

Hyperplane Margin Classifiers on the Multinomial Manifold

Guy Lebanon

Joint work with John Lafferty

Linear Classifiers and Euclidean Geometry

- Linear classifiers are a mainstay of machine learning algorithms including SVM, logistic regression, AdaBoost and the perceptron. Best classification results in text classification.
- Their motivation stems from an implicit assumption of Euclidean geometry

Linear Classifiers and Euclidean Geometry

- Linear classifiers are a mainstay of machine learning algorithms including SVM, logistic regression, AdaBoost and the perceptron. Best classification results in text classification.
- Their motivation stems from an implicit assumption of Euclidean geometry

In the absence of Euclidean geometry, a generalization of linear classifiers, tuned to the geometry of the data, outperforms its Euclidean counterpart

Outline

- Properties and motivations of linear classifiers
- Hyperplanes and margins in the multinomial simplex
- Logistic regression for multinomial geometry
- Text classification experiments
- Conclusions

Linear Classifiers

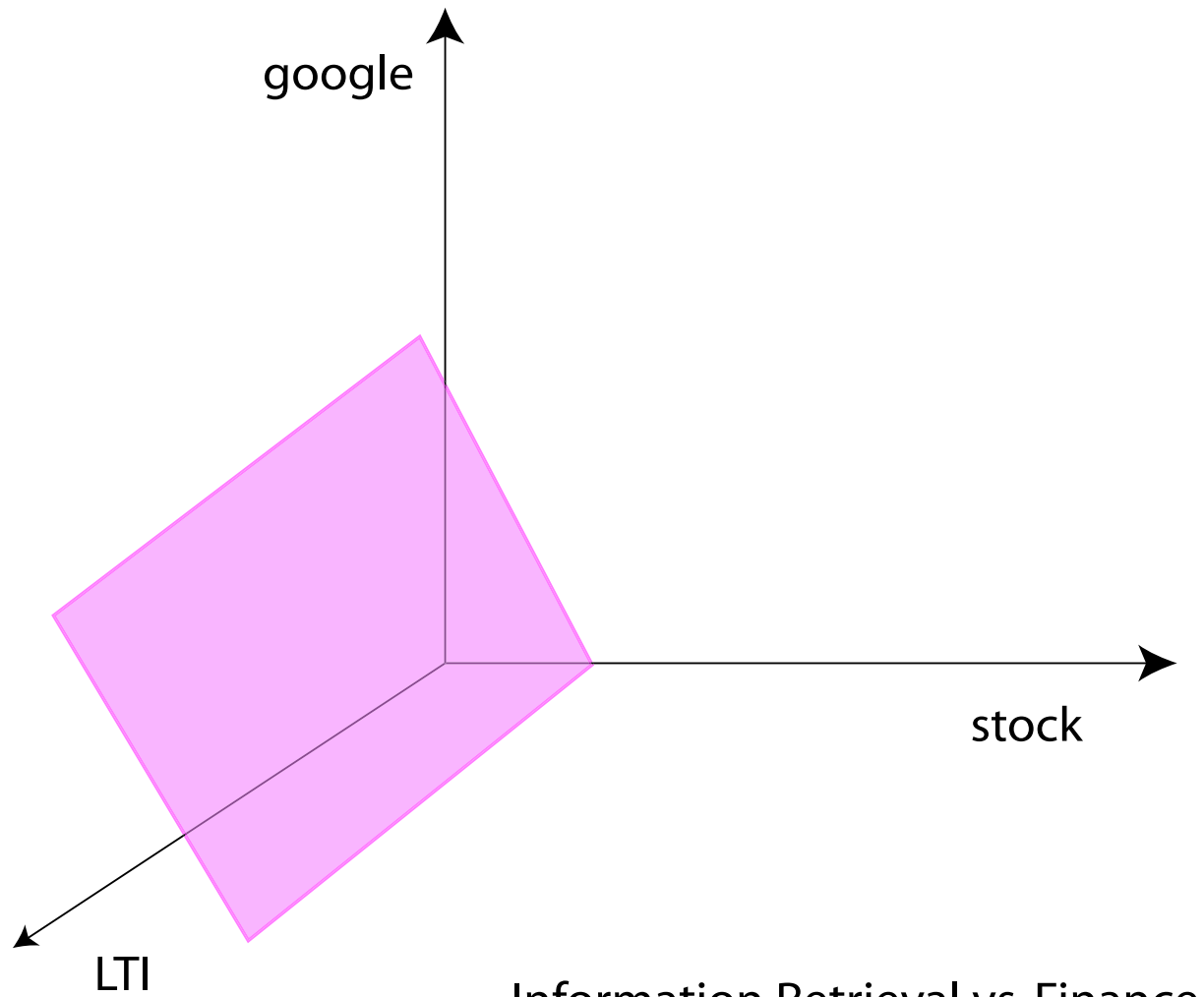
- Algebraic form

$$\hat{y}(x) = \text{sign} \left(\sum_i w_i x_i \right) = \text{sign}(\langle w, x \rangle) \in \{-1, +1\}$$

- Geometrically, the decision surface is a hyperplane or an affine subspace

$$\{x \in \mathbb{R}^n : \langle x, w \rangle = 0\}$$

- Examples: support vector machine, AdaBoost, logistic regression, perceptron etc.



