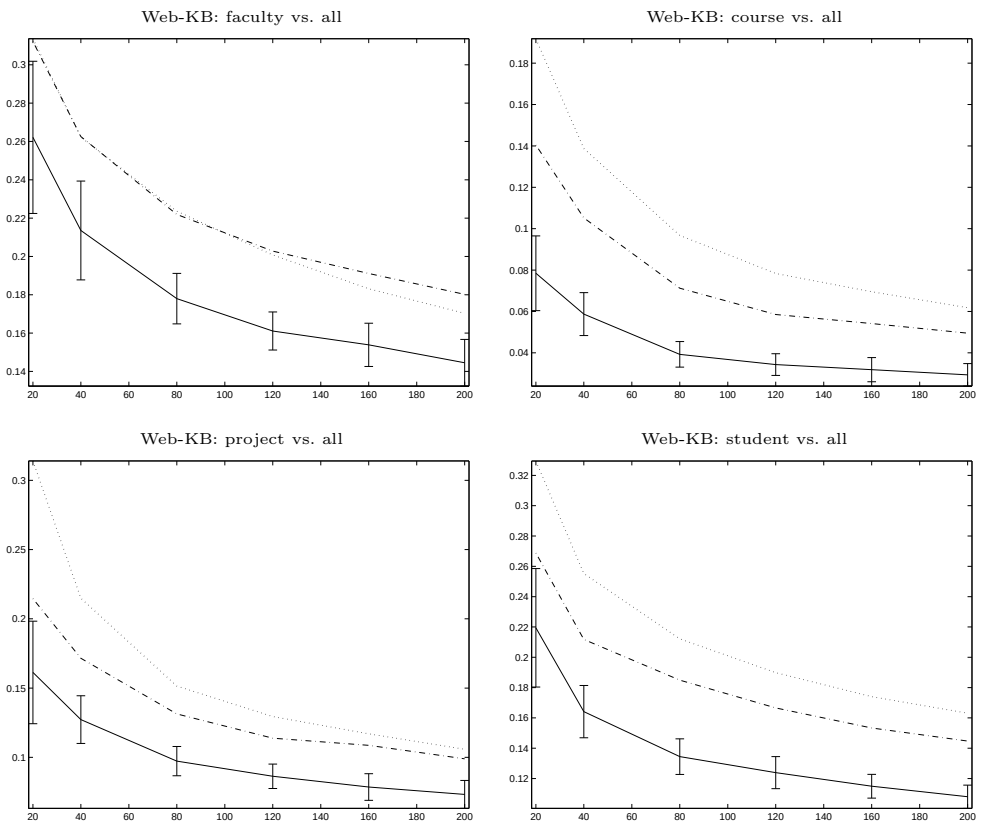

Figure 27: Test error accuracy of spherical logistic regression (solid), and linear logistic regression using tf representation with L_1 normalization (dashed) and L_2 normalization (dotted). The task is Web-KB binary "one vs. all" classification, where the name of the topic is listed above the individual plots. Error bars represent one standard deviation over 20-fold cross validation for spherical logistic regression. The error bars of the other classifiers are of similar sizes and are omitted for clarity.

Experiments were conducted on both the Web-KB (Craven et al., 1998) and the Reuters-21578 (Lewis & Ringuette, 1994) datasets. In the Web-KB dataset, the classification task that was tested was each of the classes faculty, course, project and student vs. the rest. In the Reuters dataset, the task was each of the 8 most popular classes vs. the rest. The test error rates as a function of randomly sampled training sets of different sizes are shown in Figures 27-29. In both cases, the positive and negative example sets are equally distributed, and the results were averaged over a 20-fold cross validation with the error bars indicating one standard deviation. As mentioned in Section 9.2, we assume that the boundary set of $q = \arg\min_{y \in \mathbb{S}^n_+ \cap E_u} d(y, x)$ is equal to $A = \{i : (p_u(x))_i \leq 0\}$.

The experiments show that the new linearity and margin concepts lead to more powerful classifiers than their Euclidean counterparts, which are commonly used in the literature regardless of the geometry of the data.

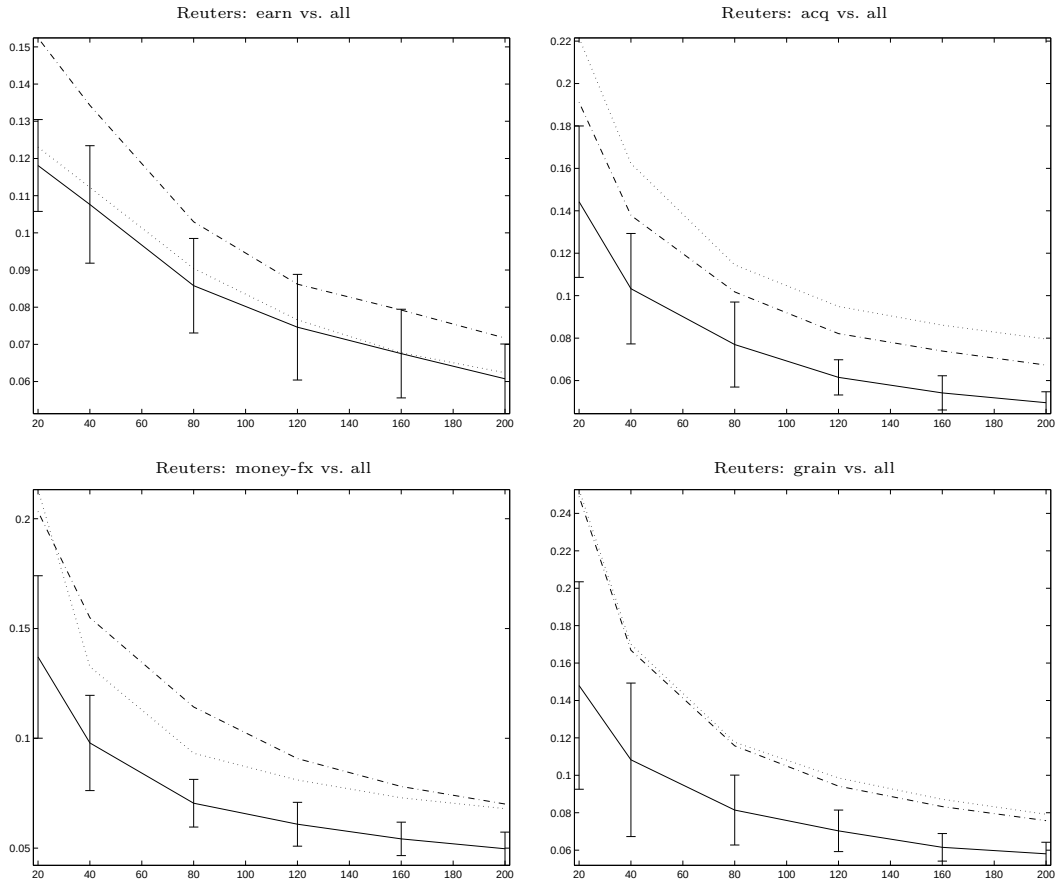

Figure 28: Test error accuracy of spherical logistic regression (solid), and linear logistic regression using tf representation with L_1 normalization (dashed) and L_2 normalization (dotted). The task is Reuters-21578 binary "one vs. all" classification, where the name of the topic is listed above the individual plots. Error bars represent one standard deviation over 20-fold cross validation for spherical logistic regression. The error bars of the other classifiers are of similar sizes and are omitted for clarity.

