Sample complexity and effective dimension for regression on manifolds

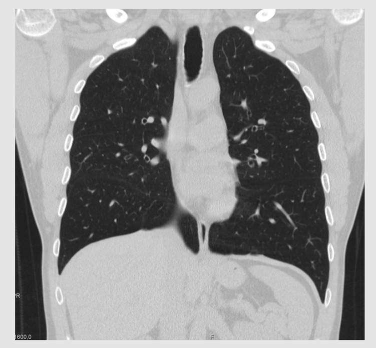
Andrew D. McRae, Justin Romberg, Mark Davenport

School of Electrical and Computer Engineering, Georgia Tech

The problem

In many modern real-world tasks, the data are very **high-dimensional**.

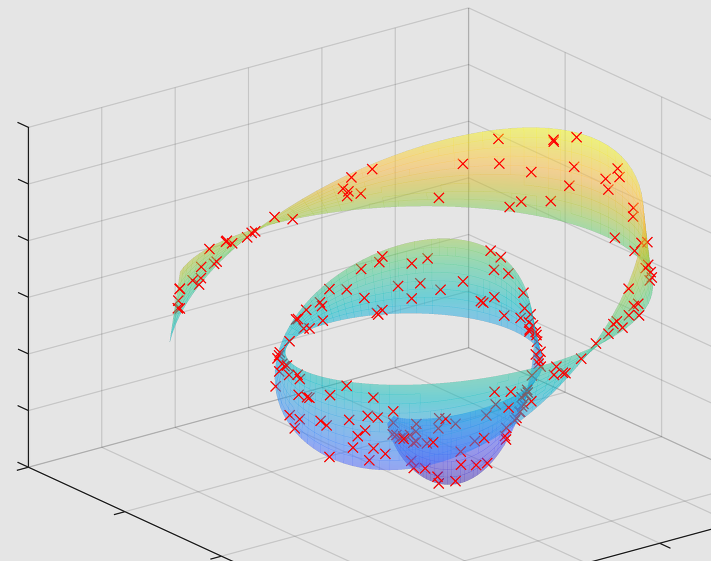




Traditional learning theory says that the number of samples needed to learn a function in d dimensions grows **exponentially** in d...

Manifold models

A common model is that all of the data lie on a low-dimensional *manifold* embedded in higher-dimensional space.



Question: If manifold dimension $m \ll$ ambient dimension d, can we get away with only using $O(C^m)$ data points instead of $O(C^d)$?

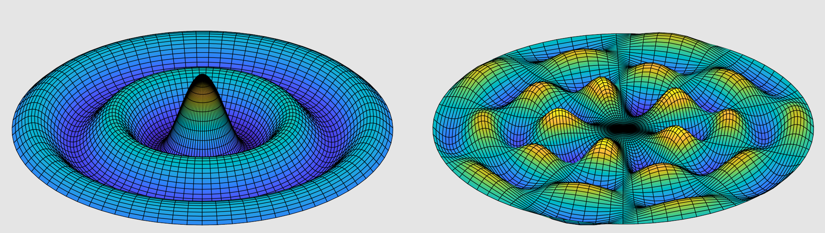
Key tool: spectral analysis of manifolds

We analyze functions on $\mathcal{M}$ via the spectral decomposition of the (positive semidefinite) Laplace differential operator $\Delta_{\mathcal{M}}$:

$$\Delta_{\mathcal{M}} f = \sum_{\ell=0}^{\infty} \omega_\ell^2 \langle f, v_\ell \rangle_{L_2} v_\ell.$$

Each v_ℓ is a vibrating mode of $\mathcal{M}$, and ω_ℓ is the corresponding vibrational frequency.

Modes of a vibrating string (1-d manifold)



Modes of a vibrating drum (2-d manifold)

The *Weyl law* from differential geometry says that, asymptotically,

$$|\{\ell : \omega_\ell \le \Omega\}| \sim c_m \operatorname{vol}(\mathcal{M})\, \Omega^m \text{ as } \Omega \to \infty,$$

where c_m is a dimension-dependent constant.

