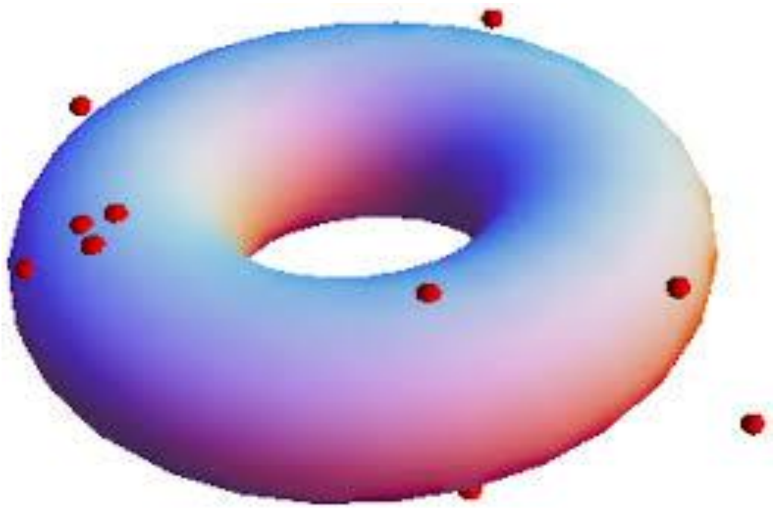
Fitting a manifold of large reach to noisy data

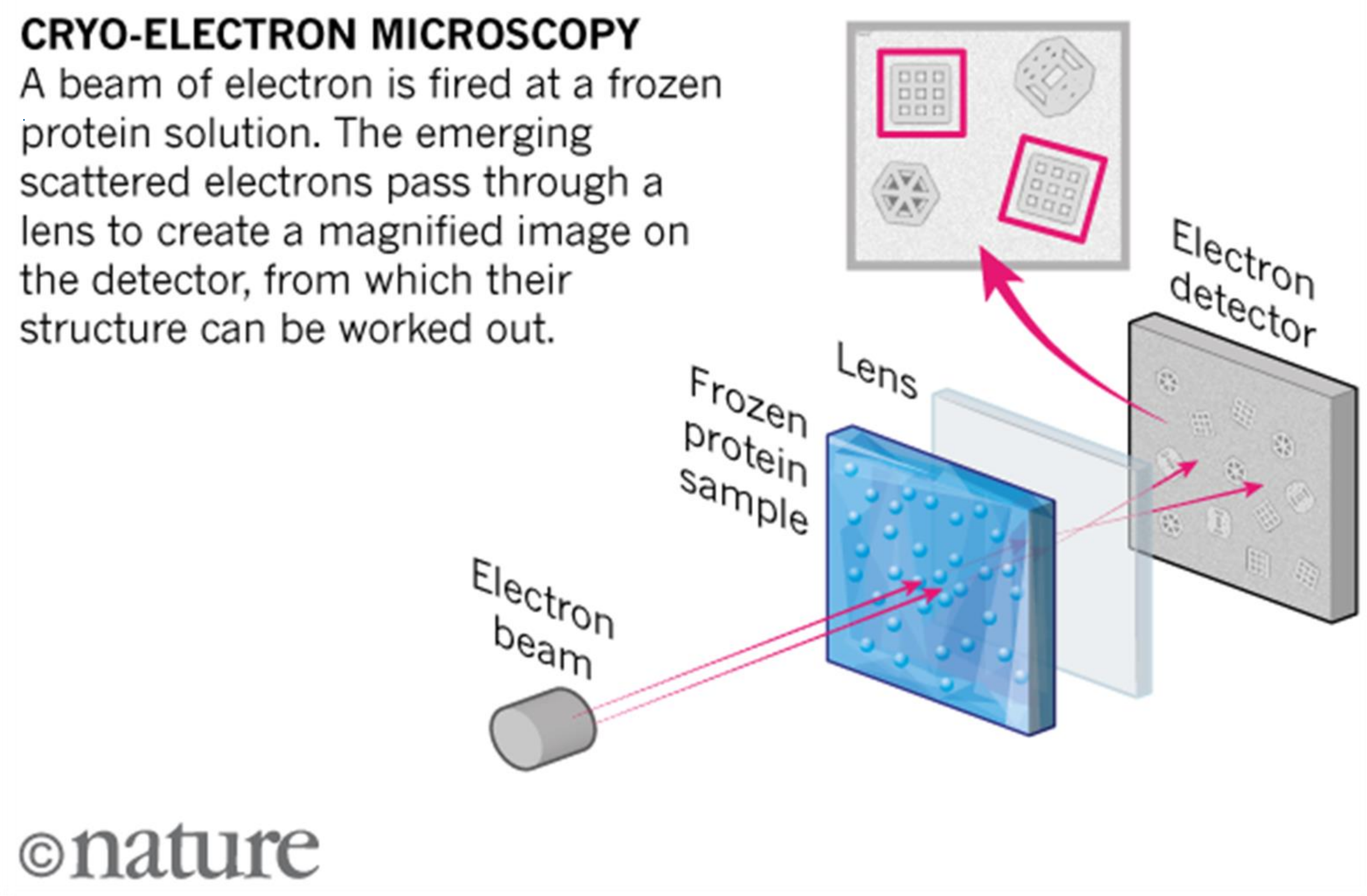
Charles Fefferman, Sergei Ivanov, Matti Lassas, Hari Narayanan

Manifold Hypothesis

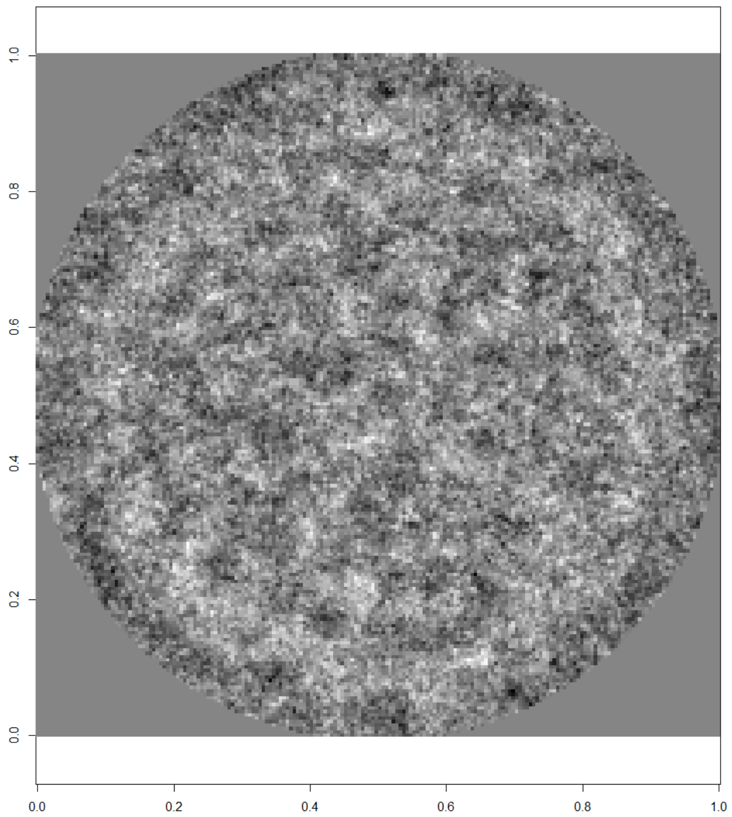
- Manifold learning is based on the manifold hypothesis, which states that high dimensional data can be modeled to lie in the vicinity of a low dimensional manifold.

