2nd International Symposium on
Information Geometry and its Applications
December 12-16, 2005, Tokyo
Pages 000–000

# INFORMATION GEOMETRY, THE EMBEDDING PRINCIPLE, AND DOCUMENT CLASSIFICATION

GUY LEBANON

Abstract. High dimensional structured data such as text and images is often poorly understood and misrepresented in statistical modeling. Typical approaches to modeling such data involve, either explicitly or implicitly, arbitrary geometric assumptions. In this paper, we review a framework introduced by Lebanon and Lafferty that is based on Čencov's theorem for obtaining a coherent geometry for data. The framework enables adaptation of popular models to the new geometry and in the context of text classification yields superior performance with respect to classification error rate on held out data. The framework demonstrates how information geometry may be applied to modeling high dimensional structured data and points at new directions for future research.

## 1. Introduction

With the dramatic growth of databases such as the internet, there has been a surge in applications that model high dimensional structured data. Such data is structured in the sense that it appears in a form that is different from points in $\mathbb{R}^n$. The typical approach to modeling such data is to use features $f = (f_1, \ldots, f_n)$, $f : \mathcal{X} \to \mathbb{R}^n$ or sufficient statistics to embed the data in $\mathbb{R}^n$. For example, typical features for text documents are the relative frequency of dictionary words in the document.

The statistical approach to modeling such data continues with hypothesizing a parametric family of distributions $\{p_\theta : \theta \in \Theta\}$ where $p_\theta$ is either generative $p_\theta(x)$ or conditional $p_\theta(y|x)$ as is the case in classification, and depends on the data $x$ through the features $f(x)$. The classical approach often proceeds with selecting a single model $p_{\hat{\theta}}$ that corresponds to a point estimate $\hat{\theta}$ such as the maximum likelihood estimator.

As representation of structured data is far from being a solved problem, it is not clear what form should the model $p_\theta(y|x) = p_\theta(y|f(x))$ take (in this paper we concentrate on conditional modeling or classification). In most cases, standard models such as a mixture of Gaussians or logistic regression for $p_\theta(y|x)$ are used. A key observation is that in forming such models, assumptions are being made, either explicitly or implicitly concerning the geometry of the data $x$ (or of the features $f(x)$). For example, it is relatively obvious that using a mixture of Gaussians with $\Sigma_1 = \Sigma_2 = \sigma^2 I$ as the conditional model assumes Euclidean geometry for the data. If the covariance matrix $\Sigma$ is not a multiple of the identity matrix, an alternative geometry expressed as a normed vector space is assumed. It is less obvious that similar assumptions are implicitly made for popular conditional models such as logistic regression, support vector machines (SVM), and boosting[1]. In fact, logistic regression, boosting, linear kernel SVM, and RBF kernel SVM all assume Euclidean geometry on the data (this statement will be partially motivated in Section 3).

[1]While SVM and boosting are, strictly speaking, not conditional distributions, we view them as non-normalized conditional models. See for example [8] for more details.

The assumption of Euclidean geometry mentioned above is rather arbitrary. There is no reason for believe that word frequencies in documents or pixel brightness values in images should be modeled using Euclidean geometry. Such an assumption is driven by the familiarity of Euclidean models and perhaps computational efficiency, rather than being a statement about the data. This approach, sometimes called "the dirty laundry of machine learning", treats popular models such as SVM or logistic regression as a black box that processes structured data regardless of the data origin.

A more motivated approach is to obtain a domain dependent geometry for the data that would lead to alternative models. This geometry may be obtained in several ways. It could be specified by experts familiar with the specific domain. However, it is typically a hard task, even for experts, to specify such a geometry. Alternatively, the geometry can be adapted based on known a data set, as in [7, 10] or on domain-dependent axiomatic arguments as in [4, 5].

In this paper we review a framework described in a series of papers by Lebanon and Lafferty [9, 4, 6] for obtaining domain-dependent geometry and adapting existing classification models for this geometry. Section 2 discusses the embedding principle for obtaining the geometry. Section 3 describes adapting the conditional models of radial basis kernel SVM and logistic regression to alternative geometries. Section 4 concludes the paper with a brief discussion.

## 2. The Embedding Principle

As mentioned in the introduction, we would like to avoid arbitrary assumptions on the geometry of the data. Čencov's theorem [3] (see also its extensions in [2, 5]) provides a theoretical motivation for the use of the Fisher information metric on a manifold $\Theta$ of distributions. At first glance it is not clear how this can contribute towards obtaining a well-motivated data geometry. Data such as documents and images are not distributions and therefore their space is not similar to a statistical manifold $\Theta$. The embedding principle, formulated in [6], gets around this difficulty by embedding the data in a statistical manifold as follows.

