Information Geometric Perspectives to ML

Georg Schonherr **Mrinmaya Sachan** **Shashank Srivastava**
{georg, mrinmays, shashans}@cs.cmu.edu

1 Introduction

Statistical learning entails the choice of an optimal model θ from a larger ambient model space Θ based on training samples from a data space X. Many learning algorithms make implicit assumptions about the geometry of these spaces. Some learning techniques assume spaces to be Euclidean. However, in real life, data and model spaces rarely behave as Euclidean spaces [13]. For example, vector space intuitions of addition and scaling of vectors don't often make sense when vectors denote data-points. For example, gene expression values in microarray experiments can be represented as numeric vectors, but don't scale linearly. Additionally these values might be constrained to lie in a bounded space rather than the entire real line (for example, in the microarray example, expression readings are non-negative), and these notions are flouted by vector space notions of vector addition and scaling (say by a negative scalar). Similarly, parametric families of models typically are not Euclidean spaces. Rather, different geometries could be appropriate for different models and data spaces and by studying these geometries, we can develop a better understanding of various learning algorithms. Information geometry provides an attractive alternative theory of analysis for statistical-spaces that are not naturally Euclidean.

Despite the fact that most data and model spaces are not Euclidean, they share two important properties: they are smooth and are locally Euclidean. This paves the way for a Riemannian geometric approach [7] [21] [8] [3] (which is a natural way to investigate the differential geometric structure of families of probability distributions) to study the geometries of X and Θ.

We begin the review with an introduction to Riemannian Geometry. In Section 3, we relate Riemannian Geometry to Statistical model spaces. In Section 4, we review some of the implications of information theoretic perspectives to specific learning algorithms. Here, we see that logistic regression and AdaBoost solve the same primal optimization problem with the only difference being that AdaBoost lacks normalization constraints, hence resulting in non-normalized models. We also look at statistical inference in general, and then try to understand the reasoning behind conjugate priors in bayesian statistics using information geometry. Next, we study a generalization of the notion of linear classifiers to non-Euclidean geometries[18]. Finally, we wrap up with a discussion on metric learning from data using the locally Euclidean assumption and conclude.

2 Differential Geometry Fundamentals

Here, we will introduce some key concepts of differential Geometry. Note that some of these concepts are not given in their most general form. We will restrict ourselves to definitions which are useful in the applications discussed in this document.

A *coordinate system* can be viewed as a function from a set to an open subset of $\mathbb{R}^n$, for some n. Intuitively, thiis can be seen as taking a set and arranging its points inside $\mathbb{R}^n$ in a smooth fashion.

A *manifold* is a set with a coordinate system. Given a manifold $\mathcal{M}$ and some coordinate system ϕ mapping into $\mathbb{R}^n$, we are interested in the set of all coordinate systems ψ for $\mathcal{M}$ such that ψ also maps into $\mathbb{R}^n$ and $\psi \circ \phi^{-1}$ is a smooth bijection between the images of ϕ and ψ. These coordinate systems form an equivalence class. The best way to think about a manifold is simply as an open subset of $\mathbb{R}^n$ that can be reparametrized smoothly. We are interested in studying properties that are

invariant under such reparametrization. One such property is its topology. If we view $\mathcal{M}$ as a subset of $\mathbb{R}^n$, it inherits the canonical topology. Continuous (and hence smooth) transformations preserve the topology. From now on, we will always assume that the manifold is given by coordinates with respect to some coordinate system.

A *diffeomorphism* F is a smooth coordinate transformation (such as the $\psi \circ \phi^{-1}$ described earlier).

A *curve* is a smooth function from an interval of $\mathbb{R}$ to the manifold. It is not quite enough to view a curve as the "squiggly line" that is defined defined by the image of that function. The function also contains information about the "density" of that squiggly line, i.e. the speed at which we move along it when we trace out the interval.

The *tangent space* $T(p)$ at a point p in a manifold is the set of all differential operators t of form $t(f) = \lim_{h\to 0} \frac{1}{h}(f(p + hd) - f(p))$ for some d. We see that $T(p)$ is a vector space. Since the manifold is an open subset of $\mathbb{R}^n$, these operators are all well-defined for smooth f because letting h go to 0 will ensure that $p + hd$ with lie inside the manifold for small enough h. Diffeomorphisms induce a full-rank linear transformation on the tangent space at each point through the chain rule.

$$(x_i)_p^{\psi} = (\frac{\partial \phi_j}{\partial \psi_i})_p (x_j)_p^{\phi}$$

Here the partial derivatives represent the local shift in one coordinate system with respect to a shift in the other. We also use the Einstein summation convention.