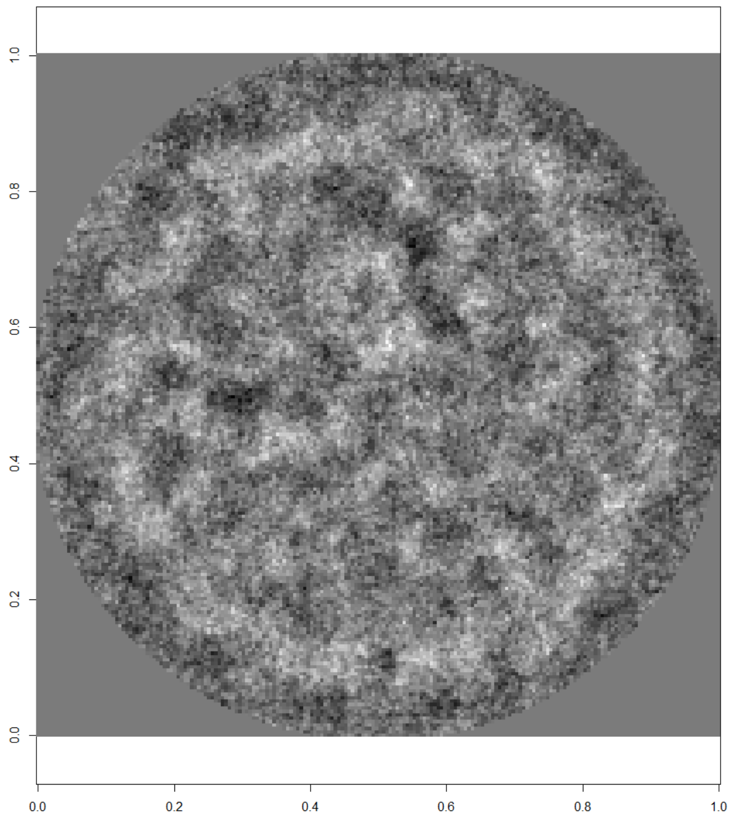


Measurements can lie near a manifold



Typical preprocessed cryo-EM images





Cryo-EM images

A single Cryo-EM image has 40,000 pixels (shown in the previous slide). Thus each data point is a vector in 40,000 dimensional Euclidean space.

Assuming the noise is small, these points approximately lie in an orbit of SO_3 which is a 3 dimensional Lie group.

Assuming the molecule is "generic", the orbit would be a 3-dimensional manifold diffeomorphic to SO_3 .

Assumptions on the target manifold

- We would like to infer the manifold from noisy samples.

In order for this to be possible, we need to place restrictions on the manifold $\mathcal{M}$.

Assumptions:

1. $\mathcal{M} \subseteq \mathbb{R}^n$ has no boundary and is $d-$dimensional and C^2.

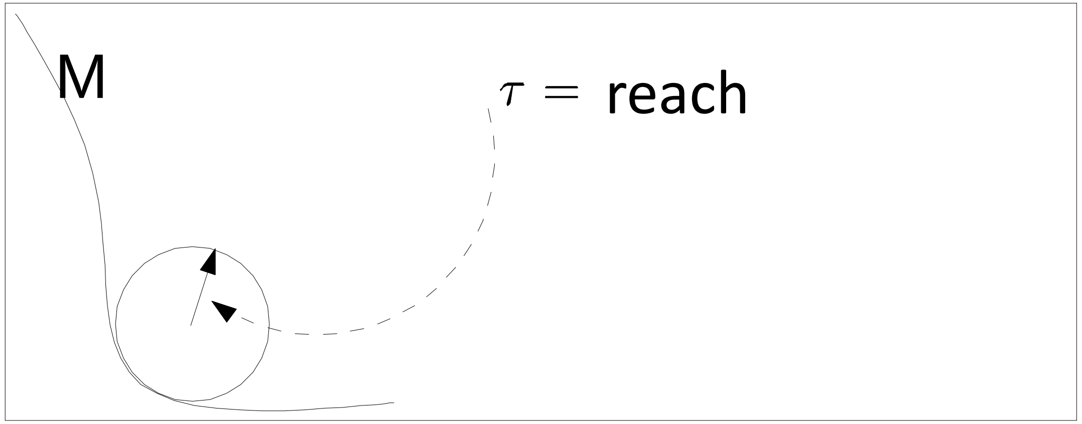
2. The reach of $\mathcal{M}$ is at least τ.

