Diffusion Kernels on Statistical Manifolds

John Lafferty Guy Lebanon

January 16, 2004

CMU-CS-04-101

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

A family of kernels for statistical learning is introduced that exploits the geometric structure of statistical models. The kernels are based on the heat equation on the Riemannian manifold defined by the Fisher information metric associated with a statistical family, and generalize the Gaussian kernel of Euclidean space. As an important special case, kernels based on the geometry of multinomial families are derived, leading to kernel-based learning algorithms that apply naturally to discrete data. Bounds on covering numbers and Rademacher averages for the kernels are proved using bounds on the eigenvalues of the Laplacian on Riemannian manifolds. Experimental results are presented for document classification, for which the use of multinomial geometry is natural and well motivated, and improvements are obtained over the standard use of Gaussian or linear kernels, which have been the standard for text classification.

This research was partially supported by the Advanced Research and Development Activity in Information Technology (ARDA), contract number MDA904-00-C-2106, and by the National Science Foundation (NSF), grants CCR-0122581 and IIS-0312814.

Keywords: Kernels, heat equation, diffusion, information geometry, text classification.

1 Introduction

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 and 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 and 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 paper, 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 and 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 paper 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 and Croft, 1998, Zhai and 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 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 remaining sections are organized as follows. In Section 2 we review the relevant concepts that are required from Riemannian geometry and define the heat kernel for a general Riemannian manifold, together with its parametrix expansion. In Section 3 we define the Fisher metric associated with a statistical manifold of distributions, and examine in some detail the special cases of the multinomial and spherical normal families; the proposed use of the heat kernel or its parametrix approximation on the statistical manifold is the main contribution of the paper. Section 4 derives 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 5 describes the results of applying the multinomial diffusion kernels to text classification, and we conclude with a discussion of our results in Section 6.

2 Riemannian Geometry and the Heat Kernel

We begin by briefly reviewing some of the elementary concepts from Riemannian geometry that will be used in the construction of information diffusion kernels, since these concepts are not widely used in machine learning. We refer to Spivak (1979) for details and further background, or Milnor (1963) for an elegant and concise overview; however most introductory texts on differential geometry include this material. The basic properties of the heat kernel on a Riemannian manifold are then presented in Section 2.3. An excellent introductory account of this topic is given by Rosenberg (1997), and an authoritative reference for spectral methods in Riemannian geometry is Schoen and Yau (1994). Readers whose differential geometry is in good repair may wish to proceed directly to Section 2.3.1 or to Section 3.

2.1 Basic Definitions

An n-dimensional differentiable manifold M is a set of points that is locally equivalent to $\mathbb{R}^n$ by smooth transformations, supporting operations such as differentiation. Formally, a *differentiable manifold* is a set M together with a collection of *local charts* $\{(U_i, \varphi_i)\}$, where $U_i \subset M$ with $\cup_i U_i = M$, and $\varphi_i : U_i \subset M \longrightarrow \mathbb{R}^n$ is a bijection. For each pair of local charts (U_i, φ_i) and (U_j, φ_j), it is required that $\varphi_j(U_i \cap U_j)$ is open and $\varphi_{ij} = \varphi_i \circ \varphi_j^{-1}$ is a diffeomorphism.

The tangent space $T_pM \cong \mathbb{R}^n$ at $p \in M$ can be be thought of as directional derivatives operating on $C^\infty(M)$, the set of real valued differentiable functions $f : M \to \mathbb{R}$. Equivalently, the tangent space T_pM can be viewed in terms of an equivalence class of curves on M passing through p. Two curves $c_1 : (-\epsilon, \epsilon) \longrightarrow M$ and $c_2 : (-\epsilon, \epsilon) \longrightarrow M$ are equivalent at p in case $c_1(0) = c_2(0) = p$ and $\varphi \circ c_1$ and $\varphi \circ c_2$ are tangent at p for some local chart φ (and therefore all charts), in the sense that their derivatives at 0 exist and are equal.

In many cases of interest, the manifold M is a submanifold of a larger manifold, often $\mathbb{R}^m$, $m \geq n$. For example, the open n-dimensional simplex, defined by

$$\mathcal{P}_n \;=\; \left\{\theta \in \mathbb{R}^{n+1} : \quad \textstyle\sum_{i=1}^{n+1} \theta_i = 1,\; \theta_i > 0\right\} \tag{1}$$

is a submanifold of $\mathbb{R}^{n+1}$. In such a case, the tangent space of the submanifold T_pM is a subspace of $T_p\mathbb{R}^m$, and we may represent the tangent vectors $v \in T_pM$ in terms of the standard basis of the tangent space $T_p\mathbb{R}^m \cong \mathbb{R}^m$, $v = \sum_{i=1}^m v_i\, e_i$. The open n-simplex is a differential manifold with a single, global chart.

A *manifold with boundary* is defined similarly, except that the local charts (U, φ) satisfy $\varphi(U) \subset \mathbb{R}^{n+}$, thus mapping a patch of M to the half-space $\mathbb{R}^{n+} = \{x \in \mathbb{R}^n \mid x_n \geq 0\}$. In general, if U and V are open sets in $\mathbb{R}^{n+}$ in the topology induced from $\mathbb{R}^n$, and $f : U \longrightarrow V$ is a diffeomorphism, then f induces diffeomorphisms $\mathrm{Int}f : \mathrm{Int}U \longrightarrow \mathrm{Int}V$ and $\partial f : \partial U \longrightarrow \partial V$, where $\partial A = A \cup (\mathbb{R}^{n-1} \times \{0\})$ and $\mathrm{Int}A = A \cup \{x \in \mathbb{R}^n \mid x_n > 0\}$. Thus, it makes sense to define the *interior* $\mathrm{Int}M = \cup_U \varphi^{-1}(\mathrm{Int}(\varphi(U)))$ and *boundary* $\partial M = \cup_U \varphi^{-1}(\partial(\varphi(U)))$ of M. Since $\mathrm{Int}M$ is open it is an n-dimensional manifold without boundary, and ∂M is an $(n-1)$-dimensional manifold without boundary.

If $f : M \to N$ is a diffeomorphism of the manifold M onto the manifold N, then f induces a *push-foward mapping* f_* of the associated tangent spaces. A vector field $X \in TM$ is mapped to the push-forward $f_* X \in TN$, satisfying $(f_* X)(g) = X(g \circ f)$ for all $g \in C^\infty(N)$. Intuitively, the push-forward mapping transforms velocity vectors of curves to velocity vectors of the corresponding curves in the new manifold. Such a mapping is of use in transforming metrics, as described next.

2.2 The Geometric Laplacian

The construction of our kernels is based on the geometric Laplacian[1]. In order to define the generalization of the familiar Laplacian $\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$ on $\mathbb{R}^n$ to manifolds, one needs a notion of geometry, in particular a way of measuring lengths of tangent vectors. A *Riemannian manifold* (M, g) is a differentiable manifold M with a family of smoothly varying positive-definite inner products $g = g_p$ on T_pM for each $p \in M$. Two Riemannian manifolds (M, g) and (N, h) are *isometric* in case there is a diffeomorphism $f : M \longrightarrow N$ such that

$$g_p(X, Y) = h_{f(p)}(f_* X, f_* Y) \tag{2}$$

for every $X, Y \in T_pM$ and $p \in M$. Occasionally, hard computations on one manifold can be transformed to easier computations on an isometric manifold. Every manifold can be given a Riemannian metric. For example, every manifold can be embedded in $\mathbb{R}^m$ for some $m \geq n$ (the Whitney embedding theorem), and the Euclidean metric induces a metric on the manifold under the embedding. In fact, every Riemannian metric can be obtained in this way (the Nash embedding theorem).

In local coordinates, g can be represented as $g_p(v, w) = \sum_{i,j} g_{ij}(p)\, v_i\, w_j$ where $g(p) = [g_{ij}(p)]$ is a non-singular, symmetric and positive-definite matrix depending smoothly on p, and tangent vectors v and w are represented in local coordinates at p as $v = \sum_{i=1}^n v_i\, \partial_{i|p}$ and $w = \sum_{i=1}^n w_i\, \partial_{i|p}$. As an example, consider the open n-dimensional simplex defined in (1). A metric on $\mathbb{R}^{n+1}$ expressed by the symmetric positive-definite matrix $G = [g_{ij}] \in \mathbb{R}^{(n+1)\times(n+1)}$

[1] 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."

induces a metric on $\mathcal{P}_n$ as

$$g_p(v,u) = g_p\left(\textstyle\sum_{i=1}^{n+1} u_i e_i, \sum_{i=1}^{n+1} v_i e_i\right) = \sum_{i=1}^{n+1}\sum_{j=1}^{n+1} g_{ij}\, u_i v_j \tag{3}$$

The metric enables the definition of lengths of vectors and curves, and therefore distance between points on the manifold. The length of a tangent vector at $p \in M$ is given by $\|v\| = \sqrt{\langle v, v\rangle_p}$, $v \in T_pM$ and the length of a curve $c : [a,b] \to M$ is then given by $L(c) = \int_a^b \|\dot{c}(t)\| \mathrm{d}t$ where $\dot{c}(t)$ is the velocity vector of the path c at time t. Using the above definition of lengths of curves, we can define the distance $d(x,y)$ between two points $x, y \in M$ as the length of the shortest piecewise differentiable curve connecting x and y. This *geodesic distance* d turns the Riemannian manifold into a metric space, satisfying the usual properties of positivity, symmetry and the triangle inequality. Riemannian manifolds also support convex neighborhoods. In particular, if $p \in M$, there is an open set U containing p such that any two points of U can be connected by a unique minimal geodesic in U.