Even though the tangent space is a set of operators, all vector space properties hold regardless of what each point in that vector space represents. Hence, we see that this space also has a canonical orthonormal basis $(\mathbf{e}_1, .., \mathbf{e}_n)$. Elements of the tangent space are called *tangent vectors*.

In particular, we can define, in each tangent space, an inner product $< \cdot, \cdot >_p$. Any inner product on a finite-dimensional vector space has a bilinear form $x^T G_p y$, with x and y tangent vectors and G_p a positive semi-definite, symmetric matrix. A *Riemannian metric* is such an inner product defined at each p such that G_p is also positive definite and varies smoothly with p. Again, a diffeomorphism affects G_p through the chain rule: $G_p^{\psi} = F'^T(p) G_p^{\phi} F'(p)$, with F' being the matrix of partial derivatives of the diffeomorphism. One major implication of a Riemannian metric is to redefine the notion of length on the manifold. Given a curve γ from the interval $[a, b]$ to $\mathcal{M}$, we can define its length asthe integral of the length of the infinitesimal bits of the curve measured in its local Riemannian metric:

$$||\gamma|| = \int_a^b \sqrt{(\frac{d\gamma}{dt}(t))^T G_{\gamma(t)} \frac{d\gamma}{dt}(t)} dt$$

We give an example to illustrate this notion. Given the upper half of the unit sphere in $\mathbb{R}^3$, we can index it with a two-dimensional coordinate vector by projecting it onto the xy-plane. This turns it into a 2-dimensional manifold that takes the shape of the unit disk. Now take a curve on the unit disk. We would like its length in the manifold to be equal to its Euclidean length when projected back onto the half-sphere. To make this true, we introduce a Riemannian metric on the disk. In fact, there exists a unique such metric for this example and any other embedding of a manifold in a higher-dimensional space. Intuitively, we would like the inner product near the boundary of the disk to be larger than near the center because the sphere is "steeper" there, elongating curves more when projected back.

We can now define the *distance* of two points under a Riemannian metric to be the length of the shortest curve joining these two points.

A *vector field* is a function that maps each point in a manifold to a vector in its tangent space. This function also has to be smooth in the sense that each component function is smooth.

A *tensor field* of type (q, r), $q = 0, 1$, $r = 1, 2, ..$ is a function that maps each point in a manifold to an operator that takes r vectors in its tangent space as input and maps them to another vector (if $q = 1$) or a real number (if $q = 0$). This function is also smooth in the sense that each component function, when viewing the tensor as an $r + q$-dimensional number field, is smooth across the manifold.

Next, we seek a way to relate tangent spaces at different points to each other. Consider points p and p'. We would like there to be a linear bijection between $T(p)$ and $T(p')$. We would also like that bijection to vary smoothly across the manifold. Let $p' = p + h$ for small vector h. Then $(x)_p$ should correspond to $(f(x))_{p'} = (x + g_p(x, h))_{p'}$. Smoothness further implies that g_p is linear in h and linearity implies that g_p is linear in x. Hence, g_p is a bilinear function from $\mathbb{R}^n * \mathbb{R}^n$ to $\mathbb{R}^n$, and hence it can be equivalently represented as a tensor of dimension 3. Expanding with respect to basis we obtain $(x_j)_p \to (x_j + \sum_i \Gamma^k_{i,j} h_i \mathbf{e}_k)_{p+h}$. We call $(\Gamma^k_{i,j})_p$ an *affine connection* on the manifold. Note that even though this defines a tensor at each point, it is not a tensor field as defined above because $g_p(x, h)$ is not $T(p) * T(p) \to T(p)$ (h, for example, is a vector in the manifold, not in the tangent space). Using the definition of f and our knowledge of how diffeomorphisms affect tangent vectors, we find that a diffeomorphism affects an affine connection as follows:

$$(\Gamma^t_{r,s})^\psi = ((\Gamma^k_{i,j})^\phi \frac{\partial \phi_i}{\partial \psi_r} \frac{\partial \phi_j}{\partial \psi_s} + \frac{\partial^2 \phi_k}{\partial \psi_r \partial \psi_s}) \frac{\partial \psi_t}{\partial \phi_k}$$

