Hyperplane Margin Classifiers on the Multinomial Manifold

Guy Lebanon LEBANON@CS.CMU.EDU
John Lafferty LAFFERTY@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh PA, USA

Abstract

The assumptions behind linear classifiers for categorical data are examined and reformulated in the context of the multinomial manifold, the simplex of multinomial models furnished with the Riemannian structure induced by the Fisher information. This leads to a new view of hyperplane classifiers which, together with a generalized margin concept, shows how to adapt existing margin-based hyperplane models to multinomial geometry. Experiments show the new classification framework to be effective for text classification, where the categorical structure of the data is modeled naturally within the multinomial family.

1. Introduction

Linear classifiers are a mainstay of machine learning algorithms, forming the basis for techniques such as the perceptron, logistic regression, boosting, and support vector machines. A linear classifier, parameterized by a vector $w \in \mathbb{R}^n$, classifies examples according to the decision rule $\hat{y}(x) = \text{sign}\left(\sum_i w_i \phi_i(x)\right) = \text{sign}(\langle w, x \rangle) \in \{-1, +1\}$, following the common practice of identifying x with the feature vector $\phi(x)$. The differences between different linear classifiers lie in the criteria and algorithms used for selecting the parameter vector w based on a training set.

Geometrically, the decision surface of a linear classifier is formed by a hyperplane or linear subspace in n-dimensional Euclidean space, $\{x \in \mathbb{R}^n : \langle x, w \rangle = 0\}$ where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. (In both the algebraic and geometric formulations, a bias term is sometimes added; we prefer to absorb the bias into the notation given by the inner product, by setting $x_n = 1$ for all x.) The linearity assumption made by such classifiers can be justified on purely computational grounds; linear classifiers are generally easy to train, and the linear form is simple to analyze and compute.

Modern learning theory emphasizes the tension between fitting the training data well and the more desirable goal of achieving good generalization. A common practice is to choose a model that fits the data closely, but from a restricted class of models. The model class needs to be sufficiently rich to allow the choice of a good hypothesis, yet not so expressive that the selected model is likely to overfit the data. Hyperplane classifiers are attractive for balancing these two goals. Indeed, linear hyperplanes are a rather restricted set of models, but they enjoy many unique properties. For example, given two points $x, y \in \mathbb{R}^n$, the set of points equidistant from x and y is a hyperplane; this lies behind the intuition that a hyperplane is the correct geometric shape for separating sets of points. Similarly, a hyperplane is the best decision boundary to separate two Gaussian distributions of equal covariance. Another distinguishing property is that a hyperplane in $\mathbb{R}^n$ is isometric to $\mathbb{R}^{n-1}$, and can therefore be thought of as a reduced dimension version of the original feature space. Finally, a linear hyperplane is the union of straight lines, which are distance minimizing curves, or geodesics, in Euclidean geometry.

However, a fundamental assumption is implicitly associated with linear classifiers, since they are based crucially on the use of the Euclidean geometry of $\mathbb{R}^n$. If the data or features at hand lack a Euclidean structure, the arguments above for linear classifiers break down; arguably, there is lack of Euclidean geometry for the feature vectors in most applications. This paper studies analogues of linear hyperplanes as a means of obtaining simple, yet effective classifiers when the data can be represented in terms of a natural geometric structure that is only locally Euclidean. This is the case for categorical data that is represented in terms of multinomial models, for which the associated geometry is spherical.

Because of the complexity of the notion of linearity in general Riemannian spaces, we focus our attention on the multinomial manifold, which permits a relatively simple analysis. The multinomial manifold and its hyperplanes are the topics of Sections 2-4. The construction and training of margin based models is discussed in Section 5, with an em-

Appearing in *Proceedings of the 21^{st} International Conference on Machine Learning*, Banff, Canada, 2004.

phasis on spherical logistic regression. A brief examination of linear hyperplanes in general Riemannian manifolds appears in Section 6 followed by experimental results for text classification given in Section 7. Concluding remarks are made in Section 8.

2. The Multinomial Manifold

The multinomial manifold is the parameter space of the multinomial distribution

$$\mathbb{P}^n = \left\{ x \in \mathbb{R}^{n+1} : \ \forall j \ \ x_j \geq 0, \ \sum_{i=1}^{n+1} x_i = 1 \right\} \qquad (1)$$

equipped with the Fisher information metric g

$$g_x(u, v) = \sum_{i=1}^{n+1} \frac{u_i v_i}{x_i} \qquad x \in \mathbb{P}^n \quad u, v \in T_x\mathbb{P}^n$$

where u, v are vectors tangent to $\mathbb{P}^n$ at x, represented in the standard basis of $\mathbb{R}^{n+1}$. Note that unlike conventional notation in statistics we denote the multinomial parameters by x. The reason for doing so is that we identify multinomial parameters with text documents, as described in further detail in Section 7.

