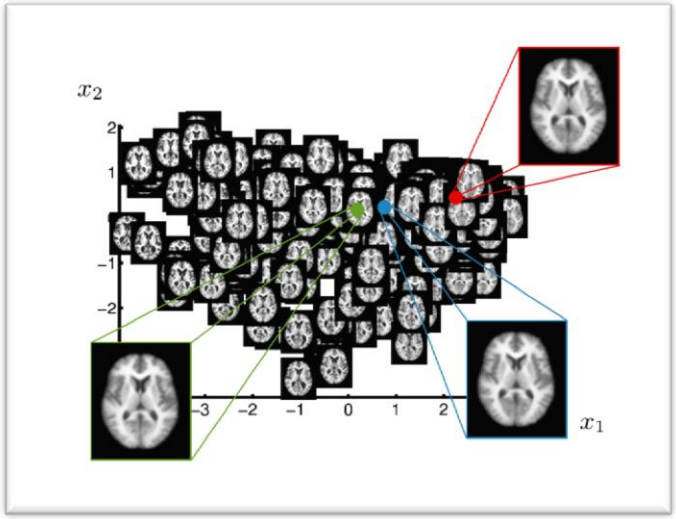
Testing the Manifold Hypothesis

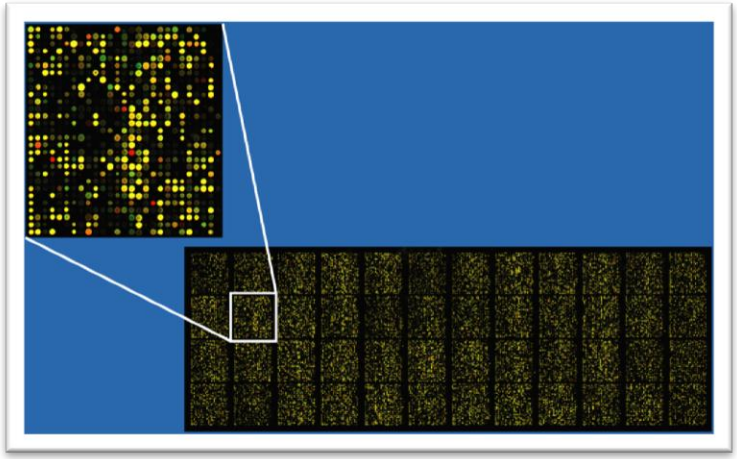
Hariharan Narayanan
Laboratory for Information and Decision Systems
MIT

Based on work with Charles Fefferman
and Sanjoy Mitter

High dimensional data

Gerber et al, On the manifold structure of the space of brain images





Number of dimensions is comparable or larger than number of samples

Curse

Sample complexity of function approximation can grow exponentially

Blessings

Concentration of measure

Asymptotic analysis

[David Donoho, AMS 2000]

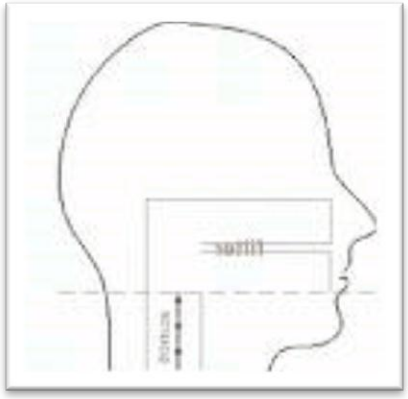
Manifold learning and manifold hypothesis

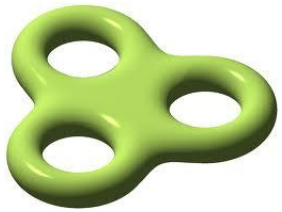
Manifold learning is a collection of methodologies for analyzing data

which are motivated by the manifold hypothesis:

high dimensional data tend to lie near a low dimensional manifold

The hypothesis is a way of avoiding the curse of dimensionality





[Hastie-Stuetzle' 89, Kambhatla-Leen'93, Tannenbaum et al'00, Roweis-Saul'00, Belkin-Niyogi'03, Donoho-Grimes'04]

