Harmless interpolation in regression and classification with structured features

Andrew D. McRae, Santhosh Karnik, Mark A. Davenport, and Vidya Muthukumar

School of Electrical and Computer Engineering, Georgia Tech

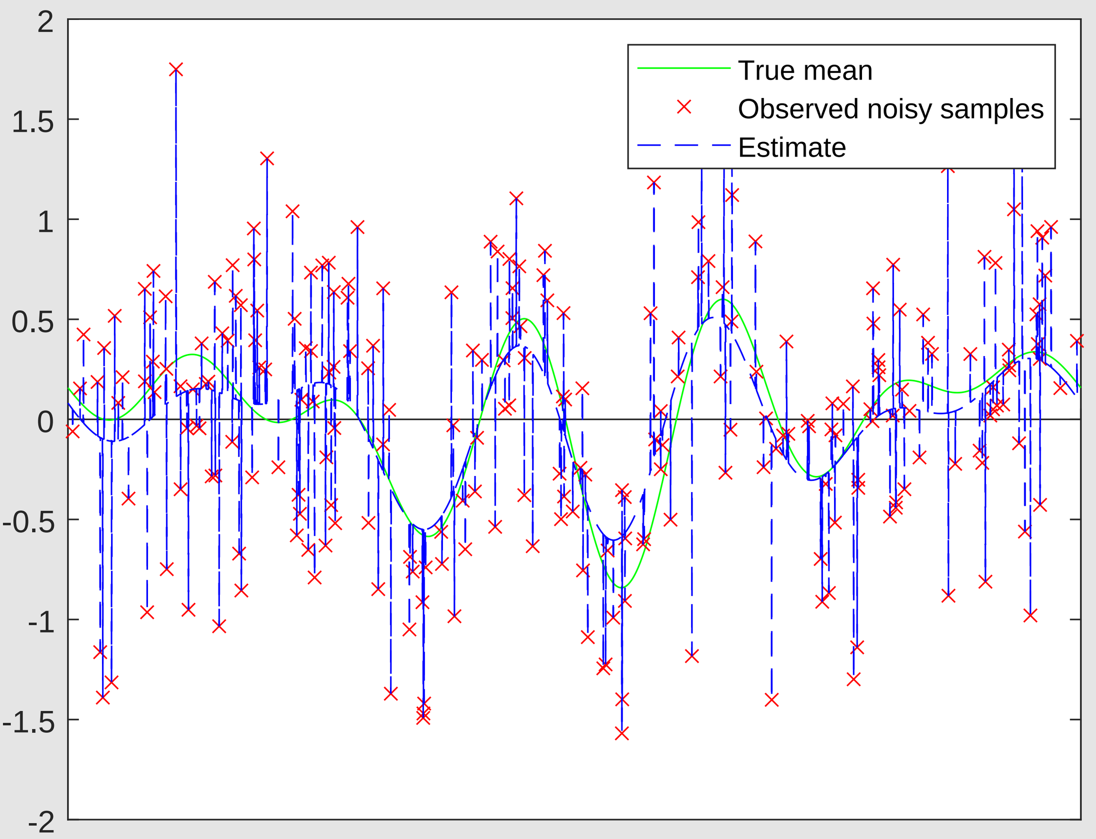
Motivation: the interpolation phenomenon

Classically, allowing machine learning models to interpolate noisy data is a bad idea.

Empirically, in **highly overparametrized** settings, it still works reasonably well.

- Arose in deep learning studies
- For simplicity, most theoretical results study linear models

Why does this occur?



Setup: linear regression with feature maps and kernels

Linear regression model with feature map $\phi(x) = (\phi_1(x), \dots, \phi_d(x))$:

$$f(x, \beta) = \langle \phi(x), \beta \rangle = \sum_\ell \beta_\ell \phi_\ell(x)$$

Suppose $f^*(x) = f(x, \beta^*)$ and $y_i = f^*(x_i) + \xi_i$ for $i = 1, \dots, n$. In matrix form,

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} \phi_1(x_1) & \cdots & \phi_d(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_n) & \cdots & \phi_d(x_n) \end{bmatrix}}_{\mathcal{A}\ (n \times d \text{ matrix})} \beta^* + \underbrace{\begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix}}_{\xi}$$

