Text Classification with Kernels on the Multinomial Manifold

Dell Zhang[1,2]
[1]Department of Computer Science
School of Computing
S16-05-08, 3 Science Drive 2
National University of Singapore
Singapore 117543
[2]Singapore-MIT Alliance
E4-04-10, 4 Engineering Drive 3
Singapore 117576
+65-68744251
dell.z@ieee.org

Xi Chen
Department of Mathematical and Statistical Sciences
University of Alberta
10350 122 St Apt 205
Edmonton, AB T5N 3W4
Canada
+1-780-492-1704
xichen@math.ualberta.ca

Wee Sun Lee[1,2]
[1]Department of Computer Science
School of Computing
SOC1-05-26, 3 Science Drive 2
National University of Singapore
Singapore 117543
[2]Singapore-MIT Alliance
E4-04-10, 4 Engineering Drive 3
Singapore 117576
+65-68744526
leews@comp.nus.edu.sg

ABSTRACT
Support Vector Machines (SVMs) have been very successful in text classification. However, the intrinsic geometric structure of text data has been ignored by standard kernels commonly used in SVMs. It is natural to assume that the documents are on the multinomial manifold, which is the simplex of multinomial models furnished with the Riemannian structure induced by the Fisher information metric. We prove that the Negative Geodesic Distance (NGD) on the multinomial manifold is conditionally positive definite (cpd), thus can be used as a kernel in SVMs. Experiments show the NGD kernel on the multinomial manifold to be effective for text classification, significantly outperforming standard kernels on the ambient Euclidean space.

Categories and Subject Descriptors
H.3.1. [**Content Analysis and Indexing**]; H.3.3 [**Information Search and Retrieval**]; I.2.6 [**Artificial Intelligence**]: Learning; I.5.2 [**Pattern Recognition**]: Design Methodology – *classifier design and evaluation*.

General Terms
Algorithms, Experimentation, Theory.

Keywords
Text Classification, Machine Learning, Support Vector Machine, Kernels, Manifolds, Differential Geometry.

1. INTRODUCTION
Recent research works have established the Support Vector Machine (SVM) as one of the most powerful and promising machine learning methods for text classification [8, 15, 16, 36].

"The crucial ingredient of SVMs and other kernel methods is the so-called kernel trick, which permits the computation of dot products in high-dimensional feature spaces, using simple functions defined on pairs of input patterns. This trick allows the formulation of nonlinear variants of any algorithm that can be cast in terms of dot products, SVMs being but the most prominent example." [32]

However, standard kernels commonly used in SVMs have neglected a-priori knowledge about the intrinsic geometric structure of text data. We think it makes more sense to view document feature vectors as points in a Riemannian manifold, rather than in the much larger Euclidean space. This paper studies kernels on the multinomial manifold that enable SVMs to effectively exploit the intrinsic geometric structure of text data to improve text classification accuracy.

In the rest of this paper, we first examine the multinomial manifold (§2), then propose the new kernel based on the geodesic distance (§3) and present experimental results to demonstrate its effectiveness (§4), later review related works (§5), finally make concluding remarks (§6).

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

2. THE MULTINOMIAL MANIFOLD
This section introduces the concept of the multinomial manifold and the trick to compute geodesic distances on it, followed by how documents can be naturally embedded in it.

2.1 Concept
Let $\mathcal{S} = \{p(\cdot|\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$ be an n-dimensional regular statistical model family on a set $\mathcal{X}$. For each $\mathbf{x}\in\mathcal{X}$ assume the mapping $\boldsymbol{\theta}\mapsto p(\mathbf{x}|\boldsymbol{\theta})$ is C^∞ at each point in the interior of Θ. Let ∂_i denote $\frac{\partial}{\partial\theta_i}$ and $\ell_{\boldsymbol{\theta}}(\mathbf{x})$ denote $\log\big(p(\mathbf{x}|\boldsymbol{\theta})\big)$. The Fisher information metric [1, 19, 21] at $\boldsymbol{\theta}\in\Theta$ is defined in terms of the matrix given by

$$g_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\partial_i \ell_{\boldsymbol{\theta}} \partial_j \ell_{\boldsymbol{\theta}}\right] = \int_{\mathcal{X}} p(\mathbf{x}\,|\,\boldsymbol{\theta}) \partial_i \log\left(p(\mathbf{x}\,|\,\boldsymbol{\theta})\right) \partial_j \log\left(p(\mathbf{x}\,|\,\boldsymbol{\theta})\right) d\mathbf{x} \quad (1)$$

or equivalently as

$$g_{ij}(\boldsymbol{\theta}) = 4\int_{\mathcal{X}} \partial_i \sqrt{p(\mathbf{x}\,|\,\boldsymbol{\theta})}\, \partial_j \sqrt{p(\mathbf{x}\,|\,\boldsymbol{\theta})}\, d\mathbf{x}. \quad (2)$$

Note that $g_{ij}(\boldsymbol{\theta})$ can be thought of as the variance of the score $\partial_i \ell_{\boldsymbol{\theta}}$. In coordinates θ_i, $g_{ij}(\boldsymbol{\theta})$ defines a Riemannian metric on Θ, giving $\mathcal{S}$ the structure of an *n*-dimensional Riemannian manifold.

Intuitively the Fisher information may be seen as the amount of information a single data point supplies with respect to the problem of estimating the parameter $\boldsymbol{\theta}$. The choice of the Fisher information metric is motivated by its attractive properties in theory and good performances in practice [21, 23, 24].