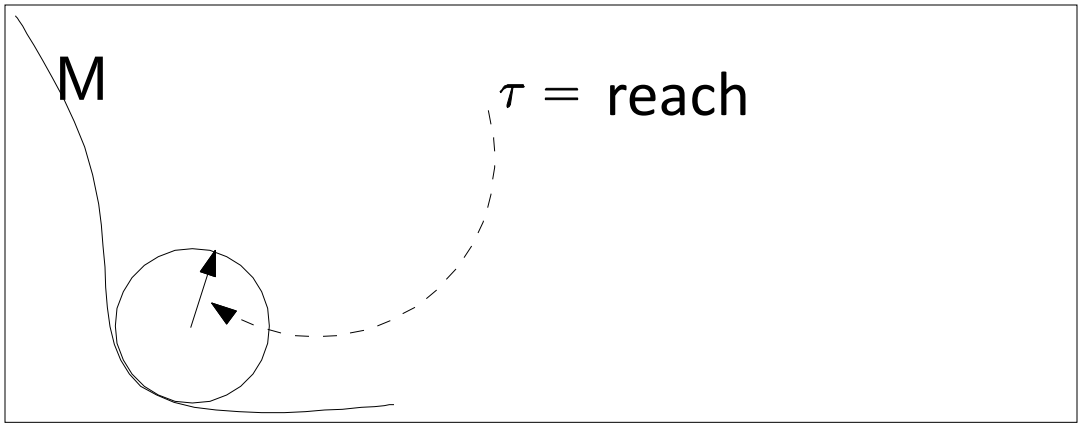
This talk

Testing the Manifold Hypothesis (MH) [Fefferman-Mitter-N'11]

Improved sample complexity analysis of k-means [Fefferman-Mitter-N'11]

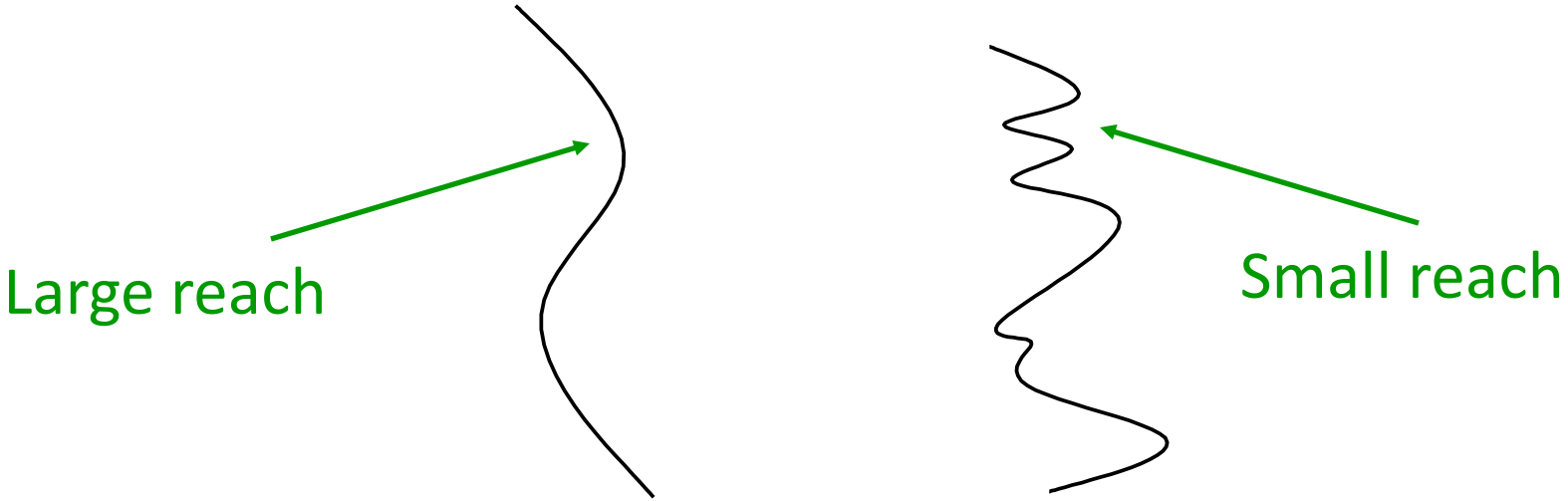
The first sample complexity analysis for k d-planes [Fefferman-Mitter-N'11]

