T-122.102 Special Course in Information Technology: Co-occurence methods in analysis of discrete data. Information diffusion kernels

Sven Laur
swen@math.ut.ee, slaur@tcs.hut.fi

April 2, 2004

1 Introduction

Information geometry is a highly abstract branch of mathematics that unifies statistical and geometrical concepts. The discipline provides general framework of many statistical phenomenas. The main drawbacks of the theory is generality—the approach seldom provides novel ideas and techniques. The survey covers one of the novelties—information diffusion kernels [6, 5]. Of course, we cannot exhibit the whole internal beauty of information geometry, but we encourage the interested reader to browse detailed monographs [7, 1].

The idea behind the information diffusion kernels is surprisingly simple. Each data point is associated with a distribution from some predefined family of distributions. The proximity measure in the feature space is defined rather over distributions themselves than on the parameterization. The association and the proximity measure together form a kernel. The modular definition allows reuse notions from differential geometry to analyze statistical properties like asymptotic consistency.

The brief outline of the survey is following. First, we discuss how to bind data with probability distributions. Next we cover the concepts and motivation behind heat-diffusion kernels and define heat-diffusion over probability distributions. Finally, we cite results of experiments [6] that show effectiveness of the approach.

2 From data to distribution

Traditional kernel methods are biased towards continuous data, i.e. most of the kernels assume that data comes from a subset of $\mathbb{R}^n$. As the real-world data is often varying and discrete like text documents, DNA and protein sequences, one has to use some *ad hoc* technique to find an embedding into $\mathbb{R}^n$. Frequently used embeddings are based on generative models, i.e. a discrete data object is assumed to be generated by a stochastic process. Two natural embeddings that map data to probability distribution are maximum likelihood (ML) and maximum a posteriori (MAP) estimates. Traditionally the corresponding parameter vector θ is interpreted as an element of $\mathbb{R}^n$ and continuous kernels are used for further inference. Hence, different parameterizations can lead to different results, although the corresponding model itself is same. Recently, several parameterization independent techniques like Fisher kernels [2] and mutual information kernels [8] have been proposed besides information diffusion kernels. Nevertheless, they can be viewed in the information geometry framework.

Text classification will be our central example. As usual we employ a bag of words approach. Namely, each document is represented by keyword count vector $\boldsymbol{x}$. The simplest generative model of text is multinomial distribution with parameters θ. Clearly, the term frequency representation

$$\widehat{\theta}_{\mathrm{tf}}(\boldsymbol{x}) = \frac{1}{x_1 + \cdots + x_n}(x_1, \ldots, x_n)$$

corresponds to the ML estimate. Second frequently used embedding is inverse document frequency weighting

$$\widehat{\theta}_{\text{tfidf}}(\boldsymbol{x}) = \frac{1}{x_1 w_i + \cdots + x_n w_n}(x_1 w_i, \ldots, x_n w_n)$$

where the weight $w_i = \log(1/f_i)$ is logarithm of the inverse term frequency in document collection.

3 Statistical manifolds

The main quest of information geometry is to derive parameterization independent quantities over families of probability distributions. But first we have to define the notion of statistical manifold. Let $\mathcal{X}$ be the domain of all possible values. Then a statistical manifold $\mathcal{P}$ is a parameterized family of probability distributions

$$\mathcal{P} = \{p(\cdot|\theta) : \mathcal{X} \to R \mid \theta \in \Theta\},$$

where Θ is open subset of $\mathbb{R}^n$. The parameterization must be unique: $p(\cdot|\theta_1) \equiv p(\cdot|\theta_2) \Rightarrow \theta_1 = \theta_2$. Hence, θ can be treated as the coordinate vector of $p(\cdot|\theta)$. We say that parameterization ψ is admissible iff ψ as a function of primary parameters θ is C^∞ smooth. It is easy to see that the set of admissible parameterizations does not change, if we take arbitrary admissible parameterization as a primary.