Consider multinomial distributions that model mutually independent events $X_1,...,X_{n+1}$ with $\Pr[X_i] = \theta_i$. Obviously $\boldsymbol{\theta} = (\theta_1,...,\theta_{n+1})$ should be on the *n*-simplex defined by $\sum_{i=1}^{n+1} \theta_i = 1$. The probability that X_1 occurs x_1 times, ..., X_{n+1} occurs x_{n+1} times is given by

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) = \frac{N!}{\prod_{i=1}^{n+1} x_i!} \prod_{i=1}^{n+1} \theta_i^{x_i} \quad (3)$$

where $N = \sum_{i=1}^{n+1} x_i$.

The multinomial manifold is the parameter space of the multinomial distribution

$$\mathbb{P}^n = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{n+1} : \sum_{i=1}^{n+1} \theta_i = 1; \forall i, \theta_i \geq 0 \right\} \quad (4)$$

equipped with the Fisher information metric, which can be shown to be

$$g_{\boldsymbol{\theta}}(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{n+1} \frac{u_i v_i}{\theta_i} \quad (5)$$

where $\boldsymbol{\theta} \in \mathbb{P}^n$, and $\mathbf{u}, \mathbf{v} \in T_{\boldsymbol{\theta}}\mathbb{P}^n$ are vectors tangent to $\mathbb{P}^n$ at $\boldsymbol{\theta}$ represented in the standard basis of $\mathbb{R}^{n+1}$.

2.2 Geodesic

It is a well-known fact that the multinomial manifold $\mathbb{P}^n$ is isometric to the positive portion of the *n*-sphere of radius 2 [18]

$$\mathbb{S}_+^n = \left\{ \boldsymbol{\psi} \in \mathbb{R}^{n+1} : \|\boldsymbol{\psi}\| = 2; \forall i, \psi_i \geq 0 \right\} \quad (6)$$

through the diffeomorphism $F : \mathbb{P}^n \to \mathbb{S}_+^n$,

$$F(\boldsymbol{\theta}) = (2\sqrt{\theta_1},...,2\sqrt{\theta_{n+1}}). \quad (7)$$

Therefore the geodesic distance between $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{P}^n$ can be computed as the geodesic distance between $F(\boldsymbol{\theta}), F(\boldsymbol{\theta}') \in \mathbb{S}_+^n$, i.e., the length of the shortest curve on $\mathbb{S}_+^n$ connecting $F(\boldsymbol{\theta})$ and $F(\boldsymbol{\theta}')$ that is actually a segment of a great circle. Specifically, the geodesic distance between $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{P}^n$ is given by

$$d_G(\boldsymbol{\theta}, \boldsymbol{\theta}') = 2\arccos\left(\left\langle F(\boldsymbol{\theta}), F(\boldsymbol{\theta}')\right\rangle\right) = 2\arccos\left(\sum_{i=1}^{n+1} \sqrt{\theta_i \theta_i'}\right). \quad (8)$$

2.3 Embedding

In text retrieval, clustering and classification, a document is usually considered as a "bag of words" [2]. It is natural to assume that the "bag of words" of a document is generated by independent draws from a multinomial distribution $\boldsymbol{\theta}$ over vocabulary $V = \{w_1,...,w_{n+1}\}$. In other words, every document is modeled by a multinomial distribution, which may change from document to document. Given the feature representation of a document, $\mathbf{d} = (d_1,...,d_{n+1})$, it can be embedded in the multinomial manifold $\mathbb{P}^n$ by applying L_1 normalization,

$$\hat{\boldsymbol{\theta}}(\mathbf{d}) = \left(\frac{d_1}{\sum_{i=1}^{n+1} d_i},...,\frac{d_{n+1}}{\sum_{i=1}^{n+1} d_i} \right). \quad (9)$$

The simple TF representation of a document D sets $d_i = tf(w_i, D)$ which means the term frequency (TF) of word w_i in document D, i.e., how many times w_i appears in D. The embedding that corresponds to the TF representation is theoretically justified as the maximum likelihood estimator for the multinomial distribution [15, 21, 24].

The popular TF×IDF representation [2] of a document D sets $d_i = tf(w_i, D) \cdot idf(w_i)$, where the TF component $tf(w_i, D)$ is weighted by $idf(w_i)$, the inverse document frequency (IDF) of word w_i in the corpus. If there are m documents in the corpus and word w_i appears in $df(w_i)$ documents, then $idf_i = \log\left(m/df(w_i)\right)$. The embedding that corresponds to the TF×IDF representation can be interpreted as a pullback metric of the Fisher information through the transformation

$$G_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \left(\frac{\theta_1 \lambda_1}{\sum_{i=1}^{n+1} \theta_i \lambda_i},...,\frac{\theta_{n+1} \lambda_{n+1}}{\sum_{i=1}^{n+1} \theta_i \lambda_i} \right) \quad (10)$$

with $\lambda_i = \frac{idf_i}{\sum_{j=1}^{n+1} idf_j}$.

How to find the optimal embedding is an interesting research problem to explore. It is possible to learn a Riemannian metric even better than using the TF×IDF weighting [23].

Under the above embeddings, the kernel (that will be discussed later) between two documents $\mathbf{d}$ and $\mathbf{d}'$ means $k\left(\hat{\boldsymbol{\theta}}(\mathbf{d}), \hat{\boldsymbol{\theta}}(\mathbf{d}')\right)$.