Arguments for Linearity

To avoid overfitting in choosing a classifier $f \in \mathcal{F}$ based on the training data, the candidate family \mathcal{F} has to be

1. rich enough to allow a good description of the data
2. simple enough to avoid overfitting

This is a fundamental tradeoff in which the class of linear decision surfaces strikes a good balance.

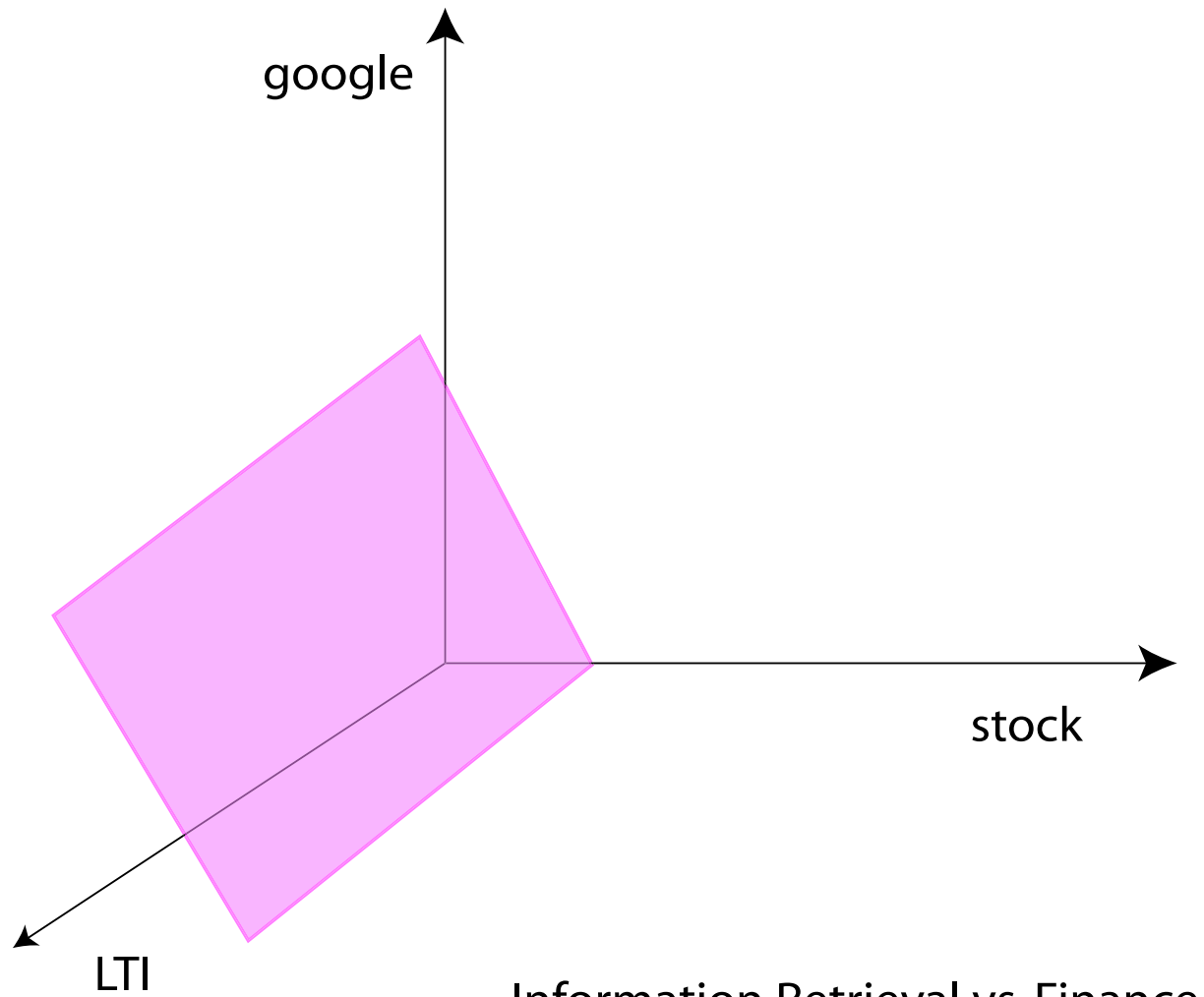
Distinguishing Properties of a Hyperplane