It is a well known fact that the multinomial manifold is isometric to the positive n-sphere

$$\mathbb{S}^n_+ = \left\{ x \in \mathbb{R}^{n+1} : \forall j \ x_j \geq 0, \ \sum_{i=1}^{n+1} x_i^2 = 1 \right\}$$

with the metric inherited from the embedding Euclidean space (Kass, 1989). The isometry $\pi : \mathbb{P}^n \to \mathbb{S}^n_+$, $\pi(x) = \sqrt{x_1}, \ldots, \sqrt{x_{n+1}})$, allows us to perform our calculations on the positive sphere and apply them to $\mathbb{P}^n$ through π^{-1}. Of particular interest is the fact that the geodesic distance between $x, y \in \mathbb{P}^n$ may be now computed as the Euclidean length of the great circle connecting $\pi(x)$ and $\pi(y)$, specifically,

$$d(x, y) = \arccos\left(\sum_{i=1}^{n+1} \sqrt{x_i y_i} \right).$$

Using the above isometry we focus our attention, in the next few sections, on hyperplanes and margins on the n-sphere $\mathbb{S}^n$ and the positive n-sphere $\mathbb{S}^n_+$. The results developed there will apply directly to the multinomial manifold when followed by π^{-1}.

It is worth mentioning that the singular boundary of $\mathbb{P}^n$ and $\mathbb{S}^n_+$ prevents them from being differentiable manifolds or even differentiable manifolds with boundary. However, this is a technical issue that can be overcome by taking the interior of $\mathbb{P}^n$ and $\mathbb{S}^n_+$. Instead of points on the boundary, we can now take points arbitrarily close to it, in both the Euclidean and the geodesic metric.

We do not explore here the many interesting and motivating properties of the Fisher information metric. For details on this topic see Kass and Voss (1997) and Amari and Nagaoka (2000). Spivak (1975) contains a comprehensive introduction to Riemannian geometry.

3. Hyperplanes and Margins on $\mathbb{S}^n$

This section generalizes the notion of linear hyperplanes and margins to the n-sphere $\mathbb{S}^n = \{x \in \mathbb{R}^{n+1} : \sum_i x_i^2 = 1\}$. A similar treatment on the positive n-sphere $\mathbb{S}^n_+$ is more complicated, and is postponed to the next section. In the remainder of the paper we denote points on $\mathbb{P}^n, \mathbb{S}^n$ or $\mathbb{S}^n_+$ as vectors in $\mathbb{R}^{n+1}$ using the standard basis of the embedding space. The notation $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ will be used for the Euclidean inner product and norm.

A hyperplane on $\mathbb{S}^n$ is defined as $H_u = \mathbb{S}^n \cap E_u$ where E_u is an n-dimensional linear subspace of $\mathbb{R}^{n+1}$ associated with the normal vector u. We occasionally need to refer to the unit normal vector and denote it by $\hat{u}$. H_u is an $n-1$ dimensional submanifold of $\mathbb{S}^n$ which is isometric to $\mathbb{S}^{n-1}$ (Bridson & Haefliger, 1999). Using the common notion of the distance of a point from a set $d(x, S) = \inf_{y \in S} d(x, y)$ we make the following definitions.

Definition 1. *Let X be a metric space. A decision boundary is a subset of X that separates X into two connected components. The margin of x with respect to a decision boundary H is $d(x, H) = \inf_{y \in H} d(x, y)$.*

Note that this definition reduces to the common definition of margin for Euclidean geometry and affine hyperplanes.

In contrast to Gous (1998), our submanifolds are intersections of the sphere with linear subspaces, not affine sets. One motivation for the above definition of hyperplane as the correct generalization of a Euclidean hyperplane is that H_u is the set of points equidistant from $x, y \in \mathbb{S}^n$ in the spherical metric. Further motivation is given in Section 6.

Before we can obtain a closed form expression for margins on $\mathbb{S}^n$ we need the following definitions.

Definition 2. *Given a point $x \in \mathbb{R}^{n+1}$, we define its reflection with respect to E_u as*

$$r_u(x) = x - 2\langle x, \hat{u} \rangle \hat{u}.$$

Note that if $x \in \mathbb{S}^n$ then $r_u(x) \in \mathbb{S}^n$ as well, since $\|r_u(x)\|^2 = \|x\|^2 - 4\langle x, \hat{u} \rangle^2 + 4\langle x, \hat{u} \rangle^2 = 1$.

Definition 3. *The projection of $x \in \mathbb{S}^n \setminus \{\hat{u}\}$ on H_u is defined to be*

$$p_u(x) = \frac{x - \langle x, \hat{u} \rangle \hat{u}}{\sqrt{1 - \langle x, \hat{u} \rangle^2}}.$$

Note that $p_u(x) \in H_u$, since $\|p_u(x)\| = 1$ and $\langle x - \langle x, \hat{u}\rangle\hat{u}, \hat{u}\rangle = \langle x, \hat{u}\rangle - \langle x, \hat{u}\rangle \|\hat{u}\|^2 = 0$. The term projection is justified by the following proposition.

Proposition 1. *Let $x \in \mathbb{S}^n \setminus (H_u \cup \{\hat{u}\})$. Then*

$$(a) \quad d(x, q) = d(r_u(x), q) \quad \forall q \in H_u$$
$$(b) \quad d(x, p_u(x)) = \arccos\left(\sqrt{1 - \langle x, \hat{u}\rangle^2}\right)$$
$$(c) \quad d(x, H_u) = d(x, p_u(x)).$$

