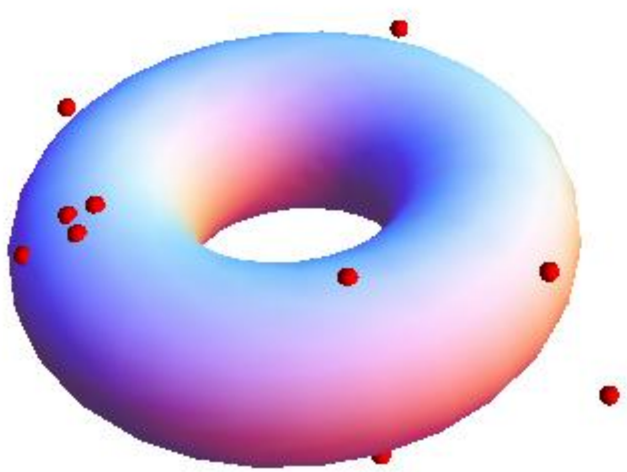
# Sample Complexity of Testing the Manifold Hypothesis

Hariharan Narayanan and Sanjoy Mitter

MIT



Manifold learning is based on the hypothesis that data in high dimensional Euclidean spaces usually lie in the *vicinity of a low dimensional submanifold*.

Can this hypothesis be tested using limited data?

# Low dimensional manifolds with bounded volume and curvature

Let $\mathcal{G}_e = \mathcal{G}_e(k, V, \tau)$ be the family of Riemannian $k-$submanifolds of the unit ball in $\mathbb{R}^n$, with volume $\leq V$ and curvature $\leq \kappa$.

Low curvature

high curvature

# A positive result

Let $\mathcal{P}$ be a probability distribution supported in the unit ball from which data $x_1, \ldots, x_s$ is drawn i.i.d. If s is greater than

$$C\left(\min\left(\left(\frac{1}{\epsilon^2}\right)\log^4\left(\frac{N_p}{\epsilon}\right), N_p\right)\frac{N_p}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$$

where $N_p$ is $V\left(C\left(k \max(\frac{1}{\epsilon}, \kappa)\right)\right)^k$

Then, independent of ambient dimension n,

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

# K-means

In particular, this improves the best known upper bound on the sample complexity of k-means from $O(\frac{k^2+\log\frac{1}{\delta}}{\epsilon^2})$ to

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

Aspects of the proof:

1. Estimates for the volumes of balls in Riemannian manifolds
2. Bounding the Fat-Shattering dimension using Random Projections onto a low-dimensional subspace.