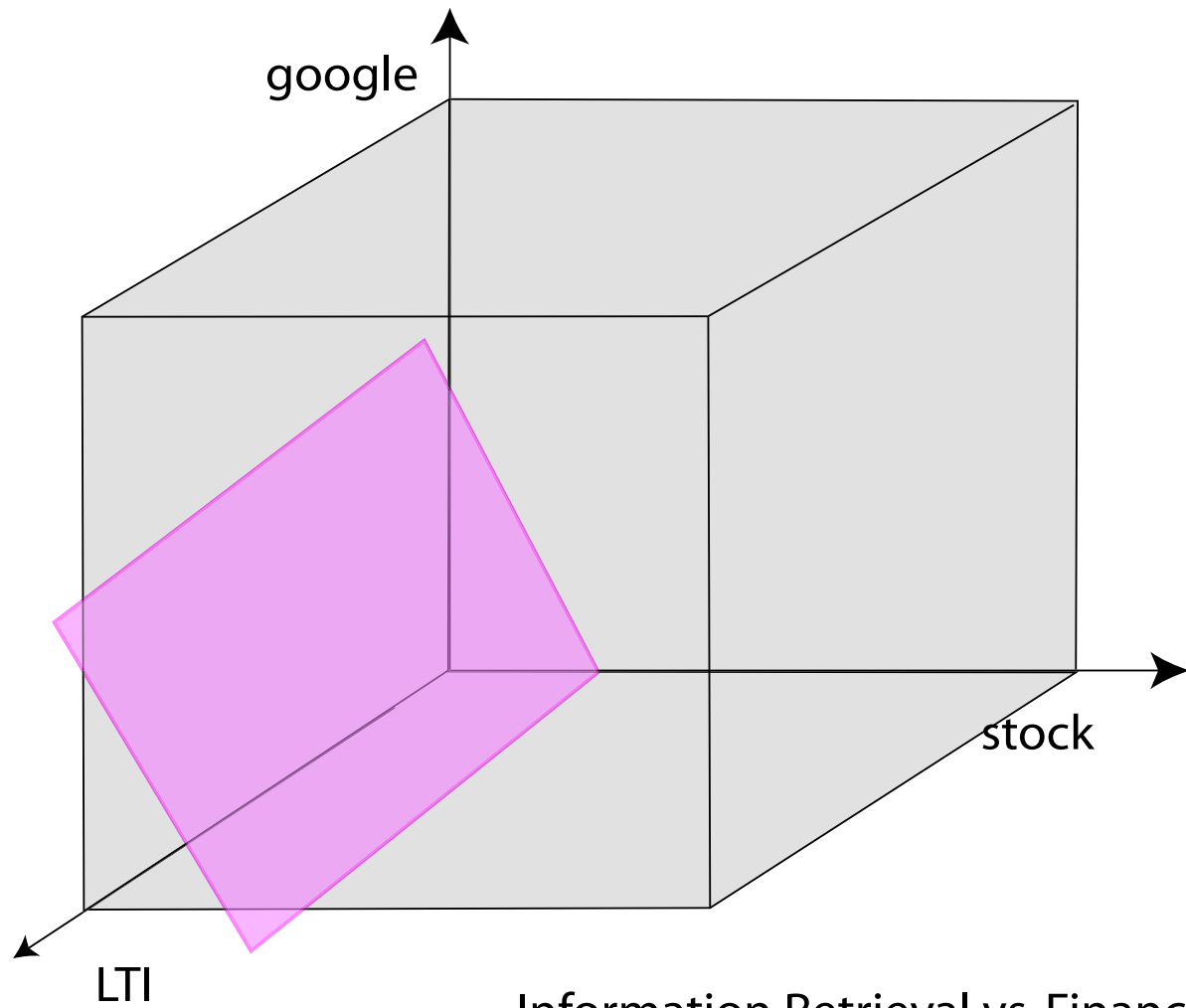
- The set of points equidistant from $x, y \in \mathbb{R}^n$
- Optimal classifier between $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$
- Isometric to a reduced dimension version of the space
- A union of distance minimizing curves (geodesics)