We consider only manifolds where log-likelihood function $\ell(\boldsymbol{x}|\theta) = \log p(\boldsymbol{x}|\theta)$ is C^∞ differentiable w.r.t. θ. Note that the property is invariant under all admissible parameterizations. For example, multinomial family satisfies the C^∞ requirement, since

$$\ell(\boldsymbol{x}|\theta) = \log \prod_{j=1}^{m} \theta_{x_j} = \sum_{j=1}^{m} \log \theta_{x_j}.$$

For proximity, we need also vectors and distance measure. The distance between two points is defined as length of the shortest path and the length itself is defined via integral and Euclidean norm

$$d(p,q) = \int_0^1 \|\dot{\gamma}(t)\|\, dt = \int_0^1 \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t)\rangle} dt,$$

where $\gamma : [0,1] \to \mathcal{P}$ corresponds to a (simple) path and $\dot{\gamma}(t)$ to a tangent vector. Since the family $\mathcal{P}$ *a priori* does not have any geometrical structure, we need an abstract definition of vectors. Formally, a vector will be a function that maps functions with the type $\mathcal{P} \to \mathbb{R}$ to real numbers. For fixed coordinates θ and point p natural maps $\left(\frac{\partial}{\partial \theta_i}\right)_p$ emerge

$$\left(\frac{\partial}{\partial \theta_i}\right)_p (f) = \left.\frac{\partial f}{\partial \theta_i}\right|_p.$$

Let us denote lines with only varying coordinate $\gamma_i(t) = p(\cdot|\theta_1, \ldots, \theta_i + t, \ldots, \theta_n)$. Then the derivative of $f(\gamma_i(t))$ at point p will be exactly $\left(\frac{\partial}{\partial \theta_i}\right)_p(f)$. Moreover, if the path γ is differentiable, we can always decompose $f(\gamma(t))'$ as linear combination of partial derivatives

$$f(\gamma(t))' = \left[\theta_1(t)' \left(\frac{\partial}{\partial \theta_1}\right)_{\gamma(t)} + \cdots \theta_n(t)' \left(\frac{\partial}{\partial \theta_n}\right)_{\gamma(t)}\right](f)$$

As the operator in the square brackets does not depend on f and has a right type, we baptize it as the speed vector $\dot{\gamma}(t)$. For a fixed coordinate system θ, we can express $\gamma(t) = \big(\theta_1(t), \ldots, \theta_n(t)\big)$ and we have a natural representation

$$\dot{\gamma}(t) \longmapsto \big(\dot{\theta}_1(t), \ldots, \dot{\theta}_n(t)\big) \in \mathbb{R}^n,$$

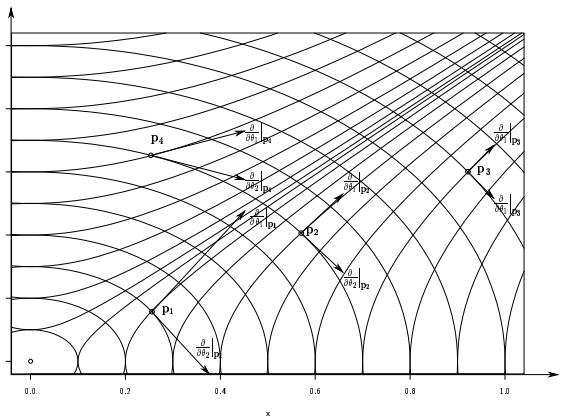

Figure 1: Non-Cartesian coordinate system and corresponding tangent vectors

but this is not coordinate independent. Moreover, the basis $\left(\frac{\partial}{\partial\theta_i}\right)_p$ may change along the manifold (see Figure 1). Remarkably, the abstract tangent vector does not depend on coordinates. In case of $\mathbb{R}^n$, we can interpret tangent vectors $\dot\gamma(t)$ as usual vectors of $\mathbb{R}^n$. But without supporting geometrical structure the generalized speed has retained only the most important feature—for each function $f : \mathcal{P} \to \mathbb{R}$, the speed uniquely characterizes the rate of change at point $\gamma(t)$.