Proof. Since $q \in H_u$,

$$\begin{aligned}\cos d(r_u(x), q) &= \langle x - 2\langle x, \hat{u}\rangle\hat{u}, q\rangle \\ &= \langle x, q\rangle - 2\langle x, \hat{u}\rangle\langle \hat{u}, q\rangle \\ &= \langle x, q\rangle = \cos d(x, q)\end{aligned}$$

and (a) follows. Assertion (b) follows from

$$\cos d(x, p_u(x)) = \left\langle x, \frac{x - \langle x, \hat{u}\rangle\hat{u}}{\sqrt{1 - \langle x, \hat{u}\rangle^2}} \right\rangle = \frac{1 - \langle x, \hat{u}\rangle^2}{\sqrt{1 - \langle x, \hat{u}\rangle^2}}.$$

Finally, to prove (c) note that by the identity $\cos 2\theta = 2\cos^2\theta - 1$,

$$\begin{aligned}\cos(2d(x, p_u(x))) &= 2\cos^2(d(x, p_u(x))) - 1 \\ &= 1 - 2\langle x, \hat{u}\rangle^2 = \cos(d(x, r_u(x)))\end{aligned}$$

and hence $d(x, p_u(x)) = \frac{1}{2} d(x, r_u(x))$. The distance $d(x, q), q \in H_u$ cannot be any smaller than $d(x, p_u(x))$ since this would result in a path from x to $r_u(x)$ of length shorter than the geodesic $d(x, r_u(x))$. $\square$

Parts (b) and (c) of Proposition 1 provide a closed form expression for the $\mathbb{S}^n$ margin analogous to the Euclidean unsigned margin $|\langle x, \hat{u}\rangle|$. Similarly, the $\mathbb{S}^n$ analogue of the Euclidean signed margin $y\langle \hat{u}, x\rangle$ is

$$y \frac{\langle x, \hat{u}\rangle}{|\langle x, \hat{u}\rangle|} \arccos\left(\sqrt{1 - \langle x, \hat{u}\rangle^2}\right).$$

A plot of the signed margin as a function of $\langle x, \hat{u}\rangle$ and a geometric interpretation of the spherical margin appear in Figure 2.

4. Hyperplanes and Margins on $\mathbb{S}^n_+$

A hyperplane on the positive n-sphere $\mathbb{S}^n_+$ is defined as $H_{u+} = E_u \cap \mathbb{S}^n_+$, assuming it is non-empty. This definition leads to a margin concept $d(x, H_{u+})$ different from the $\mathbb{S}^n$ margin $d(x, H_u)$

$$\begin{aligned}d(x, H_{u+}) &= \inf_{y \in E_u \cap \mathbb{S}^n_+} d(x, y) \\ &\geq \inf_{y \in E_u \cap \mathbb{S}^n} d(x, y) = d(x, H_u).\end{aligned}$$

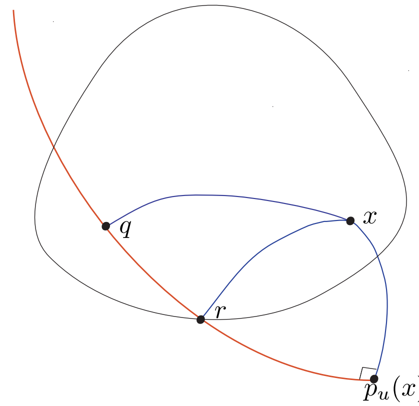

Figure 1. The spherical law of cosines implies $d(r, x) \leq d(q, x)$.

The infimum above is attained by the continuity of d and compactness of $E_u \cap \mathbb{S}^n_+$ justifying the notation $q = \arg\min_{y \in E_u \cap \mathbb{S}^n_+} d(x, y)$ as a point realizing the margin distance $d(x, H_{u+})$.

The following theorem will be useful in computing $d(x, H_{u+})$. For a proof see Bridson and Haefliger (1999) page 17.

Theorem 1. (**The Spherical Law of Cosines**) *Consider a spherical triangle with geodesic edges of lengths a, b, c, where γ is the vertex angle opposite to edge c. Then*

$$\cos c = \cos a \cos b + \sin a \sin b \cos \gamma.$$

We have the following corollaries of Proposition 1.

Proposition 2. *If $x \in \mathbb{S}^n_+$ and $p_u(x) \in \mathbb{S}^n_+$ then*

$$p_u(x) = \arg\min_{y \in \mathbb{S}^n_+ \cap E_u} d(x, y)$$
$$d(x, H_u) = d(x, H_{u+})$$

Proof. This follows immediately from the fact that $p_u(x) = \arg\min_{y \in \mathbb{S}^n \cap E_u} d(x, y)$ and from $\mathbb{S}^n_+ \cap E_u \subset \mathbb{S}^n \cap E_u$. $\square$

Proposition 3. *For $x \in \mathbb{S}^n_+$ and $p_u(x) \notin \mathbb{S}^n_+$ we have*

$$q = \arg\min_{y \in \mathbb{S}^n_+ \cap E_u} d(x, y) \in \partial\mathbb{S}^n_+$$

where $\partial\mathbb{S}^n_+$ is the boundary of $\mathbb{S}^n_+$.