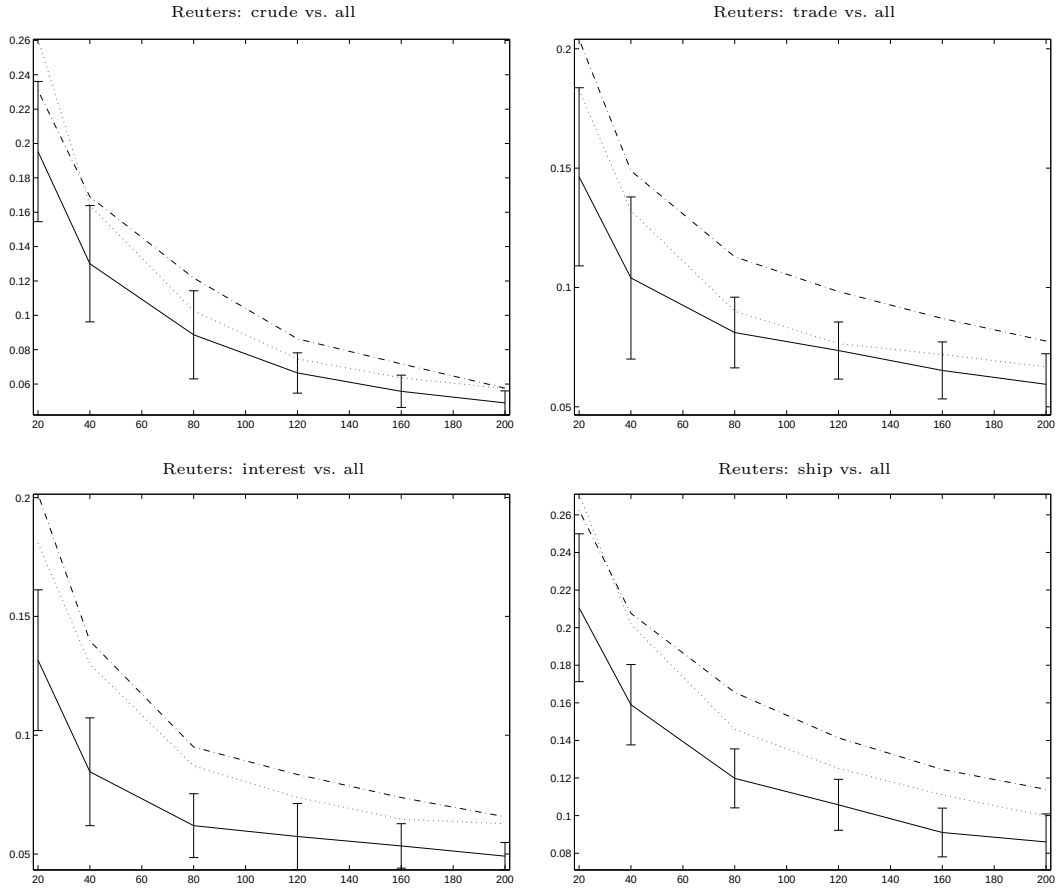

Figure 29: Test error accuracy of spherical logistic regression (solid), and linear logistic regression using tf representation with L_1 normalization (dashed) and L_2 normalization (dotted). The task is Reuters-21578 binary "one vs. all" classification, where the name of the topic is listed above the individual plots. Error bars represent one standard deviation over 20-fold cross validation for spherical logistic regression. The error bars of the other classifiers are of similar sizes and are omitted for clarity.

10 Metric Learning

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. The popular assumption of a Euclidean metric structure 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 data (if available). After obtaining the metric structure, it may be passed to a learning algorithm for use in tasks such as classification and clustering.

Several attempts have recently been made to learn the metric structure of the embedding space from a given data set. Saul and Jordan (1997) use geometrical arguments to learn optimal paths connecting two points in a space. Xing et al. (2003) learn a global metric structure. that is able to capture non-Euclidean geometry, but only in a restricted manner since the metric is constant throughout the space. Lanckriet et al. (2002) 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 dataset. Work in this area includes linear methods such as principal component analysis and nonlinear methods such as spherical subfamily models (Gous, 1998) or locally linear embedding (Roweis & Saul, 2000) and curved multinomial subfamilies (Hall & Hofmann, 2000). 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 10.1, this approach is a limiting case of learning a Riemannian metric for the embedding high-dimensional 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 in high dimensional sparse datasets. 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.

In Section 10.1 we discuss our formulation of the Riemannian metric problem. Section 10.2 describes the set of metric candidates as pull-back metrics of a group of transformations followed by a discussion of the resulting generative model in Section 10.3. In Section 10.4 we apply the framework to text classification and report experimental results on the WebKB data.

10.1 The Metric Learning Problem

The metric learning problem may be formulated as follows. Given a differentiable manifold $\mathcal{M}$ and a dataset $D = \{x_1, \ldots, x_N\} \subset \mathcal{M}$, choose a Riemannian metric g from a set of metric candidates $\mathcal{G}$. As in statistical inference, $\mathcal{G}$ may be a parametric family

$$\mathcal{G} = \{g^\theta : \theta \in \Theta \subset \mathbb{R}^k\} \tag{123}$$

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 in high dimensional sparse data such as text documents. The reason we use a superscript g^θ is that the subscript of the metric is reserved for its value at a particular point of the manifold.

We propose to choose the metric based on maximizing the following objective function $\mathcal{O}(g, D)$

$$\mathcal{O}(g, D) = \prod_{i=1}^{N} \frac{(\operatorname{dvol} g(x_i))^{-1}}{\int_{\mathcal{M}} (\operatorname{dvol} g(x))^{-1} \mathrm{d}x} \tag{124}$$

where $\operatorname{dvol} g(x) = \sqrt{\det G(x)}$ is the differential volume element, and $G(x)$ is the Gram matrix of the metric g at the point x. Note that $\det G(x) > 0$ since $G(x)$ is positive definite.

The volume element $\operatorname{dvol} g(x)$ summarizes the size of the metric g at x in one scalar. Intuitively, paths crossing areas with high volume will tend to be longer than the same paths over an area with low volume. Hence maximizing the inverse volume in (124) 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.

The normalization in (124) is necessary since the problem is clearly unidentifiable without it. Metrics cg with $0 < c < 1$ will always a have higher inverse volume element than g. The normalized inverse volume element may be seen as a probability distribution over the manifold. As a result, we may cast the problem of maximizing $\mathcal{O}$ as a maximum likelihood problem.

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 small set of relatively 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 $\mathcal{O}(g)$ is equivalent to the loglikelihood $\ell(\lambda)$ under the model

$$p(x\,;\lambda) = \frac{1}{Z}\left(\sqrt{\det G^\lambda(x)}\right)^{-1}.$$

If G is the Gram matrix of the Fisher information, the above model is the inverse of Jeffreys' prior $p(x) \propto \sqrt{\det G(x)}$. 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 (Murray & Rice, 1993).

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 $\mathcal{J}$.

If $F : (\mathcal{M}, g) \to (\mathcal{N}, \delta)$ is an isometry (recall that δ is the metric inherited from an embedding Euclidean space) we call it a flattening transformation. 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. F thus takes a locally distorted space and converts it into a subset of $\mathbb{R}^n$ equipped with the flat Euclidean metric.

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.

10.2 A Parametric Class of Metrics

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

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

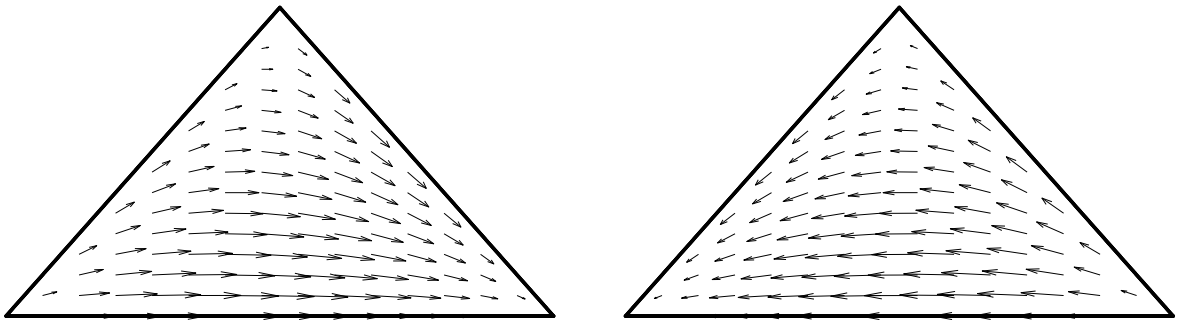

Figure 30: F_λ acting on $\mathbb{P}_2$ for $\lambda = (\frac{2}{10}, \frac{5}{10}, \frac{3}{10})$ (left) and F_λ^{-1} (right) acting on $\mathbb{P}_2$.

where $\langle x, \lambda \rangle$ is the scalar product $\sum_{i=1}^{n+1} x_i \lambda_i$. The family F_λ is a Lie group of transformations under composition whose parametric space is $\Lambda = \mathbb{P}_n$. 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. See Figure 10.2 for an illustration of 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 (see Section 4) we have that $F_\lambda^* \mathcal{J}$ is also the pull-back metric of $(\mathbb{S}_+^n, \delta)$ through the transformation

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

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

$$d_{F_\lambda^* \mathcal{J}}(x, y) = \operatorname{acos} \left(\sum_{i=1}^{n+1} \sqrt{\frac{x_i \lambda_i}{\langle x, \lambda \rangle} \frac{y_i \lambda_i}{\langle y, \lambda \rangle}} \right). \tag{125}$$

Note the only difference between (125) and tf-idf cosine similarity measure (Salton & McGill, 1983) is the square root and the choice of the λ parameters.

To apply the framework described in Section 10.1 to the metric $F_\lambda^* \mathcal{J}$ we need to compute the volume element given by $\sqrt{\det F_\lambda^* \mathcal{J}}$. 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 $T_x\mathbb{P}_n$ given by the rows of the matrix

$$U = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & 0 & \ddots & 0 & -1 \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \in \mathbb{R}^{n\times n+1}. \tag{126}$$

and computing $\det G$ in Propositions 11-12 below.