Reach of a submanifold of R^n



τ is the largest number such that for any $r < \tau$

any point at a distance r of $\mathcal{M}$ had a unique nearest point on $\mathcal{M}$



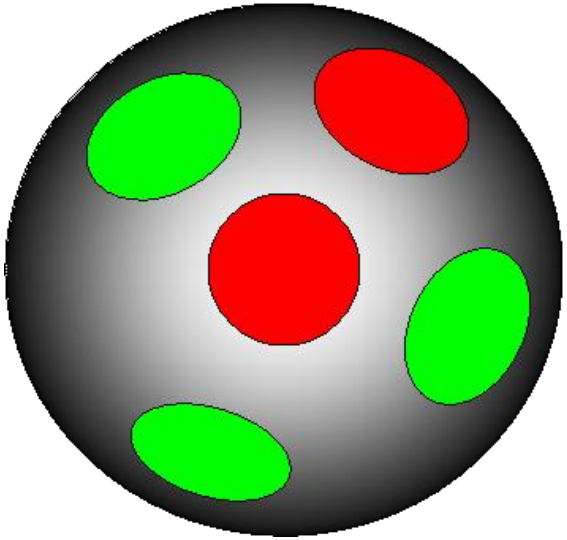
Low dimensional manifolds with bounded volume and curvature

Let $\mathcal{G}_e = \mathcal{G}_e(d, V, \tau)$ be the family of

$d-$submanifolds of the unit ball in $\mathbb{R}^n$, with

volume $\leq V$ and reach $\geq \tau$.

Packing number



$N_p(\epsilon)$ is the largest N s.t. $\mathcal{M}$ contains N disjoint geodesic balls of radius ϵ

In our setting, N_p is bounded above by $VC^d(\frac{d}{\min(\epsilon,\tau)})^d$

Testing the Manifold Hypothesis

Suppose $\mathcal{P}$ is an unknown probability distribution
supported in the unit ball $\mathbb{R}^m$, $m >> 1$
and $x_1, x_2, ...$ are i.i.d random samples from $\mathcal{P}$

Given error ϵ, dimension d, volume V, reach τ and confidence $1 - \delta$
is there an algorithm that takes a number of samples
that is independent of m and outputs whether or not there is

$$\mathcal{M} \in \mathcal{G}_e = \mathcal{G}_e(d, V, \tau)$$

such that w.p $\geq 1 - \delta$, $\mathcal{L}(\mathcal{M}, \mathcal{P}) := \int \mathbf{d}(\mathcal{M}, x)^2 d\mathcal{P}(x) < \epsilon$?

Sample Complexity of testing the manifold hypothesis

What is the number of samples needed for testing

the hypothesis that data lie near a low dimensional manifold?

---

the sample complexity of the task depends only on

the intrinsic dimension, volume and reach, but

not ambient dimension

[N-Mitter NIPS 2010], [Fefferman-Mitter-N 2011]

Sample complexity of testing the Manifold Hypothesis

Loss

$\mathcal{L}(\mathcal{M}, \mathcal{P}) =$ expected squared distance of a random point to $\mathcal{M}$