Proof. Assume that $q \notin \partial\mathbb{S}^n_+$ and connect q and $p_u(x)$ by a minimal geodesic α. Since $p_u(x) \notin \mathbb{S}^n_+$, the geodesic α intersects the boundary $\partial\mathbb{S}^n_+$ at a point r. Since $q, p_u(x) \in H_u$ and H_u is geodesically convex, $\alpha \subset H_u$. Now, since $p_u(x) = \arg\min_{y \in \alpha} d(y, x)$, the geodesic from x to $p_u(x)$

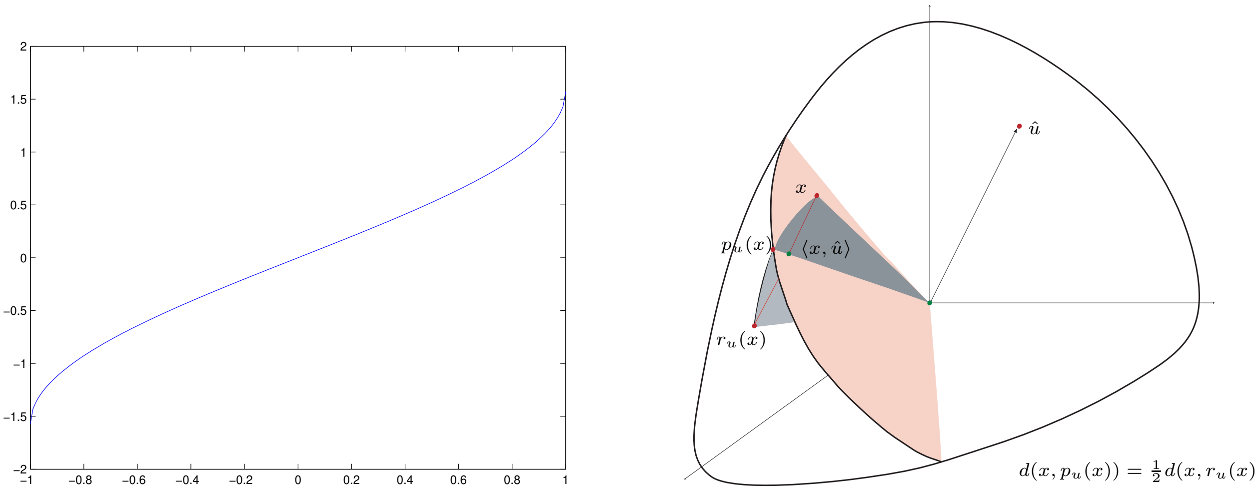

Figure 2. The signed margin $\text{sign}(\langle x, \hat{y} \rangle) d(x, H_u)$ as a function of $\langle x, \hat{u} \rangle$, which lies in the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$ (left) and a geometric interpretation of the spherical margin (right).

and α intersect orthogonally (this is an elementary result in Riemannian geometry, e.g. Lee (1997) p. 113). Using the spherical law of cosines, applied to the spherical triangles $(q, x, p_u(x))$ and $(r, x, p_u(x))$ (see Figure 1), we deduce that

$$\begin{aligned} \cos d(x, q) &= \cos d(q, p_u(x)) \cos d(x, p_u(x)) \\ &\leq \cos d(r, p_u(x)) \cos d(x, p_u(x)) \\ &= \cos d(x, r) \end{aligned}$$

Hence r is closer to x than q. This contradicts the definition of q; thus q can not lie in the interior of $\mathbb{S}^n_+$. $\square$

Before we proceed to compute $d(x, H_{u+})$ for $p_u(x) \notin \mathbb{S}^n_+$ we define the following concepts.

Definition 4. *The boundary of $\mathbb{S}^n$ and $\mathbb{S}^n_+$ with respect to $A \subset \{1, \ldots, n+1\}$ is*

$$\begin{aligned} \partial_A \mathbb{S}^n &= \mathbb{S}^n \cap \{x \in \mathbb{R}^{n+1} : \forall i \in A,\ x_i = 0\} \cong \mathbb{S}^{n-|A|} \\ \partial_A \mathbb{S}^n_+ &= \mathbb{S}^n_+ \cap \{x \in \mathbb{R}^{n+1} : \forall i \in A,\ x_i = 0\} \cong \mathbb{S}^{n-|A|}_+ \end{aligned}$$

Note that if $A \subset A'$ then $\partial_{A'} \mathbb{S}^n_+ \subset \partial_A \mathbb{S}^n_+$. We use the notation $\langle \cdot, \cdot \rangle_A$ and $\|\cdot\|_A$ to refer to the Euclidean inner product and norm, where the summation is restricted to indices *not* in A.

Definition 5. *Given $x \in \mathbb{S}^n$ we define $x|_A \in \partial_A \mathbb{S}^n$ as*

$$(x|_A)_i = \begin{cases} 0 & i \in A \\ x_i / \|x\|_A & i \notin A\,. \end{cases}$$

We abuse the notation by identifying $x|_A$ also with the corresponding point on $\mathbb{S}^{n-|A|}$ under the isometry $\partial_A \mathbb{S}^n \cong \mathbb{S}^{n-|A|}$ mentioned in Definition 4. Note that if $x \in \mathbb{S}^n_+$ then $x|_A \in \partial_A \mathbb{S}^n_+$. The following proposition computes the $\mathbb{S}^n_+$ margin $d(x, H_{u+})$ given the boundary set of $q = \arg\min_{y \in \mathbb{S}^n_+ \cap E_u} d(y, x)$.