The Embedding Principle: Assume that the data $x_1, \ldots, x_n$ is drawn from $n$ distinct distributions $p(x\,;\theta_1^{\text{true}}), \ldots, p(x\,;\theta_n^{\text{true}})$ that lie in the same family $\theta_1^{\text{true}}, \ldots, \theta_n^{\text{true}} \in \Theta$. As the sampling simulates noisy corruption, we can replace the data $x_1, \ldots, x_n$ with the underlying distributions $\theta_1^{\text{true}}, \ldots, \theta_n^{\text{true}}$ thus obtaining an embedding of the data in a Riemannian manifold $(\Theta, g)$ where $g$ is the Fisher information metric.

In most cases, we do not know the underlying distributions $\theta_1^{\text{true}}, \ldots, \theta_n^{\text{true}}$. An approximate version of the embedding principle is to embed the data in $(\Theta, g)$ using estimates $\hat{\theta}_1, \ldots, \hat{\theta}_n$ obtained by estimators such as MLE, MAP or empirical Bayes. Accounting for embedding error resulting from inaccurate estimates and a Bayesian version of the embedding principles are interesting directions for future research. A diagram illustrating the embedding principle appear in Figure 1.

A straightforward example for the embedding principle is the popular term frequency (tf) representation for documents. In this representation, the word order is ignored and a document is represented by a vector of the relative frequencies of words (or histogram)

$$x_i = \frac{\#\text{ times word } i \text{ appeared in document}}{\sum_j \#\text{ times word } j \text{ appeared in document}}.$$

This representation may be interpreted as the MLE approximate embedding of documents under a multinomial distribution and embeds the data in the multinomial

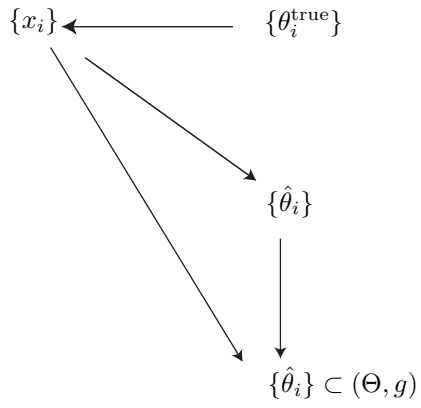


FIGURE 1. Motivation for the embedding principle. See text for more details.

simplex

$$\mathbb{P}_m = \left\{ \theta \in \mathbb{R}^{m+1} : \forall i \ \theta_i > 0, \sum_j \theta_j = 1 \right\}.$$

The case of zero word occurrences presents some technical difficulties as the resulting embedding is in the closure of the simplex which is not technically a manifold or even a manifold with a boundary, but rather a manifold with corners. However, this may be solved easily in a number of ways, for example by replacing the MLE with the MAP under a Dirichlet prior which results in smoothing the zero values.

Applying the embedding principle to data generated from Gaussian distributions is less straightforward. The MLE, for example, would result in a degenerate embedding possessing zero variance everywhere. A possible solution, described in [6], is to consider embedding functions that operate on the entire data set $\hat{\theta} : \mathcal{X}^n \to \Theta^n$ in a way that is not decomposable to independent embeddings. One example, illustrated in Figure 2, describes embedding in the hyperbolic upper half space representing spherical Gaussians using sampling from the posterior of a Dirichlet process mixture model (DPMM) (see [6] for more details).

## 3. MODELING

Once the question of which geometry to use is resolved we need to focus on ways of using it in classification. We describe here the geometric adaptation of three models described in [4, 9, 6]. We start with the simplest case of nearest neighbor, then proceed to radial basis function support vector machines and logistic regression.

Adapting a $k$-nearest neighbor classifier simply requires switching the way of computing the distances. Instead of Euclidean distance $\|x-y\|^2$ we use the geodesic distance, for example in the case of the Fisher geometry on the multinomial simplex

$$d(\hat{\theta}(x), \hat{\theta}(y)) = \arccos\left( \sum_i \sqrt{[\hat{\theta}(x)]_i [\hat{\theta}(y)]_i} \right).$$

The simplex is illustrated in Figure 3 while the decision boundaries for both geometries on $\mathbb{P}_m$ is illustrated in Figure 4.