To fix a geometry, we need also reasonable definition of metric. A usual derivation of metric on statistical manifolds is too involved and abstract for the survey. Hence, we take a shortcut. What should be the distance of two adjacent distributions p and q? The most reasonable answer is the weighted Kullback-Leibler divergence

$$\begin{aligned} J(p,q) &= D_{p\|q} + D_{q\|p} \\ &= \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} + \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}, \end{aligned}$$

since it quantifies average additional utility if we swap the distributions. Now consider an infinitesimal movement along the curve $\gamma(t)$. The corresponding change of coordinates is from θ to $\theta + \dot\theta\Delta t$ and the distance formula gives

$$d(p,q)^2 \approx \Delta t^2 \left\|\dot\gamma(t)\right\|^2 = \Delta t^2 \sum_{i,j=1}^{n} \dot\theta_i \dot\theta_j \left\langle \frac{\partial}{\partial\theta_i}, \frac{\partial}{\partial\theta_j} \right\rangle$$

On the other hand, we expect $d(p,q)^2 = J(p,q)$ and under mild regularity conditions (see [4, p.26–28])

$$J(p,q) \approx \Delta t^2 \sum_{i,j=1}^{n} \dot\theta_i \dot\theta_j g_{ij},$$

where g_{ij} are the Fisher information matrix entries

$$g_{ij} = \int p(\boldsymbol{x}) \cdot \frac{\partial \ell(\boldsymbol{x}|\theta)}{\partial\theta_i} \cdot \frac{\partial \ell(\boldsymbol{x}|\theta)}{\partial\theta_j} d\boldsymbol{x}.$$

Hence, the scalar product is defined via Fisher information matrix

$$\left\langle \tfrac{\partial}{\partial\theta_i}, \tfrac{\partial}{\partial\theta_j} \right\rangle = g_{ij}.$$

Now, the geometry of the statistical manifold is complete and we proceed with a discussion about heat diffusion kernels.

4 Why heat diffusion?

The most suitable kernels in statistics are Mercer kernels. A Mercer is defined via (continuous) transformation ϕ from a data space to a high dimensional Euclidean space. Mercer kernels are used together with support vector machines. Recall, that support vector machines use a transformation ϕ to convert data to linearly separable sets and the Mercer kernel $K(\boldsymbol{x},\boldsymbol{y}) = \langle\phi(\boldsymbol{x}),\phi(\boldsymbol{y})\rangle$ allows to calculate necessary quantities without computing the map ϕ.

In first glance, the geodesic distance $d(p,q)$ seems a natural starting point for a suitable Mercer kernel. Namely, the map ϕ should preserve distances

$$d(\boldsymbol{x},\boldsymbol{y})^2 = \langle\phi(\boldsymbol{x}) - \phi(\boldsymbol{y}), \phi(\boldsymbol{x}) - \phi(\boldsymbol{y})\rangle$$

or equivalently

$$K(\boldsymbol{x},\boldsymbol{y}) = \frac{d^2(\boldsymbol{x},\boldsymbol{y}) - d^2(\boldsymbol{x},\boldsymbol{o}) - d^2(\boldsymbol{y},\boldsymbol{o})}{2}, \quad \phi(\boldsymbol{o}) = \mathbf{0}.$$

But generally there are no continuous transformations ϕ that would embed the manifold into Euclid space so that the distances are preserved. For example consider a sphere. There are infinitely many shortest paths between antipodes, but in the Euclidean space there is only one shortest path between two points. Hence, no continuous transformation to Euclid space can preserve metrics.

A heat diffusion kernel seems a good alternative. First, kernels still capture the metric of the manifold. Secondly, they coincide with Gaussian kernels in $\mathbb{R}^n$ and thirdly kernels that are closely related with heat diffusion have been used before [3]. If a heat diffusion kernel is used with support vector machines, we get a remarkably clear interpretation. A temperature of a manifold point is defined as a result of heat diffusion process, where small areas around training points have initial temperature proportional to labeling and importance in classification. The initial temperature is set to zero in all other areas. The classification rule divides points to two classes: "hot" and "cool" points. Intuitively, close points in the manifold have very similar temperature and will have same labels. Since the temperature equalizes over the time, a wide spectrum of kernels from very sensitive to completely robust are available.

