Risk bounds for regression and classification with structured feature maps

Andrew D. McRae

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

IFDS-MADLab Workshop, UW Madison
August 2021

Setup: feature maps for linear regression

Linear regression model with feature map $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_d(x))$:

$$f(x, \boldsymbol{w}) = \langle \boldsymbol{\phi}(x), \boldsymbol{w} \rangle = \sum_{\ell} w_\ell \phi_\ell(x)$$

Suppose $f^*(x) = f(x, \boldsymbol{w}^*)$, and observe $y_i = f^*(x_i) + \xi_i$ for $i = 1, \ldots, n$. In matrix form,

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \phi_1(x_1) & \cdots & \phi_d(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_n) & \cdots & \phi_d(x_n) \end{bmatrix}}_{\boldsymbol{\Phi}} \boldsymbol{w}^* + \underbrace{\begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix}}_{\boldsymbol{\xi}}$$

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

$$\widehat{\boldsymbol{w}} = (\delta \boldsymbol{I}_d + \boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{y} = \boldsymbol{\Phi}^T (\delta \boldsymbol{I}_n + \underbrace{\boldsymbol{\Phi}\boldsymbol{\Phi}^T}_{\text{Gram matrix}})^{-1} \boldsymbol{y}$$

Noise requires regularization—right?

$$\boldsymbol{y} = \underbrace{\boldsymbol{\Phi}}_{n \times d} \boldsymbol{w}^* + \boldsymbol{\xi}$$

$$\widehat{\boldsymbol{w}} = \boldsymbol{\Phi}^T (\delta \boldsymbol{I}_n + \boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1} (\boldsymbol{\Phi}\boldsymbol{w}^* + \boldsymbol{\xi})$$

If $\delta = 0$ and $d \geq n$, $f(\cdot, \widehat{\boldsymbol{w}})$ will **interpolate** the samples

$$\boldsymbol{\Phi}\widehat{\boldsymbol{w}} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T (\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1} \boldsymbol{y} = \boldsymbol{y}$$

Noise requires regularization—right?

$$\boldsymbol{y} = \underbrace{\boldsymbol{\Phi}}_{n\times d} \boldsymbol{w}^* + \boldsymbol{\xi}$$

$$\widehat{\boldsymbol{w}} = \boldsymbol{\Phi}^T(\delta \boldsymbol{I}_n + \boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1}(\boldsymbol{\Phi}\boldsymbol{w}^* + \boldsymbol{\xi})$$

If $\delta = 0$ and $d \geq n$, $f(\cdot, \widehat{\boldsymbol{w}})$ will **interpolate** the samples

$$\boldsymbol{\Phi}\widehat{\boldsymbol{w}} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T(\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1}\boldsymbol{y} = \boldsymbol{y}$$

Overparametrization...

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

- Why?

Split the features into two groups:

$$\boldsymbol{\phi}(x) = (\underbrace{\phi_1(x), \ldots, \phi_p(x)}_{\boldsymbol{\phi}_H(x)}, \underbrace{\phi_{p+1}(x), \ldots, \phi_d(x)}_{\boldsymbol{\phi}_R(x)}), \qquad \boldsymbol{\Phi} = [\underbrace{\boldsymbol{\Phi}_H}_{n\times p} \quad \underbrace{\boldsymbol{\Phi}_R}_{n\times(d-p)}]$$

Then

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^T = \boldsymbol{\Phi}_H\boldsymbol{\Phi}_H^T + \boldsymbol{\Phi}_R\boldsymbol{\Phi}_R^T$$

...can give implicit regularization

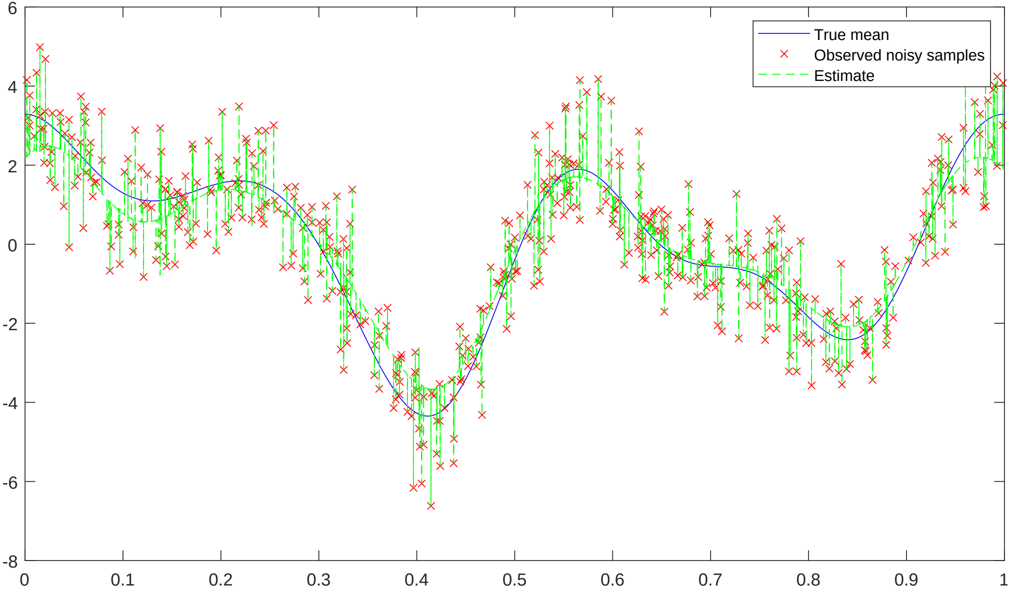
Gram matrix: $\boldsymbol{\Phi}\boldsymbol{\Phi}^T = \boldsymbol{\Phi}_H\boldsymbol{\Phi}_H^T + \boldsymbol{\Phi}_R\boldsymbol{\Phi}_R^T$