Distinguishing Properties of a Hyperplane

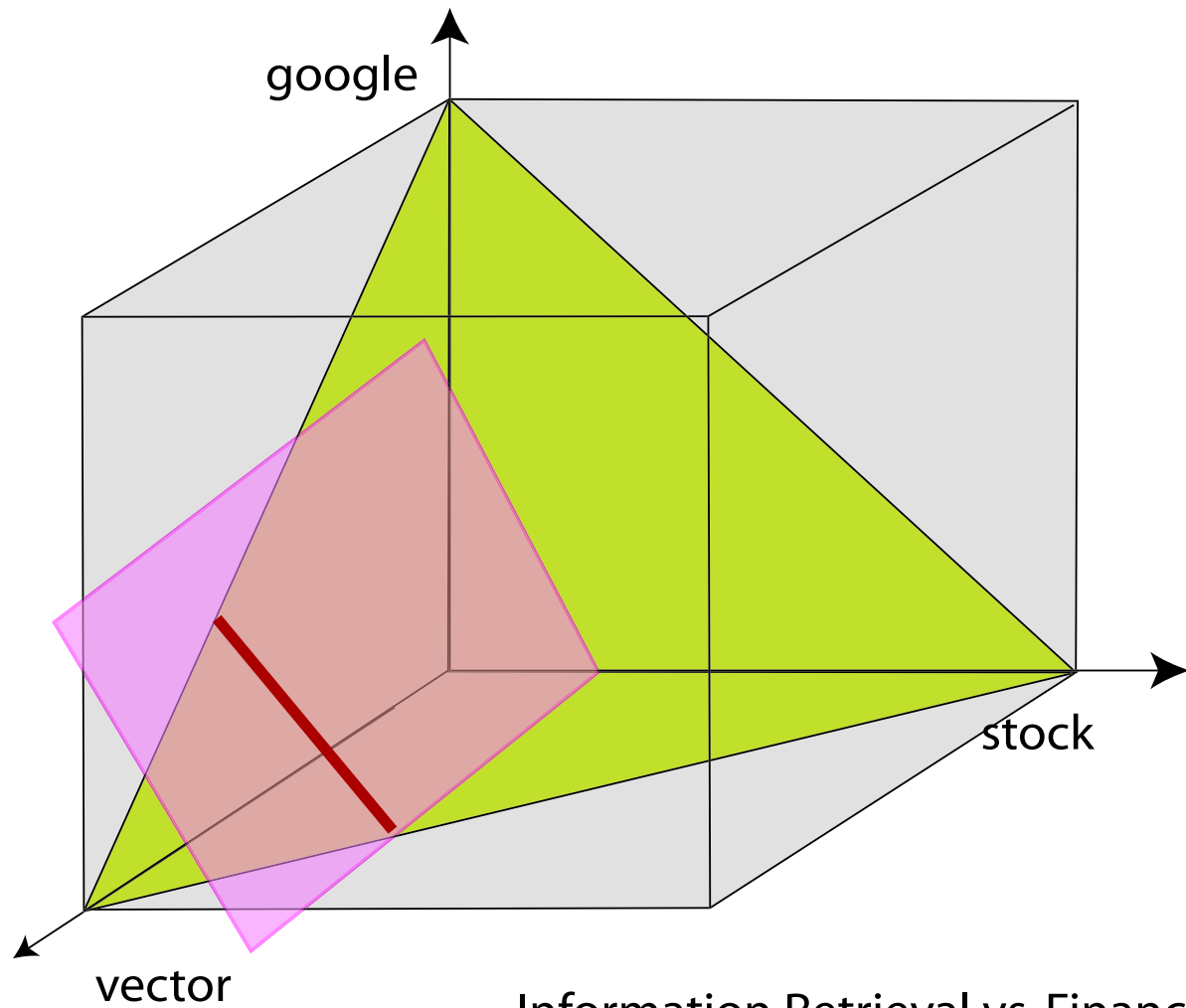
- The set of points equidistant from $x, y \in \mathbb{R}^n$
- Optimal classifier between $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$
- Isometric to a reduced dimension version of the space
- A union of distance minimizing curves (geodesics)

Euclidean geometry is implicit in these properties





Information Retrieval vs. Finance



Information Retrieval vs. Finance

Objections to Euclidean Geometry

Data is often embedded in a Euclidean geometry without careful considerations

- Topological Objection: Discrete data is only artificially viewed as a subset of \mathbb{R}^n
- Geometric Objection: Distances between objects are often not Euclidean

Objections to Euclidean Geometry

Data is often embedded in a Euclidean geometry without careful considerations

- Topological Objection: Discrete data is only artificially viewed as a subset of \mathbb{R}^n
- Geometric Objection: Distances between objects are often not Euclidean

We generalize the idea of margin based hyperplane classifiers to Riemannian manifolds. We treat in detail the analogue of logistic regression in the multinomial manifold with the Fisher geometry.

Hyperplanes and Margins in Riemannian Manifolds

Definition: A hyperplane in a manifold M is an **autoparallel** submanifold N such that $M \setminus N$ has **two connected components**

The first condition guarantees flatness of the hyperplane and the second guarantees that it is a decision boundary

Definition: The margin of $x \in M$ with respect to a hyperplane N is $d(x, N) = \inf_{y \in N} d(x, y)$

In the general case hyperplanes may not exist and the margin may be difficult to compute

The Multinomial Manifold and the Fisher metric (\mathbb{P}^n, g)

$$\mathbb{P}^n = \left\{ x \in \mathbb{R}^{n+1} : \forall j \ x_j \geq 0, \sum_{i=1}^{n+1} x_i = 1 \right\} \quad g_x(u, v) = \sum_{i=1}^{n+1} \frac{u_i v_i}{x_i}$$

where u, v are vectors tangent to \mathbb{P}^n represented in \mathbb{R}^{n+1}

- \mathbb{P}^n is the natural space for considering frequencies of categorical data e.g. word counts in text classification
- g is the unique metric invariant under congruent embeddings

The Simplex and the Positive Sphere

The simplex (\mathbb{P}^n, g) 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 δ of the embedding Euclidean space.