The heat diffusion is governed by partial differential equations

$$\begin{aligned} &\frac{\partial f}{\partial t} - \Delta f = 0 \\ &f(x,0) = f(x) \end{aligned} \tag{1}$$

augmented with suitable boundary conditions, if the manifold has a boundary. The general Laplace operator is rather complex

$$\Delta f = \det G^{-1/2} \sum_{i,j=1}^{n} \frac{\partial}{\partial \theta_j}\Big[g^{ij} \det G^{1/2} \frac{\partial f}{\partial \theta_i}\Big]$$

where g^{ij} are elements of the inverse Fisher matrix G. Note that the operator Δ is completely determined by the geometry of the manifold. In the Euclidean space $\mathbb{R}^n$ the definition simplifies to usual notation

$$\Delta f = \frac{\partial^2 f}{\partial x_1^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2}.$$

and it is straightforward to verify

$$f(x,t) = (4\pi)^{-n/2} \int \exp\Big(-\tfrac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{4t}\Big) f(\boldsymbol{y}) d\boldsymbol{y}$$

is a solution to heat diffusion problem with the initial temperature f. The inner term

$$K_t(\boldsymbol{x},\boldsymbol{y}) = \exp\Big(-\tfrac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{4t}\Big)$$

corresponds to Gaussian kernel and the discriminant function of SVM corresponds to heat diffusion from point sources $\boldsymbol{x_i}$. In that sense, heat diffusion on statistical manifolds is simple generalization of Gaussian kernels.

Formally, a heat diffusion kernel is a parametric function $K : \mathcal{P} \times \mathcal{P} \times \mathbb{R} \to \mathbb{R}$ such that for all initial conditions f the solution of equations (1) can be obtained as an integral

$$f(\boldsymbol{x},t) = \int K(\boldsymbol{x},\boldsymbol{y},t) f(\boldsymbol{y}) d\boldsymbol{y}.$$

The existence of such kernel is guaranteed.

Theorem 1. *Let M be a complete Riemannian manifold. Then there exists a kernel function K (heat kernel), which satisfies the following properties:*
(1) $K(\boldsymbol{x},\boldsymbol{y},t) = K(\boldsymbol{y},\boldsymbol{x},t)$;
(2) $\lim_{t\to 0} K(\boldsymbol{x},\boldsymbol{y},t) = \delta(\boldsymbol{x},\boldsymbol{y})$;
(3) $(\Delta - \frac{\partial}{\partial t})K(\boldsymbol{x},\boldsymbol{y},t) = 0$;
(4) $K(\boldsymbol{x},\boldsymbol{y},t) = \int K(\boldsymbol{x},\boldsymbol{z},t-s)K(\boldsymbol{z},\boldsymbol{y},s)d\boldsymbol{z}$.

The completeness assumption is equivalent to the assumption that all bounded closed sets on the manifold are compact. In other words, all limits remain in the manifold and every sequence has a convergent subsequence. Or equivalently $J(p,q) \to 0$, if q converges parameter-wise to p. The assumption is rather weak and is fulfilled by many distribution families. Properties (2) and (3) assure that the function $f(\boldsymbol{x},t)$ is a sought solution to equations (1). Properties (1) and (4) assure that $K_t(\boldsymbol{x},\boldsymbol{y}) = K(\boldsymbol{x},\boldsymbol{y},t)$ is a proper Mercer kernel, as a symmetric and positively definite operator.