Proposition 4. *Let $\hat{u} \in \mathbb{R}^{n+1}$ be a unit vector, $x \in \mathbb{S}^n_+$ and $q = \arg\min_{y \in \mathbb{S}^n_+ \cap E_u} d(y, x) \in \partial_A \mathbb{S}^n_+$ where A is the (possibly empty) set $A = \{1 \leq i \leq n+1 \,:\, q_i = 0\}$. Then*

$$d(x, H_{u+}) = \arccos\left(\|x\|_A \sqrt{1 - \langle x|_A, \hat{u}|_A \rangle^2}\right)$$

Proof. If $p_u(x) \in \mathbb{S}^n_+$ then the proposition follows from earlier propositions and the fact that when $A = \emptyset$, $\|x\|_A = \|x\| = 1$ and $v|_A = v$. We thus restrict our attention to the case of $A \neq \emptyset$.

For all $I \subset \{1, \ldots, n+1\}$ we have

$$\begin{aligned} \arg\min_{y \in \partial_I \mathbb{S}^n_+ \cap E_u} d(x, y) &= \arg\max_{y \in \partial_I \mathbb{S}^n_+ \cap E_u} \langle x, y \rangle \\ &= \arg\max_{y \in \partial_I \mathbb{S}^n_+ \cap E_u} \langle x, y \rangle_I \\ &= \arg\max_{y \in \mathbb{S}^{n-|I|}_+ \cap E_{u|I}} \|x\|_I \langle x|_I, y \rangle \\ &= \arg\min_{y \in \mathbb{S}^{n-|I|}_+ \cap E_{u|I}} d(x|_I, y)\,. \end{aligned}$$

It follows that

$$q|_A = \arg\min_{y \in \mathbb{S}^{n-|A|}_+ \cap E_{u|A}} d(x|_A, y). \qquad (2)$$

By Proposition 3 applied to $\mathbb{S}^{n-|A|}$ we have that since $q|_A$ lies in the interior of $\mathbb{S}^{n-|A|}$ then so does

$$p_{u|A}(x|_A) = \frac{x|_A - \langle x|_A, \hat{u}|_A \rangle \hat{u}|_A}{\sqrt{1 - \langle x|_A, \hat{u}|_A \rangle^2}}\,, \quad x|_A, \hat{u}|_A \in \mathbb{S}^{n-|A|}_+.$$

Using Proposition 1 applied to $\mathbb{S}^{n-|A|}$ we can compute $d(x, H_{u+})$ as

$$d(x, p_{u|A}(x|_A)) = \arccos\left\langle x, \frac{x|_A - \langle x|_A, \hat{u}|_A\rangle \hat{u}|_A}{\sqrt{1-\langle x|_A, \hat{u}|_A\rangle^2}}\right\rangle$$
$$= \arccos\frac{\|x\|_A - \langle x|_A, \hat{u}|_A\rangle\langle x, \hat{u}|_A\rangle}{\sqrt{1-\langle x|_A, \hat{u}|_A\rangle^2}}$$
$$= \arccos\left(\|x\|_A\sqrt{1-\langle x|_A, \hat{u}|_A\rangle^2}\right).$$

□

In practice the boundary set A of q is not known. In our experiments we set $A = \{i : (p_u(x))_i \leq 0\}$; in numerical simulations in low dimensions, the true boundary never lies outside of this set.

5. Logistic Regression on the Multinomial Manifold

The logistic regression model $p(y \mid x) = \frac{1}{Z}\exp(y\langle x, w\rangle)$, with $y \in \{-1, 1\}$, assumes Euclidean geometry. It can be reexpressed as

$$\begin{aligned} p(y \mid x\,; u) &\propto \exp\left(y\,\|u\|\,\langle x, \hat{u}\rangle\right) \\ &= \exp\left(y\,\mathrm{sign}(\langle x, \hat{u}\rangle)\,\theta d(x, H_u)\right) \end{aligned}$$

where d is the Euclidean distance of x from the hyperplane that corresponds to the normal vector $\hat{u}$, and where $\theta = \|u\|$ is a parameter.

The generalization to spherical geometry involves simply changing the margin to reflect the appropriate geometry:

$$p(y|x\,; \hat{u}, \theta) \propto$$
$$\exp\left(y\,\mathrm{sign}(\langle x, \hat{u}\rangle)\;\theta\arccos\left(\|x\|_A\sqrt{1-\langle x|_A, \hat{u}|_A\rangle^2}\right)\right).$$

Denoting $s_x = y\,\mathrm{sign}(\langle x, \hat{u}\rangle)$, the log-likelihood of the example (x, y) is

$$\ell(\hat{u}, \theta\,; (x, y))$$
$$= -\log\left(1 + e^{-2s_x\theta\arccos\left(\|x\|_A\sqrt{1-\langle x|_A, \hat{u}|_A\rangle^2}\right)}\right).$$

We compute the derivatives of the log-likelihood in several steps, using the chain rule and the notation $z = \langle x|_A, \hat{u}|_A\rangle$. We have

$$\frac{\partial\arccos\left(\|x\|_A\sqrt{1-z^2}\right)}{\partial z} = \frac{z\,\|x\|_A}{\sqrt{1-\|x\|_A^2(1-z^2)}\sqrt{1-z^2}}$$

and hence

$$\frac{\partial\ell(\hat{u}, \theta\,; (x, y))}{\partial z} = \tag{3}$$
$$\frac{2s_x\theta z\,\|x\|_A/(1+e^{2s_x\theta\arccos\left(\|x\|_A\sqrt{1-z^2}\right)})}{\sqrt{1-\|x\|_A^2(1-z^2)}\sqrt{1-z^2}}.$$