A manifold is said to be *geodesically complete* in case every geodesic curve $c(t)$, $t \in [a,b]$, can be extended to be defined for all $t \in \mathbb{R}$. It can be shown (Milnor, 1963), that the following are equivalent: (1) M is geodesically complete, (2) d is a complete metric on M, and (3) closed and bounded subsets of M are compact. In particular, compact manifolds are geodesically complete. The Hopf-Rinow theorem (Milnor, 1963) asserts that if M is complete, then any two points can be joined by a minimal geodesic. This minimal geodesic is not necessarily unique, as seen by considering antipodal points on a sphere. The *exponential map* $\exp_x$ maps a neighborhood V of $0 \in T_xM$ diffeomorphically onto a neighborhood of $x \in M$. By definition, $\exp_x v$ is the point $\gamma_v(1)$ where γ_v is a geodesic starting at x with initial velocity $v = \frac{d\gamma_v}{dt}|_{t=0}$. Any such geodesic satisfies $\gamma_{rv}(s) = \gamma_v(rs)$ for $r > 0$. This mapping defines a local coordinate system on M called *normal coordinates*, under which many computations are especially convenient.

For a function $f : M \longrightarrow \mathbb{R}$, the gradient $\operatorname{grad} f$ is the vector field defined by

$$\langle \operatorname{grad} f(p), X\rangle \;=\; X(f) \tag{4}$$

In local coordinates, the gradient is given by

$$(\operatorname{grad} f)_i = \sum_j g^{ij} \frac{\partial f}{\partial x_j} \tag{5}$$

where $\left[g^{ij}(p)\right]$ is the inverse of $[g_{ij}(p)]$. The divergence operator is defined to be the adjoint of the gradient, allowing "integration by parts" on manifolds with special structure. An orientation of a manifold is a smooth choice of orientation for the tangent spaces, meaning that for local charts φ_i and φ_j, the differential $D(\varphi_j \circ \varphi_i)(x) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is orientation preserving, so the sign of the determinant is constant. If a Riemannian manifold M is orientable, it is possible to define a *volume form* μ, where if $v_1, v_2, \ldots, v_n \in T_pM$ (positively oriented), then

$$\mu(v_1, \ldots, v_n) \;=\; \sqrt{\det\langle v_i, v_j\rangle} \tag{6}$$

A volume form, in turn, enables the definition of the *divergence* of a vector field on the manifold. In local coordinates, the divergence is given by

$$\operatorname{div} X = \frac{1}{\sqrt{\det g}} \sum_i \frac{\partial}{\partial x_i} \left(\sqrt{\det g}\, X_i \right) \tag{7}$$

Finally, the *Laplace-Beltrami operator* on functions is defined by

$$\Delta = \operatorname{div} \circ \operatorname{grad} \tag{8}$$

which in local coordinates is thus given by

$$\Delta f = \frac{1}{\sqrt{\det g}} \sum_j \frac{\partial}{\partial_{x_j}} \left(g^{ij} \sqrt{\det g}\, \frac{\partial f}{\partial x_i} \right) \tag{9}$$

These definitions preserve the familiar intuitive interpretation of the usual operators in Euclidean geometry; in particular, the gradient points in the direction of steepest ascent and the divergence measures outflow minus inflow of liquid or heat.

2.3 The Heat Kernel

The Laplacian is used to model how heat will diffuse throughout a geometric manifold; the flow is governed by the following second order differential equation with initial conditions

$$\frac{\partial f}{\partial t} - \Delta f = 0 \tag{10}$$

$$f(x,0) = f(x) \tag{11}$$

The value $f(x,t)$ describes the heat at location x at time t, beginning from an initial distribution of heat given by $f(x)$ 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

$$f(x,t) = \int_M K_t(x,y)\, f(y)\, dy \tag{12}$$

As a simple special case, consider heat flow on the circle, or one-dimensional sphere $M = S^1$. Parameterizing the manifold by angle θ, and letting $f(\theta,t) = \sum_{j=0}^{\infty} a_j(t) \cos(j\theta)$ be the discrete cosine transform of the solution to the heat equation, with initial conditions given by $a_j(0) = a_j$, it is seen that the heat equation leads to the equation

$$\sum_{j=0}^{\infty} \left(\frac{d}{dt}\, a_j(t) + j^2 a_j(t) \right) \cos(j\theta) = 0 \tag{13}$$

which is easily solved to obtain $a_j(t) = e^{-j^2 t}$ and therefore $f(\theta,t) = \sum_{j=0}^{\infty} a_j\, e^{-j^2 t} \cos(j\theta)$. As the time parameter t gets large, the solution converges to $f(\theta,t) \longrightarrow a_0$, which is the average value of f; thus, the heat diffuses until the manifold is at a uniform temperature. To express the solution in terms of an integral kernel, note that by the Fourier inversion formula

$$f(\theta, t) = \sum_{j=0}^{\infty} \langle f, e^{ij\theta} \rangle \, e^{-j^2 t} \, e^{ij\theta} \tag{14}$$

$$= \frac{1}{2\pi} \int_{S^1} \sum_{j=0}^{\infty} e^{-j^2 t} e^{ij\theta} \, e^{-ij\phi} \, f(\phi) \, d\phi \tag{15}$$

thus expressing the solution as $f(\theta, t) = \int_{S^1} K_t(\theta, \phi) \, f(\phi) \, d\phi$ for the heat kernel

$$K_t(\phi, \theta) = \frac{1}{2\pi} \sum_{j=0}^{\infty} e^{-j^2 t} \cos\left(j(\theta - \phi)\right) \tag{16}$$

This simple example shows several properties of the general solution of the heat equation on a (compact) Riemannian manifold; in particular, note that the eigenvalues of the kernel scale as $\lambda_j \sim e^{-j^{2/d}}$ where the dimension in this case is $d = 1$.

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

$$f(x, t) = \frac{1}{\sqrt{4\pi t}} \int_{\mathbb{R}} e^{-\frac{(x-y)^2}{4t}} f(y) \, dy \tag{17}$$

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

When M is compact, the Laplacian has discrete eigenvalues $0 = \mu_0 < \mu_1 \leq \mu_2 \cdots$ with corresponding eigenfunctions ϕ_i satisfying $\Delta \phi_i = -\mu_i \phi_i$. When the manifold has a boundary, appropriate boundary conditions must be imposed in order for Δ to be self-adjoint. Dirichlet boundary conditions set $\phi_i|_{\partial M} = 0$ and Neumann boundary conditions require $\left.\frac{\partial \phi_i}{\partial \nu}\right|_{\partial M} = 0$ where ν is the outer normal direction. The following theorem summarizes the basic properties for the kernel of the heat equation on M; we refer to Schoen and Yau (1994) for a proof.

Theorem 1 *Let 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. $K_t(x, y) = \int_M K_{t-s}(x, z) K_s(z, y) \, dz$ *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 $K_t(x, y) = \sum_{i=0}^{\infty} e^{-\mu_i t} \phi_i(x) \, \phi_i(y)$.

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

$$\begin{aligned}
\frac{\partial f(x,t)}{\partial t} &= \int_M \frac{\partial K_t(x,y)}{\partial t}\, f(y)\, dy && (18)\\
&= \int_M \Delta K_t(x,y)\, f(y)\, dy && (19)\\
&= \Delta \int_M K_t(x,y)\, f(y)\, dy && (20)\\
&= \Delta f(x) && (21)
\end{aligned}$$

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,

$$\begin{aligned}
\int_M \int_M K_t(x,y) f(x) f(y)\, dx\, dy &= \int_M f(x)\, e^{t\Delta} f(x)\, dx && (22)\\
&= \langle f, e^{t\Delta} f \rangle \geq 0 && (23)
\end{aligned}$$

Thus $K_t(x,y)$ is positive-definite. In the compact case, positive-definiteness follows directly from the expansion $K_t(x,y) = \sum_{i=0}^{\infty} e^{-\mu_i t} \phi_i(x)\, \phi_i(y)$, which shows that the eigenvalues of K_t as an integral operator are $e^{-\mu_i t}$. 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 $x, y \in M$, 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 5, 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 initial temperature $y_0(x) = 0$ elsewhere.

2.3.1 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 5 we will employ the first-order parametrix expansion for text classification.

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

$$K_t^{\text{Euclid}}(x,y) = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{\|x-y\|^2}{4t}\right) \tag{24}$$

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. To begin the definition of the parametrix, let

$$P_t^{(m)}(x,y) = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{d^2(x,y)}{4t}\right) \left(\psi_0(x,y) + \psi_1(x,y)t + \cdots + \psi_m(x,y)t^m\right) \tag{25}$$

for currently unspecified functions $\psi_k(x,y)$, but where $d^2(x,y)$ now denotes the square of the geodesic distance on the manifold. The idea is to obtain ψ_k recursively by solving the heat equation approximately to order t^m, for small diffusion time t.

Let $r = d(x,y)$ denote the length of the radial geodesic from x to $y \in V_x$ in the normal coordinates defined by the exponential map. For any functions $f(r)$ and $h(r)$ of r, it can be shown that

$$\Delta f = \frac{d^2 f}{dr^2} + \frac{d\left(\log\sqrt{\det g}\right)}{dr}\frac{df}{dr} \tag{26}$$

$$\Delta(fh) = f\Delta h + h\Delta f + 2\frac{df}{dr}\frac{dh}{dr} \tag{27}$$

Starting from these basic relations, some calculus shows that

$$\left(\Delta - \frac{\partial}{\partial t}\right) P_t^{(m)} = \left(t^m\,\Delta\psi_m\right)(4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{r^2}{4t}\right) \tag{28}$$

when ψ_k are defined recursively as

$$\psi_0 = \left(\frac{\sqrt{\det g}}{r^{n-1}}\right)^{-\frac{1}{2}} \tag{29}$$

$$\psi_k = r^{-k}\psi_0 \int_0^r \psi_0^{-1}\left(\Delta\phi_{k-1}\right) s^{k-1} ds \quad \text{for } k > 0 \tag{30}$$

With this recursive definition of the functions ψ_k, the expansion (25), which is defined only locally, is then extended to all of $M \times M$ by smoothing with a "cut-off function" η, with the specification that $\eta : \mathbb{R}_+ \longrightarrow [0,1]$ is C^∞ and

$$\eta(r) = \begin{cases} 0 & r \geq 1 \\ 1 & r \leq c \end{cases} \tag{31}$$

for some constant $0 < c < 1$. Thus, the order-m parametrix is defined as

$$K_t^{(m)}(x,y) = \eta(d(x,y))\, P_t^{(m)}(x,y) \tag{32}$$

As suggested by equation (28), $K_t^{(m)}$ is an approximate solution to the heat equation, and satisfies $K_t(x,y) = K_t^{(m)}(x,y) + O(t^m)$ for x and y sufficiently close; in particular, the parametrix is not unique. For further details we refer to (Schoen and 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

$f(t) = \min\operatorname{spec}(K_t^m)$, where min 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 $t \in [0, \epsilon)$. This fact will be used when we employ the parametrix approximation to the heat kernel for statistical learning.