A more effective classification model is support vector machine (SVM). SVM requires the definition of a function $K(x, y)$ known as a Mercer kernel function.

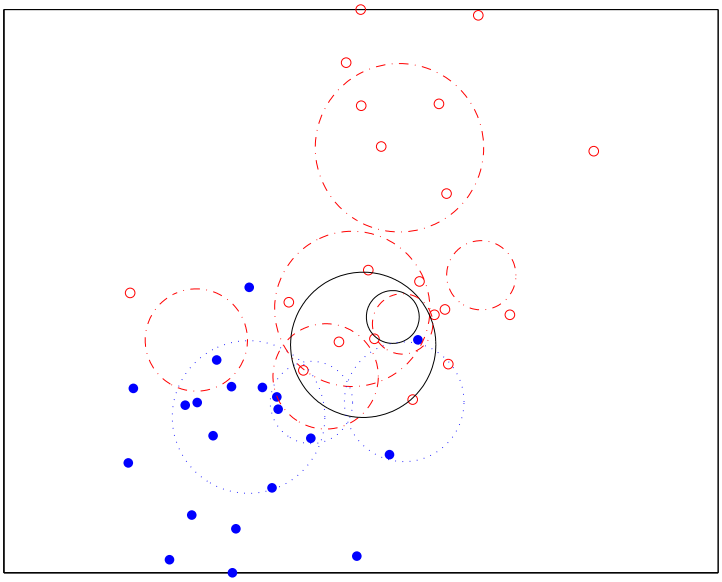

FIGURE 2. A sample from the posterior of a DPMM based on data from two Gaussians $N((-1,-1)^\top, I)$ (solid blue dots) and $N((1,1)^\top, I)$ (hollow red dots). The embedding is realized by the displayed circles representing points in the upper-half plane $\Theta = \mathbb{R}^2 \times \mathbb{R}_+$ parameterizing spherical Gaussian distributions.

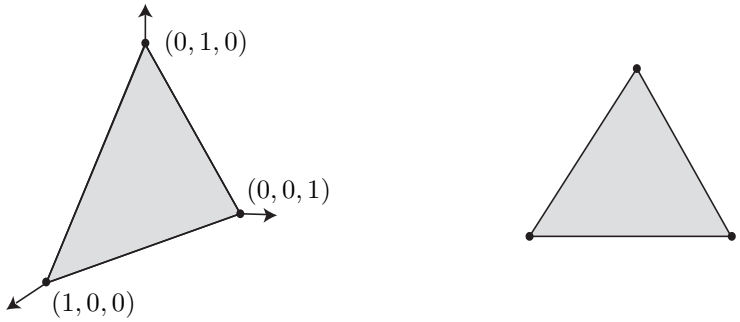


FIGURE 3. The 2-simplex $\mathbb{P}_2$ may be visualized as a surface in $\mathbb{R}^3$ (left) or as a triangle in $\mathbb{R}^2$ (right).

Mercer kernels serve as a measure of similarity between points and yet are fundamentally different from distances. They have to satisfy certain properties such as symmetry and positive definiteness. One of the most popular kernels is the radial basis function (RBF) kernel $K_\sigma(x,y) = \exp(-\|x-y\|^2/\sigma)$ which obviously incorporates the Euclidean assumption. Lafferty and Lebanon [4] generalized it to arbitrary Riemannian manifolds by observing that the RBF kernel is the heat