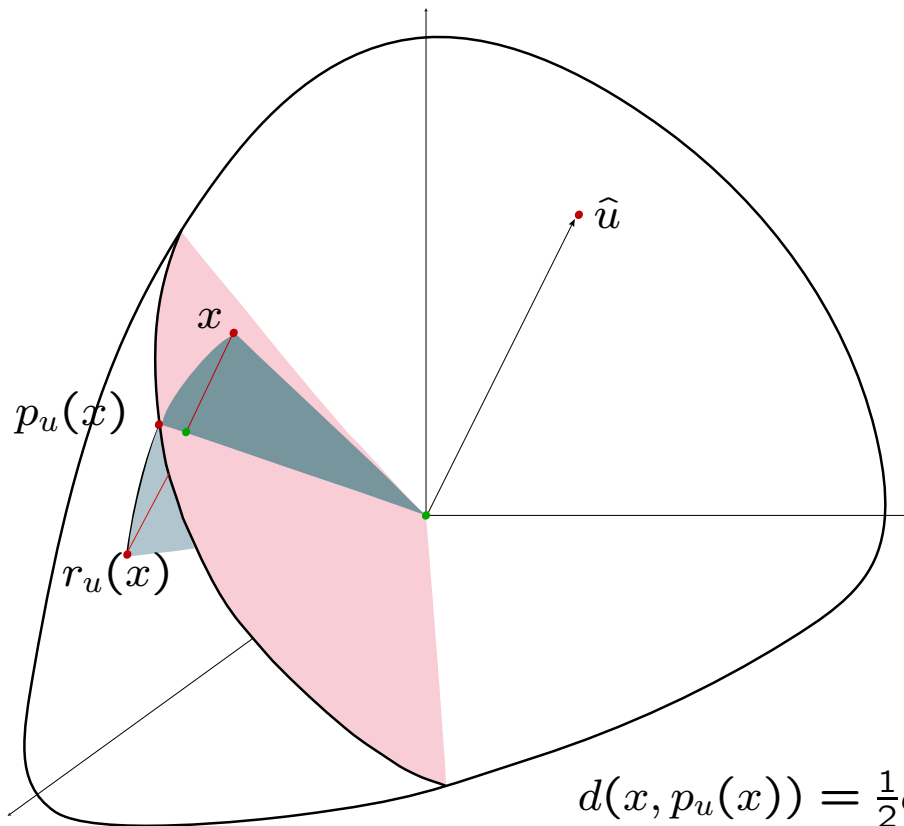
The isometry

$$\pi : (\mathbb{P}^n, g) \rightarrow (\mathbb{S}_+^n, \delta) \quad \pi(x) = (\sqrt{x_1}, \dots, \sqrt{x_{n+1}})$$

allows us to perform our calculations on (\mathbb{S}_+^n, δ) and apply them to (\mathbb{P}^n, g) through π^{-1} .

Hyperplanes and Margins on \mathbb{S}^n

Definition: A hyperplane is $H_u = \mathbb{S}^n \cap E_u$ where E_u is an n dimensional subspace of \mathbb{R}^{n+1} associated with the normal u .

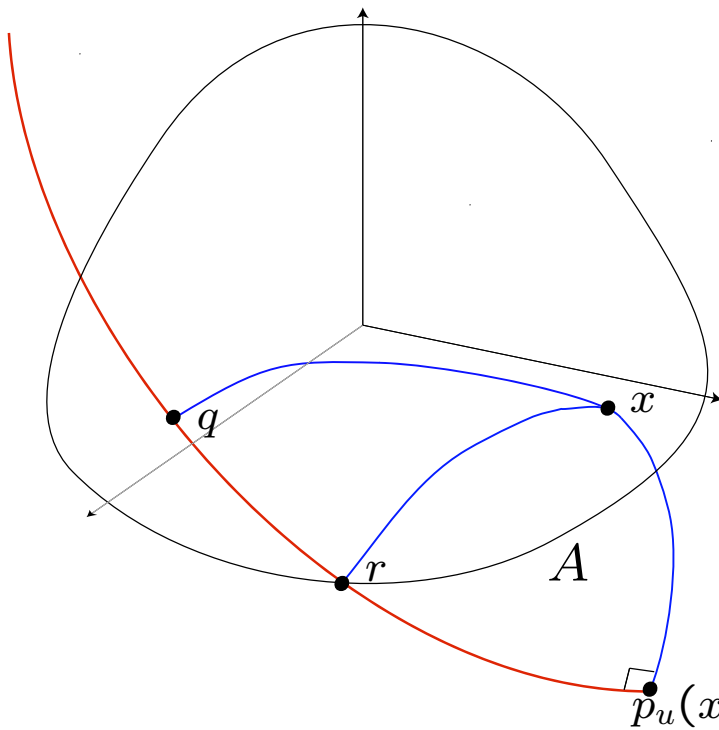


$$d(x, p_u(x)) = \frac{1}{2}d(x, r_u(x)) = \arccos(\sqrt{1 - \langle x, \hat{u} \rangle^2})$$