3. The $d-$dimensional Hausdorff measure is at most V.

Assumptions on the target manifold

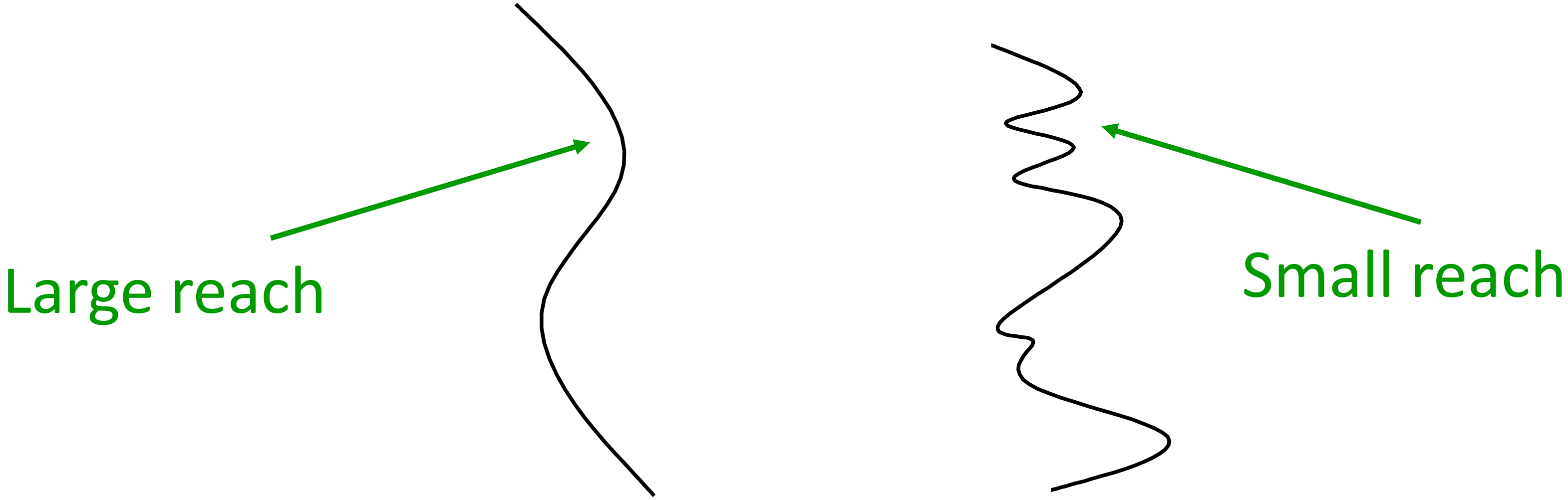
$\mathcal{M} \subseteq \mathbb{R}^n$ has
no boundary and is $d-$dimensional and C^2
means that for every point x in $\mathcal{M}$
there is $\epsilon > 0$ and a ball B_ϵ^x such that $B_\epsilon^x \cap \mathcal{M}$
is the graph of a function from a
$d-$dimensional disc $Tan_x \cap B_\epsilon^x$
to the Normal space Nor_x at x.

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 $\mathcal{M}$



Hausdorff measure on a smooth manifold

For a boundaryless C^2 manifold $\mathcal{M}$ with positive reach the $d-$dimensional Hausdorff measure of $\mathcal{M}$ is equal to

$$\lim_{\epsilon \to 0} \frac{vol(\mathcal{M}_\epsilon)}{vol(B_\epsilon)},$$

where B_ϵ is the $\epsilon-$ball of dimension $n-d$ and $\mathcal{M}_\epsilon$ is the tube of radius ϵ around $\mathcal{M}$.

Testing the Manifold Hypothesis [Fefferman-Mitter-N, JAMS'16]

Suppose $\mathcal{P}$ is an unknown probability distribution supported in the unit ball in a separable Hilbert space, and $x_1, x_2, \ldots$ 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 depending on these parameters and outputs whether or not there is

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

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

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

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

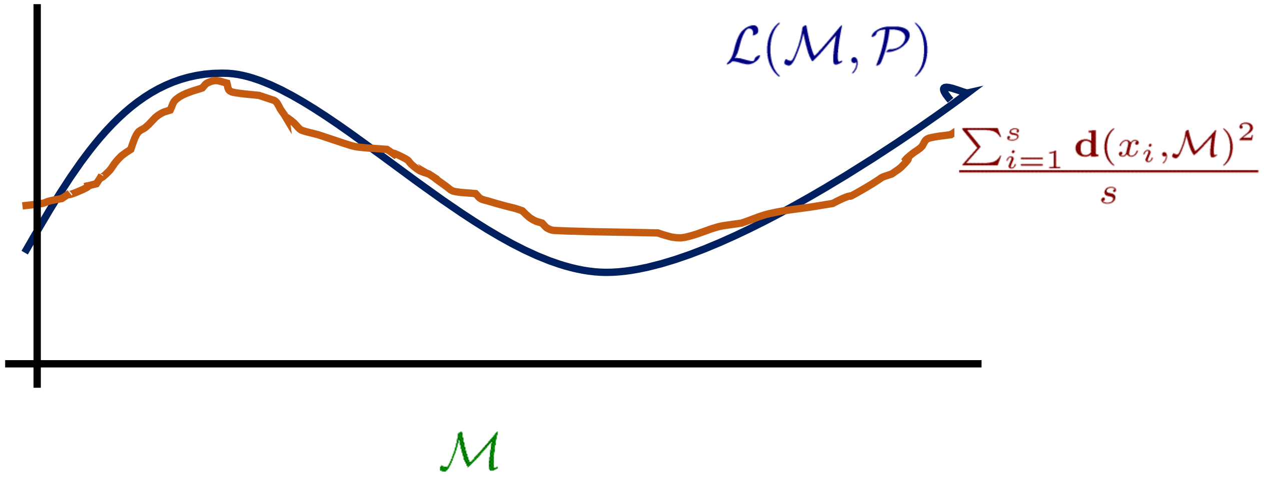
Smallest s such that $\exists$ a rule $\mathcal{A}$ given $x_1, ..., x_s$ i.i.d from $\mathcal{P}$,

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

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:

Let $x_1, \dots, x_s$ be i.i.d samples from $\mathcal{P}$, a distribution supported on the ball of radius 1 in a separable Hilbert space. If

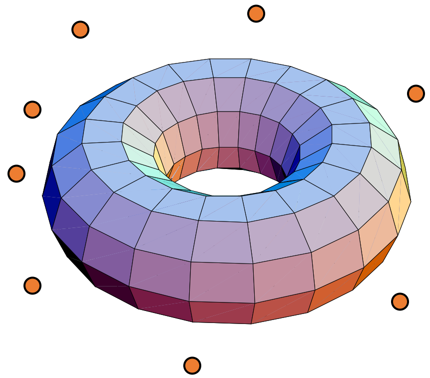
$$s \geq \frac{C\left(V\left(\frac{1}{\sqrt{\epsilon\tau}} + \frac{1}{\tau}\right)^{d+o(d)} + \log 1/\delta\right)}{\epsilon^2}$$

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 $N_p(\sqrt{\epsilon\tau})$ points



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

$$\wr\wr$$

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

Proving a Uniform bound for k-means

Proving uniform bounds for k-means

reduces to proving a uniform bound over functions of the form

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

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 $x_1, \dots, x_s$
and thresholds $t_1, \dots, t_s$ such that for every $\{-1, 1\}$ $s-$vector $(b_1, \dots, b_s)$,
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: If

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

then

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

Proof:
Put together results of Dudley, Rudelson-Vershynin and Bousquet-Boucheron-Lugosi.

