Generative Kernels for Exponential Families

Arvind Agarwal
Department of Computer Science
University of Maryland
College Park Maryland, USA 20740
hal@umiacs.umd.edu

Hal Daumé III
Department of Computer Science
University of Maryland
College Park Maryland, USA 20740
arvinda@cs.umd.edu

Abstract

In this paper, we propose a family of kernels for the data distributions belonging to the *exponential family*. We call these kernels *generative kernels* because they take into account the generative process of the data. Our proposed method considers the geometry of the data distribution to build a set of efficient *closed-form* kernels best suited for that distribution. We compare our generative kernels on multinomial data and observe improved empirical performance across the board. Moreover, our generative kernels perform significantly better when training size is small, an important property of the generative models.

1 Introduction

Generative models provide a useful statistical language for describing data; discriminative methods achieve excellent classification performance. We define *generative kernels*, a family of kernels built around generative models for use in discriminative classifiers. The key idea of generative kernels is to use the generative model to automatically define a statistical manifold, on which a particular natural divergence (based on the Fisher information metric) can be translated directly into a positive definite kernel. Our approach is applicable to any statistical model belonging to the exponential family, which includes common distributions like the Gaussian and multinomial, as well as more complex models. In order to compute the value of a generative kernel, one only needs to evaluate the Legendre dual of its log-partition function, a well-studied problem in the graphical models literature.

Apart from the geometric perspective, there are other reasons to consider the data distribution when constructing the kernels. It is commonly observed that generative models perform well when only a small amount of data is available [15], especially when the model is a true model of the data distribution [13]. However, as the model becomes less true, or as the amount of data grows, discriminative approaches prevail. Ideally, one would like to take advantage of both methods, and build a hybrid method that performs well for small training data and does not rely too much on the generative process assumption. One way to build such hybrid method is to encode the generative process information in the kernel, therefore providing the kernel method a geometry derived from the data distribution, which in turn can be derived *naturally* from the statistical manifold for that distribution family.

There have been previous efforts to build kernels from generative models: the Fisher kernel [10], the heat kernel[12], the probability product (PP) kernel [11]. The first two consider the generative process by deriving the geometry for the statistical manifold associated with the generative distribution family using the fundamental principle of the *information distance*. Unfortunately, these kernels are intractable to compute exactly even for very simple distributions due to the need to compute the Fisher information metric. The approximations required to compute these kernels result in a function that is not guaranteed to be positive definite [14]. There is another family of kernels named semi-group kernels [7] which has the same expression as generative kernels though both are derived completely differently. Unlike [7], we consider the geometry of the data distribution and reason why it is appropriate to call these kernels generative kernels. Note that semi-group kernels are not generative kernels for general probability distributions

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

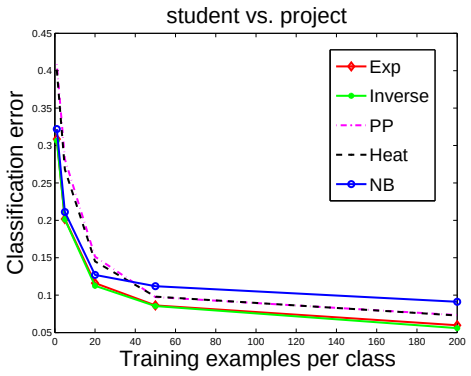

Figure 1: A typical example of relative performance on real dataset from multinomial distribution. *Exp* and *Inverse* are the generative kernels.

therefore empirical study of these kernels under the discriminative/generative paradigm is not considered in [7]. We study generative kernels under generative/discriminative paradigm, in particular, we perform experiments to see what happens when the data generation assumption is violated (noisy data).

Our generative kernels have a number of desirable properties:

- They are applicable to *any* exponential family distribution.
- They are built using the natural geometry of the distribution space.
- They are closed-form, efficient to compute, and by construction are positive definite.
- Empirical comparisons to the best published kernels using the same data and experimental setup yield improved performance.
- They demonstrate that using the geometry of the statistical manifold improves performance, which brings up many open research questions related to the use of geometry in learning algorithms.
- Empirical results with these kernels show that these kernels are able to exploit the generative properties, therefore can be called generative kernels.

Unlike other distribution based kernels, a discriminative method based on the proposed generative kernel is able to exploit the properties of the generative methods i.e. perform well when not enough data. In Figure 1, we show a typical result from a real world classification task on the text data. The blue curve which represents the generative method (Naive Bayes (NB)) performs better when training size n is small but as we increase n, discriminative methods (other curves) start to take over. Although other discriminative methods perform poorer than the NB for small n, discriminative method when used with generative kernel, perform better for all n. We (red and green curves) perform equal/better to NB when n is small, and outperform all other methods as n gets large. Generative kernel curves are lower envelopes of all other curves, giving us the best of both worlds.