Standard ridge regression estimate with regularization $\alpha \geq 0$:

$$\hat{\beta} = (\alpha I_d + \mathcal{A}^* \mathcal{A})^{-1} \mathcal{A}^* y = \mathcal{A}^* (\alpha I_n + \underbrace{\mathcal{A}\mathcal{A}^*}_{\text{Gram matrix}})^{-1} y$$

If $\alpha = 0$ and $\mathcal{A}\mathcal{A}^*$ has full rank, $\mathcal{A}\hat{\beta} = y$ (**interpolation**)

Our analysis via truncation

Split the features into two groups (truncation and residual):

$$\phi(x) = (\underbrace{\phi_1(x), \dots, \phi_p(x)}_{\phi_H(x)}, \underbrace{\phi_{p+1}(x), \dots, \phi_d(x)}_{\phi_R(x)}), \qquad \mathcal{A} = [\underbrace{\mathcal{A}_H}_{n \times p} \quad \underbrace{\mathcal{A}_R}_{n \times (d-p)}]$$

Then the data Gram matrix is

$$\mathcal{A}\mathcal{A}^* = \mathcal{A}_H \mathcal{A}_H^* + \mathcal{A}_R \mathcal{A}_R^*$$

Key idea: if $d - p \gg n$, we can have $\mathcal{A}_R \mathcal{A}_R^* \approx \bar{\alpha} I_n$ (for some $\bar{\alpha} > 0$)

- Then $\hat{\beta} \approx \mathcal{A}^* (\bar{\alpha} I_n + \mathcal{A}_H \mathcal{A}_H^*)^{-1} y$
- (Approximately) ridge regression with positive regularization!

Sampling model and main result

If x is random with some distribution μ, the feature covariance is

$$\Sigma = \mathbf{E}_x [\phi_i(x) \phi_j(x)]_{ij} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}$$

- We have assumed the features are **uncorrelated**
- Decreasing order: $\lambda_1 \geq \lambda_2 \geq \dots$
- $\|f(\cdot, \beta)\|_{L_2} = \|\Sigma^{1/2} \beta\|_{\ell_2}$

Suppose the sample locations $x_1, \dots, x_n$ are i.i.d. random according to μ, and the noise is independent and zero-mean with variance σ^2.

Condition 1: n is large enough that empirical $\approx$ actual L_2 norm on $\operatorname{span}\{\phi_1, \dots, \phi_p\}$

- Standard approximate isometry

Condition 2: $\mathcal{R}\mathcal{R}^* \approx \bar{\alpha} I_n$, where $\bar{\alpha} = \sum_{\ell > p} \lambda_\ell$

- Previous work proved this with *independent* features:
- **Our contribution**: this holds generally for large enough d

Theorem

Suppose conditions 1 and 2 hold and (for simplicity) $f^* \in \operatorname{span}\{\phi_1, \dots, \phi_p\}$. *Then*

$$\mathbf{E}\|\hat{f} - f^*\|_{L_2}^2 \lesssim \sqrt{\frac{\alpha}{n}} \|f^*\|_{\mathcal{H}} + \sigma^2 \left(\frac{p}{n} + \frac{n \sum_{\ell > p} \lambda_\ell^2}{\left(\sum_{\ell > p} \lambda_\ell\right)^2} \right).$$

Extension to classification

Now y is a label in $\{-1, 1\}$. Let

$$f^*(x) = \mathbf{E}[y \mid x] = 2\,\boldsymbol{P}[y = 1 \mid x] - 1, \qquad \xi = y - f^*(x)$$

Classifier: estimate $\hat{\beta}$ as before from samples $(x_1, y_1), \dots, (x_n, y_n)$ and set

