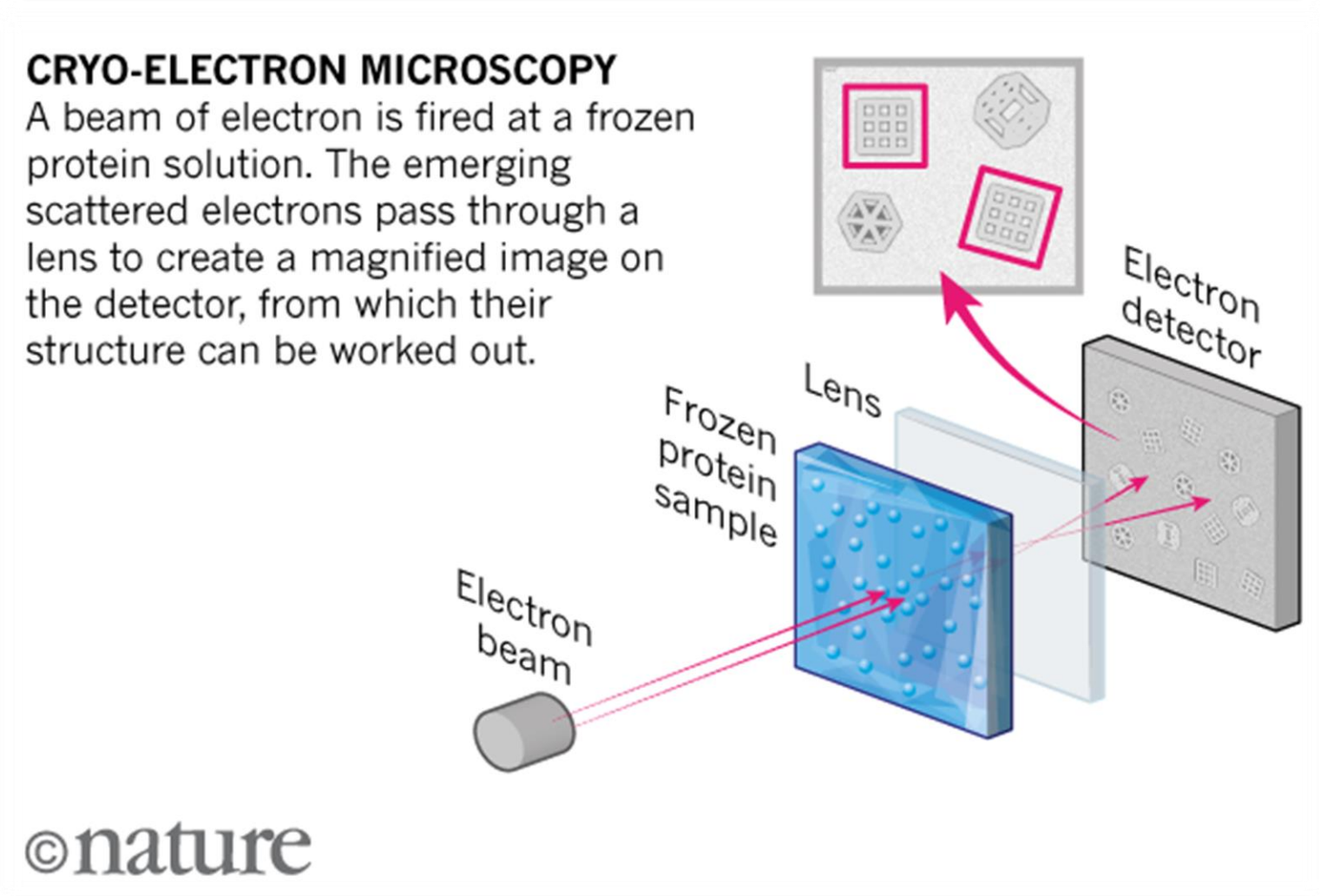
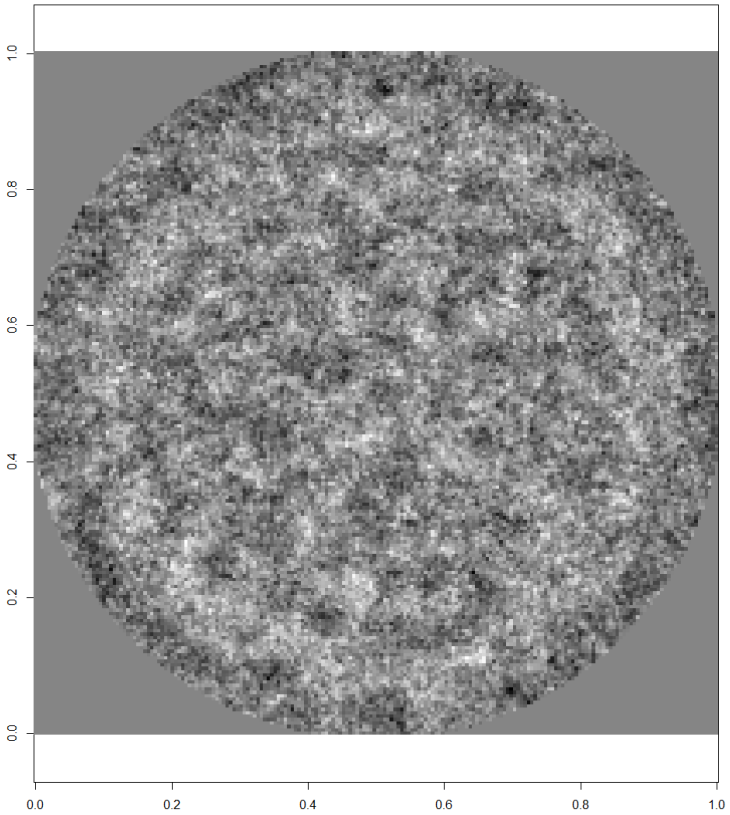
Fitting a manifold to noisy data