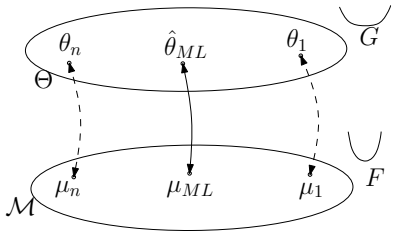

Figure 2: Duality between mean parameters and canonical parameters. Notice the convex functions defined over both spaces. these functions are dual of each other and so are the spaces.

2 Background

In this section. we give the required background, specially, we revisit the concepts related to exponential families and Bregman divergence.

2.1 Exponential Family

The exponential family is a set of distribution, whose probability density function can be expressed in the following form: $p(x;\theta) = p_o(x)\exp(\langle\theta, \phi(x)\rangle - G(\theta))$. Here $\phi(x) : \mathcal{X}^m \rightarrow \mathbb{R}^d$ is a vector *potentials* or *sufficient statistics* and $G(\theta)$ is the *log-partition function*. With the potential functions $\phi(x)$ fixed, every θ induces a particular member $p(x;\theta)$ of the family. In our framework, we deal with the exponential family that are *regular* and have *minimal representation* [17].

One important property of exponential family member is that log-partition function G is convex over the convex set $\Theta := \{\theta \in \mathbb{R}^d : G(\theta) < \infty\}$. Convexity of log-partition function G ensures that there exists a space $\mathcal{M}$, dual to Θ and a dual function F defined over $\mathcal{M}$. Here duality refers to standard Legendre duality. This convexity property also induces a Bregman divergence on both Θ and $\mathcal{M}$. For $\theta_1, \theta_2 \in \Theta$, let $B_G(\theta_1\|\theta_2)$ be the Bregman divergence defined over the space Θ, then $B_G(\theta_1\|\theta_2) = B_G(\theta_1^*\|\theta_2^*) = B_F(\mu_1|\mu_2)$, where $\mu_1, \mu_2 \in \mathcal{M}$ are the conjugate dual of θ_1 and θ_2 respectively. It is to be noted that Bregman divergence

is not symmetric i.e. in general, $B_F(p\|q) \neq B_F(q\|p)$, therefore its is important what directions these divergences are measured in.

Another important property of the exponential family is the *one-to-one* mapping between the *mean parameters* and the *canonical parameters*. For each canonical parameter $\theta \in \Theta$, there exists a corresponding mean parameter $\mu \in \{\mu \in \mathbb{R}^d : \mu = \int \phi(x)p(x;\theta)\,dx \quad \forall\theta \in \Theta\}$ such that$\nabla G(\theta) = \theta^* = \mu$. It can be easily shown that space of mean parameters and dual space of Θ, $\mathcal{M}$ are the same spaces. A pictorial representation of the duality between canonical parameter space Θ and mean parameter space $\mathcal{M}$ is given in Figure 2.

2.2 Statistical Manifolds and Dualistic Structure

We now define statistical manifolds and the dualistic structure associated with them. We in particular give reasons why it is important to choose the KL divergence to define the kernel for the exponential family. A statistical manifold S is a d-dimensional manifold $S = \{\theta \in \Theta\}$ such that every $\theta \in \Theta$ induces a probability distribution over some space $\mathcal{X}$.

Following [1], it is well known that all arbitrary divergences induce a dualistic structure and a metric. In particular, for statistical manifolds, the most natural divergence is the one that induces the Fisher information metric. A divergence function that induces the Fisher information metric on the exponential family manifolds is KL divergence. It is also known as D^{-1} divergence (a special case of D^α divergence for $\alpha = -1$). Since exponential family manifolds have dualistic structure (in fact they are dually flat), there exists a dual space, where one can define the dual divergence i.e. D^1 divergence. Following [18], this duality is called *referential duality*. In referential duality, for two points $p, q \in \Theta$ $D^1(p\|q) = D^{-1}(q^*\|p^*)$ or $KL(p\|q) = KL^*(q^*\|p^*)$. There exists another form of duality (Legendre duality) based on the convex analysis which allows us to define the Bregman divergence, is called *representational duality*, and in such duality, $B_F(p\|q) = B_G(q^*\|p^*)$, where F and G are dual of eachother.

3 Generative Kernels

In this section, we develop a family of generative kernels, In Section 3.1, we first consider the exponential family that generated the data, and transform the maximum likelihood estimation (MLE) problem for the generative model into a *Bregman median* problem in Θ-space. As discussed earlier, a natural divergence in Θ-space is KL divergence. We use this KL divergence to build a metric in Θ-space. Using duality, we project this metric into $\mathcal{M}$-space, and then in Section 3.2, we convert this projected metric into a p.d. kernel.