Next, we would like to augment the notion of partial derivatives of vector fields with affine connections. Specifically, when considering the change of vector fields across infinitesimal distances, we would like to take into account that vectors in different tangent spaces relate to each other via an affine connection. Hence, a derivative would need to take into account not just the change of the raw components of the values of the vector fields, but also the change in meaning of those components as specified by the connection. This leads to the notion of the *covariant derivative*. Given two vector fields X, Y, we define the covariant derivative $\nabla_X Y$ as this "connection-aware" derivative of Y at each point in the direction specified by X at that point. This leads to the formula:

$$\nabla_X Y = x_i(\partial_i Y_k + y_j \Gamma^k_{i,j}) \mathbf{e}_k$$

Here, lower case letters denote component values and $\partial_i Y_k$ is the partial derivative of the k^{th} component of Y as a function on the manifold in the direction of the i^{th} basis vector of the manifold.

The covariant derivative satisfies several properties we expect from differential operators (e.g. chain rule). In fact, any bilinear operator on X and Y satisfying these connections induces a unique affine connection. Hence the two concepts are equivalent.

An affine connection can be used to translate any vector along a curve by taking the limit of infinitesimal mappings defined by Γ. This is called *parallel translation*. Note that if we translate a vector along a curve and then translate it back, we do not necessary end up with the same vector. This is only achieved then $\Gamma^k_{i,j} = \Gamma^k_{j,i}$. We call such a connection *symmetric*.

If we have introduced a Riemannian metric on a manifold, it is natural to want an affine connections to preserve inner products with respect to that metric under parallel translation. It turns out we can differentiate that metric to obtain a unique symmetric connection with that property. We call this the *Riemannian connection* with respect to that metric.[1]

3 Information Geometry Fundamentals

We consider the space of *parametric models*, i.e. a set of distributions that is indexed by an n-dimensional vector parameter. This definition reveals that a parametric model is naturally an n-dimensional manifold. This becomes even more evident when we consider that the parametrization of the model is arbitrary. The only property we would like our parametrization to fulfill is that the underlying distributions vary smoothly with the parameter. Hence, given an initial such admissible parametrization, we consider all its smooth transformations as equivalent. Again, this coincides with the defining property of a manifold.

Let's now define this notion rigorously. We consider the distributions of a parametric model as points of an n-dimensional manifold. At each point in the manifold, we have a distribution defined on some set which is completely apart from the manifold. We call the space where the manifold is embedded the *model space* and the set on which the distributions are defined as the *data space*. In

[1] Note: There are also non-symmetric connections satisfying this property

practice, we will assume that the data space is either a finite set or the closure of an open set in $\mathbb{R}^d$, for some d.

We would now like to define a Riemannian metric that gives the manifold a notion of distance "inherent" to the probability distributions underlying it. Precisely, our metric should fulfill the following requirements:

1. It should be based on a notion of "rate of change of distribution". If distributions vary more in a region, curves should be longer there.
2. The metric should be well-defined in the sense that if we derive the metric under two different coordinate systems, they have to obey the chain rule with respect to the diffeomorphism.
3. The metric should be invariant to reparametrization of the data space.

Let us analyze the last requirement more carefully. We have stated that parametrizations of the model spaces are arbitrary up to smooth transformations. In the same way, we could simply change the unit of the data we measured. Hence, the parametrization of the data space is equally arbitrary up to continuous bijections. Hence we require that these changes do not affect our notion of distance as defined by the Riemannian metric.

It turns out that there exists precisely one metric fulfilling all three criteria, and it is the *Information metric* (also called *Fisher metric*):

$$(G_{ij})_\epsilon = \mathbb{E}_{p_\epsilon}[\frac{\partial \ln p_\epsilon(x)}{\partial \epsilon_i} \frac{\partial \ln p_\epsilon(x)}{\partial \epsilon_j}]$$

Here, the partial derivatives act pointwise on the log-likelihood. ϵ is the coordinate vector in the manifold and p_ϵ is the distribution at that point. The G of this metric is usually called $\mathcal{I}$ ('I' for information) and $\mathcal{I}$ itself is called the *Information matrix* or *Fisher matrix*.