3 Diffusion Kernels on Statistical Manifolds

We now proceed to the main contribution of the paper, which is the application of the heat kernel constructions reviewed in the previous section to the geometry of statistical families, in order to obtain kernels for statistical learning.

Under some mild regularity conditions, general parametric statistical families come equipped with a canonical geometry based on the Fisher information metric. This geometry has long been recognized (Rao, 1945), and there is a rich line of research in statistics, with threads in machine learning, that has sought to exploit this geometry in statistical analysis; see Kass (1989) for a survey and discussion, or the monographs by Kass and Vos (1997) and Amari and Nagaoka (2000) for more extensive treatments.

We remark that in spite of the fundamental nature of the geometric perspective in statistics, many researchers have concluded that while it occasionally provides an interesting alternative interpretation, it has not contributed new results or methods that cannot be obtained through more conventional analysis. However in the present work, the kernel methods we propose can, arguably, be motivated and derived only through the geometry of statistical manifolds.[2]

3.1 Geometry of Statistical Families

Let $\mathfrak{F} = \{p(\cdot \,|\, \theta)\}_{\theta\in\Theta}$ be an n-dimensional regular statistical family on a set $\mathcal{X}$. Thus, we assume that $\Theta \subset \mathbb{R}^n$ is open, and that there is a σ-finite measure μ on $\mathcal{X}$, such that for each $\theta \in \Theta$, $p(\cdot \,|\, \theta)$ is a density with respect to μ, so that $\int_{\mathcal{X}} p(x \,|\, \theta)\, d\mu(x) = 1$. We identify the manifold M with Θ by assuming that for each $x \in \mathcal{X}$ the mapping $\theta \mapsto p(x \,|\, \theta)$ is C^∞. Below, we will discuss cases where Θ is closed, leading to a manifold M with boundary.

Let ∂_i denote $\partial/\partial\theta_i$, and $\ell_\theta(x) = \log p(x \,|\, \theta)$. The *Fisher information metric at $\theta \in \Theta$* is defined in terms of the matrix $g(\theta) \in \mathbb{R}^{n\times n}$ given by

$$g_{ij}(\theta) \;=\; E_\theta\left[\partial_i \ell_\theta\, \partial_j \ell_\theta\right] \;=\; \int_{\mathcal{X}} p(x \,|\, \theta) \partial_i \log p(x \,|\, \theta)\, \partial_j \log p(x \,|\, \theta)\, d\mu(x) \qquad (33)$$

Since the score $s_i(\theta) = \partial_i \ell_\theta$ has mean zero, $g_{ij}(\theta)$ can be seen as the variance of $s_i(\theta)$, and is therefore positive-definite. By assumption, it is smoothly varying in θ, and therefore defines a Riemannian metric on $\Theta = M$.

[2]By a *statistical manifold* we mean simply a manifold of densities together with the metric induced by the Fisher information matrix, rather than the more general notion of a Riemannian manifold together with a (possibly non-metric) connection, as defined by Lauritzen (1987).

An equivalent and sometimes more suggestive form of the Fisher information matrix, as will be seen below for the case of the multinomial, is

$$g_{ij}(\theta) \;=\; 4\int_{\mathcal{X}} \partial_i \sqrt{p(x\,|\,\theta)}\,\partial_j \sqrt{p(x\,|\,\theta)}\,d\mu(x) \tag{34}$$

Yet another equivalent form is $g_{ij}(\theta) = -E_\theta[\partial_j \partial_i \ell_\theta]$. To see this, note that

$$\begin{aligned}
E_\theta[\partial_j \partial_i \ell_\theta] &= \int_{\mathcal{X}} p(x\,|\,\theta) \partial_j \partial_i \log p(x\,|\,\theta)\,d\mu(x) && (35)\\
&= -\int_{\mathcal{X}} p(x\,|\,\theta) \frac{\partial_j p(x\,|\,\theta)}{p(x\,|\,\theta)^2} \partial_i p(x\,|\,\theta)\,d\mu(x) - \int_{\mathcal{X}} \partial_j \partial_i p(x\,|\,\theta)\,d\mu(x) && (36)\\
&= -\int_{\mathcal{X}} p(x\,|\,\theta) \frac{\partial_j p(x\,|\,\theta)}{p(x\,|\,\theta)} \frac{\partial_i p(x\,|\,\theta)}{p(x\,|\,\theta)}\,d\mu(x) - \partial_j \partial_i \int_{\mathcal{X}} p(x\,|\,\theta)\,d\mu(x) && (37)\\
&= -\int_{\mathcal{X}} p(x\,|\,\theta) \partial_j \log p(x\,|\,\theta) \partial_i \log p(x\,|\,\theta)\,d\mu(x) && (38)\\
&= -g_{ij}(\theta) && (39)
\end{aligned}$$

Since there are many possible choices of metric on a given differentiable manifold, it is important to consider the motivating properties of the Fisher information metric. Intuitively, the Fisher information may be thought of as the amount of information a single data point supplies with respect to the problem of estimating the parameter θ. This interpretation can be justified in several ways, notably through the efficiency of estimators. In particular, the asymptotic variance of the maximum likelihood estimator $\hat{\theta}$ obtained using a sample of size n is $(ng(\theta))^{-1}$. Since the MLE is asymptotically unbiased, the inverse Fisher information represents the asymptotic fluctuations of the MLE around the true value. Moreover, by the Cramér-Rao lower bound, the variance of any unbiased estimator is bounded from below by $(ng(\theta))^{-1}$. Additional motivation for the Fisher information metric is provided by the results of Čencov (1982), which characterize it as the only metric (up to multiplication by a constant) that is invariant with respect to certain probabilistically meaningful transformations called congruent embeddings.

The connection with another familiar similarity measure is worth noting here. If p and q are two densities on $\mathcal{X}$ with respect to μ, the Kullback-Leibler divergence $D(p,q)$ is defined by

$$D(p,q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}\,d\mu(x) \tag{40}$$

The Kullback-Leibler divergence behaves at nearby points like the square of the information distance. More precisely, it can be shown that

$$\lim_{q \to p} \frac{d^2(p,q)}{2D(p,q)} = 1 \tag{41}$$

where the convergence is uniform as $d(p,q) \to 0$. As we comment below, this relationship may be of use in approximating information diffusion kernels for complex models.

The following two basic examples illustrate the geometry of the Fisher information metric and the associated diffusion kernel it induces on a statistical manifold. 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 report extensive experiments with the resulting kernels in Section 5.

3.2 Diffusion Kernels for Gaussian Geometry

Consider the statistical family given by $\mathfrak{F} = \{p(\cdot \mid \theta)\}_{\theta\in\Theta}$ where $\theta = (\mu, \sigma)$ and $p(\cdot \mid (\mu, \sigma)) = \mathcal{N}(\mu, \sigma I_{n-1})$, the Gaussian having mean $\mu \in \mathbb{R}^{n-1}$ and variance σI_{n-1}, with $\sigma > 0$. Thus, $\Theta = \mathbb{R}^{n-1} \times \mathbb{R}_+$.

To compute the Fisher information metric for this family, it is convenient to use the general expression given by equation (39). Let $\partial_i = \partial/\partial\mu_i$ for $i = 1 \ldots n-1$, and $\partial_n = \partial/\partial\sigma$. Then simple calculations yield, for $1 \le i, j \le n-1$

$$\begin{aligned}
g_{ij}(\theta) &= -\int_{\mathbb{R}^{n-1}} \partial_i\partial_j \left(-\sum_{k=1}^{n-1} \frac{(x_k-\mu_k)^2}{2\sigma^2}\right) p(x \mid \theta)\, dx && (42)\\
&= \frac{1}{\sigma^2}\,\delta_{ij} && (43)\\
g_{ni}(\theta) &= -\int_{\mathbb{R}^{n-1}} \partial_n\partial_i \left(-\sum_{k=1}^{n-1} \frac{(x_k-\mu_k)^2}{2\sigma^2}\right) p(x \mid \theta)\, dx && (44)\\
&= \frac{2}{\sigma^3}\int_{\mathbb{R}^{n-1}} (x_i-\mu_i)\, p(x \mid \theta)\, dx && (45)\\
&= 0 && (46)\\
g_{nn}(\theta) &= -\int_{\mathbb{R}^{n-1}} \partial_n\partial_n \left(-\sum_{k=1}^{n-1} \frac{(x_k-\mu_k)^2}{2\sigma^2} - (n-1)\log\sigma\right) p(x \mid \theta)\, dx && (47)\\
&= \frac{3}{\sigma^4}\int_{\mathbb{R}^{n-1}} \sum_{k=1}^{n-1}(x_k-\mu_k)^2\, p(x \mid \theta)\, dx - \frac{n-1}{\sigma^2} && (48)\\
&= \frac{2(n-1)}{\sigma^2} && (49)
\end{aligned}$$

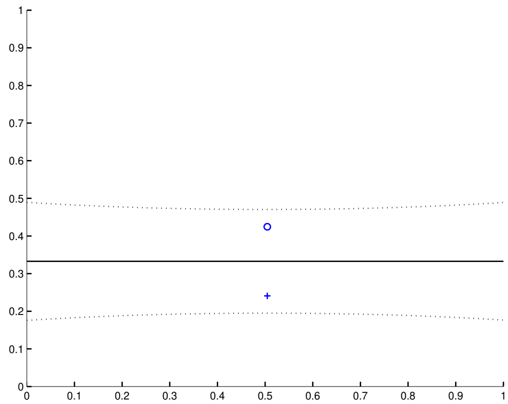
Letting θ' be new coordinates defined by $\theta'_i = \mu_i$ for $1 \le i \le n-1$ and $\theta'_n = \sqrt{2(n-1)}\,\sigma$, we see that the Fisher information matrix is given by

$$g_{ij}(\theta') = \frac{1}{\sigma^2}\,\delta_{ij} \tag{50}$$

Thus, the Fisher information metric gives $\Theta = \mathbb{R}^{n-1} \times \mathbb{R}_+$ the structure of the upper half plane in hyperbolic space. The distance minimizing or geodesic curves in hyperbolic space are straight lines or circles orthogonal to the mean subspace.