3.1 Generative Model to Metric

Here, we first consider the generative model. Let $\mathcal{X}$ be the input data space, and Θ be the parameter space such that for each $\theta \in \Theta$, $p(x;\theta)$ is the likelihood of the point $x \in \mathcal{X}$ under distribution given by θ. Now, for a set of i.i.d. observed points $X_n = \{x_1 \ldots x_n\} \subset \mathcal{X}$, the log likelihood of X_n is $\sum_{i=1}^n \log p(x_i;\theta)$. A standard estimate of the "best" parameters for this data is the MLE which solves the following problem:

Definition 1. MLE. *Given a set of data points X_n and a family of distribution $p(x;\theta)$ parametrized by $\theta \in \mathbb{R}^d$, MLE finds $\hat{\theta}_{ML} \in \Theta$ such that $\hat{\theta}_{ML} = \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^n log\, p(x_i;\theta)$.*

It turns out that for exponential family distributions, the MLE problem can transformed into a geometric problem, in particular into a Bregman median problem:

Lemma 1. *Let X_n be a set of n i.i.d. training data points drawn from the exponential family distribution with the log partition function G and F be the dual function of G. Then the dual of the MLE ($\hat{\theta}_{ML}$) of X_n under the assumed exponential family model solves the following Bregman median problem:*

$$\hat{\theta}^*_{ML} = \hat{\mu}_{ML} = \min_{\mu\in\mathcal{M}} \sum_{i=1}^n B_F(\phi(x_i)\|\mu) \tag{1}$$

Proof. The proof is simply a combination of many known facts [1, 6]. For the sake of understanding, we briefly reprove it. For an exponential family distribution, $-\log p(x;\theta) = \log (p_0(x) - \langle\phi(x),\theta\rangle + G(\theta)$ which using the relationships, $F(\nabla G(\theta)) + G(\theta) = \nabla G(\theta)\theta$ and $\nabla F(\nabla G(\theta)) = \theta$, gives $\log p(x;\theta) = \log p_o(x) + F(\phi(x)) - B_F(\phi(x)\|\nabla G(\theta))$. Now using Definition 1,

$$\begin{aligned}\hat{\theta}_{ML} &= \max_{\theta\in\Theta} \sum_{i=1}^n \Big(\log p_o(x_i) + F(\phi(x_i)) - B_F(\phi(x_i)\|\nabla G(\theta))\Big) \\ &= \min_{\theta\in\Theta} \sum_{i=1}^n B_F(\phi(x_i)\|\nabla G(\theta)) \end{aligned} \tag{2}$$

which using $\nabla G(\theta) = \mu$, takes the entire problem into the $\mathcal{M}$-space and gives the desired result. □

We now take the above Bregman median problem from $\mathcal{M}$-space to Θ-space where we will be able to use KL divergence. For this. we need the following result.

Corollary 1 (ML Estimation for Single Point). *Let x_i be the only point observed, then MLE $\hat{\mu}_{i,ML}$ under this observed point is $\phi(x_i)$*

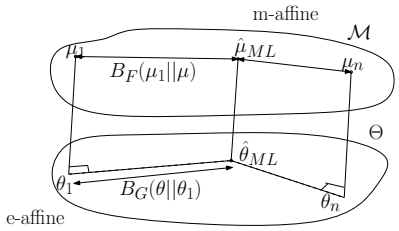

Figure 3: MLE as Bregman median problem into Θ and $\mathcal{M}$ spaces for exponential family distributions. In Θ space, it is a problem over θ_is with optimal θ appearing in the first argument of the divergence function while in $\mathcal{M}$-space, it is a problem over mean parameters with optimal μ appearing in the second argument in the divergence function.

Proof. In Lemma 1, for single point, (1) is minimized when $\hat{\mu}_{i,ML} = \phi(x_i)$. □

Unless otherwise stated, we will use μ_i instead of $\hat{\mu}_{i,ML}$ to make notations less cluttered.