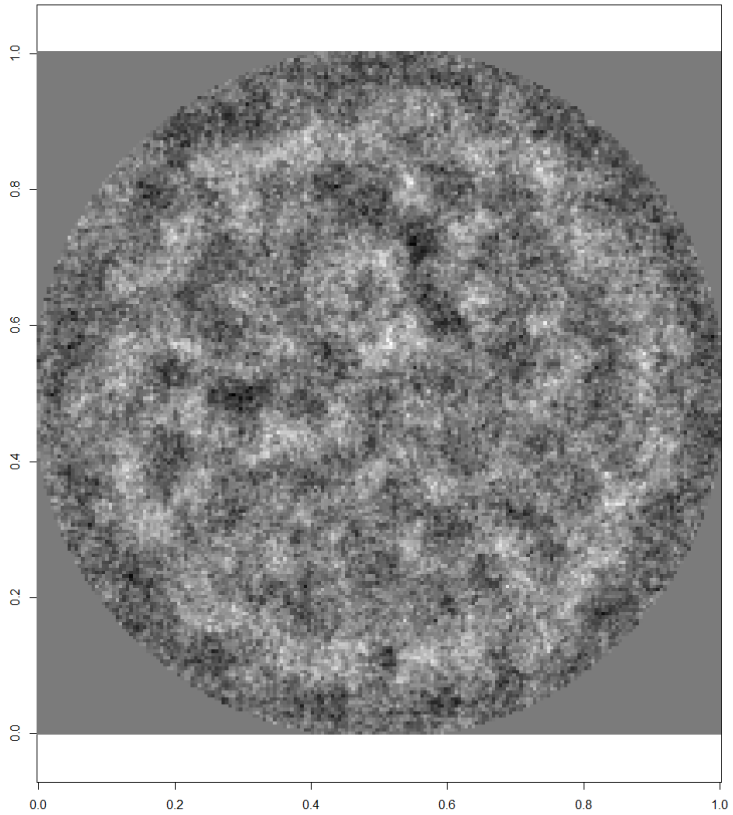
Charles Fefferman, Sergei Ivanov, Yaroslav Kurylev, Matti Lassas, Hariharan Narayanan

Measurements can lie near a manifold



Typical preprocessed cryo-EM images





Data on a manifold with additive Gaussian noise

- In the case of Cryo-Electron-Microscopy, or Cryo-EM, the manifold is

$$SO_3(\mathbb{R})$$

If the images are not centered, the relevant object is a two dimensional disc bundle over the same, corresponding to a manifold with boundary.

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

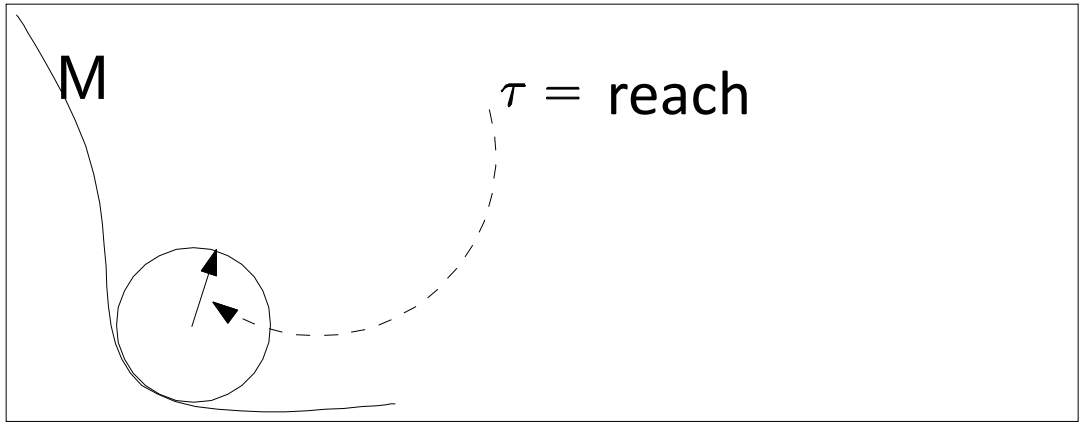
In order for this to be possible, we need to place restrictions on the manifold $\mathcal{M}$.