5 Approximation of heat kernel

We have seen that the Gaussian kernels are also heat kernels. But closed forms of heat kernels are rare. Usually, we can only seek an approximation

$$\begin{aligned} K_t(\boldsymbol{x},\boldsymbol{y}) \approx K_t^{(m)} &= (4\pi t)^{-n/2} \exp\Big(-\tfrac{d^2(\boldsymbol{x},\boldsymbol{y})}{4t}\Big) \\ &\cdot \big[\psi_0(\boldsymbol{x},\boldsymbol{y}) + \psi_1(\boldsymbol{x},\boldsymbol{y})t + \cdots + \psi_m(\boldsymbol{x},\boldsymbol{y})t^m\big]. \end{aligned}$$

Intuitively, terms $\psi_0, \ldots, \psi_m$ correct distortion from a flat Euclidean geometry. Let $r = d(\boldsymbol{x},\boldsymbol{y})$, then we can express $\psi_i(\boldsymbol{x},\cdot) = \psi_i(r)$ and the terms ψ_i can be computed by recursive equations

$$\begin{aligned} \psi_0 &= \Big(\tfrac{\sqrt{\det G}}{r^{n-1}}\Big)^{-1/2} \\ \psi_k &= r^{-k}\psi_0 \int_0^r \psi_0^{-1}(\Delta\psi_{k-1}) s^{k-1} ds, \end{aligned}$$

where the residue term is

$$\big(\Delta - \tfrac{\partial}{\partial t}\big) K_t^{(m)} = (t^m \Delta\psi_m)(4\pi t)^{-n/2} \exp\big(-r^2/4t\big).$$

Hence, first m terms allow to obtain approximation $K_t(\boldsymbol{x},\boldsymbol{y}) = K_t^{(m)}(\boldsymbol{x},\boldsymbol{y}) + O(t^m)$ provided ψ_m

is smooth. Of course, the approximation is valid for ε-neighborhood $\varepsilon < 1$ and generally the approximation $K_t^{(m)}$ is not positively definite unless $t \in [0, \varepsilon)$. Fortunately, estimates about ε can be obtained.

6 Geometry of multinomials

In principle, the sufficient approximations of heat kernel can derived. However, practical details can be quite messy. Next, we derive appropriate kernel approximation for multinomial family. We consider the multinomial distribution with $n+1$ different outcomes. The usual parameterization

$$\Theta = \{(\theta_1, \ldots, \theta_n) \in \mathbb{R}^n,\ \theta_i > 0,\ \theta_1 + \cdots + \theta_n \leq 1\}$$

corresponds to n dimensional simplex. Let $\boldsymbol{x} = (x_1, \ldots, x_{n+1})$ be the indicator of a single draw, i.e. $x_i = 1$ iff the the ith event has happened. Then

$$\frac{\partial \ell(\boldsymbol{x}|\theta)}{\partial \theta_i} = \frac{x_i}{\theta_i} - \frac{x_{n+1}}{\theta_{n+1}}$$
$$\frac{\partial^2 \ell(\boldsymbol{x}|\theta)}{\partial \theta_i \partial \theta_j} = -\frac{x_i}{\theta_i^2}\delta_{i,j} - \frac{x_{n+1}}{\theta_{n+1}^2},$$

where $\theta_{n+1} = 1 - \theta_1 - \cdots - \theta_n$. The equation

$$E_{\boldsymbol{x}}\left[\tfrac{\partial \ell(\boldsymbol{x}|\theta)}{\partial \theta_i} \cdot \tfrac{\partial \ell(\boldsymbol{x}|\theta)}{\partial \theta_j}\right] = -E_{\boldsymbol{x}}\left[\tfrac{\partial^2 \ell(\boldsymbol{x}|\theta)}{\partial \theta_i \partial \theta_j}\right]$$

allows to express Fisher matrix elements

$$g_{ij}(\theta_1, \ldots, \theta_n) = \begin{cases} 1/\theta_{n+1}, & \text{if } i \neq j, \\ 1/\theta_i + 1/\theta_{n+1}, & \text{if } i = j. \end{cases}$$

It is rather hard to compute geodesic distances on the simplex $\Theta \subset \mathbb{R}^n$ and therefore we define two mappings that ease the task.