3. DISTANCE BASED KERNELS

Good kernels should be consistent with one's intuition of pair-wise similarity/dissimilarity in the domain. The motivation of this paper is to exploit the intrinsic geometric structure of text data to design a kernel that can better capture document similarity/dissimilarity. Standard text retrieval, clustering and classification usually rely on the similarity measure defined by the dot product (inner product) of two document vectors in a Euclidean space [2]. The geometric interpretation of the dot product is that it computes the cosine of the angle between two vectors provided they are normalized to unit length. When turning to the Riemannian geometry, this similarity measure is no longer available on general manifolds, because the concept of dot product is only defined locally on the tangent space but not globally on the manifold itself. However, there exists a natural dissimilarity measure on general manifolds: *geodesic distance*.

This section first describes the concept of kernels (§3.1), then discusses the Negative Euclidean Distance kernel (§3.2) and the Negative Geodesic Distance kernel (§3.3) in detail.

3.1 PD and CPD Kernels

Definition 1 (Kernels [32]). *Let $\mathcal{X}$ be a nonempty set. A real-valued symmetric function $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a* positive definite (pd) kernel *if for all $m \in \mathbb{N}$ and all $\mathbf{x}_1, ..., \mathbf{x}_m \in \mathcal{X}$ the induced $m \times m$ Gram (kernel) matrix $\mathbf{K}$ with elements $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ satisfies $\mathbf{c}^T \mathbf{K} \mathbf{c} \geq \mathbf{0}$ given any vector $\mathbf{c} \in \mathbb{R}^m$. The function k is called a* conditionally positive definite (cpd) kernel *if $\mathbf{K}$ satisfies the above inequality for any vector $\mathbf{c} \in \mathbb{R}^m$ with $\mathbf{c}^T \mathbf{1} = \mathbf{0}$.*

As a direct consequence of Definition 1, we have

Lemma 1 ([32]).

(i) *Any pd kernel is also a cpd kernel.*

(ii) *Any constant $c \geq 0$ is a pd kernel; any constant $c \in \mathbb{R}$ is a cpd kernel.*

(iii) *If k_1 and k_2 are pd (resp. cpd) kernels, $\alpha_1, \alpha_2 \geq 0$, then $\alpha_1 k_1 + \alpha_2 k_2$ is a pd (resp. cpd) kernel.*

(iv) *If k_1 and k_2 are pd kernels, then $k_1 k_2$ defined by $k_1 k_2(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') k_2(\mathbf{x}, \mathbf{x}')$ is a pd kernel.*

Lemma 2 (Connection of PD and CPD Kernels [4, 29, 32]). Let *k be real-valued symmetric function defined on $\mathcal{X} \times \mathcal{X}$. Then we have*

(i) $\tilde{k}(\mathbf{x}, \mathbf{x}') := \frac{1}{2}\left(k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, \mathbf{x}_0) - k(\mathbf{x}_0, \mathbf{x}') + k(\mathbf{x}_0, \mathbf{x}_0)\right)$ *is pd if and only if k is cpd;*

(ii) $\exp(tk)$ *is pd for all $t > 0$ if and only if k is cpd.*

Theorem 1 (Hilbert Space Representation of PD Kernels [32]). *Let k be a real-valued pd kernel on $\mathcal{X}$. Then there exists a Hilbert space of real-valued functions on $\mathcal{X}$ and a mapping $\Phi: \mathcal{X} \to \mathcal{H}$ such that*

$$\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle = k(\mathbf{x}, \mathbf{x}'). \tag{11}$$

Theorem 2 (Hilbert Space Representation of CPD Kernels [32]). *Let k be a real-valued cpd kernel on $\mathcal{X}$. Then there exists a Hilbert space of real-valued functions on $\mathcal{X}$ and a mapping $\Phi: \mathcal{X} \to \mathcal{H}$ such that*

$$\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\|^2 = -k(\mathbf{x}, \mathbf{x}') + \frac{1}{2}\left(k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}', \mathbf{x}')\right). \tag{12}$$

The former theorem implies that pd kernels are justified to be used in all kernel methods. The latter theorem implies that cpd kernels are justified to be used in the kernel methods which are translation invariant, i.e., distance based, in the feature space. Since SVMs are translation invariant in the feature space, they are able to employ not only pd kernels but also cpd kernels [31, 32].

Standard (commonly used) kernels [32] include:

Linear $k_{LIN}(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$, (13)

Polynomial $k_{POL}(\mathbf{x}, \mathbf{x}') = \left(\langle \mathbf{x}, \mathbf{x}' \rangle + c\right)^d$ with $d \in \mathbb{N}, c \geq 0$, (14)

Gaussian $k_{RBF}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$ with $\sigma > 0$, and (15)

Sigmoid $k_{SIG}(\mathbf{x}, \mathbf{x}') = \tanh\left(\kappa \langle \mathbf{x}, \mathbf{x}' \rangle + \vartheta\right)$ with $\kappa > 0, \vartheta < 0$. (16)

The former three are pd but the last one is not.

3.2 The Negative Euclidean Distance Kernel

Lemma 3 ([31, 32]). *The negative squared Euclidean distance function $-d_E^2(\mathbf{x}, \mathbf{x}') = -\|\mathbf{x} - \mathbf{x}'\|^2$ is a cpd kernel.*

Lemma 4 (Fractional Powers and Logs of CPD Kernels [4, 32]). *If $k: \mathcal{X} \times \mathcal{X} \to (-\infty, 0]$ is cpd, then so are $-(-k)^\beta$, $0 < \beta < 1$ and $-\ln(1 - k)$.*

Proposition 1. *The* Negative Euclidean Distance (NED) *function*

$$k_{NED}(\mathbf{x}, \mathbf{x}') = -d_E(\mathbf{x}, \mathbf{x}') \tag{17}$$

is a cpd kernel.