- If $d - p \gg n$, sometimes $\boldsymbol{\Phi}_R\boldsymbol{\Phi}_R^T \approx r\boldsymbol{I}_n$ (for some $r > 0$)! So $\hat{\boldsymbol{w}} \approx \boldsymbol{\Phi}^T(r\boldsymbol{I}_n + \boldsymbol{\Phi}_H\boldsymbol{\Phi}_H^T)^{-1}\boldsymbol{y}$.
- Previous work[1 2] assumes **independent features**
  - Only requires $d - p \gtrsim n$
  - Not always realistic: kernel/RKHS regression, Fourier features, etc.
- Our work: for merely **uncorrelated** features, $d - p \gtrsim n^2$ is enough

[1] Peter L. Bartlett et al. (2020). "Benign overfitting in linear regression". In: *Proc. Natl. Acad. Sci. U.S.A.* 117.48, pp. 30063–30070.

[2] Tengyuan Liang and Alexander Rakhlin (2020). "Just interpolate: Kernel "Ridgeless" regression can generalize". In: *Ann. Stat.* 48.3, pp. 1329–1347.

Example (Fourier series)



What about classification?

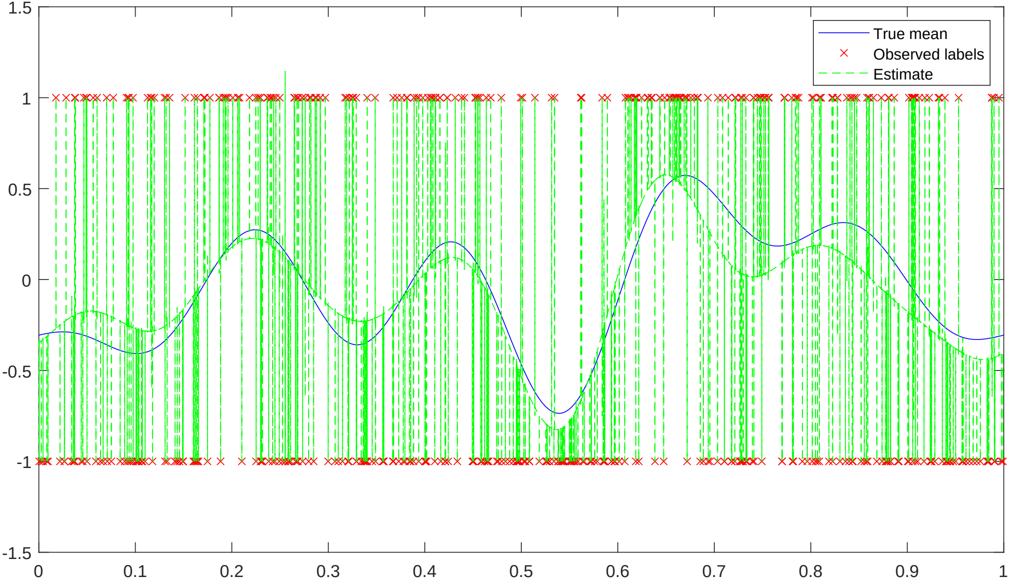
- Now y is a label in $\{-1, 1\}$. Let

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

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

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

Binary labels example



Finer analysis for classification

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

- Classification is **easier** than regression since we only need the sign!
  - $\exists$ regimes where regression error is large but classification risk is small
  - Previously shown under very special conditions[3]
  - We've proved this in more general setting (uncorrelated features, more general f^*)
  - Basic idea: if $f(x, \widehat{\boldsymbol{w}}) = af^*(x) + h(x)$, then (excess) classification risk is small as long as $a > 0$ and $\|h\|_{L_1} \ll a$, even if $a \ll 1$!

[3] Vidya Muthukumar et al. (2021). "Classification vs. regression in overparameterized regimes: Does the loss function matter?" In: *J. Mach. Learn. Res.* arXiv: 2005.08054. Forthcoming.

Large regression but small classification error