In particular, the univariate normal density has hyperbolic geometry. As a generalization in this 2-dimensional case, any location-scale family of densities is seen to have hyperbolic geometry (Kass and Vos, 1997). Such families have densities of the form

$$p(x \mid (\mu, \sigma)) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \tag{51}$$



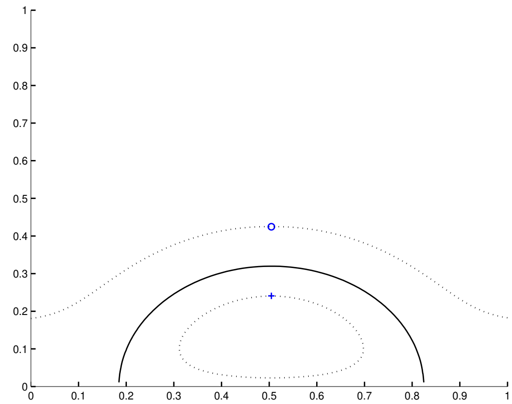


Figure 1: Example decision boundaries for a kernel-based classifier using information diffusion kernels for spherical normal geometry with $d = 2$ (right), which has constant negative curvature, compared with the standard Gaussian kernel for flat Euclidean space (left). Two data points are used, simply 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.

where $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ and $f : \mathbb{R} \to \mathbb{R}$.

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

$$K_t(x, x') = \frac{(-1)^m}{2^m \pi^m} \frac{1}{\sqrt{4\pi t}} \left(\frac{1}{\sinh r} \frac{\partial}{\partial r} \right)^m \exp\left(-m^2 t - \frac{r^2}{4t} \right) \tag{52}$$

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

$$K_t(x, x') = \frac{(-1)^m}{2^m \pi^m} \frac{\sqrt{2}}{\sqrt{4\pi t}^3} \left(\frac{1}{\sinh r} \frac{\partial}{\partial r} \right)^m \int_r^\infty \frac{s \exp\left(-\frac{(2m+1)^2 t}{4} - \frac{s^2}{4t} \right)}{\sqrt{\cosh s - \cosh r}} \, \mathrm{d}s \tag{53}$$

where $r = d(x, x')$ is the geodesic distance between the two points in $\mathbb{H}^n$. If only the mean $\theta = \mu$ is unspecified, then the associated kernel is the standard Gaussian RBF kernel.

A possible use for this kernel in statistical learning is where data points are naturally represented as sets. That is, suppose that each data point is of the form $x = \{x_1, x_2, \ldots x_m\}$ where $x_i \in \mathbb{R}^{n-1}$. Then the data can be represented according to the mapping which sends each group of points to the corresponding Gaussian under the MLE: $x \mapsto (\widehat{\mu}(x), \widehat{\sigma}(x))$ where $\widehat{\mu}(x) = \frac{1}{m} \sum_i x_i$ and $\widehat{\sigma}(x)^2 = \frac{1}{m} \sum_i (x_i - \widehat{\mu}(x))^2$.

In Figure 3.2 the diffusion kernel for hyperbolic space $\mathbb{H}^2$ is compared with the Euclidean space Gaussian kernel. The curved decision boundary for the diffusion kernel makes intuitive sense, since as the variance decreases the mean is known with increasing certainty.

Note that we can, in fact, consider M as a manifold with boundary by allowing $\sigma \geq 0$ to be non-negative rather than strictly positive $\sigma > 0$. In this case, the densities on the boundary become singular, as point masses at the mean; the boundary is simply given by $\partial M \cong \mathbb{R}^{n-1}$, which is a manifold without boundary, as required.

3.3 Diffusion Kernels for Multinomial Geometry

We now consider the statistical family of the multinomial over $n+1$ outcomes, given by $\mathfrak{F} = \{p(\cdot \,|\, \theta)\}_{\theta \in \Theta}$ where $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ with $\theta_i \in (0,1)$ and $\sum_{i=1}^n \theta_i < 1$. The parameter space Θ is the open n-simplex $\mathcal{P}_n$ defined in equation (1), a submanifold of $\mathbb{R}^{n+1}$.

To compute the metric, let $x = (x_1, x_2, \ldots, x_{n+1})$ denote one draw from the multinomial, so that $x_i \in \{0,1\}$ and $\sum_i x_i = 1$. The log-likelihood and its derivatives are then given by

$$\log p(x \,|\, \theta) = \sum_{i=1}^{n+1} x_i \log \theta_i \tag{54}$$

$$\frac{\partial \log p(x \,|\, \theta)}{\partial \theta_i} = \frac{x_i}{\theta_i} \tag{55}$$

$$\frac{\partial^2 \log p(x \,|\, \theta)}{\partial \theta_i \partial \theta_j} = -\frac{x_i}{\theta_i^2}\, \delta_{ij} \tag{56}$$

Since $\mathcal{P}_n$ is an n-dimensional submanifold of $\mathbb{R}^{n+1}$, we can express $u, v \in T_\theta M$ as $(n+1)$-dimensional vectors in $T_\theta \mathbb{R}^{n+1} \cong \mathbb{R}^{n+1}$; thus, $u = \sum_{i=1}^{n+1} u_i e_i$, $v = \sum_{i=1}^{n+1} v_i e_i$. Note that due to the constraint $\sum_{i=1}^{n+1} \theta_i = 1$, the sum of the $n+1$ components of a tangent vector must be zero. A basis for $T_\theta M$ is

$$\left\{ e_1 = (1, 0, \ldots, 0, -1)^\top, e_2 = (0, 1, 0, \ldots, 0, -1)^\top, \ldots, e_n = (0, 0, \ldots, 0, 1, -1)^\top \right\} \tag{57}$$

Using the definition of the Fisher information metric in equation (35) we then compute

$$\langle u, v \rangle_\theta = -\sum_{i=1}^{n+1} \sum_{j=1}^{n+1} u_i v_j E_\theta \left[\frac{\partial^2 \log p(x \,|\, \theta)}{\partial \theta_i \partial \theta_j} \right] \tag{58}$$

$$= -\sum_{i=1}^{n+1} u_i v_i E \left\{ -x_i / \theta_i^2 \right\} \tag{59}$$

$$= \sum_{i=1}^{n+1} \frac{u_i v_i}{\theta_i} \tag{60}$$

While geodesic distances are difficult to compute in general, in the case of the multinomial information geometry 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. This relationship is suggested by the form of the Fisher information given in equation (34).

To be concrete, 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 $\mathcal{P}_n$ onto the positive portion of the n-sphere of radius 2; denote this portion of the sphere as $\mathcal{S}_n^+ = \left\{ \theta \in \mathbb{R}^{n+1} \;:\; \sum_{i=1}^{n+1} \theta_i^2 = 2,\ \theta_i > 0 \right\}$. Given tangent vectors $u = \sum_{i=1}^{n+1} u_i e_i$, $v = \sum_{i=1}^{n+1} v_i e_i$, the pull-back of the Fisher information metric through

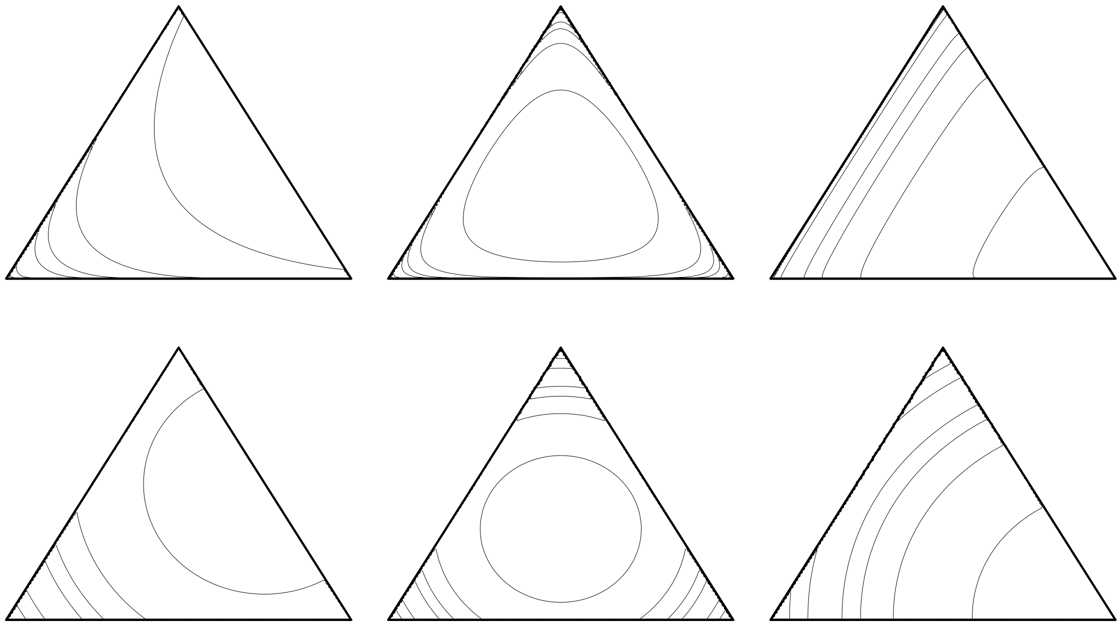


Figure 2: Equal distance contours on $\mathcal{P}_2$ from the upper right edge (left column), the center (center column), and lower right corner (right column). The distances are computed using the Fisher information metric g (top row) or the Euclidean metric (bottom row).