Theorem 1. *Let $\mathcal{M}$ and Θ be dual spaces as defined earlier, θ_i be the MLE of data point x_i under the exponential family parametric model $p(x;\theta)$. Given such $\{\theta_i, \ldots, \theta_i\}$ for all points $\{x_i, \ldots, x_n\}$, $\hat{\theta}_{ML}$ is equivalent to:*

$$\hat{\theta}_{ML} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} B_G(\theta\|\theta_i) \tag{3}$$

Proof. From Corollary 1, $\phi(x_i) = \mu_i$, now replacing this in (1) gives $\hat{\mu}_{ML} = \min_{\mu\in\mathcal{M}} \sum_{i=1}^{n} B_F(\mu_i\|\mu)$. Now for $\mu_1, \mu 2 \in \mathcal{M}$ and $\theta_1, \theta_2 \in \theta$, using the relationship $B_F(\mu_1\|\mu_2) = B_G(\theta_1\|\theta_2)$ takes the entire problem from $\mathcal{M}$-space to Θ-space giving the desired result. □

Figure 3 gives a pictorial summarization of these MLE(s) as Bregman median problems in Θ and $\mathcal{M}$ spaces. Now we transform this Bregman median problem in Θ-space into a KL minimization problem using the following Lemma.

Lemma 2 (KL and Bregman for Exponential Family). *Let $KL(\theta_1\|\theta_2)$ be the KL divergence for $\theta_1, \theta_2 \in \Theta$, then $KL(\theta_1\|\theta_2) = B_G(\theta_2\|\theta_1)$*

Proof. This directly follows from the definitions of KL divergence and exponential family; and from the relation, $\mathbb{E}_\theta(x) = \nabla G(\theta)$ for exponential family.

$$\begin{aligned} KL(\theta_1\|\theta_2) &= \mathbb{E}_{\theta_1}(\log p(x;\theta_1) - \log p(x;\theta_2) \\ &= \mathbb{E}_{\theta_1}\Big(\langle x, \theta_1 - \theta_2\rangle - (G(\theta_1) - G(\theta_2))\Big) \\ &= \langle \nabla G(\theta_1), \theta_1 - \theta_2\rangle - (G(\theta_1 - G(\theta_2)) \\ &= B_G(\theta_2\|\theta_1) \end{aligned}$$

□

Theorem 2. *MLE $\hat{\theta}_{ML}$ of data points X_n generated from exponential family is now given by:*

$$\hat{\theta}_{ML} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} KL(\theta_i\|\theta) \tag{4}$$

Proof. From Lemma 2, $KL(\theta_i\|\theta) = B_G(\theta\|\theta_i)$, substituting this in (3) gives the desired result. □

Note that the MLE $\hat{\theta}_{ML}$ can also be compute from a rather well known relation: $\hat{\theta}_{ML} = \arg\min_{\theta\in\Theta} KL(\mathcal{P}^n\|\theta)$ where $\mathcal{P}^n$ is the empirical distribution of X_n. This relationship holds true for all parametric distributions. In (4), It is worth nothing that parameter being estimated comes in the second argument of KL divergence. This observation along with the following definition is used to construct a metric in Θ-space.

Definition 2 (JS Divergence). *For $\theta_1, \theta_2 \in \Theta$ and $\tilde{\theta} = \frac{(\theta_1+\theta_2)}{2}$, $JS(\theta_1, \theta_2)$ is defined as:*

$$JS(\theta_1, \theta_2) = \frac{1}{2}\Big(KL(\theta_1\|\tilde{\theta}) + KL(\theta_2\|\tilde{\theta})\Big)$$

It is well known that $\sqrt{JS}$ is a metric. $\sqrt{JS}$ have also been shown to be Hilbertian [8, 2]. A metric $d(x, y)$ is said to be Hilbertian metric if and only if $d^2(x, y)$ is a negative definite(n.d.) [16]. Since $\sqrt{JS}$ is a Hilbertian metric, JS is n.d..

It is important to understand the purpose of the above analysis in deriving the metric based on the JS divergence (JS metric). This analysis builds a bridge between the JS metric and the generative models. The JS metric can therefore be used to build kernels that can exploit the generative properties of the data. Establishing the connection between the JS metric and the generative models in theory, and showing the efficacy of the kernels based on this metric in practice, is the main contribution of this work.

The JS metric, which is based on symmetrized KL divergence has been known for a long time [7]. However, what is not known is the generative behavior of the JS metric. In the existing literature, JS metric is usually derived for any probability distribution, and for a general probability distribution, JS metric does not consider the generative model of the data, and therefore can not be used to build the generative kernels. We, in this work, only consider the distributions belonging to the exponential families, for which, the JS metric can be shown to have been derived considering the generative model. This connection between the JS metric and the generative model allows us to build kernels that can be used to build hybrid (discriminative+generative) models.