Proposition 11. *The matrix $[G]_{ij} = F_\lambda^*\mathcal{J}(\partial_i, \partial_j)$ is given by*

$$G = JJ^\top = U(D - \lambda\alpha^\top)(D - \lambda\alpha^\top)^\top U^\top \tag{127}$$

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 $j = 1,\ldots,n+1$,

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

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 the metric on $\mathbb{S}_+^n$ that is inherited from the Euclidean space through $\hat{F}_\lambda$ we have

$$[G]_{ij} = \left\langle \hat{F}_{\lambda *}\partial_i, \hat{F}_{\lambda *}\partial_j \right\rangle$$

and hence

$$G = JJ^\top = U(D - \lambda\alpha^\top)(D - \lambda\alpha^\top)^\top U^\top.$$

□

Proposition 12. *The determinant of $F_\lambda^* \mathcal{J}$ is*

$$detF_\lambda^* \mathcal{J} \propto \frac{\prod_{i=1}^{n+1}(\lambda_i/x_i)}{\langle x, \lambda \rangle^{n+1}}. \tag{128}$$

Proof. We will factor G into a product of square matrices and compute $\det G$ as the product of the determinants of each factor. Note that $G = JJ^\top$ does not qualify as such a factorization since J is not square.

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

$$J = U\left(I - \frac{\lambda x^\top}{\langle x, \lambda \rangle}\right)\Lambda \tag{129}$$

$$G = U\left(I - \frac{\lambda x^\top}{\langle x, \lambda \rangle}\right)\Lambda^2\left(I - \frac{\lambda x^\top}{\langle x, \lambda \rangle}\right)^\top U^\top. \tag{130}$$

We proceed by studying the eigenvalues and eigenvectors of $I - \frac{\lambda x^\top}{\langle x,\lambda \rangle}$ in order to simplify (130) 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,

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

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

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

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

$$\det\left(UV^{|n}\right)^2 = \left(\frac{1}{x_1}\sum_{i=1}^{n+1} x_i\right)^2 = \frac{1}{x_1^2}. \tag{131}$$

An argument presented in Appendix C.2 shows that

$$\det V^{-1|n}\Lambda^2V^{-1|n\top} = \frac{x_1^2\left\langle x,\lambda\right\rangle^{n-1}}{4^n\left\langle x,\lambda\right\rangle^{2n}}\prod_{i=1}^{n+1}\frac{\lambda_i}{x_i}. \tag{132}$$

By multiplying (132) and (131) we obtain (128). □

Figure 31 displays the inverse volume element on $\mathbb{P}_1$ with the corresponding geodesic distance from the left corner of $\mathbb{P}_1$.

Propositions 11 and 12 reveal the form of the objective function $\mathcal{O}(g, D)$. In the next section we describe a maximum likelihood estimation problem that is equivalent to maximizing $\mathcal{O}(g, D)$ and study its properties.

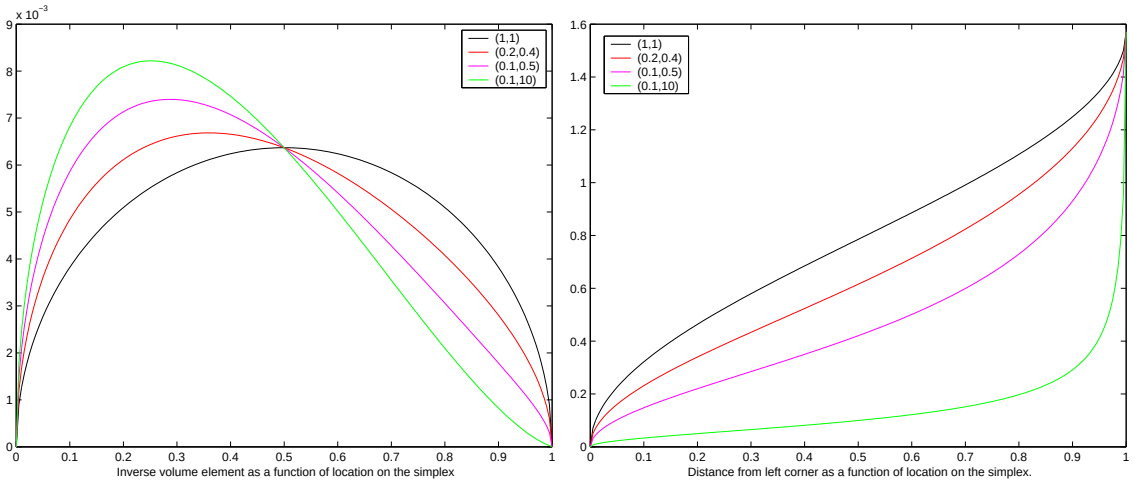

Figure 31: The inverse volume element $1/\sqrt{\det G(x)}$ as a function of $x \in \mathbb{P}_1$ (left) and the geodesic distance $d(x, 0)$ from the left corner as a function $x \in \mathbb{P}_1$ (right). Different plots represent different metric parameters $\lambda \in \{(1/2, 1/2), (1/3, 2/3), (1/6, 5/6), (0.0099, 0.9901)\}$.

10.3 An Inverse-Volume Probabilistic Model on the Simplex

Using proposition 12 we have that the objective function $\mathcal{O}(g, D)$ may be regarded as a likelihood function under the model

$$p(x\,;\lambda) = \frac{1}{Z}\,\langle x, \lambda\rangle^{\frac{n+1}{2}} \prod_{i=1}^{n+1} x_i^{1/2} \quad x \in \mathbb{P}_n,\ \lambda \in \mathbb{P}_n \tag{133}$$

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 (133) 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 loglikelihood 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

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

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 $p(x\,;\lambda)$.

10.3.1 Computing the Normalization Term

We describe an efficient way to compute the normalization term Z through the use of dynamic programming and FFT.

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

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

$$\propto \sum_{a_1+\cdots+a_{n+1}=k: a_i \geq 0} \prod_{j=1}^{n+1} \frac{\Gamma(a_j + 3/2)}{\Gamma(a_j + 1)} \lambda_j^{a_j}.$$

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

Proposition 13. *The normalization term for model* (133) *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

$$Z \propto \sum_{a_1=0}^{k} c_{a_1} \lambda_1^{a_1} \sum_{a_2=0}^{k-a_1} c_{a_2} \lambda_2^{a_2} \cdots \sum_{a_n=0}^{k-\sum_{j=1}^{n-1} a_j} c_{a_n} \lambda_n^{a_n} c_{k-\sum_{j=1}^{n} a_j} \lambda_{n+1}^{k-\sum_{j=1}^{n} a_j}. \tag{134}$$

A trivial dynamic program can compute equation (134) in $O(n^3)$ complexity.

However, each of the single subscript sums in (134) 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 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.

10.4 Application to Text Classification

In this section we describe applying the metric learning framework to document classification and report some results on the WebKB dataset (Craven et al., 1998).

We map documents to the simplex by multinomial MLE or MAP estimation. This mapping results in a the well-known term-frequency (tf) representation (see Section 7).

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 corners such as the ones that correspond to the stopwords above. To account for this in our framework we learn the metric $F_\lambda^* \mathcal{J} = (F_\theta^{-1})^* \mathcal{J}$ where θ is the MLE under model (133). 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.

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 (Salton & McGill, 1983). Despite its simplicity the tfidf representation leads to some of the best results in text classification 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 Figure 32. Note that both methods rank similar words at the bottom. These are the most common words that often carry little information for classification purposes. The top words however are completely different for the two schemes. Note the tendency of tfidf 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. Rare proper nouns such as the high scoring tfidf terms in Figure 32 appear several times in a single web page. As a result, these words will score higher with the tfidf scheme but lower with the metric learning scheme.

In Figure 33 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. Note 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 several different metrics. We compared tfidf cosine similarity and the geodesic distance under the obtained metric obtained. Figure 34 displays test-set error rates as a function of the training set size. The error rates were averaged over 20 experiments with random sampling of the training set. The λ parameter was obtained by approximated gradient descent procedure using the dynamic programming method described in Section 10.3.1. According to Figure 34 the learned metric outperforms the standard tfidf measure.

tfidf	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

Figure 32: Comparison of top and bottom valued parameters for tfidf and model (133). The dataset is the faculty vs. student webpage classification task from WebKB dataset. Note that the least scored terms are similar for the two methods while the top scored terms are completely disjoint.

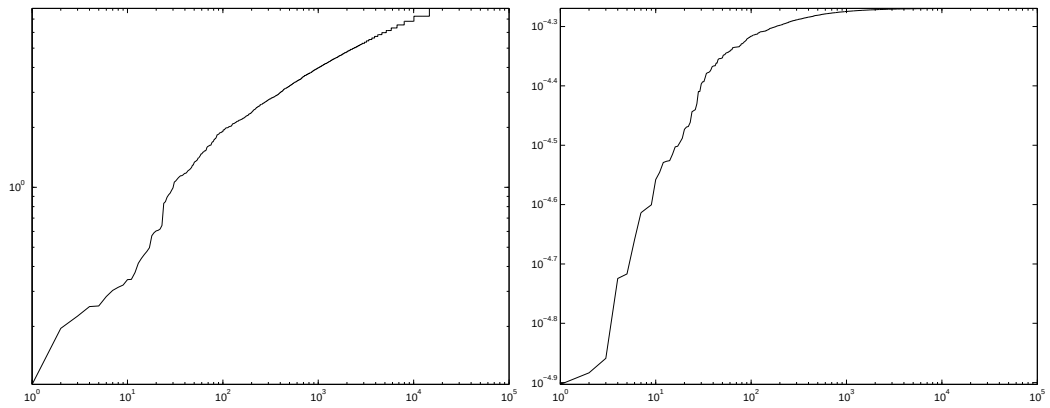

Figure 33: Log-log plots for sorted values of tfidf (top) and estimated λ values (bottom). The task is the same as in Figure 32.

