LTI Thesis Defense:

Riemannian Geometry and Statistical Machine Learning

Guy Lebanon

Committee: John Lafferty
Geoff Gordon, Michael I. Jordan, Larry Wasserman

Outline

- Motivation
- Introduction
- Previous Work
- Research Results
- Summary

Motivation

- Generative statistical learning
  Select $p(x\,;\theta), \theta \in \Theta$ based on $x_1, \ldots, x_n \subset \mathcal{X}$
- Conditional statistical learning
  Select $p(y|x\,;\theta), \theta \in \Theta$ based on $(x_1, y_1) \ldots, (x_n, y_n) \subset \mathcal{X} \times \mathcal{Y}$
- Ignore $\mathcal{Y}$ by assumption: $\mathcal{Y} = \{y_1, \ldots, y_c\}, \mathcal{X} \times \mathcal{Y} \cong \mathcal{X}^c$

- Θ,$\mathcal{X}$ are often continuous, differentiable and locally Euclidean (manifolds)
- Learning algorithms make implicit or explicit assumptions about the geometry of Θ,$\mathcal{X}$
  - For example, MLE for logistic regression assumes Θ has Fisher geometry and $\mathcal{X}$ is Euclidean (not trivial!)

Thesis Goals:

- Analyze the geometric properties of statistical learning algorithms
- Adapt learning algorithms to alternative geometries obtained through
  - expert knowledge
  - axiomatic system
  - unsupervised adaption to data

Geometric Formalism

Θ,$\mathcal{X}$ are

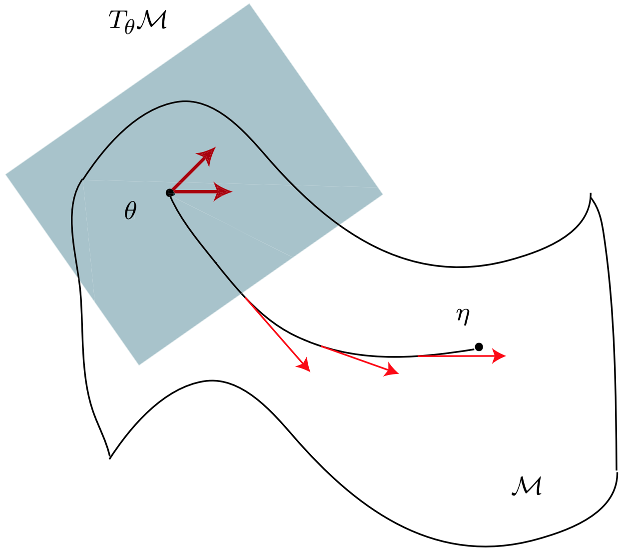
- often continuous and differentiable spaces
- often locally Euclidean
- but not always vector spaces ($\theta_1 - \theta_2$, $-3x_1$?)

$\Rightarrow$ Use Riemannian geometry formalism, which includes as special case Euclidean geometry and Fisher geometry

Riemannian Geometry

- A manifold Θ is a continuous and differentiable set of points that is locally equivalent to $\mathbb{R}^n$ (e.g. open subsets of $\mathbb{R}^n$)
- Every point $\theta \in \Theta$ is equipped with an n-dimensional vector space $T_\theta\Theta$ called the tangent space.
- Geometry is determined by a local inner product between tangent vectors $g_\theta(u, v), \;\; u, v \in T_\theta\Theta$

- Length of tangent vectors $u \in T_\theta\Theta$ defined by
$$\|u\| = \sqrt{g_\theta(u,u)}$$
- Length of paths $c : [a, b] \to \Theta$ defined by
$$L(c) = \int_a^b \|\dot{c}(t)\| \, dt$$
- Distance defined by length of shortest connecting path
$$d(x, y) = \inf_c L(c) = \inf_c \int \sqrt{g_{c(t)}(\dot{c}(t), \dot{c}(t))} \, dt$$



Geometry of Finite Dimensional Probability Spaces

- The space of positive probability distributions over $\mathcal{X}$, $|\mathcal{X}| = m+1$, is the m-simplex

$$\mathbb{P}_m = \left\{ x \in \mathbb{R}^{m+1} : x_i > 0, \sum_i x_i = 1 \right\}$$