In the case where the data space is finite, *Chentsov's theorem* states that the Fisher metric is indeed unique in fulfilling the technical conditions 2 and 3 stated above. Under some regularity conditions, it can easily be extended to the infinite case.

Let us now define the α-*connection*:

$$(\Gamma^{(\alpha)}_{ij,k})_\epsilon = \mathbb{E}_{p_\epsilon}[(\frac{\partial^2 \ln p_\epsilon(x)}{\partial \epsilon_i \partial \epsilon_j} + \frac{1-\alpha}{2} \frac{\partial \ln p_\epsilon(x)}{\partial \epsilon_i} \frac{\partial \ln p_\epsilon(x)}{\partial \epsilon_j}) \frac{\partial \ln p_\epsilon(x)}{\partial \epsilon_k}]$$

It turns out that the 0-connection (where the hyper parameter α is 0) is the Riemannian connection for the Fisher metric.

Let us assume our model is an exponential family. It turns out that if we parametrize this model by its natural parameters, the 1-connection is identically zero and hence the manifold is flat. This is easy to show by arithmetic. The 1-connection is also called *e-connection* (e for 'exponential family'). Furthermore, if we parametrize the exponential family by its mean parameters, the -1-connection is identically zero. The -1-connection is also called *m-connection*.

For this to be geometrically meaningful, we need the following statement: Given a statistical model indexed by some parameter, that model can be written as an exponential family indexed by its natural parameter if and only the same holds true under every data space transformation. Arithmetic shows that this is in fact true and the same holds for the mean parameter.

Everything that was just said about exponential families also holds true for mixture families. A mixture family is a model of the form:

$$p(x;\theta) = C(x) + \theta^T F(x)$$

Here, C is a real-valued function that integrates to 1 and F is a vector-valueed feature function where each component integrates to 0. Of course, θ must be restricted to a set where the resulting p is non-negative.

4 Connections to Statistical Learning

4.1 Statistical Estimation

We would like to look at the theory of statistical estimation from a geometric perspective. An estimator is a function of data taking values in the model space. Hence we can regard it directly as a random variable on the manifold that is induced by the true distribution. In general, this estimator has to be invariant neither with respect to data space nor with respect to model space transformations. However, arithmetic shows that the maximum likelihood estimator fulfills both of these properties, which makes it ideal for geometrical study.

The standard way to measure the quality of an estimator is through its mean square error, which is the expectation of the square of the euclidean distance between the estimator and the true parameter in the model space. Clearly, this is not invariant under model space transformation, and hence it is in a way a rather poor measure. A mathematically correct alternative would be to measure the expected square of the distance induced by the Fisher metric, which is, however, cumbersome in practice.

The Cramer-Rao bound tells us that the covariance matrix of an unbiased estimator on n data points is at least $\frac{1}{n}\mathcal{I}^{-1}$. We can also show that the there exists N such that for all $n > N$, the minimax covariance matrix of any estimator is at least $\frac{1}{n}\mathcal{I}^{-1}$. (Here, we measure the covariance about the true parameter and not the mean of the estimator.) We know that this bound is achieved asymptotically by the MLE. Furthermore, because of smoothness, asymptotically, the covariance matrix contains all information about the mean-square distance of the estimator from the true value even if distances are induced by the information metric. Hence, the problem from the last paragraph is mitigated by the fact that the MLE is asymptotically efficient even under the geometrically correct distance.

It turns out that an efficient estimator exists if and only if the model is an exponential family indexed by its mean parameter. In that case, the MLE is such an efficient estimator. This is interesting in that even though the MLE is coordinate system invariant, its efficiency is not. Exponential families in mean parametrization have further nice properties. If we parametrize the data space so that the feature function is the identity, then the MLE is the sample mean and the Fisher metric is the covariance of the distribution.

4.1.1 Projection of estimators

Let's look at the case where we know the true distribution lies in a subset of a statistical model and let's say we can easily compute an efficient estimator in the model, but not directly so in the subset. An important example is that of a non-affine subset of an exponential family, called a *curved exponential family*. The MLE of the mean parameter is easy to find in the model, but this technique cannot be used on the subset because it is not itself an exponential family.