Proof. It follows directly from Lemma 3 and 4 with $\beta = 1/2$.

3.3 The Negative Geodesic Distance Kernel

Theorem 3 (Dot Product Kernels in Finite Dimensions [30, 32]). *A function $k(\mathbf{x}, \mathbf{x}') = f(\langle \mathbf{x}, \mathbf{x}' \rangle)$ defined on the unit sphere in a finite n dimensional Hilbert space is a pd kernel if and only if its Legendre polynomial expansion has only nonnegative coefficients, i.e.,*

$$f(t) = \sum_{r=0}^{\infty} b_r \mathcal{P}_r^n(t) \text{ with } b_r \geq 0. \tag{18}$$

Theorem 4 (Dot Product Kernels in Infinite Dimensions [30, 32]). *A function $k(\mathbf{x}, \mathbf{x}') = f(\langle \mathbf{x}, \mathbf{x}' \rangle)$ defined on the unit sphere in*

an infinite dimensional Hilbert space is a pd kernel if and only if its Taylor series expansion has only nonnegative coefficients, i.e.,

$$f(t)=\sum_{r=0}^{\infty}a_r t^r \quad with \;\; a_r \geq 0 . \tag{19}$$

Since (19) is more stringent than (18), in order to prove positive definiteness for arbitrary dimensional dot product kernels it suffices to show that condition (19) holds [32].

Proposition 2. *The* Negative Geodesic Distance (NGD) *function on the multinomial manifold*

$$k_{NGD}(\boldsymbol{\theta},\boldsymbol{\theta}')=-d_G(\boldsymbol{\theta},\boldsymbol{\theta}') \tag{20}$$

for $\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathbb{P}^n$ is a cpd kernel. Moreover, the shifted NGD kernel $k_{NGD}(\boldsymbol{\theta},\boldsymbol{\theta}')+\pi$ is a pd kernel.

Proof. Plugging the formula (8) into (20), we have

$$k_{NGD}(\boldsymbol{\theta},\boldsymbol{\theta}')=-2\arccos\left(\sum_{i=1}^{n+1}\sqrt{\theta_i\theta_i'}\right). \tag{21}$$

Denote $(\sqrt{\theta_1},...,\sqrt{\theta_{n+1}})$ and $(\sqrt{\theta_1'},...,\sqrt{\theta_{n+1}'})$ by $\sqrt{\boldsymbol{\theta}}$ and $\sqrt{\boldsymbol{\theta}'}$ respectively. It is obvious that $\sqrt{\boldsymbol{\theta}}$ and $\sqrt{\boldsymbol{\theta}'}$ are both on the unit sphere. Let

$$f(t)=\pi-2\arccos(t). \tag{22}$$

Then we can rewrite $k_{NGD}(\boldsymbol{\theta},\boldsymbol{\theta}')$ as

$$k_{NGD}(\boldsymbol{\theta},\boldsymbol{\theta}')=f(\left\langle\sqrt{\boldsymbol{\theta}},\sqrt{\boldsymbol{\theta}'}\right\rangle)-\pi . \tag{23}$$

The Maclaurin series (the Taylor series expansion of a function about 0) for the inverse cosine function with $-1\leq t\leq 1$ is

$$\arccos(t)=\frac{\pi}{2}-\sum_{r=0}^{\infty}\frac{\Gamma\left(r+\frac{1}{2}\right)}{\sqrt{\pi}(2r+1)r!}t^{2r+1}, \tag{24}$$

where $\Gamma(x)$ is the gamma function. Hence

$$f(t)=\sum_{r=0}^{\infty}c_r t^{2r+1}, \tag{25}$$

and

$$c_r=\frac{2\Gamma\left(r+\frac{1}{2}\right)}{\sqrt{\pi}(2r+1)r!}. \tag{26}$$

Since $\Gamma(x)>0$ for all $x>0$, we have $c_r>0$ for all $r=0,1,2,...$. By Theorem 3 and 4, the dot product kernel $f(\left\langle\sqrt{\boldsymbol{\theta}},\sqrt{\boldsymbol{\theta}'}\right\rangle)$ is pd. Thus the NGD kernel $k_{NGD}(\boldsymbol{\theta},\boldsymbol{\theta}')$ is cpd according to Lemma 1. The shifted NGD kernel $k_{NGD}(\boldsymbol{\theta},\boldsymbol{\theta}')+\pi$ equals to $f(\left\langle\sqrt{\boldsymbol{\theta}},\sqrt{\boldsymbol{\theta}'}\right\rangle)$ which has been proved to be pd.

Definition 2 (Support Vector Machine, Dual Form [32]) *Given a set of m labeled examples* $(\mathbf{x}_1,y_1),...,(\mathbf{x}_m,y_m)$ *and a kernel* k*, the decision function of the Support Vector Machine (SVM) is*

$$f(\mathbf{x})=\operatorname{sgn}\left(\sum_{i=1}^{m}\alpha_i^* y_i k(\mathbf{x},\mathbf{x}_i)+b^*\right), \tag{27}$$

where $(\alpha_1^*,...,\alpha_m^*)=\boldsymbol{\alpha}^*$ *solves the following quadratic optimization problem:*

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^m}{\text{maximize}}\;\; W(\boldsymbol{\alpha})=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) \tag{28}$$

$$\text{subject to}\;\; 0\leq\alpha_i\leq C \;\; \text{for all} \;\; i=1,...,m \tag{29}$$

$$\text{and}\;\; \sum_{i=1}^{m}\alpha_i y_i=0 \tag{30}$$

for some $C>0$ *, and* b^* *is obtained by averaging* $y_j-\sum_{i=1}^{m}\alpha_i^* y_i k(\mathbf{x}_j,\mathbf{x}_i)$ *over all training examples with* $0<\alpha_j^*<C$.