The log-likelihood derivative with respect to $\hat{u}_i$ is equation (3) times

$$\frac{\partial\langle x|_A, \hat{u}|_A\rangle}{\partial\hat{u}_i} = \begin{cases} 0 & i \in A \\ \frac{(x|_A)_i}{\|\hat{u}\|_A} - \hat{u}_i\frac{\langle x|_A, \hat{u}|_A\rangle}{\|u\|_A^2} & i \notin A\,. \end{cases}$$

The log-likelihood derivative with respect to θ is

$$\frac{\partial\ell(\hat{u}, \theta\,; (x, y))}{\partial\theta} = \frac{2s_x\arccos(\|x\|_A\sqrt{1-z^2})}{1+e^{2s_x\theta\arccos(\|x\|_A\sqrt{1-z^2})}}.$$

Optimizing the log-likelihood with respect to $\hat{u}$ requires care. Following the gradient $\hat{u}^{(t+1)} = \hat{u}^{(t)} + \alpha\nabla\ell(\hat{u}^{(t)})$ results in a non-normalized vector. Performing the above gradient descent step followed by normalization has the effect of moving along the sphere in a curve whose tangent vector at $\hat{u}^{(t)}$ is the projection of the gradient onto the tangent space $T_{\hat{u}^{(t)}}\mathbb{S}^n$. This is the technique used in the experiments described in Section 7.

Note that the spherical logistic regression model has $n+1$ parameters in contrast to the $n+2$ parameters of Euclidean logistic regression. This is in accordance with the intuition that a hyperplane separating an n-dimensional manifold should have n parameters. The extra parameter in the Euclidean logistic regression is an artifact of the embedding of the n-dimensional multinomial space, on which the data lies, into an $(n+1)$-dimensional Euclidean space.

The derivations and formulations above assume spherical data. If the data lies on the multinomial manifold, the isometry π mentioned in Section 2 has to precede these calculations. The net effect is that x_i is replaced by $\sqrt{x_i}$ in the model equation, and in the log-likelihood and its derivatives.

Synthetic data experiments contrasting Euclidean logistic regression and spherical logistic regression on $\mathbb{S}^n_+$, as described in this section, are shown in Figure 3. The leftmost column shows an example where both models give a similar solution. In general, however, as is the case in the other two columns, the two models yield significantly different decision boundaries.

6. Hyperplanes in Riemannian Manifolds

The definition of hyperplanes in general Riemannian manifolds has two essential components. In addition to discriminating between two classes, hyperplanes should be regular in some sense with respect to the geometry. In Euclidean geometry, the two properties of discrimination and regularity coincide, as every affine subspace of dimension $n-1$ separates $\mathbb{R}^n$ into two regions. In general, however, these two properties do not necessarily coincide, and have to be considered separately.

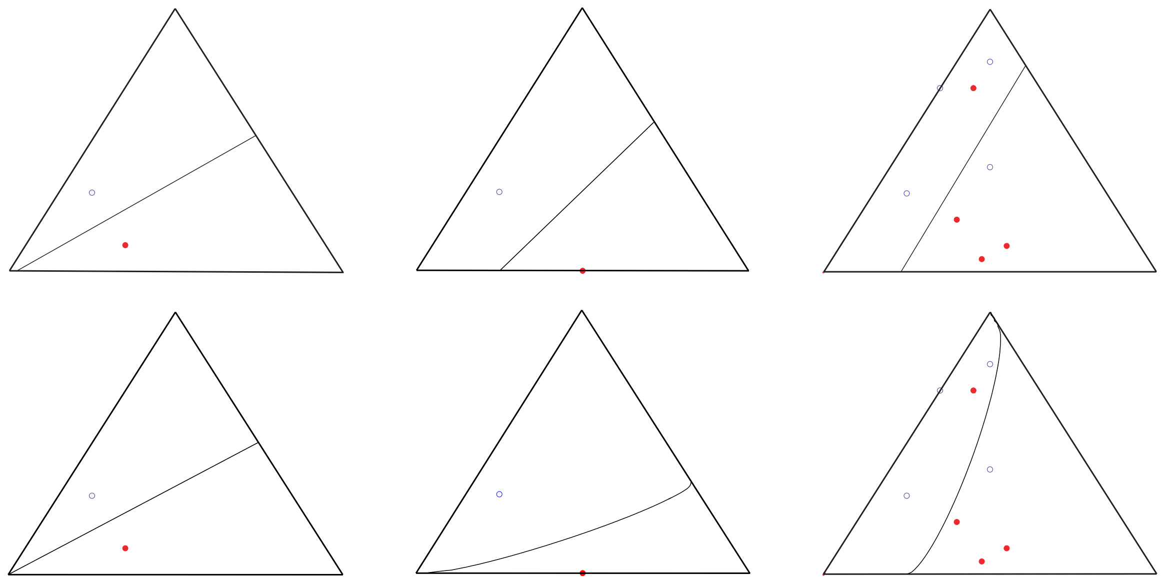

Figure 3. Experiments contrasting Euclidean logistic regression (top row) with multinomial logistic regression (bottom row) for several toy data sets in $\mathbb{P}^2$.

The separation property implies that if N is a hyperplane of M then $M \setminus N$ has two connected components. Note that this property is topological and independent of the metric. The linearity property is generalized through the notion of autoparallelism explained below. The following definitions and propositions are taken from Spivak (1975), Volume 3. We assume that ∇ is the connection inherited from the metric g.