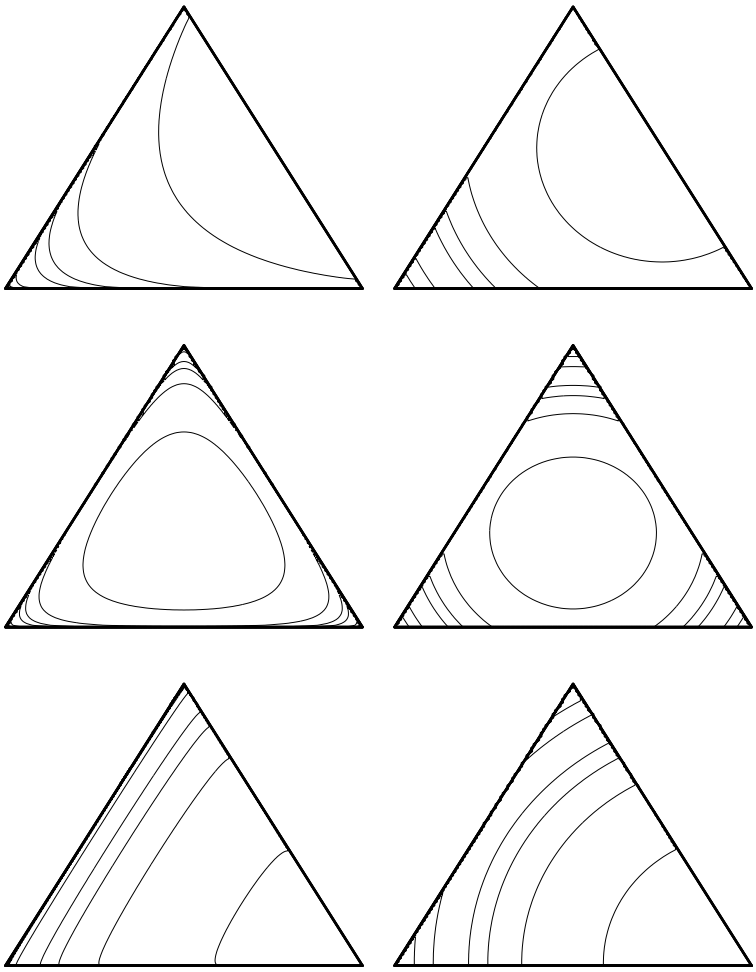

FIGURE 4. Equal distance contours on $\mathbb{P}_2$ from the upper right edge (top row), the center (center row), and lower right corner (bottom row). The distances are computed using the Fisher information metric (left) or the Euclidean metric (right).

kernel for Euclidean geometry. In other words, it is the kernel associated with the heat equation $\triangle f = \partial f / \partial t$ under the Euclidean Laplacian $\triangle$. Intuitively, $K_\sigma(x, y)$ measures the amount of heat flowing from $x$ to $y$ after time $t$, where heat flows according to Euclidean geometry. As both the Laplacian and the heat equation are well defined for any Riemannian manifold, we can easily generalize the RBF kernel (at least conceptually) and use it in a support vector machine for the embedded points $K_\sigma(x, y) = K_\sigma(\hat{\theta}(x), \hat{\theta}(y))$.

Often, the heat kernel does not have a closed form expression. Lafferty and Lebanon [4] propose to use the parametrix expansion of the heat kernel in order to obtain a closed form approximation. This leads to the following approximation in the case of the Fisher geometry on the simplex

$$K_t(x, y) = \exp\left(-\frac{1}{t} \arccos^2\left(\sum_i \sqrt{[\hat{\theta}(x)]_i [\hat{\theta}(y)]_i}\right)\right).$$

The case of linear classifiers such as logistic regression is somewhat more complex. We treat below in general the case of margin based linear classifiers and focus on the special case of logistic regression described in [9].

Linear classifiers such as logistic regression and boosting have the following linear decision rule $\hat{y} = \text{sign}(\sum_i w_i x_i) \in \{-1, +1\}$. More interestingly, from a geometric perspective the resulting decision boundary is a linear hyperplane $\{x : \sum_i w_i x_i = 0\}$ in $\mathbb{R}^n$. Such a decision boundary may be motivated by several arguments. One such argument is that the regularity of linear hyperplanes strikes a good tradeoff in the following conflict. On one side, the model family need to be rich enough to allow good fit to the data while on the other hand, it should be restricted to prevent over-fitting. A second argument for linear hyperplane classifiers is that they are the optimal classifiers for separating two Gaussians with equal covariance matrices. A problem with these and other arguments is that they all rely on the assumption of Euclidean geometry. If a different geometry is to be used, we would like to maintain the above motivations for linear classifiers, generalized to the alternative geometry. The natural generalization to Riemannian manifolds is

**Definition 1.** A linear decision boundary $N$ in a Riemannian manifold $M$ is an auto-parallel sub-manifold of $M$ such that $M \setminus N$ has two connected components.

Auto-parallelism is a geometric property that enforces flatness of $N$ with respect to the metric connection[2] of $M$. It is equivalent to the requirement that every geodesic in $M$ between two points in $N$ lie completely in $N$. The first part of the definition thus enforces a regularity condition on $N$ making it flat in an analogous way to the flatness of hyperplanes in Euclidean geometry while the second part requires $N$ to be a decision boundary.