Function spaces of spectral kernels

One model space of very smooth functions on $\mathcal{M}$ is "diffusion space"

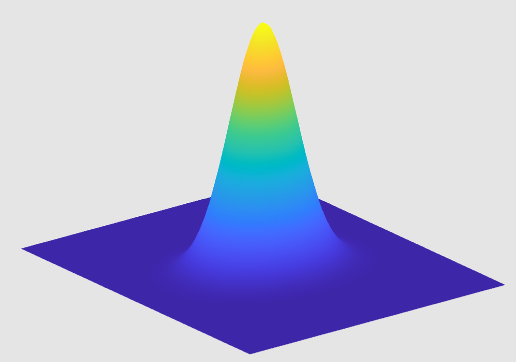
$$\mathcal{H}_t^{\mathrm{h}} = \left\{ f : \|f\|_{\mathcal{H}_t^{\mathrm{h}}}^2 := \sum_\ell e^{\omega_\ell^2 t/2} \langle f, v_\ell \rangle_{L_2}^2 < \infty \right\}$$

for $t > 0$, whose reproducing kernel is the heat kernel

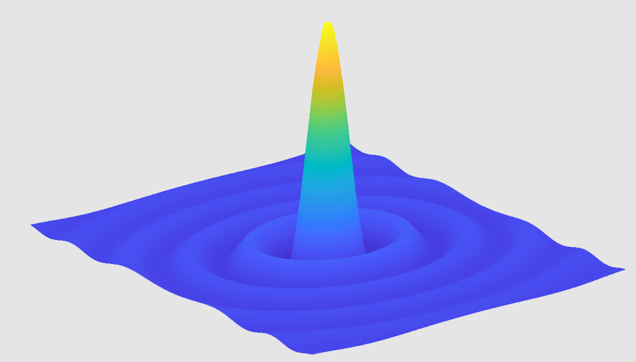
$$k_t^{\mathrm{h}}(x, y) = \sum_\ell e^{-\omega_\ell^2 t/2} v_\ell(x) v_\ell(y).$$

Another model is the space of Ω-bandlimited functions with its associated reproducing kernel:

$$\mathcal{H}_\Omega^{\mathrm{bl}} = \operatorname{span}\{v_\ell : \omega_\ell \le \Omega\}, \quad k_\Omega^{\mathrm{bl}}(x, y) = \sum_{\ell : \omega_\ell \le \Omega} v_\ell(x) v_\ell(y).$$



Heat kernel k_t^{h} on sphere



Bandlimited kernel k_Ω^{bl} on sphere

Algorithm: kernel regression (a.k.a. regularized empirical risk minimization)

Given n observations of the form $Y_i = f^*(X_i) + \xi_i$, where f^* is the function we want to learn and ξ_i is noise, our estimators have the form

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 + \alpha \|f\|_{\mathcal{H}}^2,$$

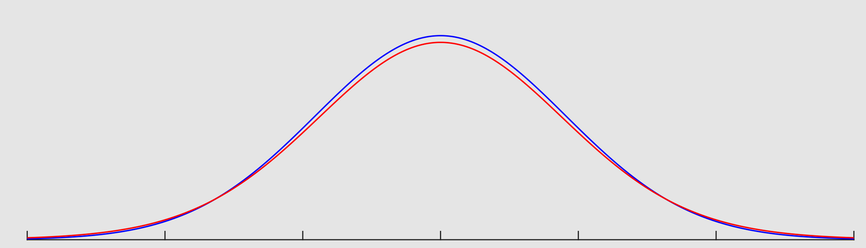
where $\mathcal{H}$ is either $\mathcal{H}_t^{\mathrm{h}}$ or $\mathcal{H}_\Omega^{\mathrm{bl}}$, and $\|\cdot\|_{\mathcal{H}_\Omega^{\mathrm{bl}}}$ is the L_2 norm.

By the kernel trick, $\hat{f}$ has a simple form in terms of the kernel function (k_t^{h} or k_Ω^{bl}) and the data.

Analysis/proof techniques

Bounding $|\{\ell : \omega_\ell \le \Omega\}|$:

- Derived from bound on heat kernel k_t^{h} for very small t via stochastic calculus



For small t, $k_t^{\mathrm{h}}(x, y) \approx \frac{1}{(2\pi t)^{m/2}} e^{-\|x-y\|_2^2/2t}$

Learning theory result:

- Standard ERM argument with finite-dimensional approximations
- Concentration inequalities on sums of random operators in L_2 and $\mathcal{H}$

Main result #1: nonasymptotic complexity

If $\mathcal{M}$ has bounded curvature, then, for large enough Ω,

$$|\{\ell : \omega_\ell \le \Omega\}| \le C_m \operatorname{vol}(\mathcal{M}) \Omega^m.$$

- First **nonasymptotic** upper bound on bandlimited function space dimension
- Lets us estimate complexity of estimation of very smooth functions