A general strategy we would like to deploy in these cases is to calculate an estimator in the model and then map that estimator into the subset with a smooth function. Let's call the parent manifold $\mathcal{S}$, the submanifold $\mathcal{M}$, the mapping f, the estimator in $\mathcal{S}$ $\hat{\epsilon}$ and the estimator in $\mathcal{M}$ $\hat{u} = f(\hat{\epsilon})$. u is some coordinate system on $\mathcal{M}$, and the true parameter ϵ^* in $\mathcal{S}$ corresponds to some parameter value u^* in $\mathcal{M}$. Let's assume $\hat{\epsilon}$ converges in probability to ϵ^*. Then, because f is smooth and hence continuous, a necessary and sufficient condition for $\hat{u}$ to converge to u^* is for $f(\epsilon^*)$ to equal u^*.

Though the arithmetic is more involved in this case, a necessary and sufficient condition for $\hat{u}$ to be asymptotically efficient is for $f(\epsilon) - \epsilon$ to be asymptotically orthogonal to $\mathcal{M}$ with respect to the Riemannian metric on $\mathcal{S}$ at $f(\epsilon)$. Essentially, this means that should pick the point in $\mathcal{M}$ that is closest to $\hat{\epsilon}$ according to the Riemannian metric to achieve efficiency, which is very sensible. One choice of projection function with this property is minimizing the KL-divergence between $p_{\hat{\epsilon}}$ and $p_{f(\hat{\epsilon})}$. It turns out that if $\hat{\epsilon}$ is the MLE on $\mathcal{S}$, then under this projection $\hat{u}$ is the MLE on $\mathcal{M}$.

4.2 A Geometric View of Conjugate Priors

We now look at some recent work on geometric interpretation of Bayesian priors. Bayesian Inference has received wide interest in recent years due to and its logical intuitivity, exibility, mathematical elegance, and an increase in available computing power. At the same time, there are clouds of issues around Bayesian Inference that has been a matter of constant debates. One of the most con-

troversial issues in Bayesian Inference is the arbitrariness in the choice of prior distributions [10]. Despite "subjectivity" being one of the basic foundations of Bayesian learning, most works simply use "non-informative" or "conjugate" priors [6].

However, Information Geometry shows that there are deeper fundamental advantages for choosing a conjugate prior. With this motivation, we look at information theoretic interpretations of such priors [2]. It looks at the relation between the geometry of the underlying likelihood model, and the geometric properties of the conjugate prior. The conjugate prior is represented in the form of Bregman divergence and it is shown that it is the inherent geometry of conjugate priors that makes them appropriate and intuitive. This geometry induces the Fisher information metric and 1-connection, which are respectively, the natural metric and connection for the exponential family, as we saw previously. This geometric interpretation allows us to view the hyperparameters of conjugate priors as the effective sample points, thus providing additional intuition. This geometric understanding of conjugate priors is also used to derive the hyperparameters and components of the prior are used to couple the generative and discriminative components of a hybrid model for semi-supervised learning. Here, we will briefly outline the proof:

Theorem 1:. Let $X = \{x_1 \dots x_n\}$ be a set of n i.i.d. training data points drawn from an exponential family distribution with the log partition function G, F be the dual function of G, then dual of ML estimate (θ_{ML}) of X under the assumed model solves the following Bregman median problem:

$$\mu_{ML} = min_{\mu \in M} \sum_{i=1}^{n} B_F(x_i||\mu)$$

Corollary 1: Let $G(\theta)$ be the log partition function of the exponential family defined over the convex set , $X = \{x_1 \dots x_n\}$ be set of n i.i.d data points from an exponential family, and θ_i be the dual of x_i, then ML estimation, θ_{ML} of X solves the following optimization problem:

$$\theta_{ML} = min_{\theta \in \Theta} \sum_{i=1}^{n} B_G(\theta||\theta_i)$$

The conjugate prior $log\ p(\theta, \alpha, \beta) = log\ m(\alpha, \beta) + \beta(< \theta, \frac{\alpha}{\beta} > -G(\theta))$ can be rewritten in the form of bregman divergence.