A map $F : \mathcal{P} \to \mathcal{Q}$ is called isometry if it satisfies two requirements: (1) F must be C^∞ differentiable w.r.t. θ; (2) all curves $\gamma(t)$ and $\gamma^*(t) = F(\gamma(t))$ must have same length. The second condition is satisfied iff $\|\dot\gamma(t)\| = \|\dot\gamma^*(t)\|$.

By adding redundant coordinate θ_{n+1} and redefining geometry

$$g_{ij}(\theta_1, \ldots, \theta_{n+1}) = \begin{cases} 0, & \text{if } i \neq j, \\ 1/\theta_i, & \text{if } i = j. \end{cases}$$

we get an isometric embedding $i : \mathbb{R}^n \to \mathbb{R}^{n+1}$, since the length of the vectors are preserved

$$\sum_{i,j=1}^{n} \dot\theta_i \dot\theta_j g_{ij} = \sum_{i=1}^{n} \frac{\dot\theta_i^2}{\theta_i} + \frac{1}{\theta_{n+1}}\left(\sum_{i=1}^{n} \dot\theta_i\right)^2 = \sum_{i=1}^{n+1} \frac{\dot\theta_i^2}{\theta_i}.$$

Intuitively, this means that tangent vectors $\frac{\partial}{\partial \theta_i}$ are orthogonal. But we can go even further and build a model of $\mathcal{P}$ in Euclidean space $\mathbb{R}^{n+1}$ that preserves distances. The target set will be $n+1$ dimensional positive orthant

$$\mathcal{S}^+ = \left\{(x_1, \ldots, x_{n+1}) : x_1^2 + \cdots + x_{n+1}^2 = 4\right\}.$$

and the isometry

$$F(\theta_1, \ldots, \theta_{n+1}) = \big(2\sqrt{\theta_1}, \ldots, 2\sqrt{\theta_{n+1}}\big).$$

Since the tangent vector is

$$f(\gamma^*(t))' = \left[\dot x_1 \cdot \big(\tfrac{\partial}{\partial x_1}\big)_{\gamma^*} + \cdots + \dot x_{n+1} \cdot \big(\tfrac{\partial}{\partial x_{n+1}}\big)_{\gamma^*}\right](f),$$

where $x_1(t), \ldots, x_{n+1}(t)$ are Cartesian coordinates of $\gamma^*(t)$, the lengths are preserved only if

$$\|\dot\gamma^*(t)\|^2 = \dot x_1(t)^2 + \cdots + \dot x_{n+1}(t)^2 = \|\dot\gamma(t)\|^2.$$

A simple calculation proves isometry

$$\|\dot\gamma^*(t)\|^2 = \left(\tfrac{\dot\theta_1(t)}{\sqrt{\theta_1(t)}}\right)^2 + \cdots + \left(\tfrac{\dot\theta_{n+1}(t)}{\sqrt{\theta_{n+1}}}\right)^2$$
$$= \tfrac{\dot\theta_1(t)^2}{\theta_1(t)} + \cdots + \tfrac{\dot\theta_{n+1}(t)^2}{\theta_n(t)} = \|\dot\gamma(t)\|^2.$$

The corresponding maps have a nice geometrical meaning, see Figure 2. The isometry provides novel insight for calculating distances between θ and θ', also depicted by Figure 2. As the shortest path between two points is an arc of the great circle, the distance becomes easily computable

$$d(\theta, \theta') = 2\arccos(\langle F(\theta), F(\theta')\rangle)$$
$$= 2\arccos\big(\sqrt{\theta_1\theta_1'} + \cdots + \sqrt{\theta_{n+1}\theta_{n+1}'}\big).$$

The corresponding metric is illustrated by Figure 3. The more rapid distance rate in the corners has intuitive explanation—slight changes in parameters cause significant change in relative entropy (utility), if some event is almost certain.