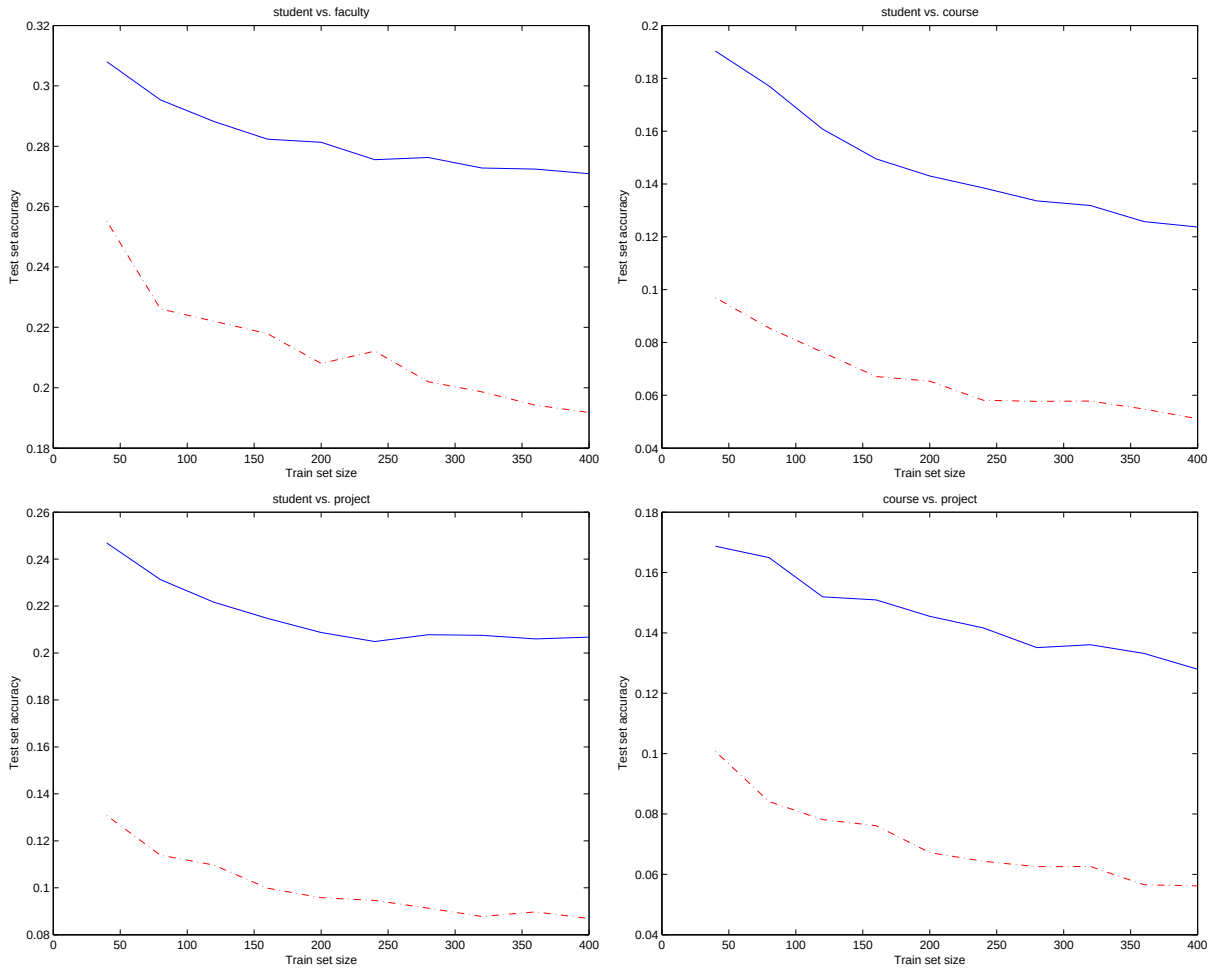

Figure 34: Test set error rate for nearest neighbor classifier on WebKB binary tasks. Distances were computed by geodesic for the learned Riemannian metric (red dashed) and tfidf with cosine similarity (blue solid). The plots are averaged over a set of 20 random samplings of training sets of the specified sizes, evenly divided between positive and negative examples.

10.5 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 transformation. When composed with a square root, the transformations are flattening transformation for the obtained metrics. The resulting optimization problem may be interpreted as maximum likelihood estimation.

Guided by the well known principle that common words should have little effect on the metric structure we learn the metric that is associated with the inverse to the transformation that maximizes the inverse volume of the training set. The resulting pull-back metric de-emphasizes common words, in a way similar to tfidf. Despite the similarity between the resulting geodesics and tfidf similarity measure, there are significant qualitative and quantitative differences between the two methods. Using a nearest neighbor classifier in a text classification experiment, the obtained metric is shown to significantly outperform the popular tfidf cosine similarity.

The framework proposed in this section is quite general and allows implementations in other domains. The key component is the specification of the set of metric candidates possibly by parametric transformations in a way that facilitates efficient computation and maximization of the volume element.

11 Discussion

The use of geometric techniques in studying statistical learning algorithms is not new. As mentioned in Section 3 the geometric view of statistics was developed over the past fifty years. The focus of research in this area has been finding connections between geometric quantities under the Fisher geometry and asymptotic statistical properties. A common criticism is that the geometric viewpoint is little more than an elegant way to explain these asymptotic properties. There has been little success in carrying the geometric reasoning further to define new models and inference methods that outperform existing models in practical situations.

The contributions of this thesis may be roughly divided into three parts. The first part contains the embedding principle and the novel algorithms of Sections 8 and 9 which are a direct response to the criticism outlined above. Based on the Fisher geometry of the embedded data, we derive generalizations of several popular state-of-the-art algorithms. These geometric generalizations significantly outperform their popular counterparts in the task of text classification.

The second part contains an extension of information geometry to spaces of conditional models. Čencov's important theorem is extended to both normalized and non-normalized conditional

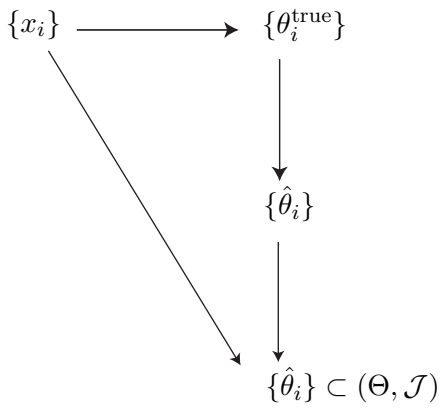

Figure 35: Motivation for the embedding principle. The data $\{x_i\}$ is assumed to be drawn from $\{p(x\,;\theta_i^{\text{true}}\}$. It is reasonable to assume that the models $\{\theta_i^{\text{true}}\}$ capture the essence of the data and may be used by the algorithms in place of the data. Since we do not know the true parameters, we replace them with an estimate $\{\hat{\theta}_i\}$ (see Section 3 for more details). The final step is to consider the estimated models $\{\hat{\theta}_i\}$ as points in a manifold endowed with the Fisher information metric, whose choice is motivated by Čencov's theorem.

models. A previously undiscovered relationship between maximum likelihood for conditional exponential models and minimum exponential loss for AdaBoost leads to a deeper understanding of both models. This relationship, together with the generalized Čencov characterization reveals an axiomatic framework for conditional exponential models and AdaBoost. An additional bi-product of this relationship is that it enables us to derive the AdaBoost analogue of maximum posterior inference. This new algorithm performs better than AdaBoost in cases where overfitting is likely to occur.

Information geometry has almost exclusively been concerned with the Fisher geometry. The third part of this thesis extends the Riemannian geometric viewpoint to data-dependent geometries. The adapted geometry exhibits a close similarity in its functional form to the tf-idf metric, but leads to a more effective metric for text classification.

It is interesting to consider the motivation leading to the algorithmic part of the thesis. The algorithms transfer the data point into a Riemannian manifold Θ with the Fisher information metric. A schematic view of this process is outlined in Figure 35. The data $\{x_i\}$ is assumed to be drawn from $\{p(x\,;\theta_i^{\text{true}}\}$. It is reasonable to assume that the models $\{\theta_i^{\text{true}}\}$ capture the essence of the data and may be used by the algorithms in place of the data. Since we do not know the true parameters, we replace them with an estimate $\{\hat{\theta}_i\}$ (See Section 3 for more details). The final step is to consider the estimated models $\{\hat{\theta}_i\}$ as points in a manifold endowed with the Fisher information metric, whose choice is motivated by Čencov's theorem.

An interesting prospect for future research is to develop further the theory behind the embedding principle, and more generally the geometric approach to machine learning. Theoretical results such

as generalization error bounds and decision theory might be able to provide a comparison of different geometries. For example, error bounds for nearest neighbor and other classification algorithms might depend on the employed geometry. A connection between such standard theoretical results and the geometry under consideration would be a significant addition to the algorithmic ideas in this thesis. Such a result may provide a more rigorous motivation for the ideas in Figure 35 that underlie a large part of this thesis.