Proposition 3. *Let* k *be a valid kernel, and* $\tilde{k}=k+\beta$ *where* $\beta\in\mathbb{R}$ *is a constant. Then* k *and* $\tilde{k}$ *lead to the identical SVM, given the same training data.*

Proof. Denote the SVMs with kernel k and $\tilde{k}$ learned from the training data $(\mathbf{x}_1,y_1),...,(\mathbf{x}_m,y_m)$ by $SVM(k)$ and $SVM(\tilde{k})$ respectively. The objective function (28) of the quadratic optimization problem for training $SVM(\tilde{k})$ is

$$W(\boldsymbol{\alpha})=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j\tilde{k}(\mathbf{x}_i,\mathbf{x}_j) \tag{31}$$

$$=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j\left(k(\mathbf{x}_i,\mathbf{x}_j)+\beta\right) \tag{32}$$

$$=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j)-\frac{\beta}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \tag{33}$$

$$=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j)-\frac{\beta}{2}\left(\sum_{i=1}^{m}\alpha_i y_i\right)^2. \tag{34}$$

In addition, for all training examples with $0<\alpha_j^*<C$,

$$y_j-\sum_{i=1}^{m}\alpha_i^* y_i\tilde{k}(\mathbf{x}_j,\mathbf{x}_i)=y_j-\sum_{i=1}^{m}\alpha_i^* y_i\left(k(\mathbf{x}_j,\mathbf{x}_i)+\beta\right) \tag{35}$$

$$=y_j-\sum_{i=1}^{m}\alpha_i^* y_i k(\mathbf{x}_j,\mathbf{x}_i)-\beta\left(\sum_{i=1}^{m}\alpha_i^* y_i\right). \tag{36}$$

Furthermore, the decision function of $SVM(\tilde{k})$ is

$$f(\mathbf{x})=\operatorname{sgn}\left(\sum_{i=1}^{m}\alpha_i^* y_i\tilde{k}(\mathbf{x},\mathbf{x}_i)+b^*\right) \tag{37}$$

$$=\operatorname{sgn}\left(\sum_{i=1}^{m}\alpha_i^* y_i k(\mathbf{x},\mathbf{x}_i)+\beta\left(\sum_{i=1}^{m}\alpha_i^* y_i\right)+b^*\right). \tag{38}$$

Making use of the equality constraint (30), all terms containing β vanish in (34), (36) and (38). Therefore $SVM(k)$ and $SVM(\tilde{k})$ arrive at equivalent $\boldsymbol{\alpha}^*$ and b^*, and their decision functions for classification are identical.

Proposition 2 and Proposition 3 reveal that although the NGD kernel $k_{NGD}(\boldsymbol{\theta},\boldsymbol{\theta}')$ is cpd (not pd), it leads to the identical SVM as a pd kernel $k_{NGD}(\boldsymbol{\theta},\boldsymbol{\theta}')+\pi$. This deepens our understanding of the NGD kernel.

Using the NGD kernel as the starting point, a family of cpd and pd kernels can be constructed based on the geodesic distance on the multinomial manifold using Lemma 1, Lemma 2, Lemma 4, etc [32]. In particular, we have the following pd kernel which will be discussed later in the section of related works.

Proposition 4. *The kernel*

$$k_{NGD-E}(\boldsymbol{\theta},\boldsymbol{\theta}') = (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{1}{t}\arccos\left(\sum_{i=1}^{n+1}\sqrt{\theta_i\theta_i'}\right)\right) \qquad (39)$$

with $t>0$ *for* $\boldsymbol{\theta},\boldsymbol{\theta}' \in \mathbb{P}^n$ *is pd.*

Proof. It is not hard to see that $k_{NGD-E}(\boldsymbol{\theta},\boldsymbol{\theta}')$ equals to $(4\pi t)^{-\frac{n}{2}} \exp\left(\frac{1}{2t}k_{NGD}(\boldsymbol{\theta},\boldsymbol{\theta}')\right)$, whose positive definiteness is a trivial consequence of Proposition 2, Lemma 1 and Lemma 2.

4. EXPERIMENTS

We have conducted experiments on two real-world datasets, WebKB [1] and 20NG [2], to evaluate the effectiveness of the proposed NGD kernel for text classification using SVMs.

The WebKB dataset contains manually classified Web pages that were collected from the computer science departments of four universities (Cornell, Texas, Washington and Wisconsin) and some other universities (misc.). Only pages in the top 4 largest classes were used, namely student, faculty, course and project. All pages were pre-processed using the Rainbow toolkit [25] with the options "--skip-header --skip-html --lex-pipe-command=tag-digits --no-stoplist --prune-vocab-by-occur-count=2". The task is sort of a four-fold cross validation in a leave-one-university-out way: training on three of the universities plus the misc. collection, and testing on the pages from a fourth, held-out university. This way of train/test split is recommended because each university's pages have their idiosyncrasies. The performance measure is the multi-class classification accuracy averaged over these four splits.

