Effective dimension in sample-complexity bounds for Hilbert space regression

Andrew D. McRae

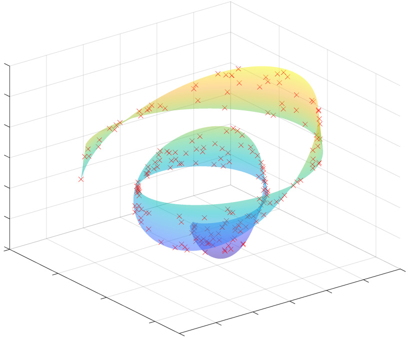
Georgia Tech School of Electrical and Computer Engineering
admcrae@gatech.edu
Joint work with Mark Davenport and Justin Romberg

High Dimensional Probability

June 15, 2020

Motivation: learning function on manifold domain

- Common machine learning model
- m-dimensional (Riemannian) manifold domain $\mathcal{M}$ embedded in R^d ($d \gg m$)
- Does difficulty scale with d or m?
- Sample complexity:
  - Effective dimension of function spaces on manifold
  - **Learning theory results that respect effective dimension**



Concrete example

- Fourier series on circle:

$$f(x) = a_0 + \sum_{\ell \geq 1} (a_\ell \cos(\ell x) + b_\ell \sin(\ell x))$$

- (Reproducing kernel) Hilbert space $\mathcal{H}$ of smooth functions:

$$\|f\|_{\mathcal{H}}^2 = \frac{a_0^2}{t_0} + \sum_{\ell \geq 1} \frac{a_\ell^2 + b_\ell^2}{t_\ell}$$

- Bounded $\mathcal{H}$-norm $\implies$ fast decay of Fourier coefficients determined by $\{t_\ell\}$
- $O(\Omega)$ coefficients below cutoff frequency Ω
- More generally, functions on $\mathcal{M}$ decompose into vibrational modes v_ℓ and frequencies ω_ℓ
- *Weyl* law says $\left|\{\ell : \omega_\ell \leq \Omega\}\right| \leq C_m \operatorname{vol}(\mathcal{M})\Omega^m$

General problem: overview

- Main problem: Hilbert space regression with i.i.d. linear measurements
- Sample complexity for low prediction error: effective rank of measurement covariance
- Key tools: empirical covariance and empirical process bounds

Framework

- $\mathcal{H}$ arbitrary separable Hilbert space
- Take n i.i.d. samples of $Y = \langle X, \beta^* \rangle + \xi$
  - $X \in \mathcal{H}$ random
  - ξ zero-mean noise
  - RKHS example: $\beta^* \longleftrightarrow f^*$, $\langle X, \beta^* \rangle \longleftrightarrow f^*(x)$
- Want small prediction error (L_2 error in RKHS):

$$R(\hat{\beta}, \beta^*) = \mathrm{E}\langle X, \hat{\beta} - \beta^* \rangle^2$$

- We analyze regularized empirical risk minimizer (usual kernel estimate in RKHS):

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

Key quantities

- Assume $\mathrm{E}\|X\|^2 < \infty$
- Difficulty of problem characterized by spectral decomposition of covariance Σ of X:

$$\langle \beta_1, \beta_2 \rangle_\Sigma := \langle \Sigma\beta_1, \beta_2 \rangle := \mathrm{E}[\langle X, \beta_1 \rangle \langle X, \beta_2 \rangle] = \sum_{\ell \geq 1} \sigma_\ell \langle \beta_1, v_\ell \rangle \langle \beta_2, v_\ell \rangle$$

- Fourier series: sampling operator covariance in $\mathcal{H}$ has eigenvalues $\approx t_\ell$ if $\|f\|^2 = a_0^2/t_0 + \sum_\ell (a_\ell^2 + b_\ell^2)/t_\ell$
- Eigenvalues $\sigma_\ell \downarrow 0$, $\{v_\ell\}$ orthonormal basis for $\mathcal{H}$
- Risk $R(\hat\beta, \beta^*) = \|\hat\beta - \beta^*\|_\Sigma^2 = \sum_\ell \sigma_\ell \langle \hat\beta - \beta^*, v_\ell \rangle^2$
  - If σ_ℓ decay quickly, prediction error approximated by finite-dimensional inner product

Notation and assumptions

- Notation: $p \geq 1$ fixed dimension, $G = \text{span}\{v_1, \dots, v_p\}$
- Boundedness of X w.r.t. G: almost surely,

$$\sum_{\ell=1}^{p} \langle X, v_\ell \rangle^2 \lesssim p$$

- Boundedness of X w.r.t. $G^\perp$: almost surely,

$$\sum_{\ell > p} \langle X, v_\ell \rangle_\Sigma^2 \lesssim p \sigma_{p+1}$$

Main result (no noise)

Theorem

If $\delta \in (0,1)$ and $n \gtrsim p \log \frac{p}{\delta}$, and there is no noise ($\xi = 0$), then, with probability at least $1-\delta$,

$$R(\beta^*, \hat{\beta}) \lesssim (\alpha + \sigma_{p+1})\|\beta^*\|^2.$$

- "Bias" error $(\alpha + \sigma_{p+1})\|\beta^*\|^2$ depends on regularization and p-dimensional approximation error
- Can even take $\alpha \downarrow 0$ (interpolation)
  - Not possible in previous results[1] that depend on "regularized dimension" $d_\alpha = \sum_\ell \frac{\sigma_\ell}{\alpha+\sigma_\ell}$
- $n \gtrsim p \log \frac{p}{\delta}$ standard for random design if we only assume bounded measurements[2]

---

[1] E.g., Daniel Hsu, Sham M. Kakade, and Tong Zhang (2014). "Random Design Analysis of Ridge Regression". In: *Found. Comput. Math.* 14, pp. 569–600.