A second issue that we need to address is how to select a specific classifier from the class of linear decision boundaries. Margin classifiers, including linear SVM, boosting and logistic regression make use of the concept of margins in their training. A geometry generalization of the margin concept leads to the following definition

**Definition 2.** The margin of a point $x$ with respect to a sub-manifold $N$ is

$$d(x, N) = \inf_{y \in N} d(x, y).$$

Conceptually all the linear margin classifiers described above may be generalized to arbitrary manifolds using Definitions 1 and 2 above. However, computing the margin and identifying linear decision boundaries may be a complicated or impractical matter. We proceed with some details concerning the adaptation of logistic regression to the Fisher geometry on $\mathbb{P}_m$.

A well-known device for working with the Fisher geometry of the simplex is the local isometry $\iota : \mathbb{P}^m \to \mathbb{S}^+_m$, $\iota(x) = (\sqrt{x_1}, \ldots, \sqrt{x_{m+1}})$ where $\mathbb{S}^+_m$ is the positive sphere

$$\mathbb{S}^+_m = \left\{ x \in \mathbb{R}^{m+1} : \forall i \; x_i > 0, \sum_i x_i^2 = 1 \right\}$$

equipped with the local Euclidean metric $g(u, v) = \langle u, v \rangle = \sum_i u_i v_i$. Using the above observation we can easily identify the linear decision surfaces on the simplex as intersections of $\mathbb{S}^+_m$ and $m$-dimensional subspaces $E$ of $\mathbb{R}^{m+1}$ – pulled back through the inverse isometry $\iota^{-1}$ to the simplex. Moreover, using techniques such as the spherical law of cosines the margin $d(x, \iota^{-1}(\mathbb{S}^+_m \cap E))$ may be efficiently approximated [9].

[2]Auto-parallelism may also be enforced with respect to other connections, for example $\alpha$ connections [1]. We concentrate here on the metric connection as in this case the motivation from Čencov's theorem is strongest.

To apply the above to logistic regression, we need to present it in a form that exposes it as a linear margin classifier. Denoting the parameter vector by $\eta$ the parametric form of logistic regression, with $y \in \{-1, 1\}$, is

$$\begin{aligned} p_\eta(y \mid x) \propto \exp(y\langle x, \eta\rangle) &= \exp\left(y \left\|\eta\right\| \langle x, \hat{\eta}\rangle\right) = \exp\left(\left\|\eta\right\| \cdot \ y\mathrm{sign}(\langle x, \hat{\eta}\rangle) \cdot \ |\langle x, \hat{\eta}\rangle|\right) \\ &= \exp\left(\left\|\eta\right\| \ y\mathrm{sign}(\langle x, \hat{\eta}\rangle) \ d(x, H_{\hat{\eta}})\right) \end{aligned}$$

where $\hat{\eta}$ is the unit norm vector proportional to $\eta$, $H_{\hat{\eta}}$ is the decision boundary which is the subspace perpendicular to $\hat{\eta}$, and $d(x, H_{\hat{\eta}})$ is the Euclidean margin. Furthermore, $s(x, \hat{\eta}) = y\mathrm{sign}(\langle x, \hat{\eta}\rangle)$ is $-1$ if $x$ is misclassified (lies on the wrong side of $H_{\hat{\eta}}$) and $+1$ if $x$ is correctly classified. Finally, we can decompose the parameter vector $\eta$ into a unit-length parameter vector $\hat{\eta}$ and a positive parameter $\theta$ leading to a geometric viewpoint of logistic regression

$$p_{\hat{\eta},\theta}(y \mid x) = \exp\left(\theta s(x, \hat{\eta}) d(x, H_{\hat{\eta}}) - \log \phi\right)$$

where $\phi$ is the normalization term. The above representation exposes several important facts concerning logistic regression. The conditional probability decreases or increases exponentially with the margin. The parameter $(\hat{\eta}, \theta)$ is composed of a unit vector which defines the decision boundary and a positive parameter $\theta$ that controls the aggressiveness of the exponential increase of decrease. The above forms finds its way into the likelihood function and demonstrates the two Euclidean-geometric assumptions of logistic regression: the Euclidean subspace decision boundaries and the Euclidean margin. Generalizing the above to arbitrary manifolds is straightforward. We simply replace $\hat{\eta}$ with whatever parameterizes the class of linear decision boundaries in the Riemannian manifold $M$ (see Definition 2) and $d(x, H_{\hat{\eta}})$ with the Riemannian version of the margin (see Definition 1). With the the above modifications, logistic regression on a Riemannian manifold becomes a parametric family of conditional distributions that may be treated using standard techniques such as conditional MLE for point estimation. An illustration of the obtained MLE for logistic regression with both Euclidean geometry and the Fisher geometry on the simplex is provided in Figure 5. Consult [9] for more details concerning Fisher geometry-logistic regression on the simplex.