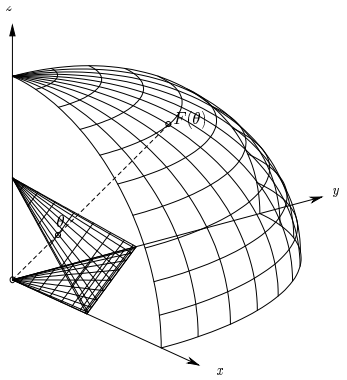

Figure 2: Geometrical interpretation of isometry

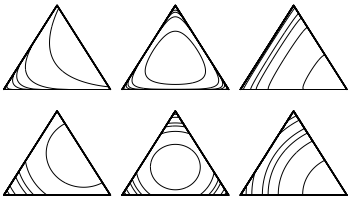

Figure 3: Iso-distant lines: above multinomial geometry, below Euclidean geometry.

7 Kernel construction

Finally, we are in position to define an approximation of the heat diffusion kernel. Nontrivial derivation shows

$$\psi_0(r) = 1 + \frac{(n-1)}{12}r^2 + \frac{(n-1)(5n-1)}{1440}r^4 + O(r^6)$$

and the first order approximation is

$$K_t^{(0)}(\theta, \theta') = (4\pi t)^{-n/2} \exp\left(-\frac{\arccos^2(\sqrt{\theta}, \sqrt{\theta'}}{t}\right).$$

The approximation has good properties compared with higher order approximations $K_t^{(m)}$, $m \geq 1$, since $K_t^{(0)}$ remains small for large distances.

As a theoretical side-mark the simplex is not a complete Riemann manifold. Moreover, the border has non-differentiable angles. Therefore, for theoretical investigation these angles must be rounded—certain and impossible events are removed from model.

This allows to find bounds of covering numbers $\mathcal{N}(\epsilon, \mathcal{F}_R)$, where $\mathcal{F}_R$ corresponds to linear classifier in high dimensional space with bounded weight vector $\|\boldsymbol{w}\| \leq R$. The cover numbers allow to estimate the difference of average classification risk and empirical risk (training error). These results are based on advanced issues of statistical learning theory and functional analysis. Compared with the usual VC-dimension approach, the result are much sharper but in the same time more obscure. Thus we cite only the main result.

Informal statement *The heat-kernel have essentially the same asymptotic generalization performance as Gaussian kernels with the same dimension.*

8 Experimental results

We cite here results of the report [6]. A support vector machine with approximation of multinomial kernel was applied to Reuters-21570 and WebKB data. Both frequency representation and inverse document frequency weighting was used. In all reported instances heat kernel gave better results than Gaussian kernel. Some results are given in Figure 4.

9 Conclusions

Good kernels for real valued vectors have been known for years. Therefore, a big effort has been made to find natural ways to define kernels for discrete data generated by random processes. Information geometry gives a theoretically justified approach. The main advantages are independence of parameterization and theoretical machinery for estimating generalization properties. Heat kernels are also natural generalizations of Gaussian kernels. Empirical results with multinomial distributions indicate that heat diffusion kernel out-performs Gaussian kernels in text classification tasks.

The most serious downside is conceptual complexity. The general concept of statistical manifolds is not easy to grasp. Secondly, only few instances of closed form solutions of heat diffusion equation is known.

Hence, generalizing the results to more complicated models requires *significant* amount of mathematical knowledge.

References

[1] Shun-ichi Amari. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

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

[3] Risi Imre Kondor and John D. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 315–322. Morgan Kaufmann Publishers Inc., 2002.

[4] Solomon Kullback. *Information Theory and Statistics*. Dover, 1997.

[5] John Lafferty and Guy Lebanon. Information diffusion kernels. In *Advances in Neural Information Processing Systems*, 2002.

[6] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. Technical Report CMU-CS-04-101, School of Computer Science Carnegie Mellon University, 2004.

[7] Michael K. Murray and John W. Rice. *Differential Geometry and Statistics*, volume 48 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1993.

[8] M. Seeger. Covariance kernels from bayesian generative models. In *Advances in Neural Information Processing Systems*, Cambridge, MA, 2002. MIT Press.