[2] See, e.g., Chapter 12 in Simon Foucart and Holger Rauhut (2013). *A Mathematical Introduction to Compressive Sensing*. New York: Birkhäuser.

Main result (noisy)

Theorem

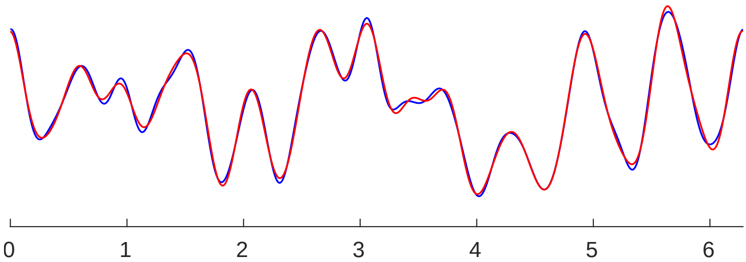
If $\delta \in (0,1)$ *and* $n \gtrsim p \log \frac{p}{\delta}$, ξ *is subexponential with variance* σ^2, *and* $\frac{n}{\log^2 n} \gtrsim \frac{\|\xi\|_{\psi_1}^2}{\sigma^2}$ *and* $\alpha \gtrsim \sigma_{p+1}$, *then, with probability at least* $1-\delta$,

$$R(\beta^*, \hat{\beta}) \lesssim \frac{p}{n}\sigma^2 + (\alpha + \sigma_{p+1})\|\beta^*\|^2.$$

- "Variance" error $\frac{p}{n}\sigma^2$ standard from p-dimensional regression

Ingredient I: covariance approximation

- Ideally, $\frac{1}{n}\sum_i \langle X_i, \beta\rangle^2 \gtrsim \|\beta\|_\Sigma^2$ uniformly in $\beta \in \mathcal{H}$...
- ...but this isn't possible with finite samples if $\text{rank}(\Sigma) = \infty$
- Instead: prove for finite-dimensional G...
- ...then show remainder is $O(\sigma_{p+1}\|\beta\|^2)$



Blue: function with $\sum e^{(\ell/10)^2}(a_\ell^2 + b_\ell^2) < \infty$
Red: approximation with 20 Fourier series frequencies

Ingredient I: covariance approximation

- Actual bound:

$$\frac{1}{n}\sum_{i=1}^{n}\langle X_i, \beta\rangle_{\mathcal{H}}^2 \gtrsim \|\beta\|_{\Sigma}^2 - \sigma_{p+1}\|\beta\|^2$$

- Proof method: concentration bound on p-dimensional random operators[3] (need $n \gtrsim p \log \frac{p}{\delta}$)

$$\frac{1}{n}\sum_{i=1}^{n}(\mathcal{P}_G X_i) \otimes (\mathcal{P}_G X_i) \succeq c\, \mathcal{P}_G \Sigma \mathcal{P}_G$$

- $\sum_{\ell=1}^{p}\langle X, v_\ell\rangle^2 \lesssim p$ a.s. for all $\beta \in G \implies (\mathcal{P}_G X_i) \otimes (\mathcal{P}_G X_i)$ are bounded
  - Other conditions on X?
  - E.g., X Gaussian would only need $n = O(p)$ samples
  - Manifold example: vibrational modes at random points?

[3] For example, matrix Chernoff bound in Joel Tropp (2015). "An Introduction to Matrix Concentration Inequalities". In: *Found. Trends Mach. Learn.* 8.1-2, pp. 1–230.

Ingredient II: empirical process bound

- Need uniform bound on $\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\langle X_i,\beta\rangle\right|$
- Our approach (Cauchy-Schwartz):

$$\mathrm{E}\sup_{\substack{\beta\in G\\ \|\beta\|_\Sigma\le 1}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\langle X_i,\beta\rangle\right|^2=\frac{p}{n}\sigma^2$$

$$\mathrm{E}\sup_{\substack{\beta\in G^\perp\\ \|\beta\|\le 1}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\langle X_i,\beta\rangle\right|^2=\frac{\sum_{\ell>p}\sigma_\ell}{n}\sigma^2\lesssim\sigma_{p+1}\frac{p}{n}\sigma^2$$

- Combine and add empirical process concentration[4]:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\langle X_i,\beta\rangle\right|^2\lesssim\frac{p}{n}\sigma^2(\|\beta\|_\Sigma^2+\sigma_{p+1}\|\beta\|^2)$$

[4] Radosław Adamczak (2008). "A Tail Inequality for Suprema of Unbounded Empirical Processes with Applications to Markov Chains". In: *Electron. J. Probab.* 13, pp. 1000–1034.

Room for further work

- Alternative assumptions on random design variable X
- Need something like $\frac{1}{n}\sum_i \langle X, \beta\rangle^2 \gtrsim \|\beta\|_\Sigma^2$
- Relax boundedness assumptions on X (particularly $\sum_{\ell=1}^{p} \langle X, v_\ell\rangle_\Sigma^2 \lesssim p$)?
  - Also used it in empirical process bound
  - Similar assumptions in other works[5] which rely on operator concentration bounds
- When could we get away with $n \gtrsim p$ (no log factor)?

[5] Such as, again, Daniel Hsu, Sham M. Kakade, and Tong Zhang (2014). "Random Design Analysis of Ridge Regression". In: *Found. Comput. Math.* 14, pp. 569–600.

Summary

- New dimension-based sample complexity results for Hilbert space regression
- Via RKHS, important applications to learning on manifolds
- Potential room for improved/more general results with other probabilistic methods
- See preprint for machine learning treatment (arXiv link coming soon)