## 4. Discussion

The motivation for the above work is the standard practice of using Euclidean geometry for data modeling. Although unmotivated and arbitrary, this approach has been the dominating one in data modeling. In proposing an alternative, two questions have to be addressed. The first question, which geometry to use, is answered to some extent by the embedding principle. The second question, what models to use, is answered by adapting existing models to the obtained geometry.

We covered the adaptation of RBF kernels SVM and logistic regression to the Fisher geometry on $\mathbb{P}_m$. By treating text documents as multinomial distributions, we can apply these extension to the problem of classification of text documents. Extensive experimental results, conducted by Lebanon and Lafferty [4, 9, 6], conclude that the Fisher adapted models significantly outperform their Euclidean counterparts thus leading to superior text classifiers.

The work described in this paper leads to practical applications of information geometry to real-world data. In general, modeling of high-dimensional structured data is poorly understood and performed. Additional research in the direction of linking information geometry with modeling would result in two important goals. It would provide new effective modeling techniques for the application areas, and it would introduce information geometry to a new arena where it can display its effectiveness.

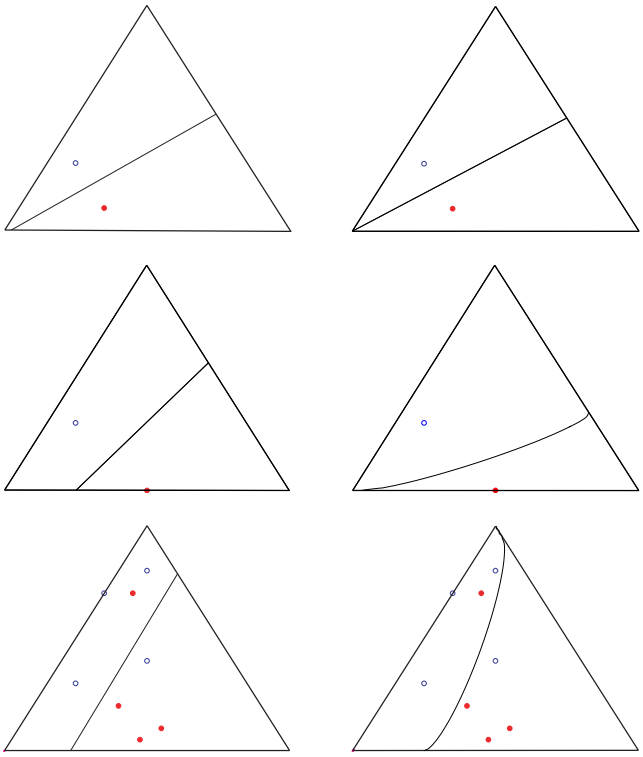

FIGURE 5. Experiments contrasting the decision boundary obtained by MLE for Euclidean logistic regression (left column) with multinomial logistic regression (right column) for toy data in $\mathbb{P}^2$.

## REFERENCES


[1] Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2000.

[2] L. L. Campbell. An extended Čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98(1):135–141, 1986.

[3] N. N. Čencov. *Statistical Decision Rules and Optimal Inference*. American Mathematical Society, 1982.

[4] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6:129–163, 2005.

[5] Guy Lebanon. Axiomatic geometry of conditional models. *IEEE Transactions on Information Theory*, 51(4):1283–1294, 2005.

[6] Guy Lebanon. *Riemannian Geometry and Statistical Machine Learning*. PhD thesis, Carnegie Mellon University, Technical Report CMU-LTI-05-189, 2005.

[7] Guy Lebanon. Metric learning for text classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.

[8] Guy Lebanon and John Lafferty. Boosting and maximum likelihood for exponential models. In *Advances in Neural Information Processing Systems, 14*. MIT press, 2002.

[9] Guy Lebanon and John Lafferty. Hyperplane margin classifiers on the multinomial manifold. In *Proc. of the 21st International Conference on Machine Learning*. ACM press, 2004.

[10] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russel. Distance metric learning with applications to clustering with side information. In *Advances in Neural Information Processing Systems, 15*. MIT Press, 2003.



DEPARTMENT OF STATISTICS, AND SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING, PURDUE UNIVERSITY - WEST LAFAYETTE, IN, USA

*E-mail address*: lebanon@stat.purdue.edu

*URL*: http://www.stat.purdue.edu/~lebanon