- Similarly, the space of all positive conditional models for $\mathcal{X}, |\mathcal{X}| = k$ and $\mathcal{Y}, |\mathcal{Y}| = m+1$ is
  - $\mathbb{P}_m \times \cdots \times \mathbb{P}_m = \mathbb{P}_m^k$ (normalized)
  - $\mathbb{R}_+^{m+1\times k}$ (non-normalized)

- Fisher geometry is given by the metric

$$g_\theta(u,v) = \sum_{i=1}^n \sum_{j=1}^n u_i v_j \int p(x\,;\theta) \frac{\partial \log p(x\,;\theta)}{\partial \theta_i} \frac{\partial \log p(x\,;\theta)}{\partial \theta_j} \,\mathrm{d}x$$

- Resulting distance is

$$d(p(x\,;\theta), p(x\,;\eta)) = d(\theta,\eta) = 2\arccos(\sum \sqrt{\theta_i \eta_i})$$

Previous Work (milestones)

- Connections between asymptotic statistics and Fisher geometry on Θ (Rao '45, Efron '75, Dawid '75)

- Axiomatic derivation of the Fisher geometry (Čencov '82, Campbell '86)

- Relations between I-divergence, KL-divergence, Hellinger distance and distance under Fisher geometry (Kullback '68, Csiszár '75, '91)

- Majority of research traditionally focused on a new interpretation of existing results from asymptotic statistics
- However, some recent algorithmic research, for which the geometric viewpoint is crucial
  - Natural gradient (Amari '98)
  - Fisher kernel (Jaakkola & Haussler, '98)
  - Spherical subfamily regression (Gous, '98)

Contributions, Part I:

Geometry of Spaces of Conditional Models $\Theta = \mathbb{P}^k_m$ and $\Theta = \mathbb{R}^{m+1\times k}_+$

- Geometry of Conditional Exponential Models and AdaBoost
- Axiomatic Geometry for Conditional Models

Geometry of Conditional Exponential Models and AdaBoost

- By using the concept of non-normalized conditional models we can view both algorithms in the same framework

$$q_{\text{mle}}(y|x\,;\theta) = \frac{1}{Z}e^{\langle f(x,y),\theta\rangle} \qquad q_{\text{ada}}(y|x\,;\theta) = e^{\langle f(x,y),\theta\rangle}$$

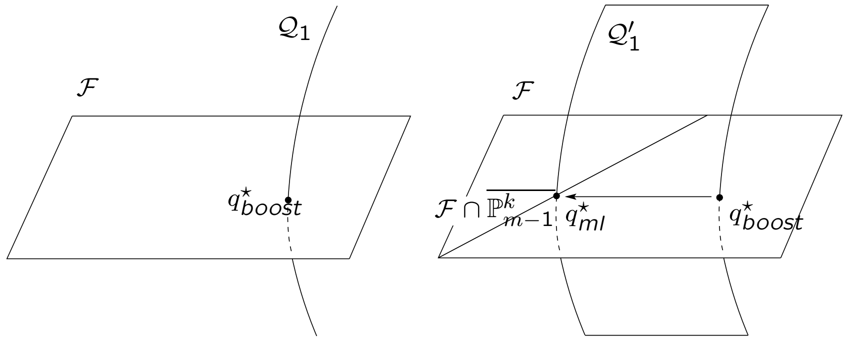
- Several connections shown between MLE for logistic regression and AdaBoost (Friedman et al. '00, Collins et al. '02)

★ We show the strongest connection yet: both problem minimize the I-divergence subject to expectation constraints, except that AdaBoost requires the model to be normalized.

$$\mathcal{F}(\tilde{p}, f) = \left\{ p \in \overline{\mathbb{R}^{k\times m}_+} : \sum_x \tilde{p}(x) \sum_y p(y|x) \left(f_j(x,y) - E_{\tilde{p}}[f_j|x]\right) = 0,\ \forall j \right\}$$

$$D(p,q) = \sum_{i=1}^{n} \sum_y \left(p(y|x_i) \log \frac{p(y|x_i)}{q(y|x_i)} - p(y|x_i) + q(y|x_i) \right)$$