Definition 3 (Dual JS Divergence). *For $\mu_1, \mu_2 \in \mathcal{M}$ and $\tilde{\mu} = \frac{(\mu_1+\mu_2)}{2}$, let $KL^*(.\|.)$ be the dual of KL divergence then the dual JS divergence $DJS(\mu_1, \mu_2)$ is defined as:*

$$DJS(\mu_1, \mu_2) = \frac{1}{2}\Big(KL^*(\tilde{\mu}\|\mu_1) + KL^*(\tilde{\mu}\|\mu_2)\Big)$$

Theorem 3. *DJS is negative definite.*

Proof. Result is direct consequence of the fact that JS divergence is symmetric. Using the relationship between KL and KL^*, one can simply take the dual of JS which is $DJS(\theta_1, \theta_2) = \frac{1}{2}(KL(\theta_1\|\tilde{\theta}) + KL(\theta_2\|\tilde{\theta})) = JS(\theta_1, \theta_2)$ which is n.d.. □

Theorem 4. *Let $\psi(\mu_1, \mu_2) = DJS(\mu_1, \mu_2)$ be a n.d. function on $\mathcal{M}$, then*

$$\psi(\mu_1, \mu_2) = \frac{F(\mu_1) + F(\mu_2)}{2} - F(\frac{\mu_1 + \mu_2}{2}) \quad (5)$$

Proof. Using the duality, $KL^*(\mu_1\|\mu_2) = B_F(\mu_2\|\mu_1)$; $DJS(\mu_1, \mu_2) = \frac{1}{2}(B_F(\mu_1\|\tilde{\mu}) + B_F(\mu_2\|\tilde{\mu}))$. Expanding the expression for the Bregman divergence and using some algebra yields the result. □

Though not analyzed, this expression is also observed in [5]. It is to be noted that this metric (5) is defined over the mean parameters, so in order to define the metric over the data points X_n, one can use Corollary 1, according to which, $\psi(\mu_1, \mu_2) = \psi(\phi(x_1), \phi(x_2))$

3.2 Metric to Kernel

In this section, we convert the previously constructed Hilbertian metric (n.d. function) into a family of kernels that we will use later in our experiments. We now state some results that can be used along with (5) to build the kernels called "generative kernels". Although there could be many ways [2, 16] to transform a metric into a kernel, we mention a few here:

Proposition 1 (Centering). *Let function $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric function, and $x_0 \in \mathcal{X}$. Let $\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be*

$$\varphi(x, y) = \psi(x, x_0) + \psi(y, x_0) - \psi(x, y) - \psi(x_0, x_0),$$

then φ is n.d. if and only if ψ is positive definite (p.d.).

Proposition 2 (Exponentiated(Exp)). *The function $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is n.d. if and only if $exp(-t\psi)$ is p.d. for all $t > 0$.*

Proposition 3 (Inverse). *The function $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is n.d. if and only if $(t + \psi)^{-1}$ is p.d. for all $t > 0$.*

4 Related Work

In this section, we briefly discuss the kernels that are close to generative kernels. One of the earliest distribution-based kernels is Fisher kernels [10]. These kernels are constructed by taking the inner product in the tangent space of Θ using the Fisher information metric. Although constructed in a principled manner, these kernels are impractical due to the intractability of the Fisher information metric computation.

Other kernels (heat kernels) based on the principle of the Fisher information metric are proposed by [12]. Similar to generative kernels, these kernels also consider the geometry of the statistical manifold. Like Fisher kernels, these kernels also suffer from the intractability issue, and so for even most simple geometries, there is closed-form solution.

[11] propose a family of kernels which is most similar to generative kernels, and at the same time, is fundamentally very different. Similar to our work, they define kernels by considering the data distribution. Their kernel is defined on the parameters space Θ while our kernel is defined on dual of Θ. For exponential family their results reduce to the Bhattacharyya kernels [3] $K(\theta_1, \theta_2) = \exp\Big(G(\frac{\theta_1+\theta_2}{2}) - \frac{G(\theta_1)+G(\theta_2)}{2}\Big)$ while our kernel (Exp form) look like $K(\mu_1, \mu_2) = \exp\Big(F(\frac{\mu_1+\mu_2}{2}) - \frac{F(\mu_1)+F(\mu_2)}{2}\Big)$. In our formulation, we use the dual of the log-partition function F while they use the log-partition function G. These two kernels are very different, one because of the space they are defined over, and other because, there is no explanation as such what divergence (or metric) Bhattacharyya kernels induce on the statistical manifold while generative kernels by construction, induce the Fisher information metric.

There has also been some recent work on the kernels on the probability measures [14, 9] which are very restricted because they are only defined on the probability measures not on the arbitrary spaces. Although we take a completely different approach to derive the generative kernels, it can be shown that for the exponential family distributions, Jensen Tsallis q-kernels [14] reduce to generative kernels for $q = 1$.