Definition 6. *Let M be a Riemannian manifold with connection ∇. A submanifold $N \subset M$ is* auto-parallel *if parallel translation in M along a curve $C \subset N$ takes vectors tangent to N to vectors tangent to N.*

Proposition 5. *A submanifold $N \subset M$ is auto-parallel if and only if*

$$X, Y \in T_pN \Rightarrow \nabla_X Y \in T_pN.$$

Definition 7. *A submanifold N of M is* totally geodesic *at $p \in N$ if every geodesic γ in M with $\gamma(0) = p, \gamma'(0) \in T_pN$ remains in N on some interval $(-\epsilon, \epsilon)$. The submanifold N is said to be totally geodesic if it is totally geodesic at every point.*

As a consequence, we have that N is totally geodesic if and only if every geodesic in N is also a geodesic in M.

Proposition 6. *Let N be a submanifold of (M, ∇). Then*

1. *If N is auto-parallel in M then N is totally geodesic.*
2. *If M is totally geodesic and ∇ is symmetric then M is autoparallel.*

Since the metric connection is symmetric, the last proposition gives a complete equivalence between auto-parallelism and totally geodesic submanifolds.

We can now define linear hyperplanes on Riemannian manifolds.

Definition 8. *A* linear decision boundary *N in M is an autoparallel submanifold of M such that $M \setminus N$ has two connected components.*

Several observations are in order. First note that if M is an n-dimensional manifold, the separability condition requires N to be an $(n-1)$-dimensional submanifold. It is easy to see that every affine subspace of $\mathbb{R}^n$ is totally geodesic and hence autoparallel. Conversely, since the metric connection is symmetric, every auto-parallel submanifold of Euclidean space that separates it is an affine subspace. As a result, we have that our generalization does indeed reduce to affine subspaces under Euclidean geometry. Similarly, the above definition reduces to spherical hyperplanes $H_u \cap \mathbb{S}^n$ and $H_u \cap \mathbb{S}^n_+$. Another example is the hyperbolic half plane $\mathbb{H}^2$ where the linear decision boundaries are half-circles whose centers lie on the x axis.

Hyperplanes on $\mathbb{S}^n$ have the following additional nice properties. They are the set of equidistant points from $x, y \in \mathbb{S}^n$ (for some x, y), they are isometric to $\mathbb{S}^{n-1}$ and they are parameterized by n parameters. These properties are particular to the sphere and do not hold in general (Bridson & Haefliger, 1999).