F^{-1} is

$$\begin{aligned}
h_\theta(u,v) &= g_{\theta^2/4}\left(F_*^{-1}\sum_{k=1}^{n+1} u_k e_k, F_*^{-1}\sum_{l=1}^{n+1} v_l e_l\right) && (61)\\
&= \sum_{k=1}^{n+1}\sum_{l=1}^{n+1} u_k v_l \, g_{\theta^2/4}(F_*^{-1}e_k, F_*^{-1}e_l) && (62)\\
&= \sum_{k=1}^{n+1}\sum_{l=1}^{n+1} u_k v_l \sum_i \frac{4}{\theta_i^2}\,(F_*^{-1}e_k)_i\,(F_*^{-1}e_l)_i && (63)\\
&= \sum_{k=1}^{n+1}\sum_{l=1}^{n+1} u_k v_l \sum_i \frac{4}{\theta_i^2}\,\frac{\theta_k\delta_{ki}}{2}\,\frac{\theta_l\delta_{li}}{2} && (64)\\
&= \sum_{i=1}^{n+1} u_i v_i && (65)
\end{aligned}$$

Since the transformation $F : (\mathcal{P}_n, g) \to (\mathcal{S}_n^+, h)$ is an isometry, the geodesic distance $d(\theta,\theta')$ on $\mathcal{P}_n$ may be computed as the shortest curve on $\mathcal{S}_n^+$ connecting $F(\theta)$ and $F(\theta')$. These shortest curves are portions of great circles—the intersection of a two dimensional plane and $\mathcal{S}_n^+$—and their length is given by

$$d(\theta,\theta') = 2\arccos\left(\sum_{i=1}^{n+1}\sqrt{\theta_i\,\theta_i'}\right) \tag{66}$$

In Section 3.1 we noted the connection between the Kullback-Leibler divergence and the information distance. In the case of the multinomial family, there is also a close relationship with the Hellinger distance. In particular, it can be easily shown that the Hellinger distance

$$d_H(\theta,\theta') = \sqrt{\sum_i\left(\sqrt{\theta_i}-\sqrt{\theta_i'}\right)^2} \tag{67}$$

is related to $d(\theta,\theta')$ by

$$d_H(\theta,\theta') = 2\sin\big(d(\theta,\theta')/4\big) \tag{68}$$

Thus, as $\theta'\to\theta$, d_H agrees with $\frac{1}{2}d$ to second order:

$$d_H(\theta,\theta') = \frac{1}{2}d(\theta,\theta') + O(d^3(\theta,\theta')) \tag{69}$$

The Fisher information metric places greater emphasis on points near the boundary, which is expected to be important for text problems, which typically have sparse statistics. Figure 2 shows equal distance contours on $\mathcal{P}_2$ using the Fisher information and the Euclidean metrics.

While the spherical geometry has been derived for a finite multinomial, the same geometry can be used non-parametrically for an arbitrary subset of probability measures, leading to spherical geometry in a Hilbert space (Dawid, 1977).

3.3.1 The Multinomial Diffusion Kernel

Unlike the explicit expression for the Gaussian geometry discussed above, there is not an 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.

Recall from Section 2.3.1 that the parametrix is obtained according to the local expansion given in equation (25), and then extending this smoothly to zero outside a neighborhood of the diagonal, as defined by the exponential map. As we have just derived, this results in the following parametrix for the multinomial family:

$$P_t^{(m)}(\theta,\theta') = (4\pi t)^{-\frac{n}{2}}\exp\left(-\frac{\arccos^2(\sqrt{\theta}\cdot\sqrt{\theta'})}{t}\right)\big(\psi_0(\theta,\theta')+\cdots+\psi_m(\theta,\theta')t^m\big) \tag{70}$$

The first-order expansion is thus obtained as

$$K_t^{(0)}(\theta,\theta') = \eta(d(\theta,\theta'))\,P_t^{(0)}(\theta,\theta') \tag{71}$$

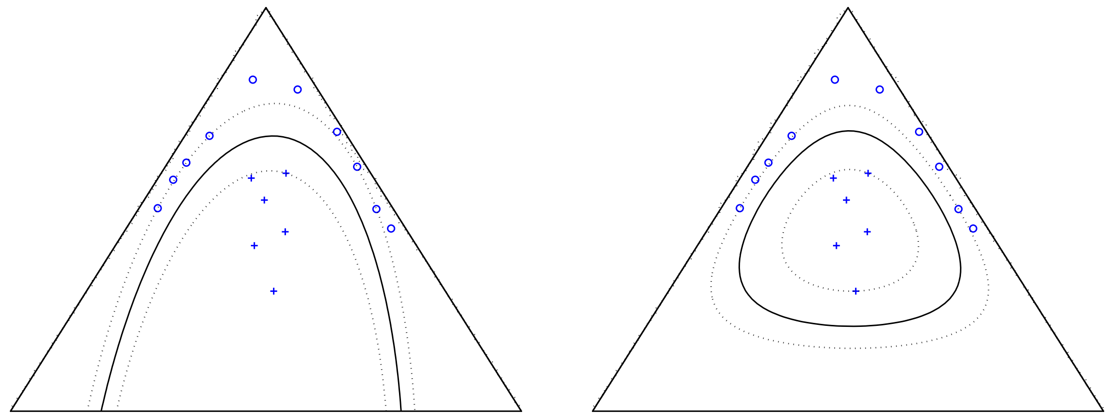


Figure 3: Example decision boundaries using support vector machines with information diffusion kernels for trinomial geometry on the 2-simplex (top right) compared with the standard Gaussian kernel (left).

Now, for the n-sphere it can be shown that the function ψ_0 of (29), which is the leading order correction of the Gaussian kernel under the Fisher information metric, is given by

$$\psi_0(r) = \left(\frac{\sqrt{\det g}}{r^{n-1}}\right)^{-\frac{1}{2}} \tag{72}$$

$$= \left(\frac{\sin r}{r}\right)^{-\frac{(n-1)}{2}} \tag{73}$$

$$= 1 + \frac{(n-1)}{12} r^2 + \frac{(n-1)(5n-1)}{1440} r^4 + O(r^6) \tag{74}$$

(Berger et al., 1971). Thus, the leading order parametrix for the multinomial diffusion kernel is

$$P_t^{(0)}(\theta, \theta') = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{1}{4t} d^2(\theta, \theta')\right) \left(\frac{\sin d(\theta, \theta')}{d(\theta, \theta')}\right)^{-\frac{(n-1)}{2}} \tag{75}$$

In our experiments we approximate this kernel further as

$$P_t^{(0)}(\theta, \theta') = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{1}{t} \arccos^2(\sqrt{\theta} \cdot \sqrt{\theta'})\right) \tag{76}$$

by appealing to the asymptotic expansion in (74); note that $(\sin r/r)^{-n}$ blows up for large r. In Figure 3 the kernel (76) is compared with the standard Euclidean space Gaussian kernel for the case of the trinomial model, $d = 2$, using an SVM classifier.

3.3.2 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

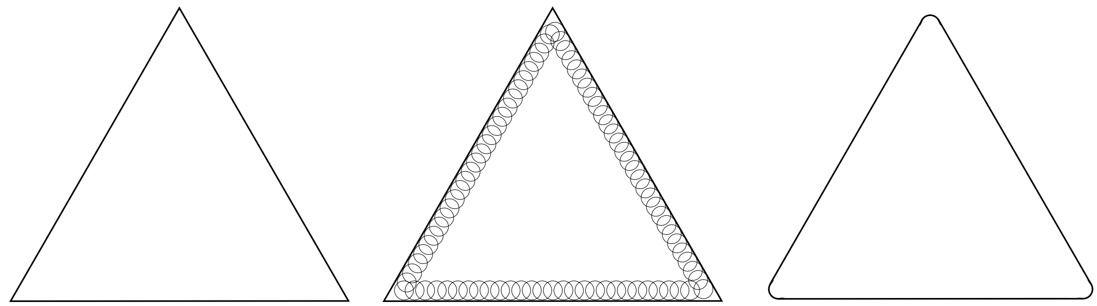


Figure 4: Rounding the simplex. Since the closed simplex is not a manifold with boundary, we carry out a "rounding" procedure to remove edges and corners. The ϵ-rounded simplex is the closure of the union of all ϵ-balls lying within the open simplex.

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 4. 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.

Let $\Delta_n = \overline{\mathcal{P}}_n$ denote the closure of the open simplex; thus Δ_n is the usual probability simplex which allows zero probability for some items. However, it does not form a compact manifold with boundary since the boundary has edges and corners. In other words, local charts $\varphi : U \to \mathbb{R}^{n+}$ cannot be defined to be differentiable. To adjust for this, the idea is to "round the edges" of Δ_n to obtain a subset that forms a compact manifold with boundary, and that closely approximates the original simplex.