$$\begin{aligned} log\ p(\theta; \alpha, \beta) &= log\ m(\alpha, \beta) + \beta(F(\frac{\alpha}{\beta}) - B_F(\frac{\alpha}{\beta}||\nabla G(\theta))) \\ &= const - B_F(x||\mu) - \beta B_F(\frac{\alpha}{\beta}||\mu) \end{aligned}$$

Theorem 2: Given a set X of n i.i.d examples from the exponential family distribution with the log partition function G and a conjugate prior as before, MAP estimation of parameters is $\theta_{MAP} = \mu_{MAP}$ where μ_{MAP} solves the problem:

$$\mu_{MAP} = min_{\mu \in \mathcal{M}} \sum_{i=1}^{n} B_F(x_i||\mu) + \beta B_F(\frac{\alpha}{\beta}||\mu)$$

The solution to the above clearly is:

$$\begin{aligned} \mu_{MAP} &= \frac{\sum_{i=1}^{n} x_i + \alpha}{n + \beta} \\ &= \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{\beta} \frac{\alpha}{\beta}}{n + \beta} \end{aligned}$$

The above solution gives a natural interpretation of MAP estimation. One can think of the prior as β extra points at position $\frac{\alpha}{\beta}$ works as the effective sample size of the prior.

4.3 AdaBoost and Exponential Models

There have been numerous studies that describe the connection between boosting and logistic regression. Here, we describe a information theoretic perspective [17] on this correspondence. In

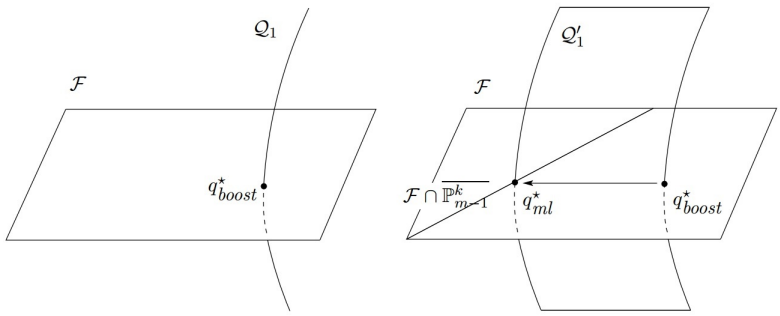

Figure 1: Geometric view of the correspondence between Adaboost & Exponential Models

this setting, it is seen that both are the same convex optimization problems and the only difference between two is that the model is normalized in the latter (leading to an additional normalization constraint). Both minimize the I-divergence $D(p, q)$ to a uniform model under the product Fisher information geometry [17] subject to feature constraints.

Information geometry results show that projecting the exponential loss model onto the simplex of conditional distributions results in the maximum likelihood exponential model with the specified sufficient statistics. Conditional I- divergence of p and q with respect to a distribution r over $\mathcal{X}$ is defined as:-

$$D_r(p, q) = \sum_r r(x) \sum_y \left(p(y|x) \log \frac{p(y|x)}{q(y|x)} - p(y|x) + q(y|x) \right)$$

In other words, if q^*_{boost} and q^*_{ml} be the optimal solution for boosting and mle then it can be shown that (see Figure 1)

$$\begin{aligned} q^*_{boost} &= \operatorname*{argmin}_{p \in \mathcal{F}} D(p, q_0) = \operatorname*{argmin}_{q \in \overline{Q_1}} D(\tilde{p}, q) \\ q^*_{ml} &= \operatorname*{argmin}_{p \in \mathcal{F} \cap \mathbb{P}^k_{m-1}} D(p, q_0) = \operatorname*{argmin}_{q \in \overline{Q_2}} D(\tilde{p}, q) = \operatorname*{argmin}_{p \in \mathcal{F} \cap \mathbb{P}^k_{m-1}} D(p, q^*_{boost}) \end{aligned}$$

We encourage the reader to look at [17] for details.