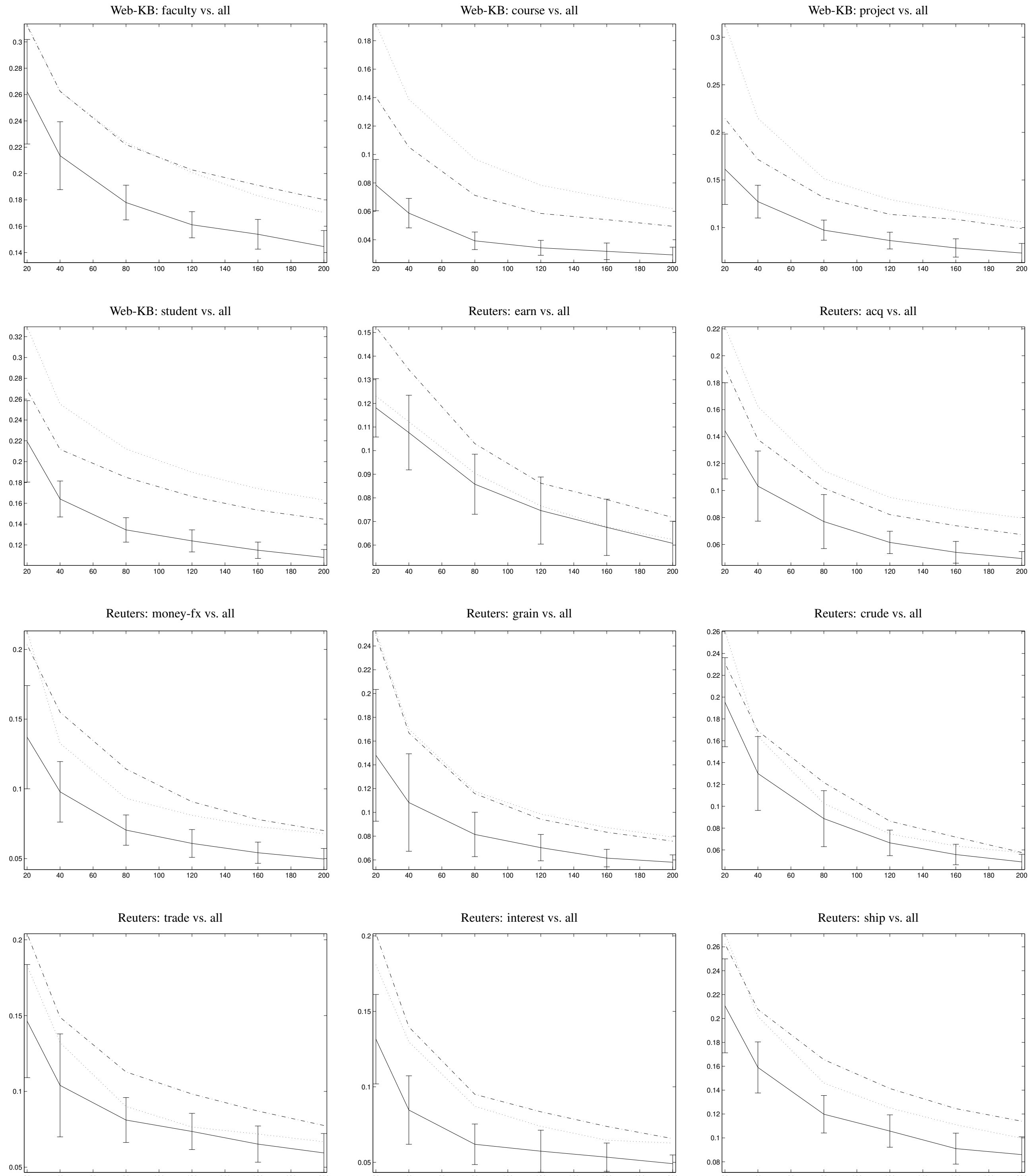

Figure 4. Test error accuracy of spherical logistic regression (solid), and linear logistic regression using tf representation with L_1 normalization (dashed) and L_2 normalization (dotted). The first four figures show Web-KB binary "one vs. all" tasks. The next 8 figures show the Reuters-21578 binary classification tasks. Error bars represent one standard deviation over 20-fold cross validation for spherical logistic regression. The error bars of the other classifiers are of similar sizes and are omitted for clarity.

7. Experiments

A natural embedding of text documents in the multinomial simplex is the L_1 normalized term frequency or tf representation (Joachims, 2000)

$$\hat{\theta}(x) = \left(\frac{x_1}{\sum_i x_i}, \dots, \frac{x_{n+1}}{\sum_i x_i} \right).$$

Using this embedding we compared the performance of spherical logistic regression with Euclidean logistic regression. Since Euclidean logistic regression often performs better with L_2 normalized tf representation, we included these results as well.

The embedding may be motivated by the following argument. Assuming that the text documents are generated by multinomial distributions $\theta \mapsto x$, the embedding $\hat{\theta}$ is theoretically justified as the maximum likelihood estimator. It makes more sense to view tf feature vectors as points in the simplex and not in the much larger Euclidean space. The choice of the Fisher information metric is motivated by the axiomatic characterization of Čencov (1982) and by the vast experimental evidence of its usefulness in statistics.

Experiments were conducted on both the Web-KB and the Reuters-21578 datasets. In the Web-KB dataset, the classification task that was tested was each of the classes faculty, course, project and student vs. the rest. In the Reuters dataset, the task was each of the 8 most popular classes vs. the rest. The test error rates as a function of randomly sampled training sets of different sizes are shown in Figure 4. In both cases, the positive and negative example sets are equally distributed, and the results were averaged over a 20-fold cross validation with the error bars indicating one standard deviation. As mentioned in Section 4, we assume that the boundary set of $q = \arg\min_{y \in \mathbb{S}^n_+ \cap E_u} d(y, x)$ is equal to $A = \{i : (p_u(x))_i \le 0\}$.

The experiments show that the new linearity and margin concepts lead to more powerful classifiers than their Euclidean counterparts, which are commonly used in the literature regardless of the geometry of the data.

8. Summary

We have presented a generalization of hyperplane margin classifiers to the multinomial manifold. In related work, Gous (1998) treats regression rather than classification, and works with affine spherical subfamilies; see also (Hall & Hofmann, 2000). Under affine subfamilies, many of the geometrical properties developed here for margin-based hyperplane models under multinomial geometry do not apply.

The point of view of treating text documents as points on the simplex and using the Fisher information for constructing new classification schemes is presented in (Lafferty & Lebanon, 2003). For categorical data, such as text, that naturally lie on the multinomial manifold, the new concepts of spherical hyperplanes and spherical margins presented here are better motivated than their Euclidean counterparts. Experimental results on the Web-KB and Reuters-21578 datasets show that the resulting geometrical approach of spherical logistic regression leads to improved performance over standard logistic regression, which assumes Euclidean geometry.

Acknowledgements

This work was supported in part by NSF grants CCR-0122581 and IIS-0312814, and by ARDA contract MDA904-00-C-2106.

References

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry*. American Mathematical Society.

Bridson, M., & Haefliger, A. (1999). *Metric spaces of non-positive curvature*, vol. 319 of *A Series in Comprehensive Studies in Mathematics*. Springer.

Čencov, N. N. (1982). *Statistical decision rules and optimal inference*. American Mathematical Society.

Gous, A. (1998). *Exponential and spherical subfamily models*. Doctoral dissertation, Stanford University.

Hall, K., & Hofmann, T. (2000). Learning curved multinomial subfamilies for natural language processing and information retrieval. *International Conference on Machine Learning*.

Joachims, T. (2000). *The maximum margin approach to learning text classifiers methods, theory and algorithms*. Doctoral dissertation, Dortmund University.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical Science*, *4*, 188–234.

Kass, R. E., & Voss, P. W. (1997). *Geometrical foundations of asymptotic inference*. John Wiley & Sons, Inc.

Lafferty, J., & Lebanon, G. (2003). Information diffusion kernels. *Advances in Neural Information Processing, 15*.

Lee, J. L. (1997). *Riemannian manifolds, an introduction to curvature*. Springer.

Spivak, M. (1975). *A comprehensive introduction to differential geometry*, vol. 1-5. Publish or Perish.