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.

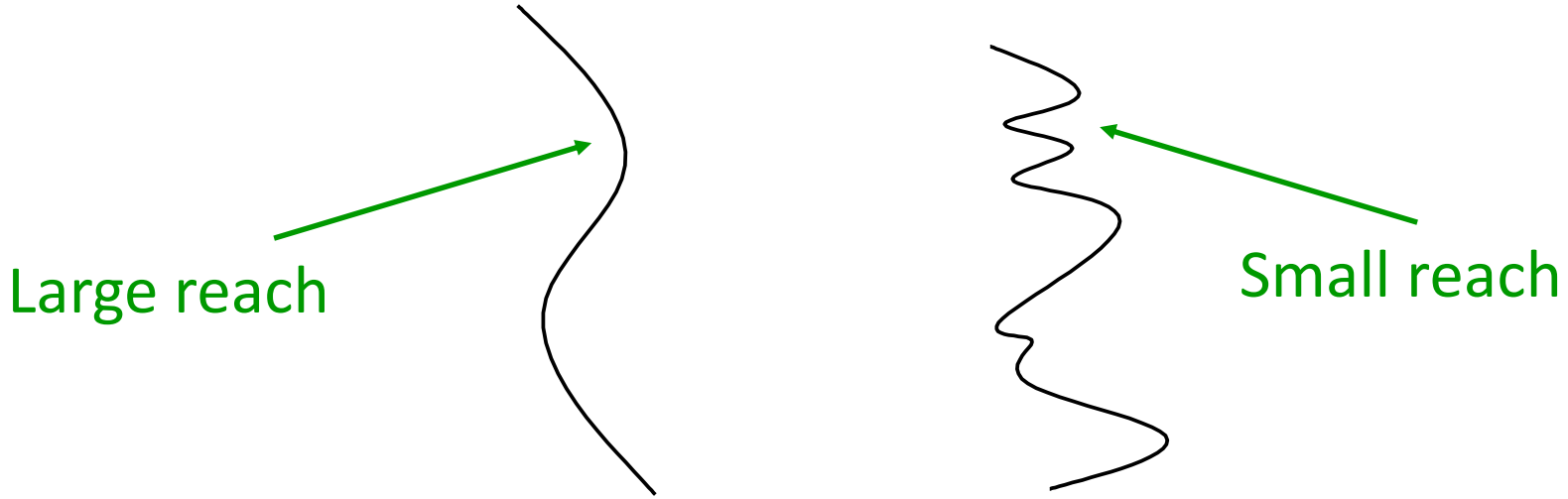
$\mathcal{M} \subseteq \mathbb{R}^n$ has
no boundary and is $d-$dimensional and C^2
means that for every point x in $\mathcal{M}$
there is $\epsilon > 0$ and a ball B_ϵ^x such that $B_\epsilon^x \cap \mathcal{M}$
is the graph of a function from a
$d-$dimensional disc $Tan_x \cap B_\epsilon^x$
to the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



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 $\mathcal{M}$.

Let $x_1, x_2, \ldots, x_N$ be i.i.d draws from a measure whose Radon-Nikodym derivative with respect to the $d-$dimensional Hausdorff measure on $\mathcal{M}$ lies between ρ_{min} and ρ_{max}.

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

$$\frac{\tau}{d^2} > C_1\sigma\sqrt{D \ln N}.$$

where D is the "effective ambient dimension" and N is the number of samples chosen (to be specified later). 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.

Suppose using $f(n)$ data points from the convolution $\mu * G(0, \sigma^2)$ we have constructed (via Principal Component Analysis) a subspace Y of dimension m. If $x \in \mathbb{R}^n$ and $S \subseteq \mathbb{R}^n$, we define

$$dist(x, S) := \inf_{y \in S} |x - y|.$$

The following theorem gives a probabilistic bound of α on the maximum distance of any point in $\tilde{M}$ to Y.