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

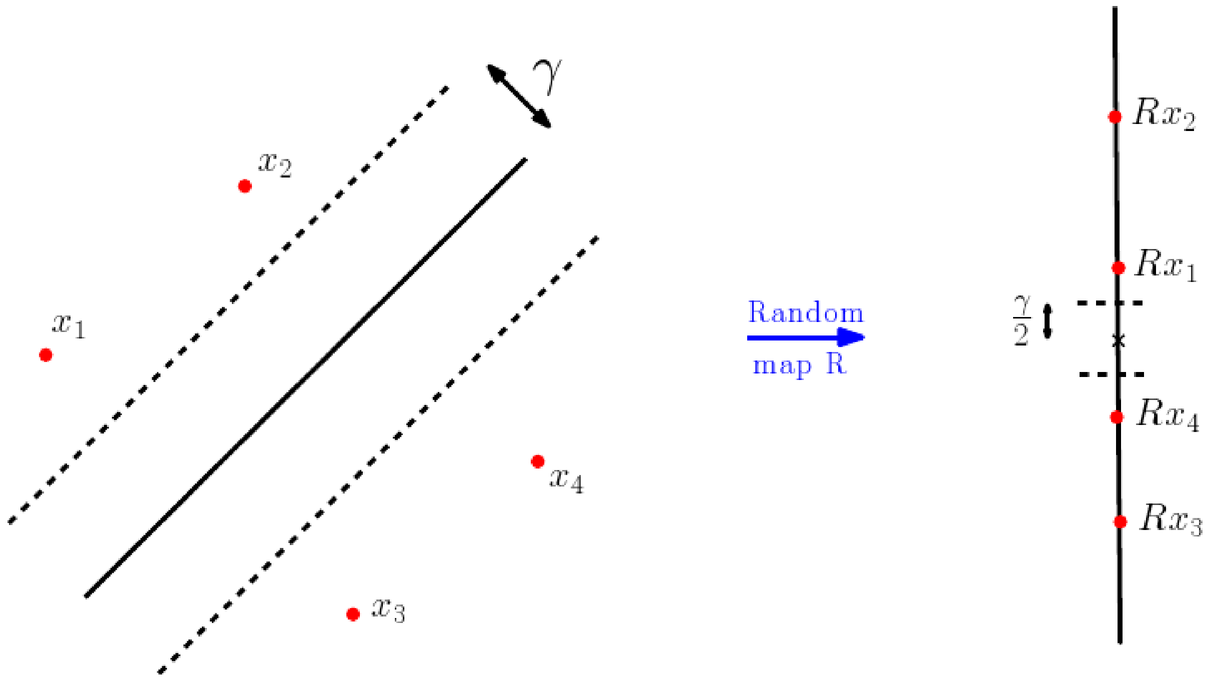
For large s, $VC(\mathcal{F}) \log s$ is roughly the

logarithm of the max number of partitions of s data

points that can be induced by functions in $\mathcal{F}$

Random projection

Thanks to the Johnson-Lindenstrauss Lemma, a 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 $\min_{1\leq i\leq k}(a_i \cdot x), \|a_i\| \leq 1$ can partition s points in $\log(s)/\epsilon^2$ dimensional image space is $O((k/\epsilon^2)\log^2(s/\epsilon))$

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

Bound on sample complexity

Theorem:

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

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

$$O\left(\frac{k}{\epsilon^2}\log^4\frac{k}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$$

Algorithmic question

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

is there a manifold $\mathcal{M} \in \mathcal{G}_e = \mathcal{G}_e(d, CV, C^{-1}\tau)$

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

Here C is some constant depending only on d.

Theorem

There is a controlled constant C depending only on d and an Algorithm that uses linear in n but doubly exponential in d number of 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

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

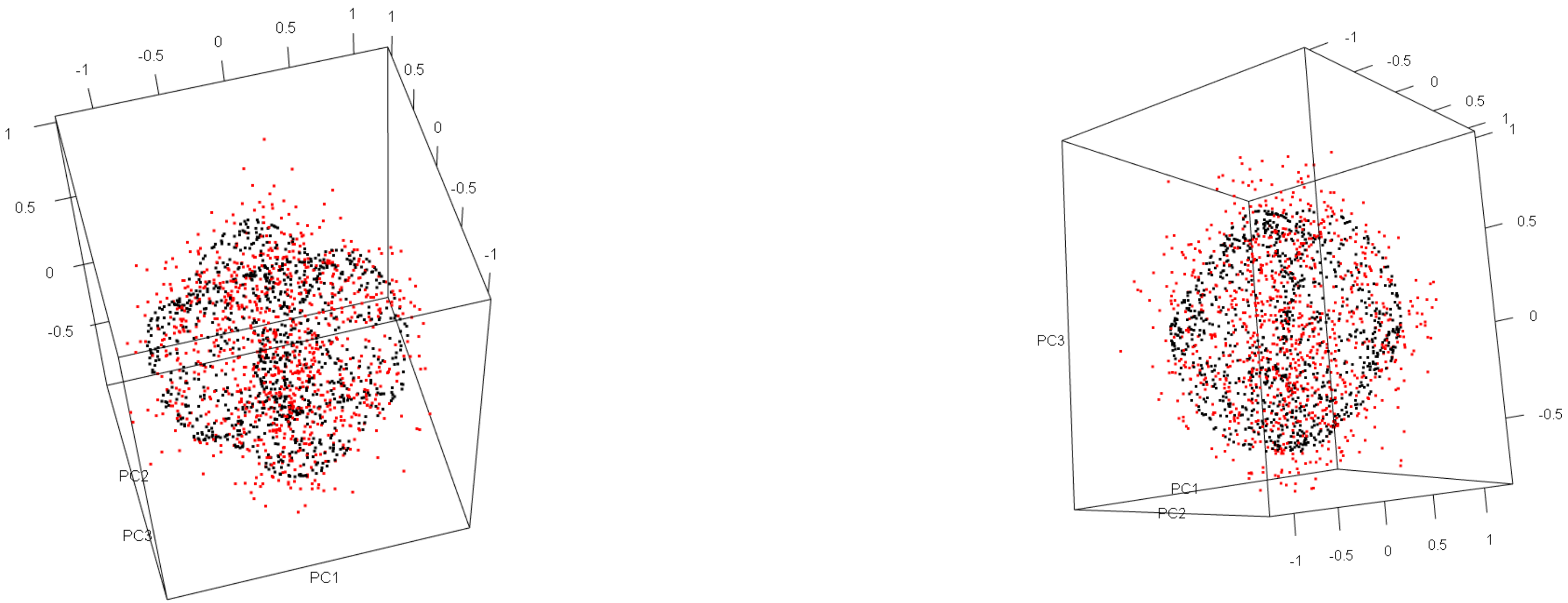
2. "No" if there exists no manifold $\mathcal{M}' \in \mathcal{G}_e(d, CV, \tau/C)$ such that

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

Next, a generative model [Fefferman-Ivanov-Lassas-N'19]

Let $x_1, x_2, \ldots, x_N$ be i.i.d draws from a measure, the logarithm of whose Radon-Nikodym derivative with respect to the $d-$dimensional Hausdorff measure on $\mathcal{M}$ is C-Lipschitz.