It is also interesting to consider the role of geometry in developing learning theory results. Replacing Euclidean geometric concepts, such as the Euclidean diameter of the data, with their non-Euclidean counterparts may lead to a new understanding of current algorithms and to alternative error bounds.

Despite the fact that information geometry is half a century old, I believe it is only beginning to affect the design of practical algorithms in statistical machine learning. Many machine learning algorithms, including some of the most popular ones, make naive unrealistic geometric assumptions. I hope that the contribution of this thesis will draw greater attention to this fact and encourage others to exploit the geometric viewpoint in the quest of designing better practical algorithms.

Appendix

A Derivations Concerning Boosting and Exponential Models

In this appendix we provide some technical details concerning Section 5. We derive update rules for minimizing the exponential loss of AdaBoost and the log-loss of exponential models, derive the dual problem of regularized I-divergence minimization and express the I-divergence between exponential models as a difference between loglikelihoods.

A.1 Derivation of the Parallel Updates

Let $x \in \mathcal{X}$ be an example in the training set, which is of size N, $\tilde{y}$ be its label and $\mathcal{Y}$ is the set of all possible labels. At a given iteration θ_j denotes the j-th parameter of the model, and $\theta_j + \Delta\theta_j$ the parameter at the following iteration.

A.1.1 Exponential Loss

The objective is to minimize $\mathcal{E}_{exp}(\theta + \Delta\theta) - \mathcal{E}_{exp}(\theta)$. In the following $h_j(x,y) = f_j(x,y) - f_j(x,\tilde{y})$, $q_\theta(y|x) = e^{\sum_j \theta_j h_j(x,y)}$, $s_j(x,y) = \text{sign}(h_j(x,y))$, $M = \max_{i,y} \sum_j |h_j(x_i,y)|$, $\omega_{i,y} = 1 - \sum_j \frac{|h_j(x_i,y)|}{M}$.

By Jensen's inequality applied to e^x we have

$$\begin{aligned}
\mathcal{E}_{exp}(\theta + \Delta\theta) - \mathcal{E}_{exp}(\theta) &= \sum_i \sum_y e^{\sum_j (\theta_j + \Delta\theta_j) h_j(x_i, y)} - \sum_i \sum_y e^{\sum_j \theta_j h_j(x_i, y)} \\
&= \sum_i \sum_y q_\theta(y|x_i) e^{\sum_j \Delta\theta_j \frac{|h_j(x_i,y)|}{M} s_j(x_i,y)\, M} - \sum_i \sum_y q_\theta(y|x_i) \\
&\leq \sum_i \sum_y q_\theta(y|x_i) \left(\sum_j \frac{|h_j(x_i, y)|}{M} e^{\Delta\theta_j s_j(x_i,y) M} + \omega_{i,y} - 1 \right) \\
&\stackrel{\text{def}}{=} \mathcal{A}(\Delta\theta, \theta). \qquad (135)
\end{aligned}$$

We proceed by finding the stationary point of the auxiliary function with respect to $\Delta\theta_j$:

$$\begin{aligned}
0 &= \frac{\partial \mathcal{A}}{\partial \Delta\theta_j} = -\sum_i \sum_y q_\theta(y|x_i) h_j(x_i, y) e^{\Delta\theta_j s_j(x_i,y)\, M} \\
&= -\sum_y \sum_{i: s_j(x_i,y)=+1} q_\theta(y|x_i) h_j(x_i, y) e^{\Delta\theta_j M} - \sum_y \sum_{i: s_j(x_i,y)=-1} q_\theta(y|x_i) h_j(x_i, y) e^{-\Delta\theta_j M} \\
&\Rightarrow e^{2M\Delta\theta_j} \sum_y \sum_{i: s_j(x_i,y)=+1} h_j(x_i, y) q_\theta(y|x_i) = \sum_y \sum_{i: s_j(x_i,y)=-1} |h_j(x_i, y)| q_\theta(y|x_i) \\
&\Rightarrow \Delta\theta_j = \frac{1}{2M} \log \left(\frac{\sum_y \sum_{i: s_j(x_i,y)=-1} |h_j(x_i, y)| q_\theta(y|x_i)}{\sum_y \sum_{i: s_j(x_i,y)=+1} |h_j(x_i, y)| q_\theta(y|x_i)} \right)
\end{aligned}$$

A.1.2 Maximum Likelihood for Exponential Models

For the normalized case, the objective is to maximize the likelihood or minimize the log-loss. In this section, the previous notation remains except for $q_\theta(y|x) = \frac{e^{\sum_j \theta_j h_j(x,y)}}{\sum_y e^{\sum_j \theta_j h_j(x,y)}}$. The log-likelihood is

$$\ell(\theta) = \sum_i \log \frac{e^{\sum_j \theta_j f_j(x_i, y_i)}}{\sum_y e^{\sum_j \theta_j f_j(x_i, y_i)}} = -\sum_i \log \sum_y e^{\sum_j \theta_j (f_j(x_i, y) - f_j(x_i, y_i))}$$

and the difference in the loss between two iterations is

$$\begin{aligned}
\ell(\theta) - \ell(\theta + \Delta\theta) &= \sum_i \log \frac{\sum_y e^{\sum_j (\theta_j + \Delta\theta_j)(f_j(x_i,y) - f_j(x_i,y_i))}}{\sum_y e^{\sum_j \theta_j (f_j(x_i,y) - f_j(x_i,y_i))}} \\
&= \sum_i \log \frac{\sum_y e^{\sum_j (\theta_j + \Delta\theta_j) h_j(x_i,y)}}{\sum_y e^{\sum_j \theta_j h_j(x_i,y)}} \\
&= \sum_i \log \sum_y q_\theta(y|x_i) e^{\sum_j \Delta\theta_j h_j(x_i,y)} \\
&\leq \sum_i \sum_y q_\theta(y|x_i) e^{\sum_j \Delta\theta_j h_j(x_i,y)} - N \qquad (136) \\
&= \sum_i \sum_y q_\theta(y|x_i) e^{\sum_j \Delta\theta_j \frac{|h_j(x_i,y)|}{M} s_j(x_i,y)\, M} - N \\
&\leq \sum_i \sum_y q_\theta(y|x_i) \left(\sum_j \frac{|h_j(x_i,y)|}{M} e^{\Delta\theta_j s_j(x_i,y)\, M} + \omega_{i,y} \right) - N \qquad (137) \\
&\stackrel{\text{def}}{=} \mathcal{A}(\theta, \Delta\theta) \qquad (138)
\end{aligned}$$

where in (136) we used the inequality $\log x \leq x - 1$ and in (137) we used Jensen's inequality. The derivative of (137) with respect to $\Delta\theta$ will be identical to the derivative of (135) and so the log-loss update rule will be identical to the exponential loss update rule, but with $q_\theta(y|x)$ representing a normalized exponential model.

A.2 Derivation of the Sequential Updates

The setup for sequential updates is similar to that for parallel updates, but now only one parameter gets updated in each step, while the rest are held fixed.

A.2.1 Exponential Loss

We now assume that that only θ_k gets updated. We also assume (with no loss of generality) that each feature takes values in $[0, 1]$, making $h_k(x_i, y) \in [-1, 1]$.

$$\begin{aligned}
\mathcal{E}_{exp}(\theta + \Delta\theta) - \mathcal{E}_{exp}(\theta) &= \sum_i \sum_y e^{\sum_j \theta_j h_j(x_i,y) + \Delta\theta_k h_k(x_i,y)} - \sum_i \sum_y e^{\sum_j \theta_j h_j(x_i,y)} \\
&= \sum_i \sum_y q_\theta(y|x_i) \left(e^{\left(\frac{1+h_k(x_i,y)}{2}\right)\Delta\theta_k + \left(\frac{1-h_k(x_i,y)}{2}\right)(-\Delta\theta_k)} - 1 \right) \qquad (139) \\
&\leq \sum_i \sum_y q_\theta(y|x_i) \left(\frac{1 + h_k(x_i,y)}{2} e^{\Delta\theta_k} + \frac{1 - h_k(x_i,y)}{2} e^{-\Delta\theta_k} - 1 \right) \\
&\stackrel{\text{def}}{=} \mathcal{A}(\theta, \Delta\theta_k) \qquad (140)
\end{aligned}$$

The stationary point of $\mathcal{A}$ (with respect to $\Delta\theta_k$) is

$$\begin{aligned} 0 &= \sum_i \sum_y q_\theta(y|x_i) \left(\frac{1+h_k(x_i,y)}{2} e^{\Delta\theta_k} + \frac{h_k(x_i,y)-1}{2} e^{-\Delta\theta_k} \right) \\ &\Rightarrow \quad e^{2\Delta\theta_k} \sum_i \sum_y q_\theta(y|x_i)(1+h_k(x_i,y)) = \sum_i \sum_y q_\theta(y|x_i)(1-h_k(x_i,y)) \\ &\Rightarrow \quad \Delta\theta_k = \frac{1}{2} \log \left(\frac{\sum_i \sum_y q_\theta(y|x_i)(1-h_k(x_i,y))}{\sum_i \sum_y q_\theta(y|x_i)(1+h_k(x_i,y))} \right) \end{aligned}$$

A.2.2 Log-Loss

Similarly, for the log-loss we have

$$\begin{aligned} \ell(\theta) - \ell(\theta + \Delta\theta_k) &= \sum_i \log \frac{\sum_y e^{\sum_j \theta_j h_j(x_i,y) + \Delta\theta_k h_k(x_i,y)}}{\sum_y e^{\sum_j \theta_j h_j(x_i,y)}} = \sum_i \log \sum_y q_\theta(y|x_i) e^{\Delta\theta_k h_k(x_i,y)} \\ &\leq \sum_i \sum_y q_\theta(y|x_i) e^{\Delta\theta_k h_k(x_i,y)} - N. \end{aligned} \tag{141}$$

Equation (141) is the same as (139), except that q_θ is now the normalized model. This leads to exactly the same form of update rule as in the previous subsubsection.