Empirical Loss

Given a set of data points $x_1, ..., x_s$

$$L_{emp}(\mathcal{M}) = \frac{\sum_i \mathbf{d}(x_i, \mathcal{M})^2}{s}$$

Sample Complexity

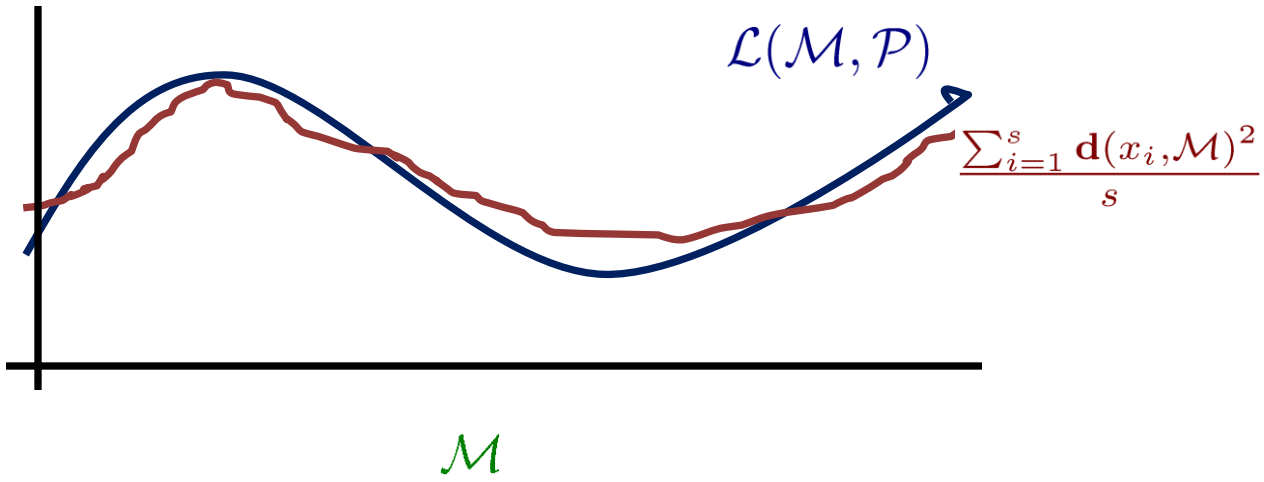
Smallest s such that $\exists$ a rule $\mathcal{A}$ given $x_1, ..., x_s$,

$$\mathbb{P}[\mathcal{L}(\mathcal{M}_{\mathcal{A}}, \mathcal{P}) - \inf_{\mathcal{M}} \mathcal{L}(\mathcal{M}, \mathcal{P}) > \epsilon] < \delta$$

Empirical Risk Minimization

How large must s be to ensure

$$\mathbf{P}\left[\sup_{\mathcal{G}_e}\left|\frac{\sum_{i=1}^{s}\mathbf{d}(\mathcal{M},x_i)^2}{s}-\mathcal{L}(\mathcal{M},\mathcal{P})\right|<\epsilon\right]>1-\delta$$



Fitting manifolds

Theorem:(Fefferman-Mitter-N.'11)
Let $x_1, \ldots, x_s$ be i.i.d samples from $\mathcal{P}$, a distribution supported on the ball of radius 1 in $\mathbb{R}^m$. If

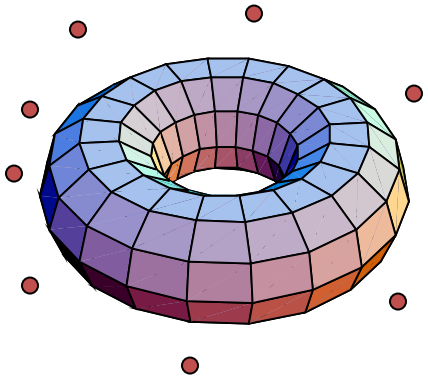
$$s \geq \frac{C\left(V\left(\frac{1}{\epsilon}+\frac{1}{\tau}\right)^{d+o(d)}+\frac{\log 1/\delta}{\epsilon^2}\right)}{\epsilon^2}$$

then $\mathbb{P}\left[\sup_{\mathcal{G}_e}\left|\frac{\sum_{i=1}^{s}\mathbf{d}(x_i,\mathcal{M})^2}{s}-\mathbb{E}_{\mathcal{P}}\mathbf{d}(x,\mathcal{M})^2\right|<\epsilon\right]>1-\delta.$

Proof: Approximates manifolds using point clouds and uses the uniform bound for $k-$means.