Let

$$\beta^2 < (1/10)\left(\frac{\alpha^2\tau}{2}\right)^2\left(\frac{\alpha^2\tau}{4}\right)^d \omega_d \rho_{min}.$$

Let

$$D := \frac{V}{\omega_d \beta^d}.$$

Let

$$N = C(n\sigma^2 + \log(Cn\sigma^2/(\epsilon\delta)))\sqrt{\log(C/\delta)}(D/\epsilon^2),$$

where C is a sufficiently large universal constant.

Lemma:

Given N data points $\{x_1, \ldots, x_N\}$ drawn i.i.d from $\tilde{\mu}$, let S be a D dimensional subspace that minimizes

$$\sum_{i=1}^{N} dist(x_i, \tilde{S})^2,$$

as $\tilde{S}$ ranges over all affine subspaces of dimension D, and $\beta < c\tau$.

Then,

$$\mathbb{P}[\sup_{x \in \mathcal{M}} dist(x, S) < \alpha^2 \tau] > 1 - \delta.$$

Suppose that $\mathcal{M} \in \mathcal{G}(d, V, \tau)$. Let

$$\hat{U} := \{y \big| |y - \Pi_x y| \leq \tau/8\} \cap \{y \big| |x - \Pi_x y| \leq \tau/8\}.$$

There exists a C^2 function $F_{x,\hat{U}}$ from $\Pi_x(\hat{U})$ to $\Pi_x^{-1}(\Pi_x(0))$ such that

$$\{y + F_{x,\hat{U}}(y) \big| y \in \Pi_x(\hat{U})\} = \mathcal{M} \cap \hat{U}.$$

Secondly, let $z \in \mathcal{M} \cap \hat{U}$ satisfy $|\Pi_x(z) - x| = \delta$. Let z be taken to be the origin and let the span of the first d canonical basis vectors be denoted $\mathbb{R}^d$ and let $\mathbb{R}^d$ be a translate of $Tan(x)$. Let the span of the last $n-d$ canonical basis vectors be denoted $\mathbb{R}^{n-d}$. In this coordinate frame, let a point $z' \in \mathbb{R}^n$ be represented as (z'_1, z'_2), where $z'_1 \in \mathbb{R}^d$ and $z'_2 \in \mathbb{R}^{n-d}$. There exists an $(n-d) \times d$ matrix A_z such that

$$Tan(z) = \{(z'_1, z'_2) | A_z z'_1 - I z'_2 = 0\}$$

where the identity matrix is $(n-d) \times (n-d)$. For $\delta < \tau/8$, let

$$z \in \mathcal{M} \cap \{z \big| |z - \Pi_x z| \leq \delta\} \cap \{z \big| |x - \Pi_x z| \leq \delta\}.$$

Then $\|A_z\|_2 \leq 20\delta/\tau$.

lemma:

Suppose $\mathcal{M}$ is a C^2 submanifold of $\mathbb{R}^n$ having reach τ and S is a $D-$dimensional linear subspace such that $\sup_{x \in \mathcal{M}} dist(x, S) < \alpha^2 \tau$ where $\alpha < \frac{1}{4}$ then

$$\Pi_S(\mathcal{M})$$

is a submanifold of $\mathbb{R}^n$ having reach at least $(1 - 4\alpha^2)\tau$.

Let X be a finite set of points in $E = \mathbb{R}^D$ and $X \cap B_1(x) := \{x, \tilde{x}_1, \ldots, \tilde{x}_s\}$ 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 $x_1, \ldots x_m$ for $m \leq d-1$, choose x_{m+1} such that

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

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

Let $\tilde{A}_x$ be the affine d-dimensional subspace containing $x, x_1, \ldots, x_d$, 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 $d_H(X_0, \tilde{D}_1(x)) \leq Cd\delta$, 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 $|z - p_i| < 3\inf_{i\neq j}, |p_i - p_j|$.

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.

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.

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 $\min_i dist(x, p_i) < r$ and $\Pi_x F(x) = 0$.

lemma: Suppose $C\sigma\sqrt{D\ln(N)}$ is less than $\frac{\tau}{Cd^2}$. The reach of $\mathcal{M}_o$ is at least $\frac{C}{d^4}\tau$ and the Hausdorff distance between $\mathcal{M}_o$ and $\Pi_{\mathbb{R}^D}\mathcal{M}$ is less or equal to $Cd\sigma\sqrt{D\ln(N)}$.

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.

lemma: Suppose $C\sigma\sqrt{D\ln(N)}$ is less than $\frac{\tau}{Cd^2}$. The reach of $\mathcal{M}_o$ is at least $\frac{C}{d^7}\tau$ and the Hausdorff distance between $\mathcal{M}_o$ and $\Pi_{\mathbb{R}^D}\mathcal{M}$ is less or equal to $Cd\sigma\sqrt{D\ln(N)}$.

Proof: Use Cauchy's Integral formula, 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.

Thank You!