A.3 Regularized Loss Functions

We derive the dual problem for the non-normalized regularized I-divergence minimization and then proceed to derive a sequential update rule.

A.3.1 Dual Function for Regularized Problem

The regularized problem $(P_{1,\text{reg}})$ is equivalent to

$$\begin{aligned} &\textit{Minimize} \qquad D(p,q_0) + U(c) = \sum_x \tilde{p}(x) \sum_y p(y|x) \left(\log \frac{p(y|x)}{q_0(y|x)} - 1 \right) + U(c) \\ &\textit{subject to} \qquad f_j(p) = \sum_{x,y} \tilde{p}(x) p(y|x) h_j(x,y) = c_j, \quad j = 1, \ldots, m \end{aligned}$$

where $c \in \mathbb{R}^m$ and $U : \mathbb{R}^m \to \mathbb{R}$ is a convex function whose minimum is at 0. The Lagrangian turns out to be

$$\mathcal{L}(p,c,\theta) = \sum_x \tilde{p}(x) \sum_y p(y|x) \left(\log \frac{p(y|x)}{q_0(y|x)} - 1 - \langle \theta, h(x,y) \rangle \right) + U(c).$$

We will derive the dual problem for $U(c) = \sum_i \frac{1}{2}\sigma_i^2 c_i^2$. The convex conjugate U^* is

$$\begin{aligned} U^*(\theta) &\stackrel{\text{def}}{=} \inf_c \sum_i \theta_i c_i + U(c) = \inf_c \sum_i \theta_i c_i + \sum_i \frac{1}{2}\sigma_i^2 c_i^2 \\ 0 &= \theta_i + \sigma_i^2 c_i \quad \Rightarrow \quad c_i = -\frac{\theta_i}{\sigma_i^2} \\ U^*(\theta) &= -\sum_i \frac{\theta_i^2}{\sigma_i^2} + \sum_i \frac{1}{2}\sigma_i^2 \frac{\theta_i^2}{\sigma_i^4} = -\sum_i \frac{\theta_i^2}{2\sigma_i^2} \end{aligned} \tag{142}$$

and the dual problem is

$$\begin{aligned}\theta^\star &= \arg\max_\theta h_{1,reg}(\theta) \\ &= \arg\max_\theta -\sum_x \tilde{p}(x) \sum_y q_0(y|x) e^{\sum_j \theta_j h_j(x,y)} + U^*(\theta) \\ &= \arg\min_\theta \sum_x \tilde{p}(x) \sum_y q_0(y|x) e^{\sum_j \theta_j h_j(x,y)} + \sum_j \frac{\theta_j^2}{2\sigma_j^2}.\end{aligned} \tag{143}$$

A.3.2 Exponential Loss–Sequential update rule

We now derive a sequential update rule for $(P_{1,\text{ref}})$. As before, $q_0 = 1$ and we replace $\frac{1}{2\sigma_k^2}$ by β.

$$\begin{aligned}&\mathcal{E}_{exp}(\theta + \Delta\theta) - \mathcal{E}_{exp}(\theta) \\ &= \sum_i \sum_y \left(e^{\sum_j \theta_j h_j(x_i,y) + \Delta\theta_k h_k(x_i,y)} - e^{\sum_j \theta_j h_j(x_i,y)} \right) + \beta(\theta_k + \Delta\theta_k)^2 - \beta\theta_k^2 \\ &= \sum_i \sum_y q_\theta(y|x_i) \left(e^{\left(\frac{1+h_k(x_i,y)}{2}\right)\Delta\theta_k + \left(\frac{1-h_k(x_i,y)}{2}\right)(-\Delta\theta_k)} - 1 \right) \\ &\quad + 2\beta\,\theta_k\,\Delta\theta_k + \beta\,\Delta\theta_k^2 \\ &\leq \sum_i \sum_y q_\theta(y|x_i) \left(\frac{1+h_k(x_i,y)}{2} e^{\Delta\theta_k} + \frac{1-h_k(x_i,y)}{2} e^{-\Delta\theta_k} - 1 \right) \\ &\quad + 2\beta\,\theta_k\,\Delta\theta_k + \beta\,\Delta\theta_k^2 \\ &\stackrel{\text{def}}{=} \mathcal{A}(\theta, \Delta\theta_k)\end{aligned} \tag{144}$$

The stationary point will be at the solution of the following equation

$$0 = \frac{\partial \mathcal{A}}{\partial \Delta\theta_k} = \frac{1}{2} \sum_i \sum_y q_\theta(y|x_i) \left((1 + h_k(x_i,y)) e^{\Delta\theta_k} + (h_k(x_i,y) - 1) e^{-\Delta\theta_k} \right) + 2\beta\,(\theta_k + \Delta\theta_k).$$

that can be readily found by Newton's method. Since the second derivative $\frac{\partial^2 \mathcal{A}}{\partial \Delta\theta_k^2}$ is positive, the auxiliary function is strictly convex and Newton's method will converge.

A.4 Divergence Between Exponential Models

In this subsection we derive the I-divergence between two exponential models, one of which is the maximum likelihood model as a difference in their loglikelihoods. The log-likelihood of an exponential model q_θ is

$$\begin{aligned}\ell(\theta) &= \frac{1}{n} \sum_i \log \frac{e^{\sum_j \theta_j f_j(x_i,y_i)}}{Z_{\theta,i}} = \frac{1}{n} \sum_j \theta_j \sum_i f_j(x_i,y_i) - \frac{1}{n} \sum_i \log Z_{\theta,i} \\ &= \sum_j \theta_j E_{\tilde{p}}[f_j] - \frac{1}{n} \sum_i \log Z_{\theta,i}\end{aligned}$$

while the I-divergence is

$$\begin{aligned} D(q_\theta, q_\eta) &= \frac{1}{n}\sum_i \sum_y q_\theta(y|x_i) \log \frac{q_\eta(y|x_i)}{q_\theta(y|x_i)} \\ &= \frac{1}{n}\sum_i \sum_y q_\theta(y|x_i) \left(\log \frac{Z_{\eta,i}}{Z_{\theta,i}} + \log \frac{e^{\sum_j \theta_j f_j(x_i,y)}}{e^{\sum_j \eta_j f_j(x_i,y)}} \right) \\ &= \frac{1}{n}\sum_i \log \frac{Z_{\eta,i}}{Z_{\theta,i}} + \frac{1}{n}\sum_i \sum_y q_\theta(y|x_i) \sum_j f_j(x_i,y)(\theta_j - \eta_j) \\ &= \frac{1}{n}\sum_i \log \frac{Z_{\eta,i}}{Z_{\theta,i}} + \frac{1}{n}\sum_j (\theta_j - \eta_j) E_{q_\theta}[f_j]. \end{aligned}$$

This corresponds to the fact that the I divergence between normalized exponential models is the Bregman divergence, with respect to the cumulant function, between the natural parameters. If q_θ is the maximum likelihood model, the moment constraints are satisfied and

$$D(q_\theta^{ml}, q_\eta) = \frac{1}{n}\sum_i \log \frac{Z_{\eta,i}}{Z_{\theta,i}^{ml}} + \sum_j (\theta_j^{ml} - \eta_j) E_{\tilde{p}}[f_j] = \ell(\theta^{ml}) - \ell(\eta).$$

B Gibbs Sampling from the Posterior of Dirichlet Process Mixture Model based on a Spherical Normal Distribution

In this appendix we derive the Gibbs sampling from the posterior of a Dirichlet Process Mixture Model (DPMM) based on a Spherical Normal Distribution. For details on DPMM see (Ferguson, 1973; Blackwell & MacQueen, 1973; Antoniak, 1974) and for a general discussion on MCMC sampling from the posterior see (Neal, 2000). As mentioned by Neal (2000) the vanilla Gibbs sampling discussed below is not the most efficient and other more sophisticated sampling scheme may be derived.