Hyperplanes and Margins on \mathbb{S}_+^n

The margin in \mathbb{S}_+^n is generally larger than in \mathbb{S}^n

$$d(x, H_{u+}) = \inf_{y \in E_u \cap \mathbb{S}_+^n} d(x, y) \geq \inf_{y \in E_u \cap \mathbb{S}^n} d(x, y) = d(x, H_u)$$



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

Logistic Regression on \mathbb{S}_+^n

Logistic regression may be re-parameterized as

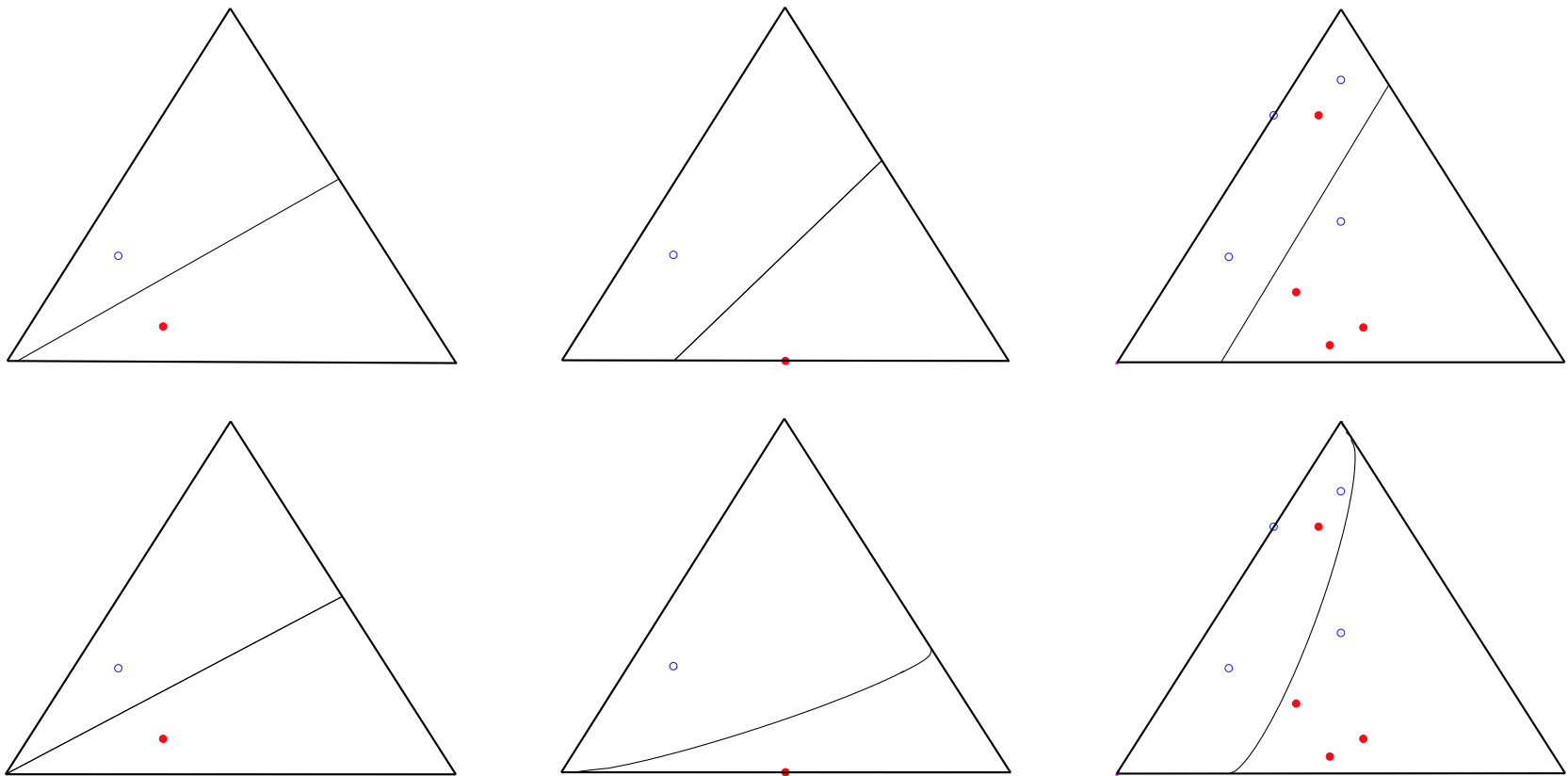
$$\begin{aligned} p(y | x; u) &\propto \exp(y \langle x, u \rangle) = \exp(y \|u\| \langle x, \hat{u} \rangle) \\ &= \exp(y \theta \operatorname{sign}(\langle x, \hat{u} \rangle) d(x, H_{\hat{u}})) = p(y | x; \hat{u}, \theta) \end{aligned}$$

where $d(x, H_{\hat{u}})$ should depend on the choice of the geometry.