Reduction to k-means

Imagine that the manifold is a dense net of $N_p(\epsilon)$ points



$$\mathbf{P}\left[\sup_{\mathcal{G}_e}\left|\frac{\sum_{i=1}^{s}\mathbf{d}(\mathcal{M},x_i)^2}{s}-\mathcal{L}(\mathcal{M},\mathcal{P})\right|<\epsilon\right]>1-\delta$$

$$\mathbf{P}\left[\sup_{\mathcal{G}_{cloud}}\left|\frac{\sum_{i=1}^{s}\mathbf{d}(\mathcal{M},x_i)^2}{s}-\mathcal{L}(\mathcal{M},\mathcal{P})\right|<\epsilon\right]>1-\delta$$

Proving a Uniform bound for k-means

Proving uniform bounds for k-means

reduces to proving a uniform bound over functions of the form

$$\min_{1 \le i \le k} (a_i \cdot x) \qquad \|a_i\| = 1$$

Fat-shattering dimension

The fat-shattering dimension $fat_\epsilon(\mathcal{F})$
of a class $\mathcal{F}$ of real-valued functions is a measure of the complexity
of the function class at a scale ϵ.

$fat_\epsilon(\mathcal{F})$ is largest s such that there exist $x_1, \ldots, x_s$
and thresholds $t_1, \ldots, t_s$ such that for every $\{-1, 1\}$ $s-$vector $(b_1, \ldots, b_s)$,
there is a function $f^b \in \mathcal{F}$ such that $\forall i, (f^b(x_i) - t_i)b_i \geq \epsilon$.

Bound on sample complexity

Theorem: (Uses Dudley's Entropy Integral)

If

$$s \geq \frac{C}{\epsilon^2}\left(\left(\int_{\epsilon/8}^{\infty}\sqrt{\mathrm{fat}_\gamma(\mathcal{F})}\log\left(\frac{\mathrm{fat}_\gamma(\mathcal{F})}{\gamma}\right)d\gamma\right)^2+\log 1/\delta\right),$$

then

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left|\frac{\sum_{i=1}^{s}f(x_i)}{s}-\mathbb{E}_{\mathcal{P}}f\right|\geq\epsilon\right]\leq 1-\delta.$$

VC dimension

The VC dimension $VC(\mathcal{F})$
of a class $\mathcal{F}$ of $\{0,1\}$-valued functions is a measure of its complexity

$VC(\mathcal{F})$ is the largest n such that there are n data
of which all 2^n partitions are induced
by class boundaries of functions in $\mathcal{F}$

If $\mathcal{F}$ consists of the indicators of halfspaces in $\mathbb{R}^d$, $VC(\mathcal{F}) = d + 1$.