The 20NG (20newsgroups) dataset is a collection of approximately 20,000 documents that were collected from 20 different newsgroups [22]. Each newsgroup corresponds to a different topic and is considered as a class. All documents were pre-processed using the Rainbow toolkit [25] with the option "--prune-vocab-by-doc-count=2". The "bydate" version of this dataset along with its train/test split is used in our experiments due to the following considerations: duplicates and newsgroup-identifying headers have been removed; there is no randomness in training and testing set selection so cross-experiment comparison is easier; separating the training and testing sets in time is more realistic. The performance measure is the multi-class classification accuracy.

LIBSVM [5] was employed as the implementation of SVM, with all the parameters set to their default values[3]. LIBSVM uses the "one-vs-one" ensemble method for multi-class classification because of its effectiveness and efficiency [12].

We have tried standard kernels, the NED kernel and the NGD kernel. The linear (LIN) kernel worked better than or as well as other standard kernels (such as the Gaussian RBF kernel) in our experiments, which is consistent with previous observations that the linear kernel usually can achieve the best performance for text classification [8, 15-17, 36]. Therefore the experimental results of standard kernels other than the linear kernel are not reported here.

The text data represented as TF or TF×IDF vectors can be embedded in the multinomial manifold by applying L_1 normalization (9), as described in §2.3. Since kernels that assume Euclidean geometry (including the LIN and NED kernel) often perform better with L_2 normalization, we report such experimental results as well. The NGD kernel essentially relies on the multinomial manifold so we stick to L_1 normalization when using it.

The experimental results[4] obtained using SVMs with the LIN, NED and NGD kernels are shown in Table 1 and 2.

Table 1, Experimental results on the WebKB dataset.

representation	normalization	kernel	accuracy
TF	L_1	LIN	72.57%
		NED	84.39%
		NGD	**91.88%**
	L_2	LIN	89.52%
		NED	90.08%
TF×IDF	L_1	LIN	55.84%
		NED	81.04%
		NGD	**91.42%**
	L_2	LIN	88.23%
		NED	87.96%

[1] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

[2] http://people.csail.mit.edu/people/jrennie/20Newsgroups

[3] We have further tried a series of values {…, 0.001, 0.01, 0.1, 1, 10, 100, 1000, …} for the parameter C and found the superiority of the NGD kernel unaffected.

[4] Our experimental results on the WebKB and 20NG datasets should not be directly compared with most published results because of the difference in experimental settings and performance measures.

Table 2, Experimental results on the 20NG dataset.

representation	normalization	kernel	accuracy
TF	L_1	LIN	24.68%
		NED	69.16%
		NGD	**81.94%**
	L_2	LIN	79.76%
		NED	79.76%
TF×IDF	L_1	LIN	21.47%
		NED	72.20%
		NGD	**84.61%**
	L_2	LIN	82.33%
		NED	82.06%

The NED kernel worked comparably to the LIN kernel under L_2 normalization, and superiorly under L_1 normalization. This observation has not been reported before.

The NGD kernel consistently outperformed its Euclidean counterpart -- the NED kernel, and the representative of standard kernels -- the LIN kernel, throughout the experiments. All improvements made by the NGD kernel over the LIN or NED kernel are statistically significant according to (micro) sign test [36] at the 0.005 level (*P-Value* < 0.005)[5], as shown in Table 3.

Table 3, Statistical significance tests about the differences between the NGD kernel (under L_1 normalization) and the LIN/NEG kernel (under L_2 normalization).

comparison	dataset	representation	*Z*
NGD vs. LIN	WebKB-top4	TF	3.4744
		TF×IDF	4.2500
	20NG-bydate	TF	7.5234
		TF×IDF	7.4853
NGD vs. NED	WebKB-top4	TF	2.6726
		TF×IDF	4.4313
	20NG-bydate	TF	7.6028
		TF×IDF	8.4718

We think the success of the NGD kernel for text classification is attributed to its ability to exploit the intrinsic geometric structure of text data.

5. RELATED WORKS

It is very attractive to design kernels that can combine the merits of generative modeling and discriminative learning. This paper lies along the line of research towards this direction.

An influential work on this topic is the *Fisher kernel* proposed by Jaakkola and Haussler [13]. For any suitably regular probability model $p(\mathbf{x}|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$, the Fisher kernel is based on the Fisher score $U_{\mathbf{x}} = \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}|\boldsymbol{\theta})$ at a single point in the parameter space:

$$k_F(\mathbf{x},\mathbf{x}') = U_{\mathbf{x}}^T I^{-1} U_{\mathbf{x}'} \tag{40}$$

where $I = E_{\mathbf{x}}[U_{\mathbf{x}} U_{\mathbf{x}}^T]$. In contrast, the NGD kernel is based on the full geometry of statistical models.

Another typical kernel of this type is the *probability product kernel* proposed by Jebara et al. [14]. Let $p(\cdot|\boldsymbol{\theta})$ be probability distributions on a space $\mathcal{X}$. Given a positive constant $\rho > 0$, the probability product kernel is defined as

$$k_{PP}(\boldsymbol{\theta},\boldsymbol{\theta}') = \int_{\mathcal{X}} p(\mathbf{x}|\boldsymbol{\theta})^{\rho} p(\mathbf{x}|\boldsymbol{\theta}')^{\rho} d\mathbf{x} = \left\langle p(\cdot|\boldsymbol{\theta})^{\rho}, p(\cdot|\boldsymbol{\theta}')^{\rho} \right\rangle_{L_2} \tag{41}$$