In (\mathbb{S}_+^n, δ) we obtain

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

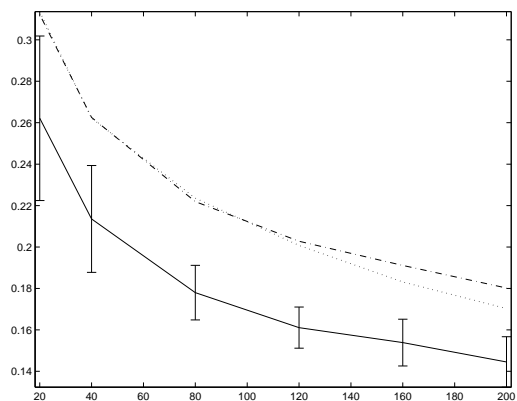
MLE for Euclidean and multinomial logistic regression



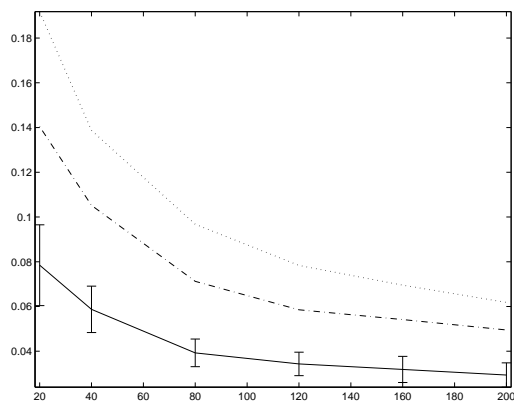
Experiments

- Term frequency representation, interpreted as a multinomial model, embeds documents in (\mathbb{P}^n, g)
- Experiments on WebKB and Reuters datasets indicate that logistic regression using the multinomial geometry consistently outperforms its Euclidean counterpart