As mentioned earlier, generative kernels have the same expression as the semi-group kernels [7] for the exponential family, though both are derived completely differently. We emphasize that semi-group kernels are defined for the general probability distributions by considering the symmetrized KL divergence, and for general probability distributions, these kernels are not the generative kernels because in general, they do not induce the natural divergence (or natural metric) on the statistical manifold. For general probability distributions, KL divergence is not the natural divergence,

therefore use of JS divergence to define the kernel is questionable. Moreover, semi-group kernels do not justify why KL divergence (with $\tilde{\theta}$ in the second argument) is the appropriate divergence. In our work, we take an information geometric approach to derive these kernels and show that these kernels are actually the natural kernel because they induce the Fisher information metric. Semi-group kernels only consider (both in theory and their evaluation) *exp* form of the kernel while we consider a different number of kernels e.g. *exp*, *inv*, *centering*, and study them in discriminative/generative framework. Note that *exp* version usually does not perform as well as others (see Tables 1 and 2).

5 Experiments

In this section we evaluate generative kernels on several text categorization tasks. We use the multinomial distribution for this evaluation for two reasons. First, multinomial is one of the most widely used distributions after Gaussian[1]. Second, other principally similar kernels which we would like to compare against, have been shown to work only for the multinomial geometry, for computational reasons. Note that for multinomial distribution, $\phi(x)$ is simply the observed frequency vector.

We have performed experiments with six kernels: generative(centering, exp and inverse), linear, Heat and PP kernels. For multinomial geometry, one can also use other probability measure based kernels for comparison (e.g. Tsallis-q kernels) but we will not use them because heat kernels are usually more effective [14]. It is emphasized here that our real competitors are heat and PP kernels, mainly because they both are distribution dependent kernels. In order to see the discriminative/generative behavior, we also include the results of a generative model (NB with α-smoothing). For each of these kernels including heat and PP, we report the results for the best parameters. Also, we run our competitors in the best possible settings i.e. heat kernel is applied on the ℓ_2 normalized frequency vector which is known to outperform ℓ_1. In order to make graphs look less cluttered, we exclude generative centering and linear kernels. In most of the cases they underperformed other kernels. For statistical significance, all of the results are averaged over 20 runs. Although in our results, we only report the mean error, variance was found to be very low (~ 0.0001), and hence not reported for the clarity. For evaluation, we use SVM toolbox[4].

[1]Gaussian is uninteresting because all kernels, heat, PP and generative reduce to RBF

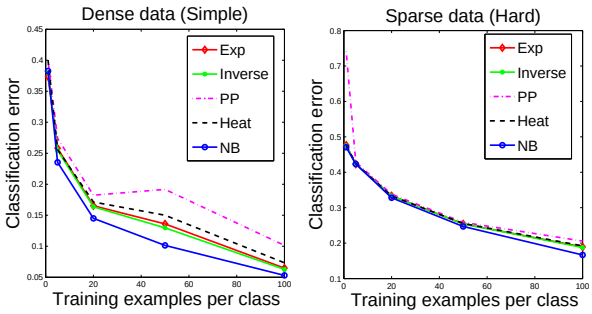

Figure 4: Performance variation with n on random multinomial dataset in *sparse* and *dense* settings

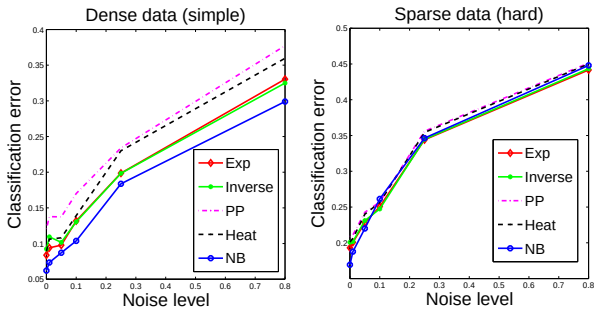

Figure 5: Performance variation with different noise levels on random multinomial dataset in *sparse* and *dense* settings

5.1 Artificial Dataset

In this section, we give a proof of concept by evaluating our kernels on synthetic data. For the multinomial distribution, the relative size of each trial w compared to the dimension of the data d is important because that's what makes problems difficult or easier. if multinomial distributions are considered to be the documents, then long documents compared to the vocabulary size(low d/w, dense setting) makes problem easier while short documents (high d/w, sparse settings) makes the problem difficult. We show results for both dense and sparse settings. For each of these settings, we perform two kind of experiments, In one we vary n (no noise), and in other we vary the noise level($n = 50$). Noise is introduced by copying the result of previous trial with probability p =noise level. In sparse setting $d/w = 100$ while in dense $d/w = 5$ with $w = 20$. These results are shown in Figure 4 and Figure 5 respectively. Since in all of these experiments, we outperform other kernel methods, an important comparison would be to see how our method perform compared to the generative method (NB) mainly because discriminative models based on generative kernels can be thought of as a mixture of discriminative/generative models.