We will assume below that the data dimensionality is 2. All the derivations may be easily extended to a higher dimensional case. We denote the data points by $y = (y_1, \ldots, y_m)$ where $y_i \in \mathbb{R}^2$ and the spherical Gaussian parameters associated with the data by $\theta = (\theta_1, \ldots, \theta_m), \theta_i \in \mathbb{H}^3$. We use the notation θ_{-i} to denote $\{\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_m\}$. In Gibbs sampling we sample from the conditionals $p(\theta_i|\theta_{-i}, y)$ repeatedly to obtain a sample from the DPMM posterior $p(\theta|y)$. The conditional is proportional to (e.g. (Neal, 2000))

$$p(\theta_i|\theta_{-i}, y) \propto p(y_i|\theta_i) p(\theta_i|\theta_{-i}) = \frac{1}{2\pi\sigma_i^2} e^{-\|y_i - \mu_i\|^2/2\sigma_i^2} \left(\frac{1}{n-1+\zeta} \sum_{j \neq i} \delta_{\theta_i,\theta_j} + \frac{\zeta}{n-1+\zeta} G_0(\theta_i) \right)$$

where ζ is the "power" parameter of the DPMM and G_0 is a conjugate prior to the spherical noral distribution

$$G_0(\mu_i, \sigma_i^2) = \text{Inv-}\Gamma^2(\sigma_i^2|\alpha, \beta) N(\mu_i|\mu_0, \sigma_i)$$

To sample from $p(\theta_i|\theta_{-i}, y)$ we need to write the product $p(y_i|\theta_i)G_0(\theta_i)$

$$\begin{aligned} p(y_i|\theta_i)G_0(\theta_i) &= \frac{1}{2\pi\sigma_i^2} e^{-\|y_i-\mu_i\|^2/2\sigma_i^2} \frac{1}{2\pi\sigma_i^2} e^{-\|\mu_i-\mu_0\|^2/2\sigma_i^2} \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma_i^2)^{-(\alpha+1)} e^{-\beta/\sigma_i^2} \\ &= \frac{\beta^\alpha}{4\pi^2\Gamma(\alpha)} (\sigma_i^2)^{-(\alpha+3)} \exp\left(-\frac{\|y_i-\mu_i\|^2 + \|\mu_i-\mu_0\|^2 + 2\beta}{2\sigma_i^2}\right) \end{aligned}$$

as a distribution times a constant. To do so, we expand the exponent's negative numerator

$$\begin{aligned} &\|\mu_i - y_i\|^2 + \|\mu_i - \mu_0\|^2 + 2\beta \\ &= \left(2(\mu_i^1)^2 - 2\mu_i^1(y_i^1 + \mu_0^1) + (y_i^1)^2 + (\mu_0^1)^2 + \beta\right) + \left(2(\mu_i^2)^2 - 2\mu_i^2(y_i^2 + \mu_0^2) + (y_i^2)^2 + (\mu_0^2)^2 + \beta\right) \\ &= 2\left((\mu_i^1 - (y_i^1 + \mu_0^1)/2)^2 + \frac{1}{2}(y_i^1)^2 + \frac{1}{2}(\mu_0^1)^2 + \frac{1}{2}\beta - (y_i^1 + \mu_0^1)^2/4\right) \\ &+ 2\left((\mu_i^2 - (y_i^2 + \mu_0^2)/2)^2 + \frac{1}{2}(y_i^2)^2 + \frac{1}{2}(\mu_0^2)^2 + \frac{1}{2}\beta - (y_i^2 + \mu_0^2)^2/4\right) \\ &= 2\|\mu_i - (y_i + \mu_0)/2\|^2 + \|y_i\|^2 + \|\mu_0\|^2 + 2\beta - \|y_i + \mu_0\|^2/2 \end{aligned}$$

to obtain

$$\begin{aligned} &p(y_i|\theta_i)G_0(\theta_i) \\ =&\frac{\beta^\alpha}{4\pi^2\Gamma(\alpha)} (\sigma_i^2)^{-(\alpha+3)} \exp\left(-\frac{\|\mu_i - (y_i+\mu_0)/2\|^2}{\sigma_i^2}\right) \exp\left(-\frac{\|y_i\|^2 + \|\mu_0\|^2 + 2\beta - \|y_i+\mu_0\|^2/2}{2\sigma_i^2}\right) \\ &= \frac{\beta^\alpha}{4\pi^2\Gamma(\alpha)} \pi N\left(\mu_i \middle| \frac{y_i+\mu_0}{2}, \frac{\sigma_i}{\sqrt{2}}\right) \frac{\Gamma(\alpha+1)}{(\beta^*)^{\alpha+1}} \text{Inv-}\Gamma\left(\sigma_i^2|\alpha+1, \beta^*\right) \\ &= \frac{\alpha\beta^\alpha}{4\pi(\beta^*)^{\alpha+1}} N\left(\mu_i \middle| \frac{y_i+\mu_0}{2}, \frac{\sigma_i}{\sqrt{2}}\right) \text{Inv-}\Gamma\left(\sigma_i^2|\alpha+1, \beta^*\right) \end{aligned}$$

where

$$\beta^* = \frac{\|y_i\|^2 + \|\mu_0\|^2 + 2\beta - \|y_i+\mu_0\|^2/2}{2}.$$

Sampling is now trivial as the conditional $p(\theta_i|\theta_{-i}, y)$ is identified as a mixture model, with known mixture coefficients, of impulses (probability concentrated on a single element) and a normal-inverse-Gamma model.

C The Volume Element of a Family of Metrics on the Simplex

This appendix contains some calculations that are used in Section 10. Refer to that section for explanations of the notation and background.

C.1 The Determinant of a Diagonal Matrix plus a Constant Matrix

We prove some basic results concerning the determinants of a diagonal matrix plus a constant matrix. These results will be useful in Appendix C.2.

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 multi-linearity property of the determinant means that the function f above is linear in each of its components

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

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 multi-linearity 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 or -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 multi-linearity argument we have

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

□

C.2 The Differential Volume Element of $F_\lambda^*\mathcal{J}$

We compute below $\det V^{-1|n}\Lambda^2 V^{-1|n\top}$. See Proposition 12 for an explanation of the notation.

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

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

where

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

$[V_n^{-1}\Lambda^2 V_n^{-1\top}]_{ij}$ is the scalar product of the ith and jth rows of the following matrix

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

We therefore have

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

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 in Section C.1 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 we obtain

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

D Summary of Major Contributions

Listed below are the major contributions of this thesis and the relevant sections and publications.

Contribution	Section	Relevant Publications
Equivalence of maximum likelihood for conditional exponential models and minimum exponential loss for AdaBoost	5, A	Lebanon and Lafferty (2002)
Axiomatic characterization of Fisher geometry for spaces of conditional probability models	6	Lebanon (2004) Lebanon (to appear)
The embedding principle and the corresponding natural geometries on the data space	7	Lafferty and Lebanon (2003)
Diffusion Kernels on Statistical Manifolds	8	Lafferty and Lebanon (2003) Lafferty and Lebanon ((accepted))
Hyperplane Margin Classifiers on the Multinomial Manifold	9	Lebanon and Lafferty (2004)
Learning framework for Riemannian metrics	10, B	Lebanon (2003a) Lebanon (2003b)

References

Aizerman, M. A., Braverman, E. M., & Rozonoér, L. I. (1964). Theoretical foundations of the potential function method in pattern recognition and learning. *Automation and Remote Control*, *25*, 821–837.

Amari, S. (1995). Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, *8*, 1379–1408.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*.

Amari, S. (1999). Natural gradient for over and under-complete bases in ICA. *Neural Computation*, *11*.

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry*. American Mathematical Society.

Antoniak, C. (1974). Mixture of dirichlet processes with applications to bayesian non-parametric problems. *The Annals of Statistics*, *2*.

Atkinson, C., & Mitchell, A. F. (1981). Rao's distance measure. *Sankhya: The Indian Journal of Statistics, A*, *43*.

Barndorff-Nilsen, O., & Blæsild, P. (1983). Exponential models with affine dual foliations. *The Annals of Statistics*, *11*.

Barndorff-Nilsen, O., & Blæsild, P. (1993). Orthogeodesic models. *The Annals of Statistics*, *21*.

Barndorff-Nilsen, O. E. (1986). Likelihood and observed geometries. *The Annals of Statistics*, *14*.

Bartlett, P. L., Bousquet, O., & Mendelson, S. (2003). Local Rademacher complexities. Manuscript.

Bartlett, P. L., & Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, *3*, 463–482.

Belkin, M., & Niyogi, P. (2003). Using manifold structure for partially labeled classification. *Advances in Neural Information Processing Systems*.

Beran, R. (1977). Minimum hellinger distance estimates for parameteric models. *Annals of Statistics*, *5*, 445–463.

Berger, M., Gauduchon, P., & Mazet, E. (1971). Le spectre d'une varieté Riemannienne. *Lecture Notes in Mathematics*, *Vol. 194,* Springer-Verlag.

Blackwell, D., & MacQueen, J. (1973). Ferguson distributions via polya urn schemes. *The Annals of Statistics*, *1*.

Blake, C., & Merz, C. (1998). The UCI repository of machine learning databases. University of California, Irvine, Department of Information and Computer Sciences, 1998. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Boser, B. E., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *Computational Learing Theory* (pp. 144–152).

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Bridson, M., & Haefliger, A. (1999). *Metric spaces of non-positive curvature*, vol. 319 of *A Series in Comprehensive Studies in Mathematics*. Springer.