Data on a manifold with additive Gaussian noise



Two views of data on an immersed Klein bottle with additive Gaussian noise

Next, a generative model

Let $\zeta_1, \ldots, \zeta_N$ be a sequence of i.i.d
spherical gaussians independent of $x_1, \ldots, x_N$
having a Gaussian distribution whose density at x is

$$\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right),$$

where we assume

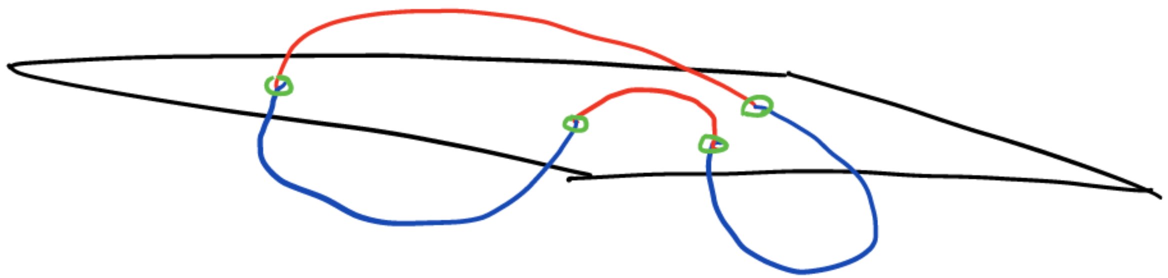
$$\frac{\tau}{Cd^C} > \sigma\sqrt{D}.$$

where D is a reduced dimension depending only on d, V and τ and N is the number of samples chosen and is roughly V/σ^d. We observe $y_i = x_i + \zeta_i$ for $i = 1, 2, \ldots$
and wish to reconstruct $\mathcal{M}$ up to a small error measured in Hausdorff distance.

Guarantees provided:

The results of this talk guarantee (for sufficiently small σ,) a Hausdorff distance of $Cd(\sigma^2/\tau) = O(\sigma^2)$ with less than $O(\sigma^{-(2d+4)})$ samples. We guarantee a reach of $\frac{\tau}{Cd^6}$.

Projection on to Principal Component Subspace



Projection on to a principal component subspace brings down the ambient dimension from n to D. The value of D depends only on intrinsic parameters d, τ and V.

Let X be a finite set of points in $E = \mathbb{R}^D$ and $X \cap B_1(x) := \{x, \tilde{x}_1, \dots, \tilde{x}_s\}$ be a set of points within a Hausdorff distance δ of some (unknown) unit d-dimensional disc $D_1(x)$ centered at x. Here $B_1(x)$ is the set of points in $\mathbb{R}^D$ whose distance from x is less or equal to 1. We give below a simple algorithm that finds a unit d-disc centered at x within a Hausdorff distance $Cd\delta$ of $X_0 := X \cap B_1(x)$, where C is an absolute constant.

The basic idea is to choose a near orthonormal basis of d vectors from X_0 where x is taken to be the origin and let the span of this basis intersected with $B_1(x)$ be the desired disc.

Algorithm FindDisc:

1. Let x_1 be a point that minimizes $|1 - |x - x'||$ over all $x' \in X_0$.

2. Given $x_1, \dots x_m$ for $m \leq d - 1$, choose x_{m+1} such that

$$\max(|1 - |x - x'||, |\langle x_1/|x_1|, x'\rangle|, \dots, |\langle x_m/|x_m|, x'\rangle|)$$

is minimized among all $x' \in X_0$ for $x' = x_{m+1}$.

Let $\tilde{A}_x$ be the affine d-dimensional subspace containing $x, x_1, \dots, x_d$, and the unit d-disc $\tilde{D}_1(x)$ be $\tilde{A}_x \cap B_1(x)$.

lemma: Suppose there exists a d-dimensional affine subspace A_x containing x such that $D_1(x) = A_x \cap B_1(x)$ satisfies $d_H(X_0, D_1(x)) \leq \delta$. Suppose $0 < \delta < \frac{1}{2d}$. Then $d_H(X_0, \tilde{D}_1(x)) \leq C d\delta$, where C is an absolute constant.

We introduce a family of n dimensional balls of radius r, $\{U_i\}_{i\in[\bar{N}]}$ where the center of U_i is p_i and a family of $d-$dimensional embedded discs of radius r $\{D_i\}_{i\in[\bar{N}]}$, $D_i \subseteq U_i$ where D_i is centered at p_i. The D_i and the p_i are chosen by a procedure described earlier. We will need the following properties of (D_i, p_i):

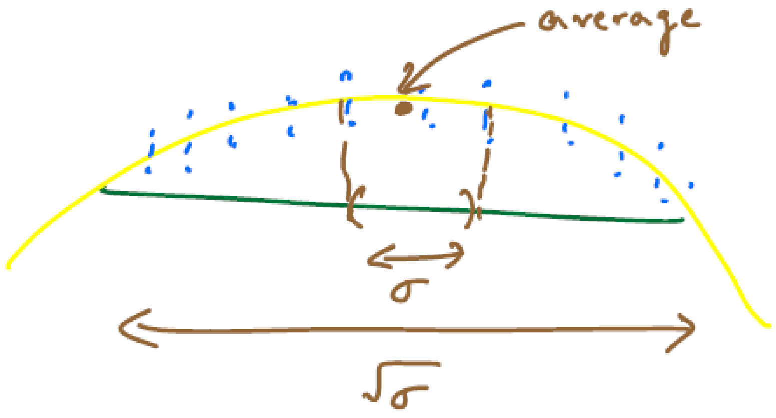
1. The Hausdorff distance between $\cup_i D_i$ and $\mathcal{M}$ is less than $\frac{Cdr^2}{\tau} = \delta$.