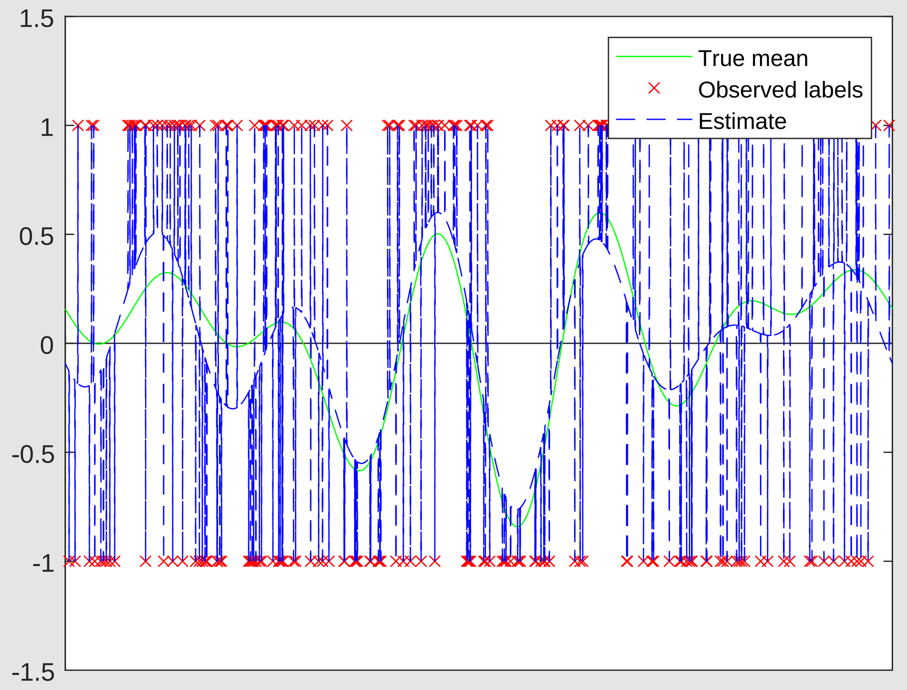
$$\hat{y}(x) = \operatorname{sign}(f(x, \hat{\beta}))$$

Classification is easier than regression since we only need the sign!

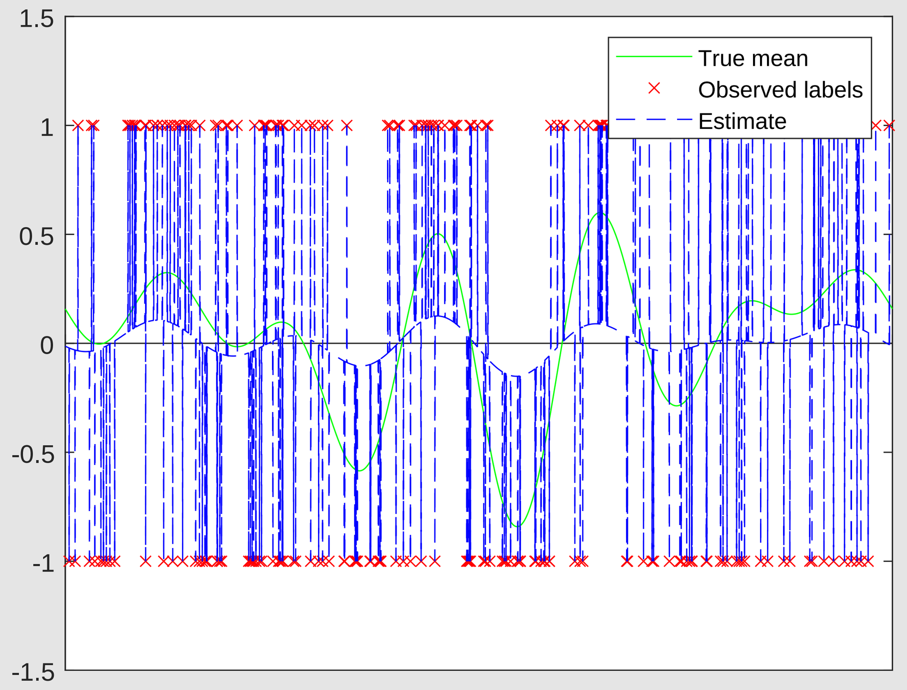
- $\exists$ regimes where regression error is large but classification risk is small
- Previous work assumes Gaussian features

Idea: if $\hat{f} = s f^* + \text{residual}$, $\hat{f}$ (mostly) has the same sign as f^* if residual is $\ll s$, even if $s \ll 1$.

- Good regression performance requires $s = 1$ and small residual.



Good regression and classification



Poor regression but good classification

Key takeaways

- General linear algebra framework for interpolation phenomenon
- Show interpolation happens in quite general settings
- Show separation between regression and classification in general settings

Reference

A. D. McRae, S. Karnik, M. A. Davenport, and V. Muthukumar, "Harmless interpolation in regression and classification with structured features," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, Virtual conference, Mar. 2022. arXiv: `2111.05198 [stat.ML]`, forthcoming