Brody, D. C., & Houghston, L. P. (1998). Statistical geometry in quantum mechanics. *Proc. of the Royal Society: Mathematical, Physical and Engineering Sciences*, *454*.

Campbell, L. L. (1986). An extended Čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, *98*, 135–141.

Casella, G. (1985). An introduction to empirical bayes data analysis. *The American Statistician*, *39*.

Čencov, N. N. (1982). *Statistical decision rules and optimal inference*. American Mathematical Society.

Chen, S., & Rosenfeld, R. (2000). A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing*, *8*.

Collins, M., Schapire, R. E., & Singer, Y. (2002). Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, *48*.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., & Slattery., S. (1998). Learning to extract symbolic knowledge from the world wide web. *Proceedings of the 15th National Conference on Artificial Intelligence*.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, *3*.

Csiszár, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, *19*.

Cutler, A., & Cordero-Brana, O. I. (1996). Minimum hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, *91*, 1716–1723.

Dawid, A. P. (1975). Discussion of Efron's paper. *The Annals of Statistics*, *3*.

Dawid, A. P. (1977). Further comments on some comments on a paper by bradley efron. *The Annals of Statistics*, *5*.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (2001). *Duality and auxiliary functions for Bregman distances* (Technical Report CMU-CS-01-109). Carnegie Mellon University.

Della-Pietra, S., Della-Pietra, V., Mercer, R., & Rukos, S. (1992). Adaptive language modeling using minimum discriminat estimation. *Proc. of the International Conference on Accoustics, Speech and Signal Processing*.

Dietterich, T. (2002). AI Seminar. Carnegie Mellon.

Duffy, N., & Helmbold, D. (2000). Potential boosters? *Advances in Neural Information Processing Systems*.

Efron, B. (1975). Defining the curvature of a statistical problem. *The Annals of Statistics*, *3*.

Efron, B., & Morris, C. N. (1977). Stein's paradox in statistics. *Scientific American*, *236*.

Efron, B., & Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Annals of Statistics*, *24*, 2431–2461.

Escobar, M., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*.

Ferguson, T. (1973). A bayesian analysis of some non-parametric problems. *The Annals of Statistics*, *1*.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *International Conference on Machine Learning*.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, *28*, 337–374.

Gous, A. (1998). *Exponential and spherical subfamily models*. Doctoral dissertation, Stanford University.

Grasselli, M. R. (2001). *Classical and quantum information geometry*. Doctoral dissertation, King's College, London.

Grigor'yan, A., & Noguchi, M. (1998). The heat kernel on hyperbolic space. *Bulletin of the London Mathematical Society*, *30*, 643–650.

Guo, Y., Bartlett, P. L., Shawe-Taylor, J., & Williamson, R. C. (2002). Covering numbers for support vector machines. *IEEE Transaction on Information Theory*, *48*.

Hall, K., & Hofmann, T. (2000). Learning curved multinomial subfamilies for natural language processing and information retrieval. *Proc. of the 17th International Conference on Machine Learning*.

Ikeda, S., Tanaka, T., & Amari, S. (2004). Stochastic reasoning, free energy, and information geometry. *Neural Computation*, *16*.

Jaakkola, T., & Haussler, D. (1998). Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems, 11*.

Joachims, T. (2000). *The maximum margin approach to learning text classifiers methods, theory and algorithms*. Doctoral dissertation, Dortmund University.

Joachims, T., Cristianini, N., & Shawe-Taylor, J. (2001). Composite kernels for hypertext categorisation. *Proceedings of the International Conference on Machine Learning (ICML)*.

Kass, R. E., & Voss, P. W. (1997). *Geometrical foudnation of asymptotic inference*. John Wiley & Sons, Inc.

Khinchin, A. I. (1957). *Mathematical foundations of information theory*. Dover Publications.

Kimeldorf, G., & Wahba, G. (1971). Some results on Tchebychean spline functions. *J. Math. Anal. Applic.*, *33*, 82–95.

Kivinen, J., & Warmuth, M. K. (1999). Boosting as entropy projection. *Proceedings of the 20th Annual Conference on Computational Learning Theory*.

Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. *Proceedings of the International Conference on Machine Learning (ICML)*. Morgan Kaufmann.

Kullback, S. (1968). *Information theory and statistics*. Dover publications.

Lafferty, J. (1988). The density manifold and configuration space quantization. *Transactions of the American Mathematical Society*, *305*.

Lafferty, J. (1999). Additive models, boosting, and inference for generalized divergences. *Proceedings of the 20th Annual Conference on Computational Learning Theory*.

Lafferty, J., & Lebanon, G. (2003). Information diffusion kernels. *Advances in Neural Information Processing, 15*. MIT press.

Lafferty, J., & Lebanon, G. ((accepted)). Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*.

Lanckriet, G. R. G., Bartlett, P., Cristianini, N., Ghaoui, L. E., & Jordan, M. I. (2002). Learning the kernel matrix with semidefinite programming. *International Conf. on Machine Learning*.

Lang, S. (1999). *Fundamentals of differential geometry*. Springer.

Lebanon, G. (2003a). *Computing the volume element of a family of metrics on the multinomial simplex* (Technical Report CMU-CS-03-145). Carnegie Mellon University.

Lebanon, G. (2003b). Learning Riemannian metrics. *Proc. of the 19th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers.

Lebanon, G. (2004). An extended Čencov characterization of conditional information geometry. *Proc. of the 20th Conference on Uncertainty in Artificial Intelligence*. AUAI press.

Lebanon, G. (to appear). Axiomatic geometry of conditional models. *IEEE Transactions on Information Theory*.

Lebanon, G., & Lafferty, J. (2002). Boosting and maximum likelihood for exponential models. *Advances in Neural Information Processing Systems, 14*. MIT press.

Lebanon, G., & Lafferty, J. (2004). Hyperplane margin classifiers on the multinomial manifold. *Proc. of the 21st International Conference on Machine Learning*. ACM press.

Lee, J. M. (1997). *Riemannian manifolds, an introduction to curvature*. Springer.

Lee, J. M. (2000). *Introduction to topological manifolds*. Springer.

Lee, J. M. (2002). *Introduction to smooth manifolds*. Springer.

Lewis, D. D., & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. *Symposium on Document Analysis and Information Retrieval* (pp. 81–93).

Li, P., & Yau, S.-T. (1980). Estimates of eigenvalues of a compact Riemannian manifold. *Geometry of the Laplace Operator* (pp. 205–239).

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum hellinger distance and related methods. *Annals of Statistics*, *22*, 1081–1114.

Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Functional gradient techniques for combining hypotheses. *Advances in Large Margin Classifiers*.

Mendelson, S. (2003). On the performance of kernel classes. *Journal of Machine Learning Research*, *4*, 759–771.

Milnor, J. W. (1963). *Morse theory*. Princeton University Press.

Moreno, P. J., Ho, P. P., & Vasconcelos, N. (2004). A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. *Advances in Neural Information Processing Systems 16*.

Murray, M. K., & Rice, J. W. (1993). *Differential geometry and statistics*. CRC Press.

Neal, R. (2000). Markov chain sampling methods for dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, *9*.

Nelson, E. (1968). *Tensor analysis*. Princeton University Press.

O'Sullivan, J. A. (1998). Alternating minimization algorithms: From Blahut-Arimoto to Expectation-Maximization. *Codes, Curves, and Signals: Common Threads in Communications* (pp. 173–192).

Pistone, G., & Sempi, C. (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The Annals of Statistics*, *23*.

Poggio, T., & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, *247*, 978–982.

Ponte, J., & Croft, W. B. (1998). A language modeling approach to information retrieval. *Proceedings of the ACM SIGIR* (pp. 275–281).

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, *37*.

Robbins, H. (1955). An empirical bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Prob.*

Rosenberg, S. (1997). *The Laplacian on a Riemannian manifold*. Cambridge University Press.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling. *Computer, Speech and Language*, *10*.

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw Hill.

Saul, L. K., & Jordan, M. I. (1997). A variational principle for model-based interpolation. *Advances in Neural Information Processing Systems 9*.

Schapire, R. E. (2002). The boosting approach to machine learning: An overview. *MSRI Workshop on Nonlinear Estimation and Classification*.

Schervish, M. J. (1995). *Theory of statistics*. Springer.

Schoen, R., & Yau, S. (1994). *Lectures on differential geometry*, vol. 1 of *Conference Proceedings and Lecture Notes in Geometry and Topology*. International Press.

Spivak, M. (1975). *A comprehensive introduction to differential geometry*, vol. 1-5. Publish or Perish.

Stein, C. (1955). Inadmissability of the usual estimator estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.*

Tamura, R. N., & Boos, D. D. (1986). Minimum hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, *81*, 223–229.

Xing, E. P., Ng, A. Y., Jordan, M. I., & Russel, S. (2003). Distance metric learning with applications to clustering with side information. *Advances in Neural Information Processing Systems, 15*.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. *Proceedings of SIGIR'2001* (pp. 334–342). New Orleans, LA.

Zhang, T., & Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Information Retrieval*, *4*, 5–31.