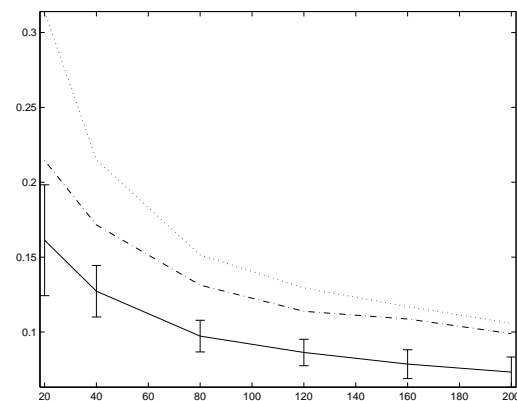
Web-KB: faculty vs. all



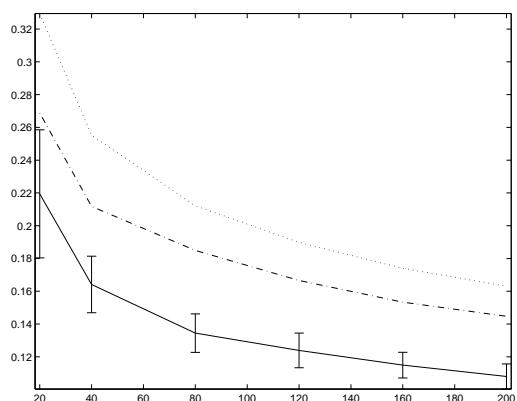
Web-KB: course vs. all



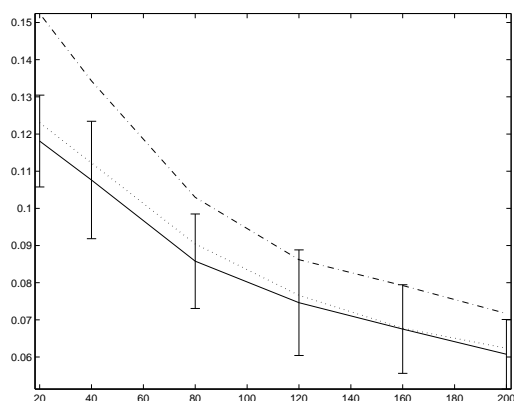
Web-KB: project vs. all



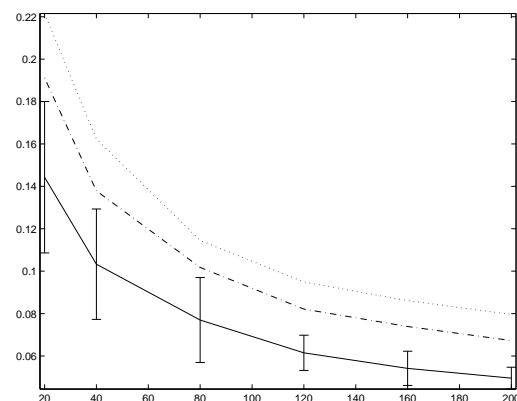
Web-KB: student vs. all



Reuters: earn vs. all

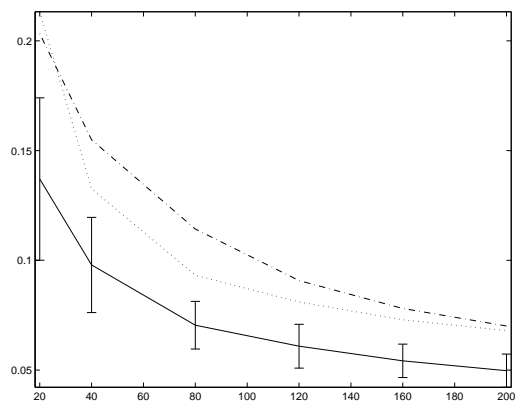


Reuters: acq vs. all

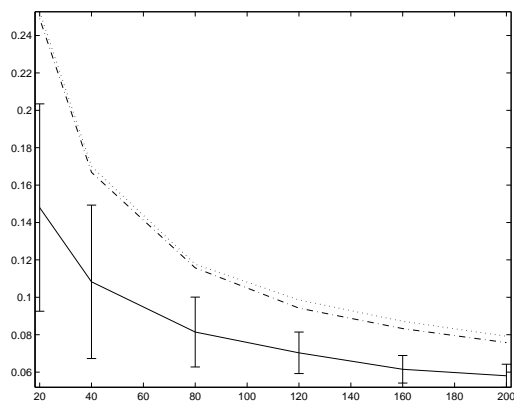


Accuracy vs. train set size for multinomial (solid) and Euclidean logistic regression using TF representation with L_1 normalization (dashed) and L_2 normalization (dotted)

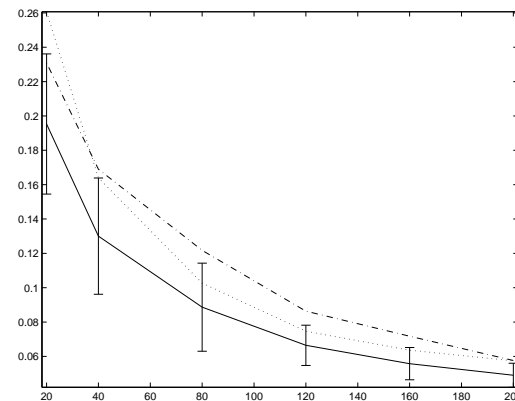
Reuters: money-fx vs. all



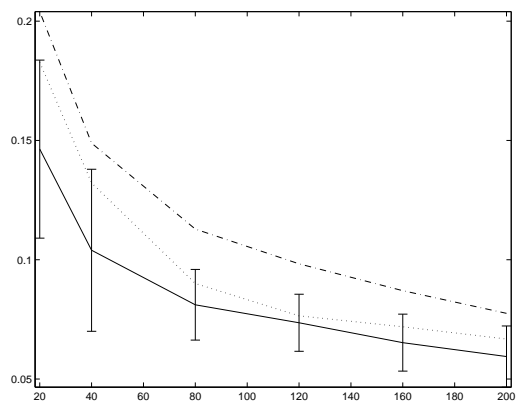
Reuters: grain vs. all



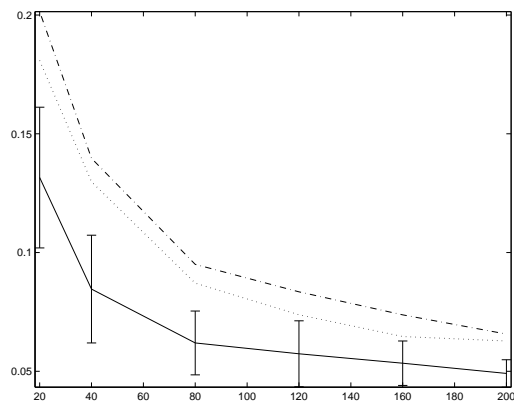
Reuters: crude vs. all



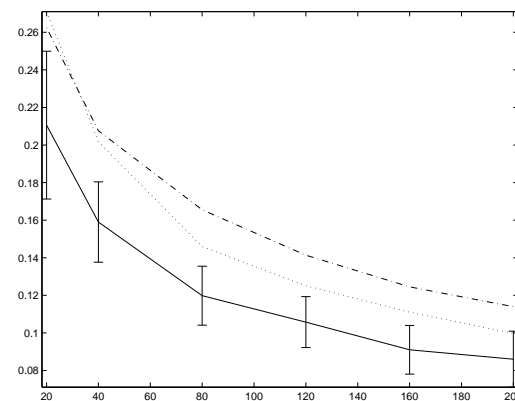
Reuters: trade vs. all



Reuters: interest vs. all



Reuters: ship vs. all



Accuracy vs. train set size for multinomial (solid) and Euclidean logistic regression using TF representation with L_1 normalization (dashed) and L_2 normalization (dotted)

Conclusions

- Linear classifiers based on margin arguments may be generalized to non-Euclidean geometries
- Logistic regression based on multinomial geometry compares favorably to Euclidean logistic regression in text classification
- Generalization to other geometries is not straightforward and remains an open question