In all of these experiments, results are found to be very interesting and consistent with previous known facts. In the non-noise dense settings, we see that we outperform all other methods except for NB. It is well known [13] that in case of correct model assumption, generative models outperform discriminative methods hence in this case, one can not hope to beat NB, though we beat all other discriminative models for all n. It is also emphasized that when n is small we perform as well as the NB because for small n, generative properties dominate discriminative properties in our method, but as we increase the data, discriminative properties tend to dominate and model starts to perform poor. Similar results are obtained for the sparse setting except that problem is now harder, and for hard problem, relative difference between generative and discriminative is not very high.

An interesting phenomena occurs when we introduce noise. Results are presented in Figure 5. For simple problem (dense setting), NB outperforms all discriminative methods for all noise levels. In simple problem, introducing noise does not make much difference, and problem is still simple enough for NB to perform better therefore, it is the hard problem that is more interesting. In hard problem (sparse setting), NB performs better when there is less noise ($\sim 10\%$), but as we increase the noise, correct model assumption breaks and our method starts to outperform.

5.2 Real Datasets

We now show the results for the real world datasets. We consider two standard datasets WebKB and Reuters[2]. Datasets were preprocessed in a standard manner (short words and stopwords removal, stemmer)[3].

WebKB dataset contains webpages classified into four categories, student, faculty, course and student. We take all four classes and construct binary classification tasks by choosing class pairs, giving us a total of six tasks. For two such tasks, performance variation with n is shown in Figure 6. These are typical results and were found to be consistent among other tasks. Our results are very promising, giving the ideal results for the discriminative/generative hybrid models. Generative kernel curves form the lower envelopes of the discriminative and generative curves. Though, we outperform all models in all cases, we perform significantly better $\sim$ 10%) when n is small. Observe that NB performs better than other discriminative models

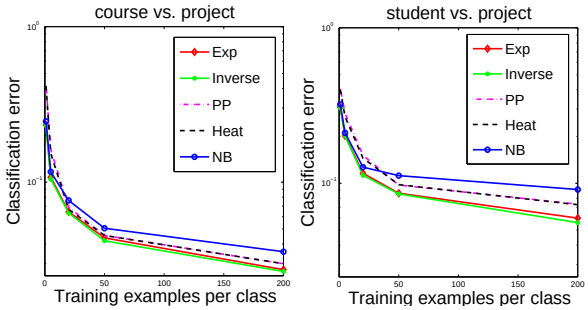

Figure 6: Performance variation with n on *WebKB* dataset. Note that Y-axis is log-scaled hence relative difference is not equal along the X-axis. Although not visible from plot, relative improvement for $n = 1$) is $\sim 10\%$

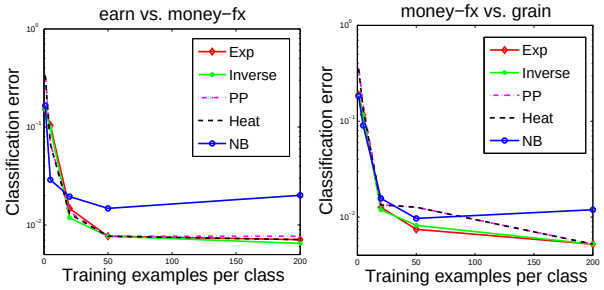

Figure 7: Performance variation with n on *Reuters* dataset. Y-axis is log-scaled.

when n is small, an important property of the generative models [15], which we are able to exploit, and perform better/equivalent to generative models for small n, and better than discriminative methods for large n. Table 5.2 shows the results for all six tasks for $n = 20$. For WebKB dataset, Generative-centering does not perform very good but other generative kernels outperform all methods. In some case i.e. *faculty vs.student* difference is as much as 5%.

Reuters dataset is a collection of newswire articles classified according to the topics. Although there are more than 140 topics, we take topics with largest number of labeled examples: acq, earn, crude, grain and money-fx. We again consider class pair as a binary classification task which gives us a total of 10 tasks. Performance variation with n for two such tasks is shown in Figure 7. Results are similar to the WebKB dataset, generative kernel curves being the lower envelopes of all other curves. Table 5.2 shows the results for all 10 tasks for $n = 20$. We see that generative kernels are able to outperform all other methos but here unlike WebKB dataset, difference among different version of generative kernels is not as high, in fact all of the generative kernels perform almost equally good.