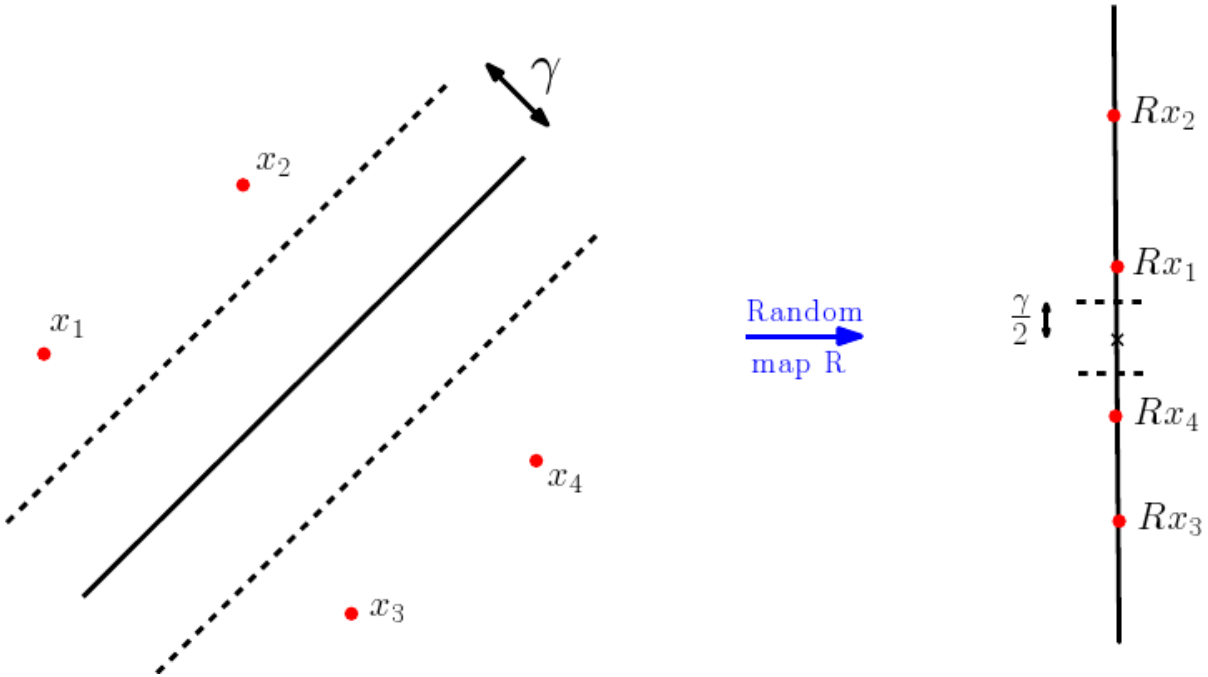
VC dimension

The VC dimension $VC(\mathcal{F})$
of a class $\mathcal{F}$ of $\{0,1\}$-valued functions is a measure of its complexity

For large s, $VC(\mathcal{F}) \log s$ is roughly the
logarithm of the max number of partitions of s data
points that can be induced by functions in $\mathcal{F}$

Random projection

Thanks to Johnson-Lindenstrauss, random projection
of robustly linearly separable s data points, is with probability at least $\frac{1}{2}$
linearly separable in the $\frac{\log s}{\epsilon^2}$ dimensional image space



Random projection

Using VC theory for halfspaces, the logarithm of the number of ways in which the level sets of functions of of the form $\min_{1 \le i \le k} (a_i \cdot x), \|a_i\| = 1$ can partition s points in $\log(s)/\epsilon^2$ dimensional image space is $O((k/\epsilon^2)\log^2(s/\epsilon))$

This gives $fat_\epsilon(\mathcal{F}) \le \frac{k}{\epsilon^2}\log^2(\frac{k}{\epsilon})$

Bound on sample complexity

Theorem:

If

$$s \geq \frac{C}{\epsilon^2}\left(\left(\int_{\epsilon/8}^{\infty}\sqrt{\mathrm{fat}_\gamma(\mathcal{F})}\log\left(\frac{\mathrm{fat}_\gamma(\mathcal{F})}{\gamma}\right)d\gamma\right)^2+\log 1/\delta\right),$$

then

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left|\frac{\sum_{i=1}^{s} f(x_i)}{s}-\mathbb{E}_{\mathcal{P}} f\right|\geq\epsilon\right]\leq 1-\delta.$$

Gives a sample complexity of

$$O\left(\frac{k}{\epsilon^2}\log^6\frac{k}{\epsilon}+\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$$

k-means Clustering

Lower bound :

$$\frac{k}{\epsilon^2} + \frac{\log \frac{1}{\delta}}{\epsilon^2}$$ [Bartlett-Linder-Lugosi'97]

Upper bound :

$$\frac{k^2}{\epsilon^2} + \frac{\log \frac{1}{\delta}}{\epsilon^2}$$ [Maurer-Pontil'08]

$$\frac{k}{\epsilon^2} \log^6 \frac{k}{\epsilon} + \frac{\log \frac{1}{\delta}}{\epsilon^2}$$ [Fefferman-Mitter-N'11]