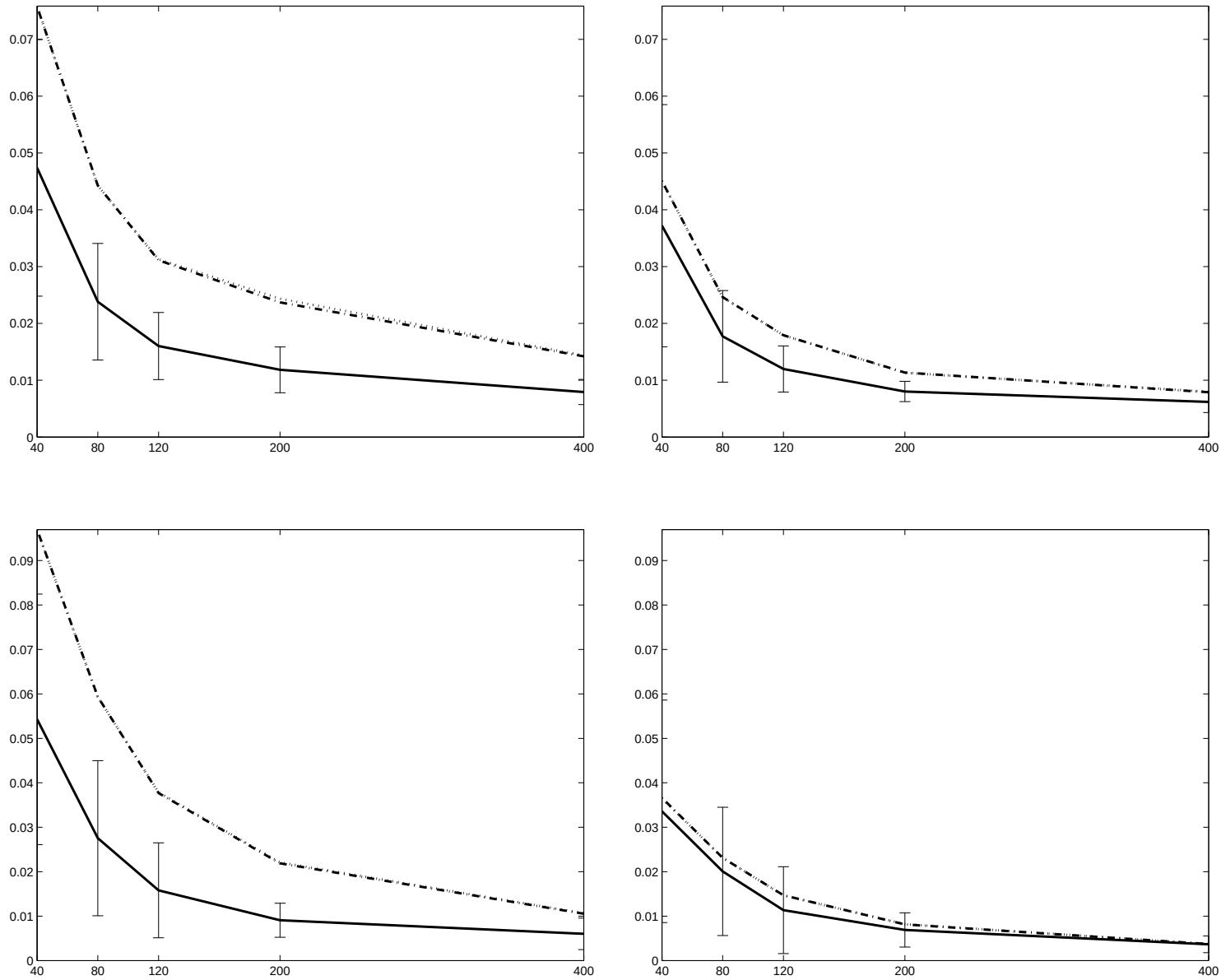

Figure 4: Experimental results on the Reuters corpus, using SVMs for linear (dotted) and Gaussian (dash-dotted) kernels, compared with the diffusion (solid). The classes moneyFx (top) and grain (bottom) are labeled as positive, and the class earn is labeled negative. The left column uses tf representation and the right column uses tfidf representation [6, p.28].