assuming that $p(\cdot|\boldsymbol{\theta})^{\rho}, p(\cdot|\boldsymbol{\theta}')^{\rho} \in L_2(\mathcal{X})$, i.e., $\int_{\mathcal{X}} p(\mathbf{x}|\boldsymbol{\theta})^{2\rho} d\mathbf{x}$ and $\int_{\mathcal{X}} p(\mathbf{x}|\boldsymbol{\theta}')^{2\rho} d\mathbf{x}$ are well-defined (not infinity). For any Θ such that $p(\cdot|\boldsymbol{\theta})^{\rho} \in L_2(\mathcal{X})$ for all $\boldsymbol{\theta} \in \Theta$, the probability product kernel defined by (41) is pd. In a special case $\rho = 1/2$, the probability product kernel is called the *Bhattacharyya kernel* because in the statistics literature it is known as the Bhattacharyya's affinity between probability distributions. When applied to the multinomial manifold, the Bhattacharyya kernel of $\boldsymbol{\theta},\boldsymbol{\theta}' \in \mathbb{P}^n$ turns out to be

$$k_B(\boldsymbol{\theta},\boldsymbol{\theta}') = \sum_{i=1}^{n+1} \sqrt{\theta_i \theta_i'} \tag{42}$$

which is closely related to the NGD kernel given by (21) through

$$k_B = \cos\left(-\frac{1}{2} k_{NGD}\right), \tag{43}$$

though they are proposed from different angles.

The idea of assuming text data as points on the multinomial manifold for constructing new classification methods has been investigated by Lebanon and Lafferty [21, 24]. In particular, they have proposed the *information diffusion kernel*, based on the heat equation on the Riemannian manifold defined by the Fisher information metric [21]. On the multinomial manifold, the information diffusion kernel can be approximated by

$$k_{ID}(\boldsymbol{\theta},\boldsymbol{\theta}') \approx (4\pi t)^{-\frac{n}{2}} \exp\left(-\frac{1}{t} \arccos^2\left(\sum_{i=1}^{n+1} \sqrt{\theta_i \theta_i'}\right)\right). \tag{44}$$

It is identical to the pd kernel $k_{NGD-E}(\boldsymbol{\theta},\boldsymbol{\theta}')$ that is constructed based on the NGD kernel (39), except that the inverse cosine component is squared. Since $-\arccos^2\left(\sum_{i=1}^{n+1} \sqrt{\theta_i \theta_i'}\right) = -d_G^2(\theta,\theta')$ looks not cpd, the kernel $k_{ID}(\boldsymbol{\theta},\boldsymbol{\theta}')$ is probably not pd, according to Lemma 2. Whether it is cpd still remains unclear. While the information diffusion kernel generalizes the Gaussian kernel of Euclidean space, the NGD kernel generalizes the NED kernel and provides more insights on this issue.

Let's look at the NGD kernel, the Bhattacharyya kernel and the information diffusion kernel on the multinomial manifold, in the

[5] Using McNemar's test also indicates that the improvements brought by the NGD kernel are statistically significant.

context of text classification using TF or TF×IDF feature representation. It is not hard to see that the above three kernels all invoke the square-root squashing function on the term frequencies, thus provides an explanation for the long-standing mysterious wisdom that preprocessing term frequencies by taking squared roots often improves performance of text clustering or classification [14].

Although the NGD kernel is not restricted to the multinomial manifold, it may be hard to have a closed-form formula to compute geodesic distances on manifolds with complex structure. One possibility to overcome this obstacle is to use manifold learning techniques [28, 33, 35]. For example, given a set of data points that reside on a manifold, the Isomap algorithm of Tenenbaum et al. [35] estimates the geodesic distance between a pair of points by the length of the shortest path connecting them with respect to some graph (e.g., the k-nearest-neighbor graph) constructed on the data points. In case of the Fisher information metric, the distance between nearby points (distributions) of $\mathcal{X}$ can be approximated in terms of the Kullback-Leibler divergence via the following relation. When $\boldsymbol{\theta}' = \boldsymbol{\theta} + \delta\boldsymbol{\theta}$ with $\delta\boldsymbol{\theta}$ a perturbation, the Kullback-Leibler divergence is proportional to the density's Fisher information [7, 20]

$$D\big(p(\mathbf{x}\,|\,\boldsymbol{\theta}')\,\|\,p(\mathbf{x}\,|\,\boldsymbol{\theta})\big) \overset{\delta\boldsymbol{\theta}\to 0}{=} \frac{1}{2}F(\boldsymbol{\theta})\big(\delta\boldsymbol{\theta}\big)^2. \qquad (45)$$

Another relevant line of research is to incorporate problem-specific distance measures into SVMs. One may simply represent each example as a vector of its problem-specific distances to all examples, or embed the problem-specific distances in a (regularized) vector space, and then employ standard SVM algorithm [9, 27]. This approach has a disadvantage of losing sparsity, consequently they are not suitable for large-scale dataset. Another kind of approach directly uses kernels constructed based on the problem-specific distance, such as the Gaussian RBF kernel with the problem-specific distance measure plugged in [3, 6, 10, 11, 26]. Our proposed NGD kernel on the multinomial manifold encodes a-priori knowledge about the intrinsic geometric structure of text data. It has been shown to be theoretically justified (cpd) and practically effective.

6. CONCLUSIONS

The main contribution of this paper is to prove that the Negative Geodesic Distance (NGD) on the multinomial manifold is a conditionally positive definite (cpd) kernel, and it leads to accuracy improvements over kernels assuming Euclidean geometry for text classification.

Future works are to extend the NGD kernel to other manifolds (particularly for multimedia tasks), and apply it in other kernel methods for pattern analysis (kernel PCA, kernel Spectral Clustering, etc.) [34].

7. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments.

8. REFERENCES

[1] Amari, S., Nagaoka, H. and Amari, S.-I. *Methods of Information Geometry*. American Mathematical Society, 2001.

[2] Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison-Wesley, 1999.

[3] Bahlmann, C., Haasdonk, B. and Burkhardt, H. On-Line Handwriting Recognition with Support Vector Machines -- A Kernel Approach. in *Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR)*, 2002, 49-54.

[4] Berg, C., Christensen, J.P.R. and Ressel, P. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer-Verlag, 1984.

[5] Chang, C.-C. and Lin, C.-J. LIBSVM: a Library for Support Vector Machines. 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] Chapelle, O., Haffner, P. and Vapnik, V.N. SVMs for Histogram Based Image Classification. *IEEE Transactions on Neural Networks*, *10* (5). 1055-1064.

[7] Dabak, A.G. and Johnson, D.H. Relations between Kullback-Leibler distance and Fisher information *Manscript*, 2002.

[8] Dumais, S., Platt, J., Heckerman, D. and Sahami, M. Inductive Learning Algorithms and Representations for Text Categorization. in *Proceedings of the 7th ACM International Conference on Information and Knowledge Management (CIKM)*, Bethesda, MD, 1998, 148-155.

[9] Graepel, T., Herbrich, R., Bollmann-Sdorra, P. and Obermayer, K. Classification on Pairwise Proximity Data. in *Advances in Neural Information Processing Systems (NIPS)*, Denver, CO, 1998, 438-444.

[10] Haasdonk, B. and Bahlmann, C. Learning with Distance Substitution Kernels. in *Proceedings of the 26th DAGM Symposium*, Tubingen, Germany, 2004, 220-227.

[11] Haasdonk, B. and Keysers, D. Tangent Distance Kernels for Support Vector Machines. in *Proceedings of the 16th International Conference on Pattern Recognition (ICPR)*, Quebec, Canada, 2002, 864-868.

[12] Hsu, C.-W. and Lin, C.-J. A Comparison of Methods for Multi-class Support Vector Machines. *IEEE Transactions on Neural Networks*, *13* (2). 415-425.

[13] Jaakkola, T. and Haussler, D. Exploiting Generative Models in Discriminative Classifiers. in *Advances in Neural Information Processing Systems (NIPS)*, Denver, CO, 1998, 487-493.

[14] Jebara, T., Kondor, R. and Howard, A. Probability Product Kernels. *Journal of Machine Learning Research*, *5*. 819-844.

[15] Joachims, T. *Learning to Classify Text using Support Vector Machines*. Kluwer, 2002.

[16] Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. in

Proceedings of the 10th European Conference on Machine Learning (ECML), Chemnitz, Germany, 1998, 137-142.

[17] Joachims, T., Cristianini, N. and Shawe-Taylor, J. Composite Kernels for Hypertext Categorisation. in *Proceedings of the 18th International Conference on Machine Learning (ICML)*, Williamstown, MA, 2001, 250-257.

[18] Kass, R.E. The Geometry of Asymptotic Inference. *Statistical Science*, *4* (3). 188-234.

[19] Kass, R.E. and Vos, P.W. *Geometrical Foundations of Asymptotic Inference*. Wiley, 1997.

[20] Kullback, S. *Information Theory and Statistics*. Wiley, 1959.

[21] Lafferty, J.D. and Lebanon, G. Information Diffusion Kernels. in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2002, 375-382.

[22] Lang, K. NewsWeeder: Learning to Filter Netnews. in *Proceedings of the 12th International Conference on Machine Learning (ICML)*, Tahoe City, CA, 1995, 331-339.

[23] Lebanon, G. Learning Riemannian Metrics. in *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, Acapulco, Mexico, 2003, 362-369.

[24] Lebanon, G. and Lafferty, J.D. Hyperplane Margin Classifiers on the Multinomial Manifold. in *Proceedings of the 21st International Conference on Machine Learning (ICML)*, Alberta, Canada, 2004.

[25] McCallum, A.K. Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering. 1996. *http://www.cs.cmu.edu/~mccallum/bow*.

[26] Moreno, P.J., Ho, P. and Vasconcelos, N. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications. in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver and Whistler, Canada, 2003.

[27] Pekalska, E., Paclik, P. and Duin, R.P.W. A Generalized Kernel Approach to Dissimilarity-based Classification. *Journal of Machine Learning Research*, *2*. 175-211.

[28] Roweis, S.T. and Saul, L.K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, *290* (5500). 2323-2326.

[29] Schoenberg, I.J. Metric Spaces and Positive Definite Functions. *Transactions of the American Mathematical Society*, *44*. 522-536.

[30] Schoenberg, I.J. Positive Definite Functions on Spheres. *Duke Mathematical Journal*, *9* (1). 96?08.

[31] Scholkopf, B. The Kernel Trick for Distances. in *Advances in Neural Information Processing Systems (NIPS)*, Denver, CO, 2000, 301-307.

[32] Scholkopf, B. and Smola, A.J. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[33] Seung, H.S. and Lee, D.D. The Manifold Ways of Perception. *Science*, *290* (5500). 2268-2269.

[34] Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[35] Tenenbaum, J.B., Silva, V.d. and Langford, J.C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, *290* (5500). 2319-2323.

[36] Yang, Y. and Liu, X. A Re-examination of Text Categorization Methods. in *Proceedings of the 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Berkeley, CA, 1999, 42-49.