[Bartlett'97, Dasgupta'02, Guha-Munagala'02, Agarwal-Mustafa'04, Bendavid'07, Maurer-Pontil'08]

Fitting manifolds

Corollary:(Fefferman-Mitter-N.'11)
If

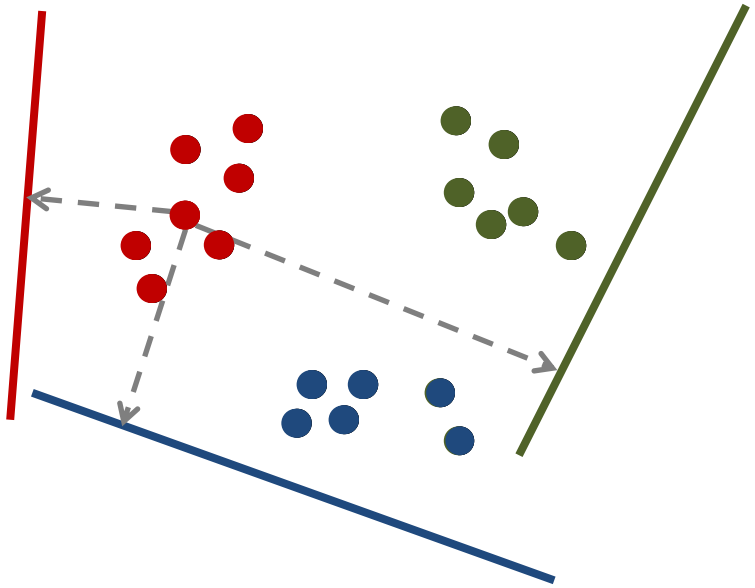
$$s \geq C\left(V\left(\frac{1}{\epsilon}+\frac{1}{\tau}\right)^{d+o(d)}+\frac{\log 1/\delta}{\epsilon^2}\right)$$

then in order to test if there exists a manifold in $\mathcal{G}_e(d,V,\tau)$, such that

$$\mathbb{E}_{\mathcal{P}}\mathbf{d}(x,\mathcal{M})^2 = O(\epsilon),$$

it suffices to take i.i.d samples $x_1,\ldots,x_s$ and test if $\{x_1,\ldots,x_s\}$ is close to such a manifold contained in the affine span of these points.

K d-planes



[Bradley-Mangasarian'99, Lerman'03, Agarwal-Mustafa'06, Zhang et al 08]

K d-planes

Theorem: Let $x_1, \ldots, x_s$ be i.i.d samples from $\mathcal{P}$, a distribution supported on the ball of radius 1 in $\mathbb{R}^m$. If

$$s \geq C\left(\frac{dk}{\epsilon^2}\log^6\left(\frac{dk}{\epsilon}\right) + \frac{d}{\epsilon^2}\log\frac{1}{\delta}\right),$$

then $\mathbb{P}\left[\sup_{F\in\mathcal{F}_{k,d}}\left|\frac{\sum_{i=1}^{s}F(x_i)}{s} - \mathbb{E}_{\mathcal{P}}F(x)\right| < \epsilon\right] > 1-\delta.$

Proof: Uses the kernel trick to map $\Phi : x \mapsto (xx^T, x, 1)$, followed by the use of the uniform bound involving functions of the form $\min_i(a_i \cdot \Phi(x))$.

Algorithmic question

Given N points $x_1, ..., x_N$ in the unit ball in $\mathbb{R}^n$

is there a manifold $\mathcal{M} \in \mathcal{G}_e = \mathcal{G}_e(d, V, \tau)$

such that $\left(\frac{1}{N}\right) \sum_{1 \leq i \leq N} \mathbf{d}(x_i, \mathcal{M})^2 \leq \epsilon$ **?**

[Forthcoming work with Charles Fefferman and Sanjoy Mitter]