2. For any $i \neq j$, $|p_i - p_j| > \frac{cr}{d}$.

3. For every $z \in \mathcal{M}$, there exists a point p_i such that $|z - p_i| < 3\inf_{i\neq j}, |p_i - p_j|$.

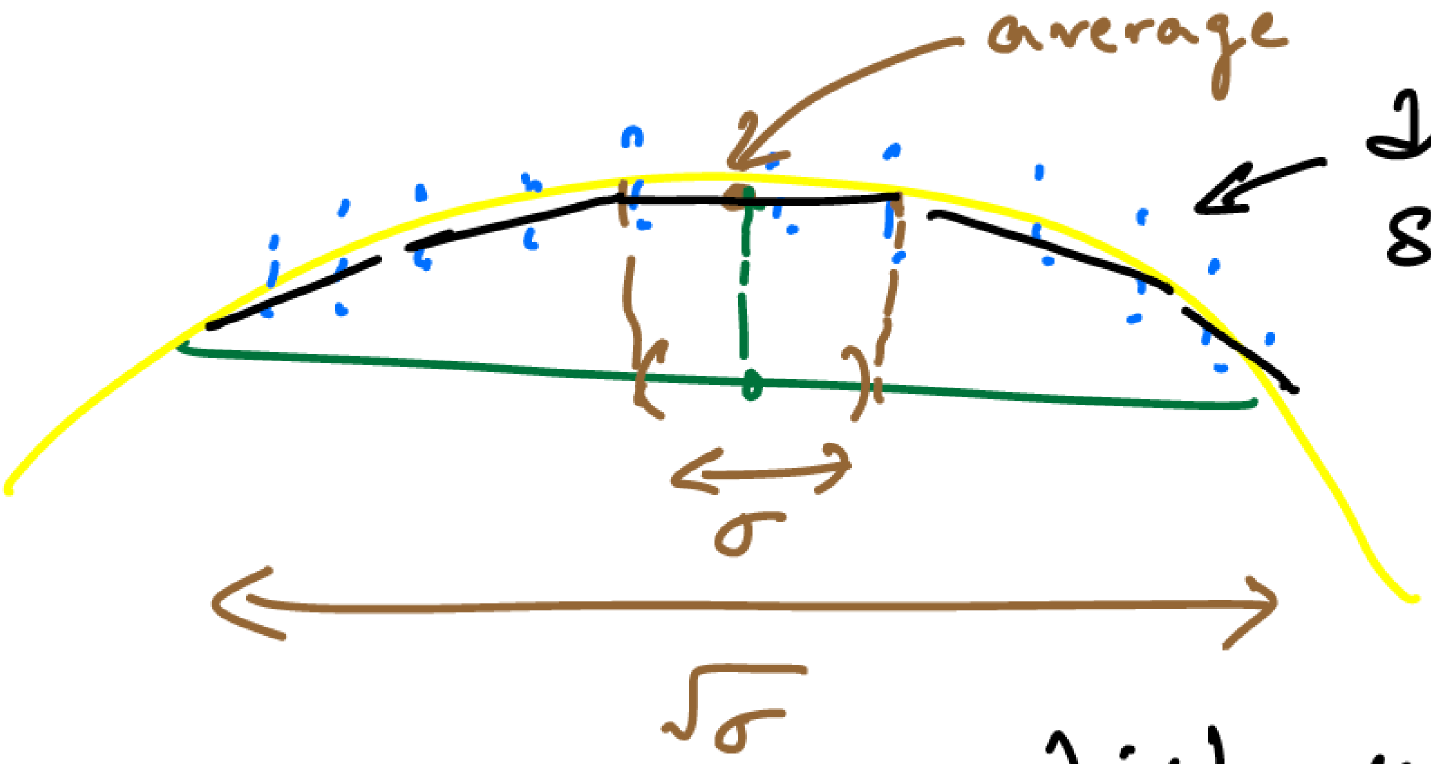
First, set $r = O(\sqrt{\sigma})$.
Find discs. Then using the discs as reference, at scale σ find the average of the displacements using $O(1/\sigma^{Cd})$ samples.

Denoising by averaging:-



If the Radon-Nikodym derivative of the measure on the manifold (wrt the Hausdorff measure) is log-Lipschitz, the average is within $O(\sigma^2)$ of the average. The set of all such averages can be arranged to be within $O(\sigma^2)$ of the true manifold, in Hausdorff distance.

Denoising by averaging:-



discs at scale $O(\sigma)$ with Hausdorff distance $O(\sigma^2)$

Bump functions

Consider the bump function $\tilde{\alpha}_i$ given by

$$\tilde{\alpha}_i(p_i + rv) = c_i(1 - \|v\|^2)^{d+2}$$

for any $v \in B_n$ and 0 otherwise. Let

$$\tilde{\alpha}(x) := \sum_i \tilde{\alpha}_i(x).$$

Let

$$\alpha_i(x) = \frac{\tilde{\alpha}_i(x)}{\sum_i \tilde{\alpha}_i(x)},$$

for each i.

Weights

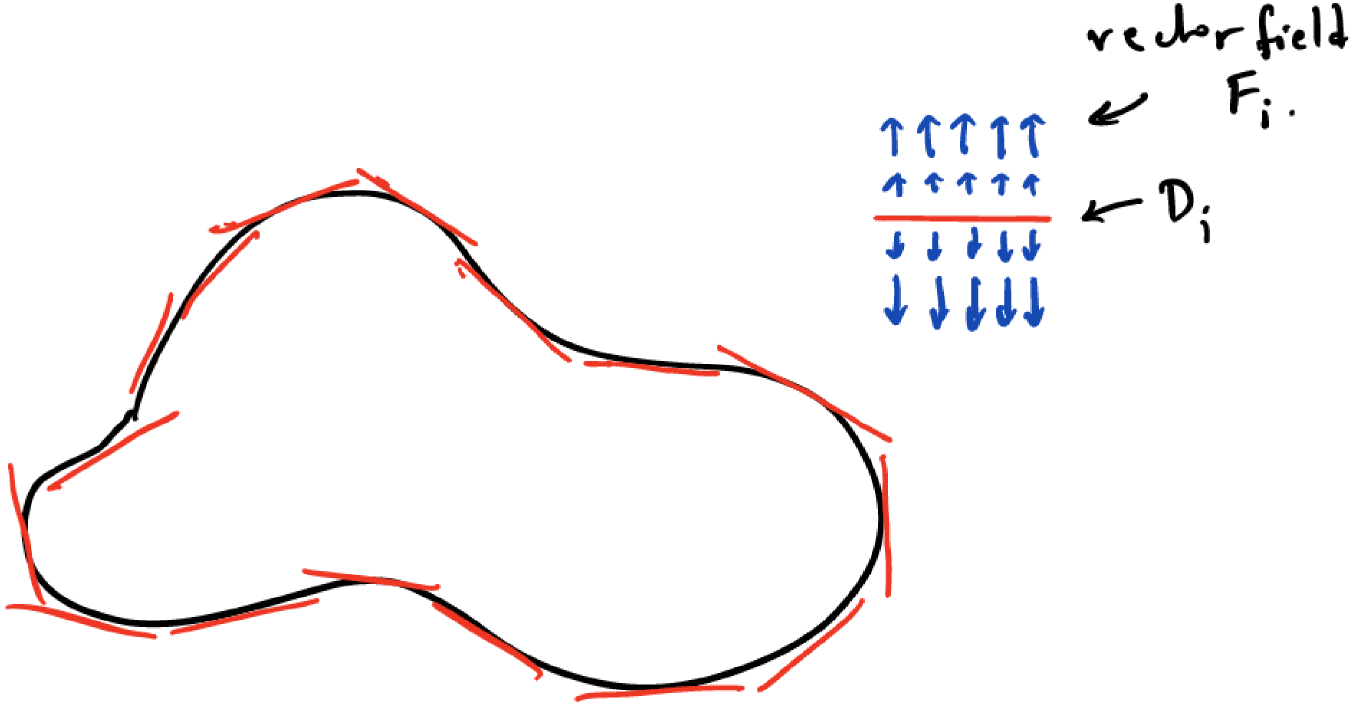
Lemma:

It is possible to choose c_i such that for any z in a $\frac{r}{4d}$ neighborhood of $\mathcal{M}$,

$$c^{-1} > \tilde{\alpha}(z) > c,$$

where c is a small universal constant. Further, such c_i can be computed using no more than $N_0(Cd)^{2d}$ operations involving vectors of dimension D.

Patching together zero sets



We wish to patch together the zero sets of the vector fields F_i in some nice way.

Projection on to the span of eigenvectors

Let Π^i be the orthogonal projection onto the $n-d-$dimensional subspace containing the origin that is orthogonal to the affine span of D_i.

We define the function $F_i : U_i \to \mathbb{R}^n$ by $F_i(x) = \Pi^i(x - p_i)$. Let $\cup_i U_i = U$. We define

$$F : U \to \mathbb{R}^n$$

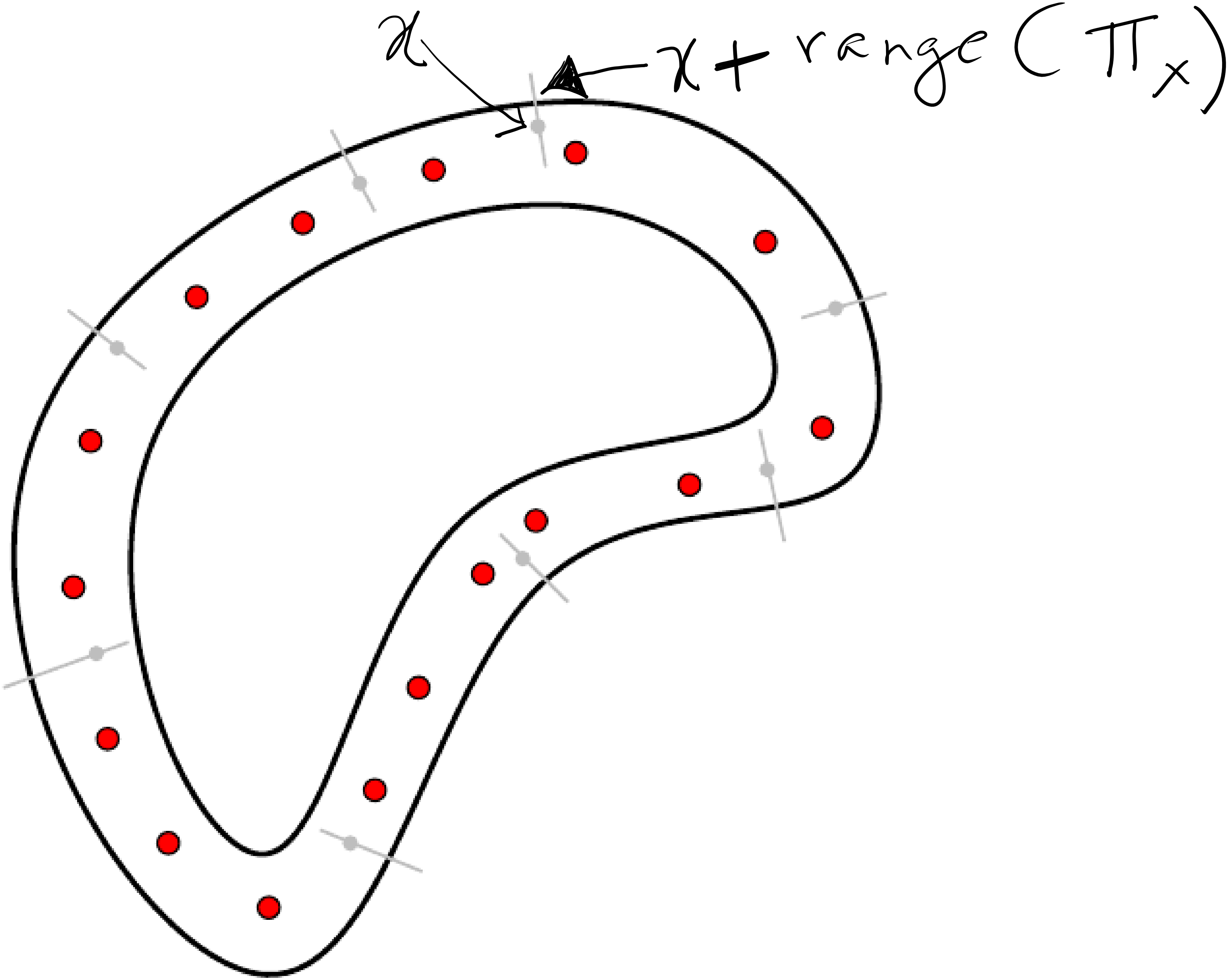
by $F(x) = \sum_i \alpha_i(x) F_i(x)$.

Given a symmetric matrix A such that A has $n-d$ eigenvalues in $(1/2, 3/2)$ and d eigenvalues in $(-1/2, 1/2)$, let $\Pi_{hi}(A)$ denote the projection onto the span of the eigenvectors corresponding to the top $n-d$ eigenvalues.

Defining the manifold

For $x \in \cup_i U_i$, we define $\Pi_x = \Pi_{hi}(A_x)$ where $A_x = \sum_i \alpha_i(x)\Pi^i$. Let $\tilde{U}_i$ be defined as the $\frac{cr}{d}$−Eucidean neighborhood of D_i inside U_i. Note that Π_x is C^2 when restricted to $\cup_i \tilde{U}_i$, because the $\alpha_i(x)$ are C^2 and when x is in this set, $c < \sum_i \tilde{\alpha}_i(x) < c^{-1}$, and for any i, j such that $\alpha_i(x) \neq 0 \neq a_j(x)$, we have $\|\Pi^i - \Pi^j\|_F < Cd\delta$.