	AdaBoost	Logistic Regression
primal	$\min_p$ $D(p,q_0)$ subject to $p \in \mathcal{F}(\tilde{p}, f)$	$\min_p$ $D(p,q_0)$ subject to $p \in \mathcal{F}(\tilde{p}, f)$, $p \in \overline{\mathbb{P}^k_{m-1}}$
dual	min exp loss for $e^{\langle f(x,y),\theta\rangle}$	MLE for $\frac{1}{Z}e^{\langle f(x,y),\theta\rangle}$



★ Both problems minimize the I-divergence, which approximates the distance under the product Fisher geometry

★ By allowing soft-constraints, the boosting analogue of MAP with Gaussian prior is obtained

$$\begin{array}{ll} \min_p & D(p, q_0) + U(c) \\ \text{subject to} & p \in \mathcal{F}(\tilde{p}, f, c) \end{array}$$

Axiomatic Geometry for Conditional Models

- The only geometry invariant under sufficient statistics transforms is the Fisher geometry (Čencov, '82)
- Extension to non-normalized models (Campbell '86)

We extend Čencov and Campbell's theorems to the conditional case, for both normalized and non-normalized models

★ A set of axioms that corresponds to sufficient statistics transformation is derived

★ A set of metrics on $\mathbb{R}_+^{k \times m}$ that satisfies the axioms is identified

★ If the conditional models are normalized, the metrics above reduce to the product Fisher geometry

★ Using the fact that the I-divergence approximates the distance under the product Fisher geometry we now have an axiomatic framework for conditional exponential models and AdaBoost

Contributions, Part 2:

Geometry of Data Spaces $\mathcal{X}$

- Diffusion Kernels on Statistical Manifolds
- Hyperplane Classifiers on the Multinomial Manifold
- Unsupervised Learning of Metrics

The Embedding Principle

What is the appropriate geometry for $\mathcal{X}$?

★ Embed the data in a manifold of statistical models and use the axiomatic Fisher geometry

- Embedding $\hat{\theta} : \mathcal{X} \to \Theta$ replaces a data point by a model that is likely to generate it
- Example: multinomial MLE or MAP embeds text documents (tf) in the multinomial simplex. Such embedding is dense $\overline{\hat{\theta}(\mathcal{X})} = \overline{\mathbb{P}_n}$.

Diffusion Kernels

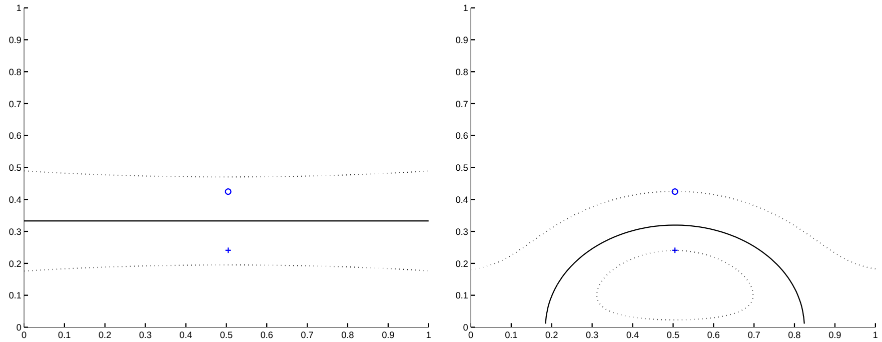
- The heat kernel on a Riemannian manifold is a natural choice for a kernel that incorporates the Riemannian metric to measure proximity between points
- $f(\theta, t) = \int K_t(\theta, \eta) u(\eta)\, d\eta$ is the solution to the heat (diffusion) equation $\frac{\partial f}{\partial t} = \triangle f$ with initial condition u
- $K_t(\theta_1, \theta_2)$ is the amount of heat arriving at θ_1 after time t if the initial heat distribution is concentrated on θ_2