[2]Available at `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/` and `http://www.daviddlewis.com/resources/testcollections/`

[3]`http://web.ist.utl.pt/~acardoso/datasets/`

Task	Generative Kernels			Linear	PP	Heat	Naive
	Centering	Exp	Inverse	Kernel	Kernel	Kernel	Bayes
faculty vs. course	0.1727	0.0443	**0.0413**	0.1268	0.0616	0.0593	0.0719
student vs. project	0.2735	0.1142	**0.1114**	0.2166	0.1386	0.1364	0.1340
course vs. project	0.2120	**0.0602**	**0.0602**	0.1747	0.0699	0.0687	0.0663
faculty vs. project	0.2012	0.1310	**0.1256**	0.2259	0.1539	0.1509	0.2036
faculty vs. student	0.2945	0.1499	**0.1476**	0.2272	0.1926	0.1896	0.2098
student vs. course	0.3165	0.0541	**0.0515**	0.1227	0.0858	0.0819	0.0590

Table 1: Error comparison of generative kernels with other kernels on *WebKB* dataset

Task	Generative Kernels			Linear	PP	Heat	Naive
	Centering	Exp	Inverse	Kernel	Kernel	Kernel	Bayes
acq vs. earn	0.0694	0.0747	**0.0664**	0.0754	0.0673	0.0671	0.0684
acq vs. crude	0.0432	0.0410	**0.0407**	0.0966	0.0472	0.0469	0.0696
acq vs. money-fx	0.0074	0.0078	**0.0071**	0.0596	0.0110	0.0106	0.0191
earn vs. grain	0.0119	0.0142	**0.0116**	0.0776	0.0142	0.0123	0.0164
grain vs. money-fx	0.0138	0.0134	**0.0131**	0.0687	0.0198	0.0194	0.0265
acq vs. grain	**0.0131**	0.0153	0.0134	0.0791	0.0168	0.0168	0.0194
crude vs. money-fx	**0.0082**	0.0096	0.0089	0.0394	0.0121	0.0117	0.0209
earn vs. money-fx	**0.0117**	0.0142	0.0121	0.0567	0.0142	0.0135	0.0199
earn vs. crude	**0.0326**	0.0366	0.0329	0.0814	0.0363	0.0348	0.0450
grain vs. crude	**0.0179**	0.0194	0.0183	0.0705	0.0246	0.0246	0.0448

Table 2: Error comparison of generative kernels with other kernels on *Reuters* dataset.

6 Conclusion and Future work

We have proposed a family of *generative kernels* for exponential family distributions, based on the principle of information distance. These kernels are simple and have a closed-form. Empirically, they outperform other kernels based on probability distributions, especially, they are able to exploit the properties of the generative process therefore, perform significantly well for small training data.

This work opens up many research questions and can be extended in many ways. One is to understand the behaviors of these kernels: There exists a unique natural Hilbertian metric for a given statistical manifold but there is no unique kernel. Because of the multiple mappings from metric to kernel, one can build many kernels and it has been observed that not all of these kernels perform equally. Another direction for the future work would be to understand the fundamental geometry of these statistical manifold and to use them in problems other than classification i.e. clustering. Although in principle, these kernels can be used for any exponential family models, one has yet to establish the practical evidence. One future work of this could be to use these kernels together with exponential families derived from complex graphics model structures in the context of classification.

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry (Translations of Mathematical Monographs)*. American Mathematical Society, April 2001.

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

[3] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by probability distributions. *Bull. Calcutta Math Soc.*, 1943.

[4] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy. SVM and kernel methods matlab toolbox. 2005.

[5] P. Chen, Y. Chen, and M. Rao. Metrics defined by Bregman divergences. *Comm. in Mathematics Sciences*, 6(4):915–926, 2008.

[6] M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal component analysis to the exponential family. In *NIPS 14*. MIT Press, 2001.

[7] M. Cuturi, K. Fukumizu, and J.-P. Vert. Semigroup kernels on measures. *J. Mach. Learn. Res.*, 6:1169–1198, December 2005.

[8] B. Fuglede and F. Topsoe. Jensen-Shannon divergence and Hilbert space embedding. In *IEEE International Symposium on Information Theory*, pages 31–31, 2004.

[9] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability. In *AISTATS*, 2005.

[10] T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS 98*, pages 487–493, Cambridge, MA, USA, 1999. MIT Press.

[11] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *J. Mach. Learn. Res.*, 5:819–844, 2004.

[12] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. *J. Mach. Learn. Res.*, 6:129–163, 2005.

[13] P. Liang and M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *ICML*, 2008.

[14] A. F. T. Martins, M. A. T. Figueiredo, P. M. Q. Aguiar, N. A. Smith, and E. P. Xing. Nonextensive entropic kernels. In *ICML*, 2008.

[15] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. 2001.

[16] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[17] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. Technical report, University of California, Berkeley, 2003.

[18] J. Zhang. Divergence function, duality, and convex analysis. *Neural Comput.*, 16(1):159–195, 2004.