We define the output manifold $\mathcal{M}_o$ to be the set of all points x such that $x \in \tilde{U}_i$ for some i and $\Pi_x F(x) = 0$.


x
x + range(Π_x)

Main theorem

Theorem: Suppose $C\sigma\sqrt{D}$ is less than $\frac{\tau}{Cd^C}$. The reach of $\mathcal{M}_o$ is at least $\frac{1}{Cd^6}\tau$ and the Hausdorff distance between $\mathcal{M}_o$ and $\mathcal{M}$ is less or equal to

$$\frac{Cd\sigma^2}{\tau}.$$

Proof: Use Cauchy's Integral formula, to write

$$\Pi_x = \frac{1}{2\pi\iota}\oint_\gamma (zI - A_x)^{-1}dz,$$

for suitable γ. Use Hölder Inequalities to get good bounds on the first and second derivatives of $\Pi_x F(x)$ and then apply a dimension-free quantitative form of the implicit function theorem.

More on bump functions:

Recall the bump function $\tilde{\alpha}_i$ given by

$$\tilde{\alpha}_i(p_i + rv) = c_i(1 - \|v\|^2)^{d+2}$$

for any $v \in B_n$ and 0 otherwise.

Observe that

$$\sum_i |\partial_v \tilde{\alpha}_i(x)|^{\frac{d+2}{d+1}} \leq Cd\|(\tilde{\alpha}_i(x))_i\|_1 \leq Cd.$$

Recall

$$\tilde{\alpha}(x) := \sum_i \tilde{\alpha}_i(x).$$

Thus by the Hölder inequality,

$$|\partial_v \tilde{\alpha}| \leq \|\partial_v \tilde{\alpha}_i(x)\|_{\frac{d+2}{d+1}} (\mathbf{1})_{d+2} \leq Cd^2.$$

Local definition of the manifold

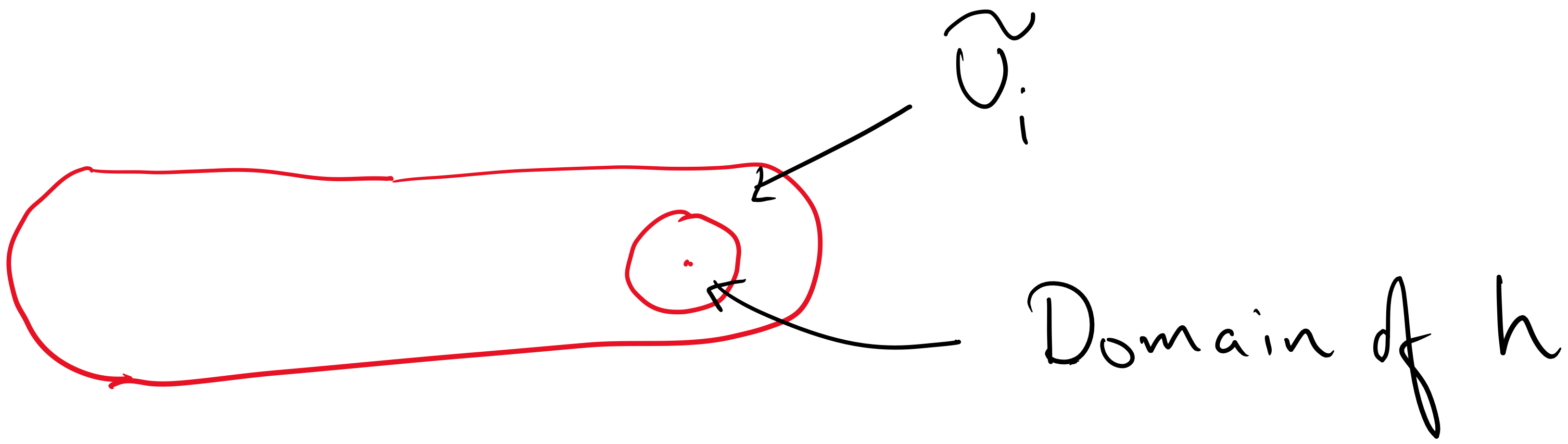
We define the output manifold $\mathcal{M}_o$ to be the set of all points x such that $x \in \tilde{U}_i$ for some i and $\Pi_x F(x) = 0$.

Restricted to $\tilde{U}_i$, if $\Pi_x F(x) \neq 0$, then $\Pi^i \Pi_x F(x) \neq 0$.

Let T denote a translation composed with a scaling, mapping from a ball of radius 1 around the origin to a ball of radius $\frac{cr}{d}$ around x_0, contained in $\tilde{U}_i$.

Let

$$h = \Pi_i \Pi_x F \circ T.$$



Quantitative implicit function Theorem:

Let $h : \mathbb{R}^{m+n} \to \mathbb{R}^n$ be a C^2–function, $h : (x, y) \mapsto h(x, y)$. Let $g : B_{m+n} \to \mathbb{R}^{m+n}$ be defined by $g : (x, y) \mapsto (x, h(x, y))$.

Suppose the Jacobian of g, Jac_g satisfies $\|Jac_g - I\| < \epsilon^2/4$ on B_{m+n} and that for any vector $v \in \mathbb{R}^{m+n}$,

$$\left\| \frac{\partial^2 g(x)}{\partial v^2} \right\| \leq \left(\frac{\epsilon^2}{4} \right) \|v\|^2$$

where $\epsilon \in [0, 1]$. Suppose also that $\|g(0)\| < \frac{\epsilon^2}{20}$ for the same ϵ.

Quantitative implicit function Theorem:

On the domain of definition of f, $g(B_{m+n})$

$$f((x,y)) = (x, e(x,y))$$

for an appropriate e and in particular, for $\|x\| \leq \frac{\eta}{2}$, where $\eta \in [0,1]$,

$$f((x,0)) = (x, e(x,0))$$

and

$$\|(x, e(x,0))\| \leq \frac{8}{5}\left(\frac{\epsilon^2}{20} + \frac{\eta}{2}\right).$$

Finally, for any $w \in \mathbb{R}^n$ such that $\|w\| = 1$, $\|Hess(e \cdot w)\| \leq \frac{16\epsilon^2}{(4-\epsilon)^3}$.

Conclusion

- We bring down the Hausdorff distance between the true and output manifolds to $O(\sigma^2 d/\tau)$ while bringing down the sample and computational complexities to depend only on the intrinsic dimension when the manifold is $\mathcal{C}^2$.

Open Problems:

- For k larger than 2, can one do better if the manifold is $\mathcal{C}^k$?
- Are there suitable conditions under which one can learn manifolds

even when $\sigma >> \tau$?

Thank You!