For $\epsilon > 0$, let $B_\epsilon(x) = \{y \mid \|x - y\| < \epsilon\}$ denote the open Euclidean ball of radius ϵ centered at x. Denote by $C_\epsilon(\mathcal{P}_n)$ the *ϵ-ball centers* of $\mathcal{P}_n$, the points of the simplex whose ϵ-balls lie completely within the simplex:

$$C_\epsilon(\mathcal{P}_n) = \{x \in \mathcal{P}_n \ : \ B_\epsilon(x) \subset \mathcal{P}_n\} \tag{77}$$

Finally, let $\mathcal{P}_n^\epsilon$ denote the *ϵ-interior* of $\mathcal{P}_n$, which we define as the union of all ϵ-balls contained in $\mathcal{P}_n$:

$$\mathcal{P}_n^\epsilon = \bigcup_{x \in C_\epsilon(\mathcal{P}_n)} B_\epsilon(x) \tag{78}$$

The *ϵ-rounded simplex* Δ_n^ϵ is then defined as the closure $\Delta_n^\epsilon = \overline{\mathcal{P}_n^\epsilon}$.

The rounding procedure that yields Δ_2^ϵ is suggested by Figure 4. Note that in general the ϵ-rounded simplex Δ_n^ϵ will contain points with a single, but not more than one component having zero probability. The set Δ_n^ϵ forms a compact manifold with boundary, and its image under the isometry $F : (\mathcal{P}_n, g) \to (\mathcal{S}_n^+, h)$ is a compact submanifold with boundary of the n-sphere.

Whenever appealing to results for compact manifolds with boundary in the following, it will be tacitly assumed that the above rounding procedure has been carried out in the case of the multinomial.

4 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, diffusion kernels behave essentially the same as the standard Gaussian kernel.

4.1 Covering Numbers

We begin by recalling the main result of Guo et al. (2002), modifying their notation slightly to conform with ours. Let $M \subset \mathbb{R}^d$ be a compact subset of d-dimensional Euclidean space, and suppose that $K : M \times M \longrightarrow \mathbb{R}$ is a Mercer kernel. Denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ the eigenvalues of K, i.e., of the mapping $f \mapsto \int_M K(\cdot, y)\, f(y)\, dy$, 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 $x_i \in M$, 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

$$\mathcal{F}_R(\boldsymbol{x}) = \{f : f(x_i) = \langle w, \Phi(x_i)\rangle \text{ for some } \|w\| \leq R\} \tag{79}$$

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

$$\lambda_{j+1} < \left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right)^{\frac{1}{j}} \tag{80}$$

and define

$$\epsilon_n^* = 6 C_K R \sqrt{j_n^* \left(\frac{\lambda_1 \cdots \lambda_{j_n^*}}{n^2}\right)^{\frac{1}{j_n^*}} + \sum_{i=j_n^*}^{\infty} \lambda_i} \tag{81}$$

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 and Yau (1980).

Theorem 3 *Let M be a compact Riemannian manifold of dimension d with non-negative Ricci curvature, and let $0 < \mu_1 \le \mu_2 \le \cdots$ denote the eigenvalues of the Laplacian with Dirichlet boundary conditions. Then*

$$c_1(d)\left(\frac{j}{V}\right)^{\frac{2}{d}} \le \mu_j \le c_2(d)\left(\frac{j+1}{V}\right)^{\frac{2}{d}} \tag{82}$$

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 4 *Let M be a compact Riemannian manifold, with volume V, satisfying the conditions of Theorem 3. Then the covering numbers for the Dirichlet heat kernel K_t on M satisfy*

$$\log \mathcal{N}(\epsilon, \mathcal{F}_R(\boldsymbol{x})) = O\left(\left(\frac{V}{t^{\frac{d}{2}}}\right)\log^{\frac{d+2}{2}}\left(\frac{1}{\epsilon}\right)\right) \tag{83}$$

Proof By the lower bound in Theorem 3, the Dirichlet eigenvalues of the heat kernel $K_t(x,y)$, which are given by $\lambda_j = e^{-t\mu_j}$, satisfy $\log\lambda_j \le -tc_1(d)\left(\frac{j}{V}\right)^{\frac{2}{d}}$. Thus,

$$-\frac{1}{j}\log\left(\frac{\lambda_1\cdots\lambda_j}{n^2}\right) \ge \frac{tc_1}{j}\sum_{i=1}^{j}\left(\frac{i}{V}\right)^{\frac{2}{d}} + \frac{2}{j}\log n \ge tc_1\frac{d}{d+2}\left(\frac{j}{V}\right)^{\frac{2}{d}} + \frac{2}{j}\log n \tag{84}$$

where the second inequality comes from $\sum_{i=1}^{j} i^p \ge \int_0^j x^p\,dx = \frac{j^{p+1}}{p+1}$. Now using the upper bound of Theorem 3, the inequality $j_n^* \le j$ will hold if

$$tc_2\left(\frac{j+2}{V}\right)^{\frac{2}{d}} \ge -\log\lambda_{j+1} \ge tc_1\frac{d}{d+2}\left(\frac{j}{V}\right)^{\frac{2}{d}} + \frac{2}{j}\log n \tag{85}$$

or equivalently

$$\frac{tc_2}{V^{\frac{2}{d}}}\left(j(j+2)^{\frac{2}{d}} - \frac{c_1}{c_2}\frac{d}{d+2}j^{\frac{d+2}{d}}\right) \ge 2\log n \tag{86}$$

The above inequality will hold in case

$$j \ge \left\lceil\left(\frac{2V^{\frac{2}{d}}}{t(c_2 - c_1\frac{d}{d+2})}\log n\right)^{\frac{d}{d+2}}\right\rceil \ge \left\lceil\left(\frac{V^{\frac{2}{d}}(d+2)}{tc_1}\log n\right)^{\frac{d}{d+2}}\right\rceil \tag{87}$$

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

$$\sum_{i=j_n^*}^{\infty} e^{-i^{\frac{2}{d}}} = O\left(e^{-j_n^{*\frac{2}{d}}}\right) \tag{88}$$

we have after some algebra that

$$\log\left(\frac{1}{\epsilon_n}\right) = \Omega\left(\left(\frac{t}{V^{\frac{2}{d}}}\right)^{\frac{d}{d+2}} \log^{\frac{2}{d+2}} n\right) \tag{89}$$

Inverting the above expression in $\log n$ gives equation (83). ■

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.

4.2 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

$$C(n) = \sqrt{j_n^* \left(\frac{\lambda_1 \cdots \lambda_{j_n^*}}{n^2}\right)^{\frac{1}{j_n^*}} + \sum_{i=j_n^*}^{\infty} \lambda_i} \tag{90}$$

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

$$C(r) = \sqrt{j_r^* \, r + \sum_{i=j_r^*}^{\infty} \lambda_i} \tag{91}$$

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 $\mathfrak{F}$ of measurable functions $f : \mathcal{X} \to \mathcal{Y}$, the objective is to minimize the expected loss $E[\ell(f(X),Y)]$. Let $E\ell_{f^*} = \inf_{f\in\mathfrak{F}} E\ell_f$, where $\ell_f(X,Y) = \ell(f(X),Y)$, and let $\hat{f}$ be any member of $\mathfrak{F}$ for which $E_n\ell_{\hat{f}} = \inf_{f\in\mathfrak{F}} E_n\ell_f$ 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 5 *Let $\mathfrak{F}$ be a convex class of functions and define ψ by*

$$\psi(r) = a\, ER_n \left\{f \in \mathfrak{F} : E(f - f^*)^2 \leq r\right\} + \frac{b\,x}{n} \tag{92}$$

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

$$E\left(\ell_{\hat{f}} - \ell_{f^*}\right) \leq c\, r + \frac{d\,x}{n} \tag{93}$$

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

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

$$\psi(r) \leq a\sqrt{\frac{2}{n}\sum_{j=1}^{\infty} \min\{r, \lambda_j\}} + \frac{bx}{n} \tag{94}$$

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

$$\begin{aligned} \widetilde{\psi}(r) &= \sqrt{\sum_{j=1}^{\infty} \min\{r, \lambda_j\}} \qquad (95)\\ &= \sqrt{j_r^*\, r + \sum_{i=j_r^*}^{\infty} \lambda_i} \qquad (96) \end{aligned}$$

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

Theorem 6 *Let M be a compact Riemannian manifold, satisfying the conditions of Theorem 3. Then the Rademacher term $\widetilde{\psi}$ for the Dirichlet heat kernel K_t on M satisfies*

$$\widetilde{\psi}(r) \leq C\sqrt{\left(\frac{r}{t^{\frac{d}{2}}}\right) \log^{\frac{d}{2}}\left(\frac{1}{r}\right)} \tag{97}$$

for some constant C depending on the geometry of M.

Proof We have that

$$\widetilde{\psi}^2(r) = \sum_{j=1}^{\infty} \min\{r, \lambda_j\} \tag{98}$$

$$= j_r^* \, r + \sum_{j=j_r^*}^{\infty} e^{-t\mu_j} \tag{99}$$

$$\leq j_r^* \, r + \sum_{j=j_r^*}^{\infty} e^{-tc_1 j^{\frac{2}{d}}} \tag{100}$$

$$\leq j_r^* \, r + Ce^{-tc_1 j_r^{*\frac{2}{d}}} \tag{101}$$

for some constant C, where the first inequality follows from the lower bound in Theorem 3. But $j_r^* \leq j$ in case $\log \lambda_{j+1} > r$, or, again from Theorem 3, if

$$t\, c_2 (j+1)^{\frac{2}{d}} \leq -\log \lambda_j < \log \frac{1}{r} \tag{102}$$

or equivalently,

$$j_r^* \leq \frac{C'}{t^{\frac{d}{2}}} \log^{\frac{d}{2}} \left(\frac{1}{r}\right) \tag{103}$$

It follows that

$$\widetilde{\psi}^2(r) \leq C'' \left(\frac{r}{t^{\frac{d}{2}}}\right) \log^{\frac{d}{2}} \left(\frac{1}{r}\right) \tag{104}$$

for some new constant C''. ■

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

$$E\left(\ell_{\hat{f}} - \ell_{f^*}\right) = O\left(\frac{\log^{\frac{d}{2}} n}{n}\right) \tag{105}$$

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 S^1. While the bounds for 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.