4.4 Hyperplane classifiers for non-Euclidean spaces

Linear classifiers are a powerful class of learning algorithms, that subsume SVMs, logistic regression and perceptrons . Linear classifiers, in general, partition the data space based on a decision boundary $< w, x >= 0$, where the learnt weight vector w depends on various optimization paradigms for different linear classifiers (max-margin for SVM's, maximizing conditional log likelihood for logistic regression etc). While linear classifiers are a restricted class of models (high bias), they have several desirable properties. In particular, linear classifiers seem extremely appropriate for Euclidean spaces in the following ways: A hyperplane decision boundary in $\mathbb{R}^n$ is isomorphic to $\mathbb{R}^{n-1}$, and can hence be thought of as a reduced version of the original space. Also, for a Euclidean space, the locus of the space equidistant from two points is a hyperplane, providing an intuitive basis of separation of points.

The notion of using hyperplanes if thus implicitly connected with Euclidean geometries. If the features spaces do not behave like Euclidean spaces, the linear classifier paradigms may be inappropriate. Since real world features are often non-Euclidean, it becomes fitting to look for analogous classes of Euclidean hyperplane classifiers for non-Euclidean geometries. We review the work in [18] [12] extending the notion of linear classifiers to general Riemannian manifolds to obtain classifiers when the data has a natural geometric structure that is only *locally Euclidean*.

In general, we can see that analogues to linear boundaries can be defined for Riemannian manifolds, with properties analogous to separating hyperplanes for Euclidean spaces. Specifically, a linear

boundary N in an n-dimensional manifold M is an auto-parallel sub-manifold of M such that M N has exactly two connected components. However, the notion of local linearity in general Riemannian spaces is complex.

We focus on the case of categorical data, which have associated tractable spherical geometries [2] . For this special case, we can observe the notion of hyperplanes and margins. We define a hyperplane on $\mathbb{S}_+^n$ as $H_{u+} = \mathbb{S}_+^n \cap E_u$ (if the intersection is not null), where E_u is an n-dimensional linear subspace in $\mathbb{R}^{n+1}$ associated with a normal vector u on the sphere. The Cosine Law of sides gives us a natural way to compute the associated margin $d(x, H_{u+})$, which is defined as the minimum distance of a point x from the set H_{u+} as:

$$d(x, H_{u+}) = \arccos(\|x\|_A \sqrt{1 - < x|_A, u|_A >})$$

where A is a boundary set of integers, with maximum value $n + 1$.

The standard logistic regression model can be seen to incorporate the Euclidean distance of a point to the normal vector $\hat{u}$. Our defined notions of hyperplanes and margins on $\mathbb{S}^n$ allows the generalization of the normal logistic regression model to a spherical geometry. The training of such a model follows a gradient descent procedure along the sphere along a curve whose tangent vector at $\hat{u}$ is the projection of the gradient vector on the tangent space at that point.

4.5 Metric Learning

Many machine learning algorithms (e.g. , K-nearest neighbors, Neural networks, Linear SVMs, etc) assume the embedding space to be $\mathbb{R}^{\aleph}$ and a Euclidean metric structure for the data and/or model spaces. In this section, we argue that in the absence of any direct evidence of Euclidean geometry, the metric structure should be inferred from data (if available). After obtaining the metric, it may then be passed on to learning algorithms for classication, clustering, etc. We see that the Fisher geometry is a natural choice of metric if the only known information is the statistical family that generates the data [13]. In the presence of actual data, it may be possible to induce a geometry (perhaps with the locally Euclidean assumption) that is better suited for this task. [13] also proposes a learning principle for the geometry of that data space that is based on maximizing the inverse volume of the given training set. It is further shown that when applied to the space of text documents in tf representation, the learned geometry is similar to, but outperforms the popular tf-idf geometry. The metric derived maximizes the following objective:

$$\mathcal{O}(g, D) = \prod_{i=1}^{N} \frac{(dvol(g(x_i)))^{-1}}{\int dvol(g(x_i)^{-1} dx}$$

Intuitively, volume element dvol g(x) summarizes the size of the metric g at x. The idea is that the paths crossing areas with high volume will tend to be longer than the same paths over an area with low volume. Hence, maximizing the inverse volume in the previous equation will result in shorter curves across densely populated regions of $\mathcal{M}$. As a result, the geodesics will tend to pass through densely populated regions. This agrees with the intuition that distances between data points should be measured on the lower dimensional data submanifold, thus capturing the geometry of the data.