Outline

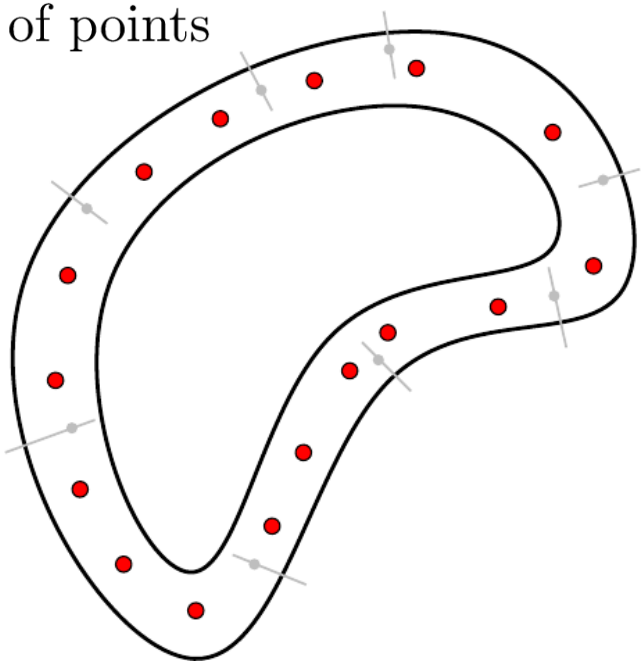
(1) Any manifold $\mathcal{M} \in \mathcal{G}_e = \mathcal{G}_e(d, V, \tau)$

is almost contained in an affine subspace W of dimension N_p

This allows us to reduce the ambient dimension m to roughly N_p

Outline

(2) Reduce the problem to the question of testing whether a discrete evenly spread set of points lie on $\mathcal{M} \in \mathcal{G}_e = \mathcal{G}_e(d, V, \tau)$



(3) Find a smooth vector bundle defined on a tubular neighborhood of data.

(4) Describe the manifold as the set of zeroes of a specific section of the vector bundle and estimate its smoothness.

Key Lemma Suppose that $f : B_d \times B_{n-d} \to R$ is such that

1. for $|\alpha| \leq k$, $|\partial^\alpha f| < C$.
2. For any $(x, y) \in B_d \times B_{n-d}$,

$$C^{-1}(|y|^2 + \rho^2) \leq f(x, y) \leq C(|y|^2 + \rho^2).$$

Then, if ρ is smaller than a controlled constant depending only on k, d, C, the following are true.

1. The set of points x at which the gradient of f is orthogonal to the subspace A_x containing the top $n - d$ eigenvectors of the Hessian of f at that point is a manifold $\mathcal{M}$ whose reach is $c\tau$.
2. Let D^{norm} be the disc bundle over $\mathcal{M}$ whose fiber over a point $x \in \mathcal{M}$ is the disc of radius c in A_x. Then, the bundle injectively embeds in R^n and the image contains a tubular neighborhood of $\mathcal{M}$ of radius c'.
3. The curvature of D^{norm} is bounded below by c'''.

Theorem

There is a controlled constant C depending only on d and an Algorithm that uses

$$n \log N \exp\left(\left(CV(\epsilon^{-d} + \tau^{-d})\right)^{1+o(1)}\right) \log \frac{1}{\delta}$$

operations on real numbers such that given $x_1, \ldots, x_N \in B_n$, with probability at least $1 - \delta$, the Algorithm outputs

1. "Yes" if there exists a manifold $\mathcal{M} \in \mathcal{G}_e(d, V, \tau)$ such that

$$\sum_{i=1}^{N} \mathbf{d}(x, \mathcal{M})^2 \leq \epsilon,$$

2. "No" if there exists no manifold $\mathcal{M}' \in \mathcal{G}_e(d, V, \tau/C)$ such that

$$\sum_{i=1}^{N} \mathbf{d}(x, \mathcal{M}')^2 \leq C\epsilon,$$

Summary

(1) The sample complexity of testing the manifold hypothesis is independent of the ambient dimension

(2) Algorithmic implications for k-means and k d-planes

Thank You!