5 Multinomial Diffusion Kernels and 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.

5.1 Representing Documents

Assuming a vocabulary V of size $n+1$, a document may be represented as a sequence of words over the alphabet V. For many classification tasks it is not unreasonable to discard word order; indeed, humans can typically easily understand the high level topic of a document by inspecting its contents as a mixed up "bag of words." Let x_v denote the number of times term v appears in a document. Then $\{x_v\}_{v\in V}$ is the sample space of the multinomial distribution, with a document modeled as independent draws from a fixed model, which may change from document to document. It is natural to embed documents in the multinomial simplex using an embedding function $\widehat{\theta} : \mathbb{Z}_+^{n+1} \to \mathcal{P}_n$. We consider several embeddings $\widehat{\theta}$ that correspond to well known feature representations in text classification (Joachims, 2000). The *term frequency* (tf) representation uses normalized counts; the corresponding embedding is the maximum likelihood estimator for the multinomial distribution

$$\widehat{\theta}_{\mathrm{tf}}(x) = \left(\frac{x_1}{\sum_i x_i}, \ldots, \frac{x_{n+1}}{\sum_i x_i}\right). \tag{106}$$

Another common representation is based on *term frequency, inverse document frequency* (tfidf). 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

$$\widehat{\theta}_{\mathrm{tfidf}}(x) = \left(\frac{x_1 \log(D/df_1)}{\sum_i x_i \log(D/df_i)}, \ldots, \frac{x_{n+1} \log(D/df_{n+1})}{\sum_i x_i \log(D/df_i)}\right) \tag{107}$$

where D is the number of documents in the corpus.

We note that in text classification applications the tf and tfidf 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

$$x \mapsto \left(\frac{x_1}{\sum_i x_i^2}, \ldots, \frac{x_{n+1}}{\sum_i x_i^2}\right) \tag{108}$$

and similarly for tfidf. When used in support vector machines with linear or Gaussian kernels, L_2-normalized tf and tfidf 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.

To be clear, we list the three kernels we compare. First, the linear kernel is given by

$$K^{\text{Lin}}(\theta, \theta') = \theta \cdot \theta' = \sum_{v=1}^{n+1} \theta_v \, \theta'_v \tag{109}$$

The Gaussian kernel is given by

$$K_\sigma^{\text{Gauss}}(\theta', \theta') \;=\; (2\pi\sigma)^{-\frac{n+1}{2}} \exp\left(-\frac{\|\theta - \theta'\|^2}{2\sigma^2}\right) \tag{110}$$

where $\|\theta - \theta'\|^2 \;=\; \sum_{v=1}^{n+1} |\theta_v - \theta'_v|^2$ is the squared Euclidean distance. The multinomial diffusion kernel is given by

$$K_t^{\text{Mult}}(\theta, \theta') = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{1}{t} \arccos^2(\sqrt{\theta} \cdot \sqrt{\theta'})\right) \tag{111}$$

as derived in Section 3.

5.2 Experimental Results

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

Figure 5 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 1. Similarly, Figure 7 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 3.

The WebKb dataset contains web pages found on the sites of four universities (Craven et al., 2000). 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

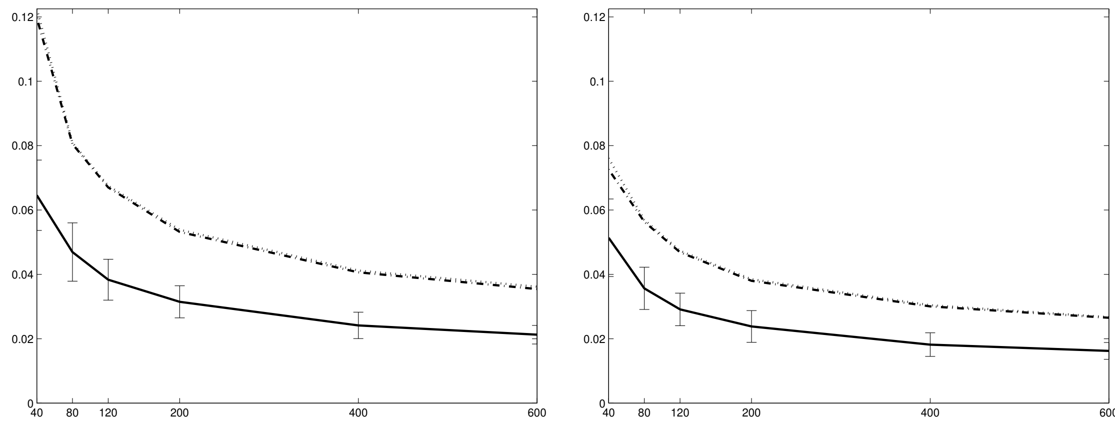


Figure 5: Experimental results on the WebKB corpus, using SVMs for linear (dotted) and Gaussian (dash-dotted) kernels, compared with the diffusion kernel for the multinomial (solid). Classification error for the task of labeling course vs. either faculty, project, or student is shown in these plots, as a function of training set size. The left plot uses tf representation and the right plot uses tfidf representation. The curves shown are the error rates averaged over 20-fold cross validation, with error bars representing one standard deviation. The results for the other "1 vs. all" labeling tasks are qualitatively similar, and are therefore not shown.

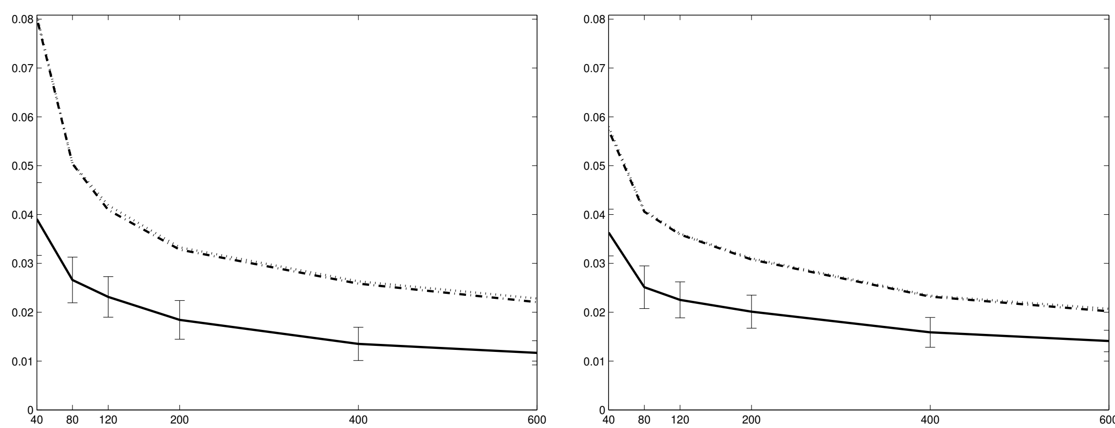


Figure 6: Results on the WebKB corpus, using SVMs for linear (dotted) and Gaussian (dash-dotted) kernels, compared with the diffusion kernel (solid). The course pages are labeled positive and the student pages are labeled negative; results for other label pairs are qualitatively similar. The left plot uses tf representation and the right plot uses tfidf representation.

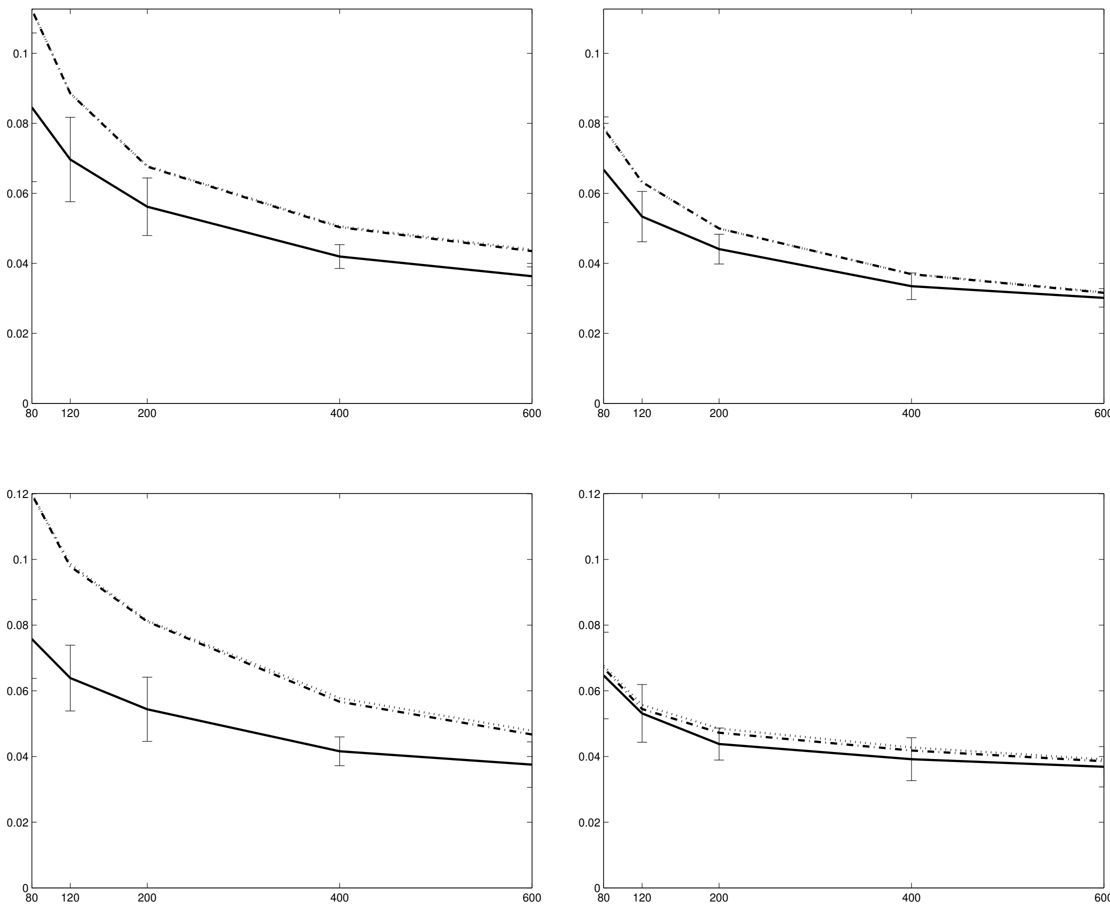


Figure 7: Experimental results on the Reuters corpus, using SVMs for linear (dotted) and Gaussian (dash-dotted) kernels, compared with the diffusion kernel (solid). The classes acq (top), and moneyFx (bottom) are shown; the other classes are qualitatively similar. The left column uses tf representation and the right column uses tfidf. The curves shown are the error rates averaged over 20-fold cross validation, with error bars representing one standard deviation.