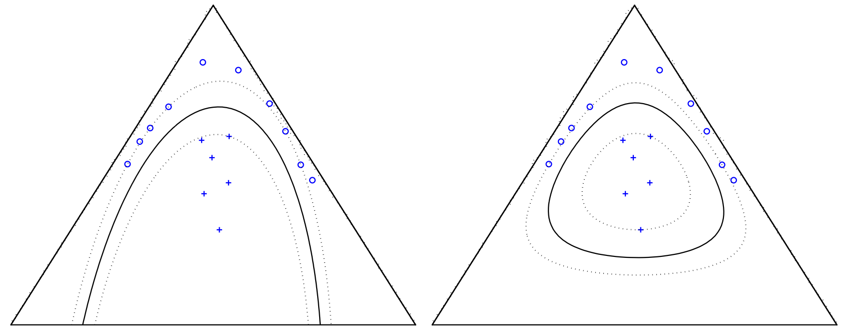
★ Construct the heat kernel for the Fisher geometry of the embedding space $K_t(x, y) = K_t(\hat{\theta}(x), \hat{\theta}(y))$

- In some cases, the heat kernel has a closed form (spherical normal parameter space)
- If closed form not available but distance is known, approximate the heat kernel with parametrix approximation

$$K_t(x, y) \approx \exp\left(-\frac{d^2(\hat{\theta}(x), \hat{\theta}(y))}{4t}\right) \psi_0(\hat{\theta}(x), \hat{\theta}(y))$$

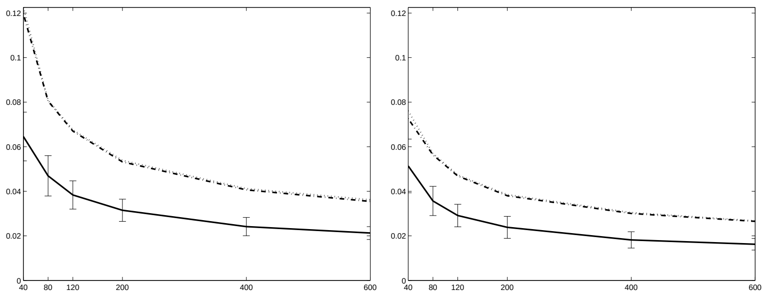
- Squared distance $d^2(x, y)$ may be further approximated as KL divergence $D(x, y)$





★ Approximated diffusion kernel $K_t(\widehat{\theta}(x), \widehat{\theta}(y))$ for text classification outperforms other standard kernels (SVM)

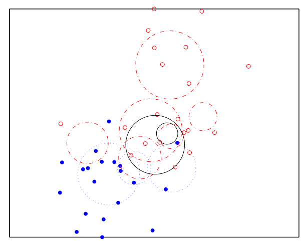
★ Obtain generalization error bounds based on eigenvalue bounds in differentiable geometry



★ Points in $\mathbb{R}^n$ may be embedded as spherical normal models using Dirichlet Process Mixture Model

★ Kernel computed by averaging posterior samples

$$\tilde{K}(x_1, x_2) = \frac{1}{N} \sum_{i=1}^{N} K(\theta^{(i)}(x_1), \theta^{(i)}(x_2)), \ \theta^{(i)} \sim p(\theta_1, \ldots, \theta_m | x_1, \ldots, x_m)$$



Hyperplane Classifiers on the Multinomial Manifold

- 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 $f \in \mathfrak{F}$ based on the training data, the candidate family $\mathfrak{F}$ 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)

Euclidean geometry is implicit in all the arguments above

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

Logistic Regression on the Multinomial Manifold

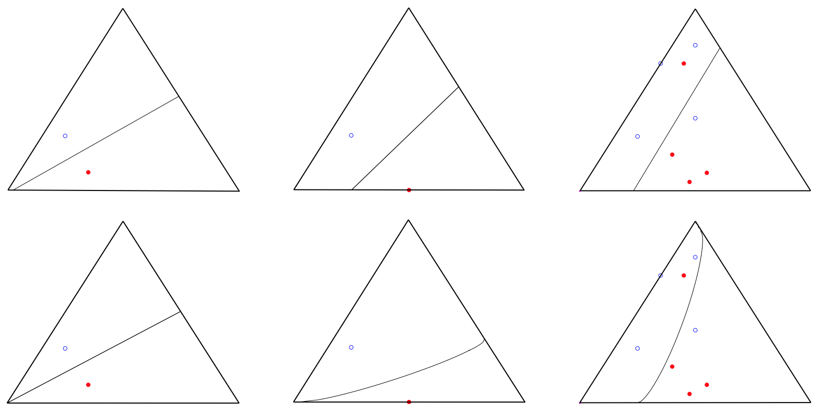
Logistic regression may be re-parameterized as

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

where $H_{\hat\theta}$ is the hyperplane specified by the unit vector $\hat\theta$.