5 Conclusion

In this project, we explored the rich connections between non-Euclidean geometries, and statistical learning models. In the initial part of the project, we developed a grounding in information geometry. Next, we reviewed existing literature for several eclectic learning scenarios, where information geometric presented new insights into the working of statistical learning methods. Our study has also helped us see that several learning algorithms make unrealistic simplifying assumptions about data and feature spaces, and that deeper geometric analysis of other learning scenarios can lead to better learning methods.

[2](This follows from the form of the Fisher Information metric for multinomials, whic implies that geodesic distances can be easily computed due to a straighforward diffeomorphism between the n-outcome simplex $\mathbb{P}^{n-1}$ and the positive orthant of the $n-1$ dimensional sphere $\mathbb{S}_+^{n-1}$)

References

[1] A. Agarwal. *Geometric Methods in Machine Learning and Data Mining*. PhD thesis, Department of Computer Science, University of Maryland, 2012.

[2] Arvind Agarwal and Hal Daumé, III. A geometric view of conjugate priors. *Mach. Learn.*, 81(1):99–113, October 2010.

[3] Shun-ichi Amari. Differential geometry of statistical inference. In JuriiV. Prokhorov and Kiyosi It, editors, *Probability Theory and Mathematical Statistics*, volume 1021 of *Lecture Notes in Mathematics*, pages 26–40. Springer Berlin Heidelberg, 1983.

[4] Shun-ichi Amari. Information geometry and its applications: Convex function and dually flat manifold. In Frank Nielsen, editor, *Emerging Trends in Visual Computing*, volume 5416 of *Lecture Notes in Computer Science*, pages 75–102. Springer Berlin Heidelberg, 2009.

[5] Mikhail Belkin and Partha Niyogi. Using Manifold Structure for Partially Labelled Classification. In *NIPS*, pages 929+, 2002.

[6] A. Gelman. Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics. *ArXiv e-prints*, January 2010.

[7] F. Gianfelici and V. Battistelli. Methods of information geometry (amari, s. and nagaoka, h.; 2000). *Information Theory, IEEE Transactions on*, 55(6):2905 –2906, june 2009.

[8] Shiro Ikeda. Stochastic reasoning, free energy, and information geometry, 2004.

[9] R. Kass and P. Vos. Geometrical Foundations of Asymptotic Inference. 1997.

[10] Robert E. Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

[11] Seungyeon Kim, Fuxin Li, Guy Lebanon, and Irfan A. Essa. Beyond sentiment: The manifold of human emotions. *CoRR*, abs/1202.1568, 2012.

[12] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds, 2004.

[13] G. Lebanon. *Riemannian Geometry and Statistical Machine Learning*. PhD thesis, School of Computer Science, Carnegie Mellon University, January 2005.

[14] Guy Lebanon. Computing the volume element of a family of metrics on the multinomial simplex. Technical report, 2003.

[15] Guy Lebanon. Learning riemannian metrics. In *In Proceedings of the 19th conference on Uncertainty in Artificial Intelligence (UAI*. Morgan Kaufmann publishers, 2003.

[16] Guy Lebanon. Information geometry, the embedding principle, and document classification. In *in Proceedings of the 2nd International Symposium on Information Geometry and its Applications*, 2005.

[17] Guy Lebanon and John Lafferty. Boosting and maximum likelihood for exponential models. In *In Advances in Neural Information Processing Systems*, pages 447–454, 2001.

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

[19] Luigi Malagò, Matteo Matteucci, and Bernardo Dal Seno. An information geometry perspective on estimation of distribution algorithms: boundary analysis. In *Proceedings of the 2008 GECCO conference companion on Genetic and evolutionary computation*, GECCO '08, pages 2081–2088, New York, NY, USA, 2008. ACM.

[20] A.M. Peter and A. Rangarajan. Information geometry for landmark shape analysis: Unifying shape representation and deformation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):337–350, 2009.

[21] I. Sciszar. i-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[22] Jangwon Seo and W. Bruce Croft. Geometric representations for multiple documents. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 251–258, New York, NY, USA, 2010. ACM.

[23] Hichem Snoussi. The geometry of prior selection. *Neurocomput.*, 67:214–244, August 2005.

[24] Hichem Snoussi and Ali Mohammad-Djafari. Information geometry and prior selection, 2002.

[25] Khoat Than and Tu-Bao Ho. A geometric interpretation of bayesian classification methods, 2011.