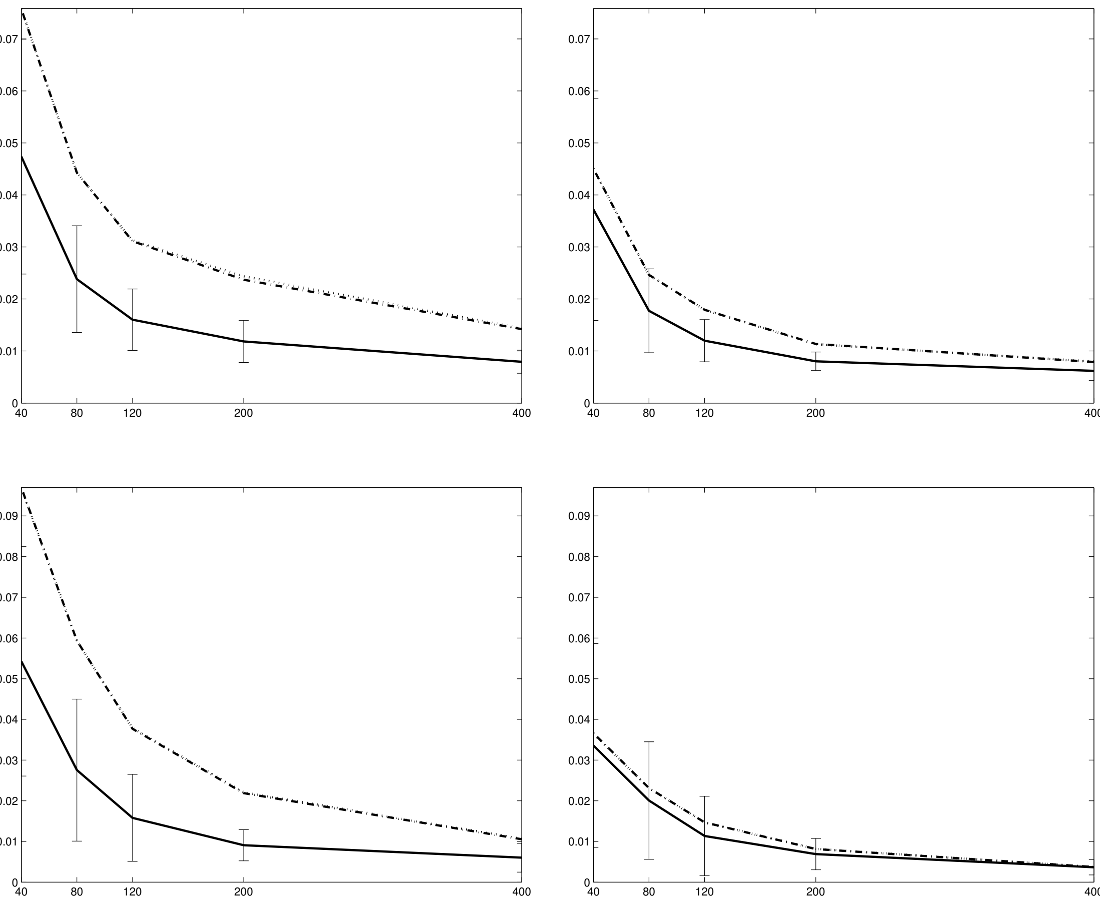


Figure 8: Experimental results on the Reuters corpus, using SVMs for linear (dotted) and Gaussian (dash-dotted) kernels, compared with the diffusion (solid). The classes moneyFx (top) and grain (bottom) are labeled as positive, and the class earn is labeled negative. The left column uses tf representation and the right column uses tfidf representation.

		tf Representation			tfidf 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 1: 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 tfidf 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.

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 6 and Figure 8 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 2 and 4.

		tf Representation			tfidf 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 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 tfidf 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.

In these figures, the leftmost plots show the performance of tf features while the rightmost plots show the performance of tfidf 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 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 5 the test set accuracies for the tf representation. For the tfidf 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

		tf Representation			tfidf 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 3: 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 tfidf 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.

with all kernels used, and the absence of cross validation data makes the results too noisy for interpretation.

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 tfidf representation, the diffusion kernel consistently

		tf Representation			tfidf 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 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 tfidf 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.

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, which are the basis for the inverse document frequency weighting in the tfidf 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.

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

6 Discussion and Conclusion

This paper has 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. A related use of spherical geometry for the multinomial has been developed by Gous (1998). 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 5 are notable since accuracies better or comparable to those obtained using heuristic weighting schemes such as tfidf are achieved directly through the geometric 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 paper differ significantly from the Fisher kernel of Jaakkola and Haussler (1998). In particular, the latter is based on the score $\nabla_\theta \log p(X \mid \hat{\theta})$ 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 reparameterization 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 (52–53) 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 distance between nearby points in terms of the Kullback-Leibler divergence, using the relation with the Fisher information that was noted in Section 3.1. 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.

Acknowledgments

We thank Rob Kass, Leonid Kontorovich and Jian Zhang for helpful discussions. This research was supported in part by NSF grants CCR-0122581 and IIS-0312814, and by ARDA contract MDA904-00-C-2106.

References

Mark A. Aizerman, Emmanuel M. Braverman, and Lev I. Rozonoér. Theoretical foundations of the potential function method in pattern recognition and learning. *Automation and Remote Control*, 25:821–837, 1964.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. Manuscript, 2003.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Mikhail Belkin and Partha Niyogi. Using manifold structure for partially labeled classification. In *Advances in Neural Information Processing Systems*, 2003.

Marcel Berger, Paul Gauduchon, and Edmond Mazet. Le spectre d'une varieté Riemannienne. *Lecture Notes in Mathematics*, Vol. 194, Springer-Verlag, 1971.

Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *Computational Learing Theory*, pages 144–152, 1992.

Nikolai Nikolaevich Čencov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translation of Mathematical Monographs*. American Mathematical Society, 1982.

Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew K. McCallum, Tom M. Mitchell, Kamal Nigam, and Seán Slattery. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, 118(1/2):69–113, 2000.

A. Philip Dawid. Further comments on some comments on a paper by Bradley Efron. *The Annals of Statistics*, 5(6):1249, 1977.

Tom Dietterich. AI Seminar. Carnegie Mellon, 2002.

Alan T. Gous. *Exponential and Spherical Subfamily Models*. PhD thesis, Stanford University, 1998.

Alexander Grigor'yan and Masakazu Noguchi. The heat kernel on hyperbolic space. *Bulletin of the London Mathematical Society*, 30:643–650, 1998.

Ying Guo, Peter L. Bartlett, John Shawe-Taylor, and Robert C. Williamson. Covering numbers for support vector machines. *IEEE Trans. Information Theory*, 48(1), January 2002.

Tommi S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, volume 11, 1998.

Thorsten Joachims. *The Maximum Margin Approach to Learning Text Classifiers Methods, Theory and Algorithms*. PhD thesis, Dortmund University, 2000.

Thorsten Joachims, Nello Cristianini, and John Shawe-Taylor. Composite kernels for hypertext categorisation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2001.

Robert E. Kass. The geometry of asymptotic inference. *Statistical Science*, 4(3), 1989.

Robert E. Kass and Paul W. Vos. *Geometrical Foundations of Asymptotic Inference*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1997.

George Kimeldorf and Grace Wahba. Some results on Tchebychean spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete input spaces. In C. Sammut and A. Hoffmann, editors, *Proceedings of the International Conference on Machine Learning (ICML)*. Morgan Kaufmann, 2002.

Stefan L. Lauritzen. Statistical manifolds. In S. I. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao, editors, *Differential Geometry in Statistical Inference*, pages 163–216. Institute of Mathematical Statistics, Hayward, CA, 1987.

David D. Lewis and Marc Ringuette. A comparison of two learning algorithms for text categorization. In *Symposium on Document Analysis and Information Retrieval*, pages 81–93, Las Vegas, NV, April 1994. ISRI; Univ. of Nevada, Las Vegas.

Peter Li and Shing-Tung Yau. Estimates of eigenvalues of a compact Riemannian manifold. In *Geometry of the Laplace Operator*, volume 36 of *Proceedings of Symposia in Pure Mathematics*, pages 205–239, 1980.

Shahar Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.

John W. Milnor. *Morse Theory*. Princeton University Press, 1963.

Pedro J. Moreno, Purdy P. Ho, and Nuno Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

Edward Nelson. *Tensor Analysis*. Princeton University Press, 1968.

Tomaso Poggio and Frederico Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

Jay Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR*, pages 275–281, 1998.

Calyampudi R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.

Steven Rosenberg. *The Laplacian on a Riemannian Manifold*. Cambridge University Press, 1997.

Richard Schoen and Shing-Tung Yau. *Lectures on Differential Geometry*, volume 1 of *Conference Proceedings and Lecture Notes in Geometry and Topology*. International Press, 1994.

Michael Spivak. *Differential Geometry*, volume 1. Publish or Perish, 1979.

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR'2001*, pages 334–342, Sept 2001.