★ replace $d(x, H_{\hat u})$ with a geometry-dependent margin

MLE for Euclidean and multinomial logistic regression



★ 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

Metric Learning

- The axiomatic framework motivates the Fisher geometry if no information other than the parametric family is known.
- If (unlabeled) data is provided, the geometry of $\mathcal{X}$ may be fit by choosing a metric g from a restricted family of metrics $\mathcal{G}$
- Alternative approaches
  - Learning a kernel matrix (Lanckriet et al. '02)
  - Learning a global distance function (Xing et al. '03)

★ A parametric family of metrics $\{g^\lambda : \lambda \in \Lambda\}$ defines a parametric family of models

$$p(x\,;\lambda) = \frac{1}{Z}\left(\sqrt{\det g_x^\lambda}\right)^{-1}$$

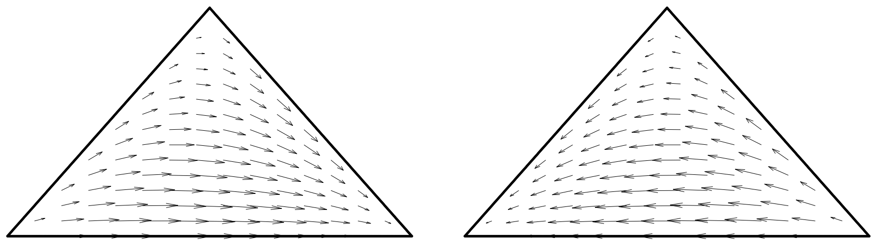
- If g^λ is the Fisher information the numerator is the inverse Jeffreys prior
- The MLE model will have high metric 'volume' in regions that are sparsely populated, hence geodesics will tend to pass along populated regions.

The Parametric Family of Metrics

★ The following Lie group of diffeomorphisms

$$F_\lambda : \mathbb{P}_n \to \mathbb{P}_n \quad F_\lambda(x) = \left(\frac{x_1\lambda_1}{x\cdot\lambda}\ ,\ \ldots\ ,\ \frac{x_{n+1}\lambda_{n+1}}{x\cdot\lambda}\right),$$

acts on the simplex by increasing the components of x with high λ_i values while remaining in the simplex.



F_λ acting on $\mathbb{P}_2$ for $\lambda = (\frac{2}{10}, \frac{5}{10}, \frac{3}{10})$ (left) and F_λ^{-1} (right)

★ The parametric family is the set of pull-back metrics of the Fisher metric through F_λ

$$\mathcal{G} = \{F_\lambda^* \mathcal{J} : \lambda \in \mathbb{P}_n\}.$$

★ The resulting geodesics (under $F_\lambda^* \mathcal{J}$) are

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

- Note the similarity of the geodesic distance to tfidf cosine similarity. The learned λ fill a role similar to idf weights.

★ To obtain a tfidf like effect we compute the MLE metric (quite complicated) and take its Lie-group inverse

★ Resulting weights are similar to tfidf, yet outperform it, when used with nearest neighbor classifier for text classification

Summary

★ A geometric analysis of log. regression and AdaBoost [NIPS'02]

★ Axiomatic framework for geometry of spaces of conditional models [UAI'04, IEEE Trans. Information Theory]

★ Embedding principle allows geometric variants of

- ★ RBF (heat) kernels [NIPS'03, JMLR]
- ★ logistic regression [ICML'04]

★ Unsupervised learning of metrics [UAI'03]

Acknowledgements

I thank

John Lafferty, Geoff Gordon, Michael I. Jordan, Larry Wasserman

Katharina Probst

Leonid Kontorovich, Luo Si, Jian Zhang