Harmless interpolation in regression and classification with structured features

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

Georgia Tech School of Electrical and Computer Engineering
admcrae@gatech.edu

AISTATS
March 2022

Setup: feature maps for linear regression

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 we observe $y_i = f^*(x_i) + \xi_i$ 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 $\alpha \geq 0$:

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

Noise requires regularization—right?

$$y = \underbrace{\mathcal{A}}_{n\times d}\beta^* + \xi$$

$$\hat{\beta} = \mathcal{A}^T(\alpha I_n + \mathcal{A}\mathcal{A}^T)^{-1}(\mathcal{A}\beta^* + \xi)$$

If $\alpha = 0$ and $\mathcal{A}$ has full row rank (requires $d \geq n$), $f(\cdot, \hat{\beta})$ will **interpolate** the samples

$$\mathcal{A}\hat{\beta} = \mathcal{A}\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}y = y$$

Noise requires regularization—right?

$$y = \underbrace{\mathcal{A}}_{n\times d}\beta^* + \xi$$

$$\hat{\beta} = \mathcal{A}^T(\alpha I_n + \mathcal{A}\mathcal{A}^T)^{-1}(\mathcal{A}\beta^* + \xi)$$

If $\alpha = 0$ and $\mathcal{A}$ has full row rank (requires $d \geq n$), $f(\cdot, \hat{\beta})$ will **interpolate** the samples

$$\mathcal{A}\hat{\beta} = \mathcal{A}\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}y = y$$

Overparametrization can make interpolation less harmful

Many recent papers show that in certain settings, interpolating noise isn't too bad

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

Why does this occur? (in linear settings)

Our new framework

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

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

Then the data Gram matrix is

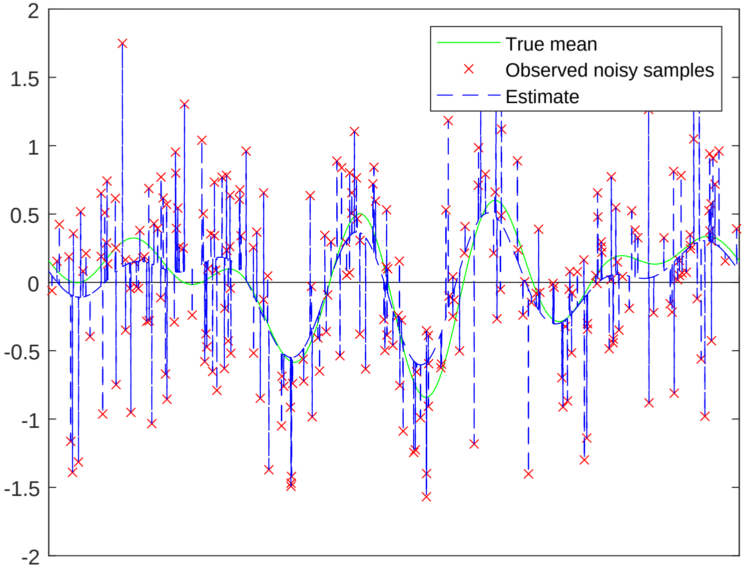
$$\mathcal{A}\mathcal{A}^* = \mathcal{A}_H \mathcal{A}_H^* + \mathcal{A}_R \mathcal{A}_R^*$$

Overparametrization $\implies$ implicit regularization

Gram matrix decomposition: $\mathcal{A}\mathcal{A}^* = \mathcal{A}_H\mathcal{A}_H^* + \mathcal{A}_R\mathcal{A}_R^*$

- If $d - p \gg n$, we can have $\mathcal{A}_R\mathcal{A}_R^* \approx \bar{\alpha} I_n$ ($\bar{\alpha} > 0$)!
  - $\implies \hat{\beta} \approx \mathcal{A}^*(\bar{\alpha} I_n + \mathcal{A}_H\mathcal{A}_H^*)^{-1} y$.
  - (Approximately) ridge regression with positive regularization!
- Previous work assumes **independent features** (or other very restrictive assumptions)
  - Only requires $d - p \gtrsim n$
  - Not very realistic: kernel/RKHS regression, Fourier features, etc.
- **Our work**: for merely **uncorrelated** features, $d - p \gtrsim n^2$ is enough

Example (Fourier basis)



Extension to classification

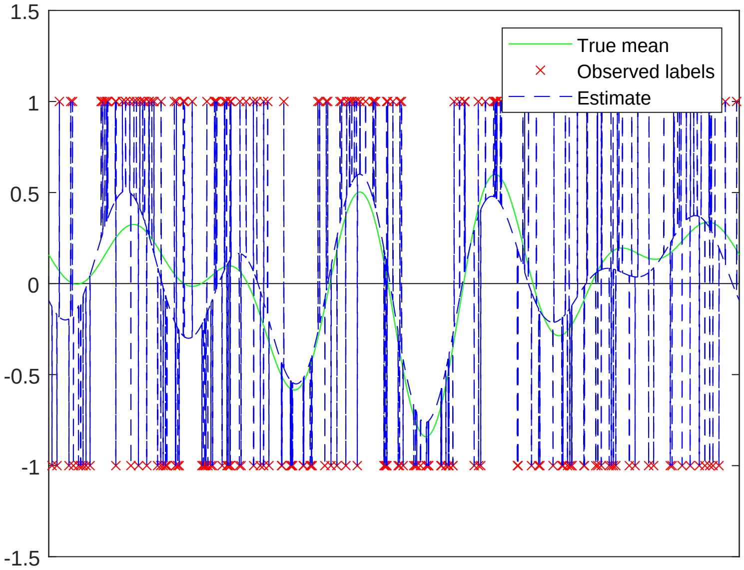
- Now y is a label in $\{-1, 1\}$. Let

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

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

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

Binary labels example



Finer analysis for classification

$$\hat{y}(x) = \text{sign}(f(x, \hat{w}))$$

- Classification is **easier** than regression since we only need the sign!
  - $\exists$ regimes where regression error is large but classification risk is small
  - Again, we show this in much more general settings than before

Large regression but small classification error