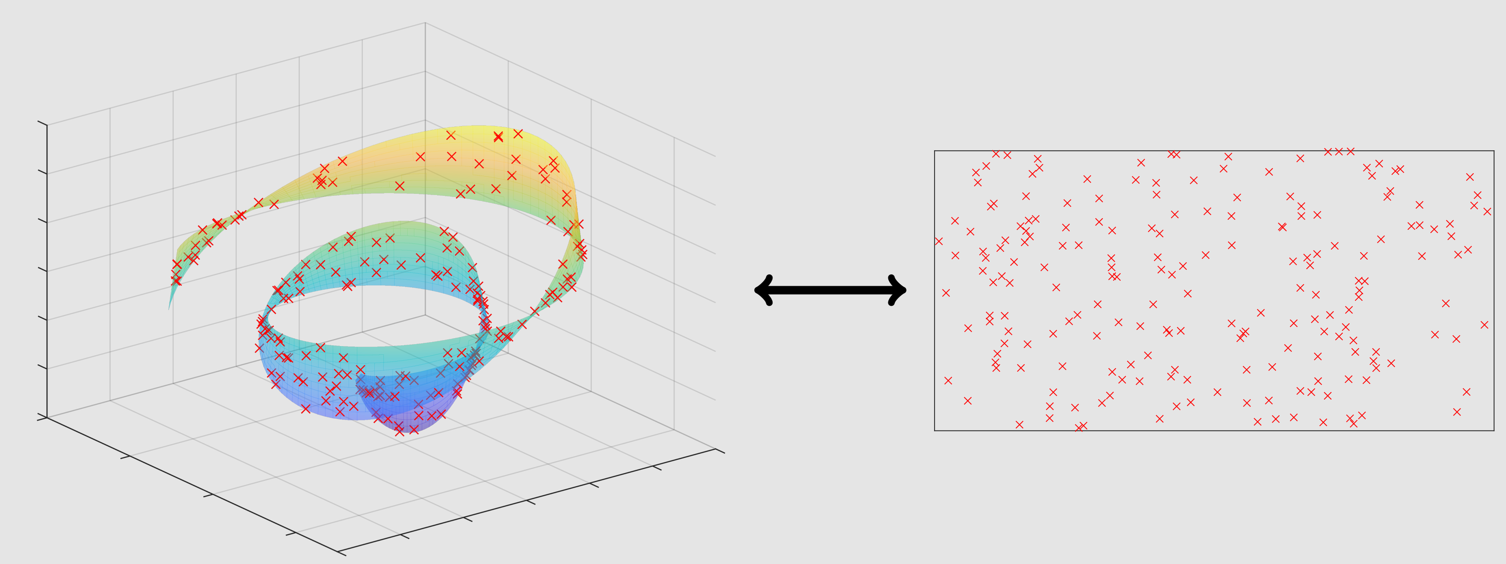
Main result #2: learning theory bounds

Let $p(\Omega) := C_m \operatorname{vol}(\mathcal{M}) \Omega^m$. Suppose we observe $n \gtrsim p(\Omega) \log p(\Omega)$ i.i.d. samples of the form $Y_i = f^*(X_i) + \xi_i$, where X_i is distributed uniformly at random over $\mathcal{M}$, and ξ_i is independent noise with variance σ^2.

1. If the true regression function $f^* \in \mathcal{H}_\Omega^{\mathrm{bl}}$, and we perform kernel regression with k_Ω^{bl}, then
$$\|\hat{f} - f^*\|_{L_2}^2 \lesssim \frac{p(\Omega)}{n} \sigma^2.$$
2. If $f^* \in \mathcal{H}_t^{\mathrm{h}}$, and we perform kernel regression with k_t^{h}, then
$$\|\hat{f} - f^*\|_{L_2}^2 \lesssim \frac{p(\Omega)}{n} \sigma^2 + e^{-\Omega^2 t/2} \|f^*\|_{\mathcal{H}_t^{\mathrm{h}}}^2$$
(same error as bandlimited case plus small residual due to error of finite-dimensional approximation).
3. These error bounds are minimax-optimal.

Key takeaways

1. Sample complexity and error due to noise scale like Ω^m: **difficulty scales with manifold dimension** m**, not ambient dimension** d



Same complexity on 2-d manifold as in $\mathbf{R}^2$!

2. Very smooth function spaces have (almost) *parametric* error rates
   - Since the space $\mathcal{H}_\Omega^{\mathrm{bl}}$ of Ω-bandlimited functions is finite-dimensional, we get parametric rate n^{-1} with dimension $p(\Omega)$
   - For $\mathcal{H}_t^{\mathrm{h}}$, optimizing Ω gives almost-parametric error rate $\frac{\log^{m/2} n}{n}$
   - By comparison, standard nonparametric rate for functions that are only s-differentiable is $n^{-2s/(m+2s)}$