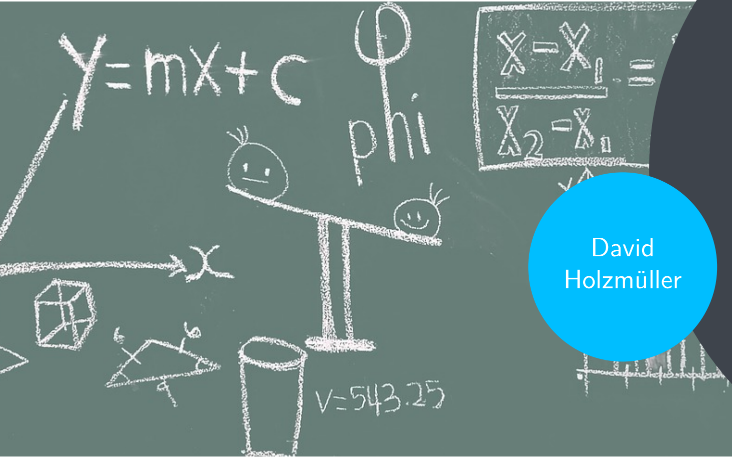
University of Stuttgart
Department of Mathematics



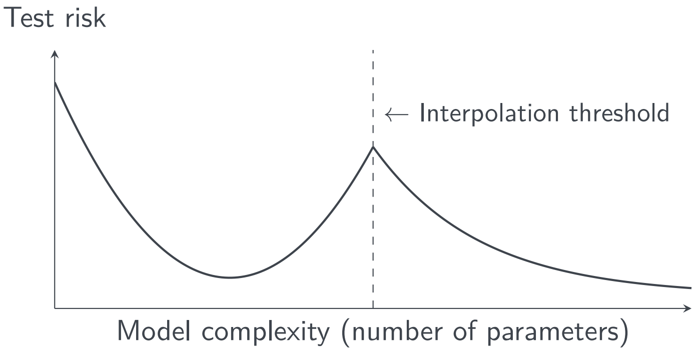
On the Universality of the Double Descent Peak in Ridgeless Regression

David Holzmüller

ICLR 2021

22nd March 2021

Double Descent: Main message



Every interpolating linear model is sensitive to label noise around the interpolation threshold.

Setting (slightly simplified)

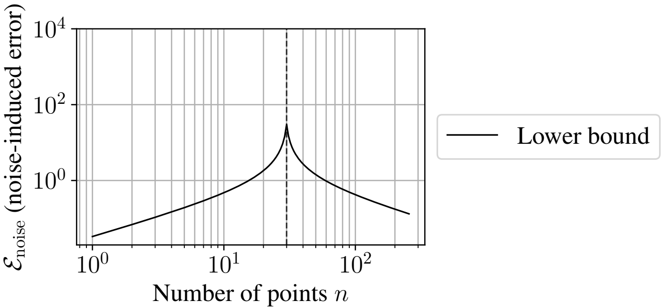
'Ridgeless' linear regression in feature space with p features and n samples:

- Inputs $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ i.i.d., $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$ with centered i.i.d. label noise ε_i
- Given a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^p$, compute

$$\hat{f}_{\boldsymbol{X}_{\text{train}}, \boldsymbol{y}_{\text{train}}}(\boldsymbol{x}) = \widehat{\boldsymbol{\beta}}^\top \phi(\boldsymbol{x}), \qquad \widehat{\boldsymbol{\beta}} = \underbrace{\phi(\boldsymbol{X}_{\text{train}})^+}_{\text{pseudoinverse}} \boldsymbol{y}_{\text{train}} \,.$$

- Expected excess risk: $\mathcal{E}(f) := \mathbb{E}_{\boldsymbol{X}_{\text{train}}, \varepsilon, \boldsymbol{x}_{\text{test}}} \left(f(\boldsymbol{x}_{\text{test}}) - \hat{f}_{\boldsymbol{X}_{\text{train}}, f(\boldsymbol{X}_{\text{train}}) + \varepsilon}(\boldsymbol{x}_{\text{test}}) \right)^2$
- Lower bound: $\mathcal{E}_{\text{noise}} := \mathcal{E}(0) = \min_{f : \mathbb{R}^d \to \mathbb{R}} \mathcal{E}(f)$

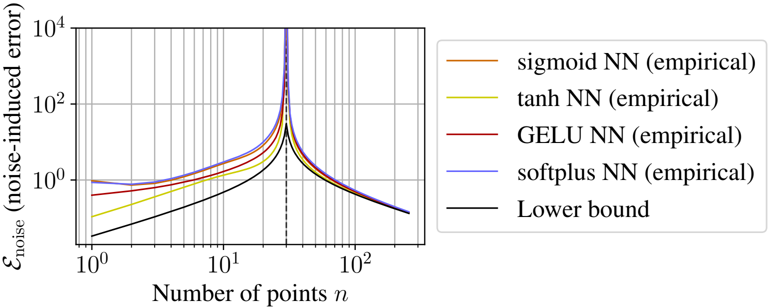
A lower bound (simplified)



Assume that $\mathrm{Var}(\varepsilon_i) \geq \sigma^2$ and that p points can be interpolated almost surely. Then:

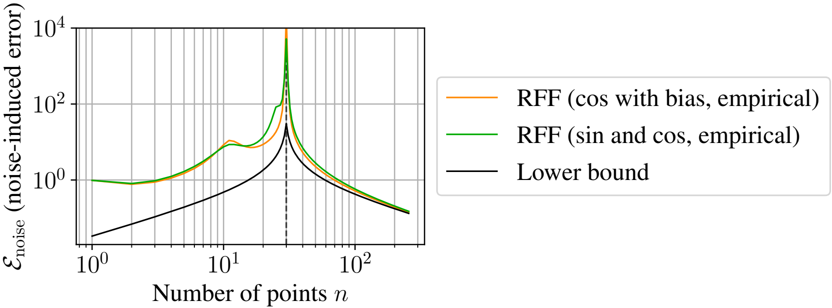
- Underparameterized case $p \leq n$: $\mathcal{E}_{\text{noise}} \geq \sigma^2 \frac{p}{n+1-p}$
- Overparameterized case $p \geq n$: $\mathcal{E}_{\text{noise}} \geq \sigma^2 \frac{n}{p+1-n}$

When are the assumptions satisfied? (1)



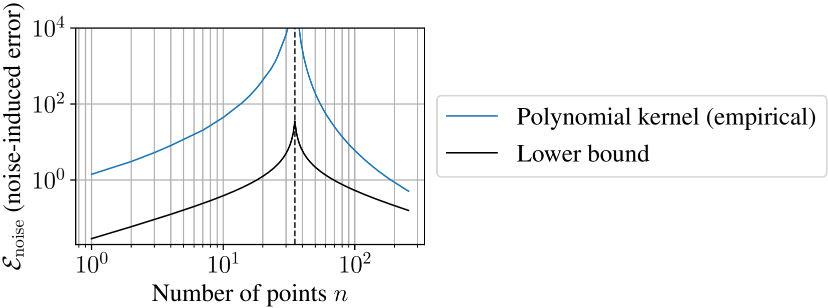
- Random deep NN feature map with non-polynomial *analytic activation function*
- Input x with non-atomic distribution (every point has probability zero)

When are the assumptions satisfied? (2)



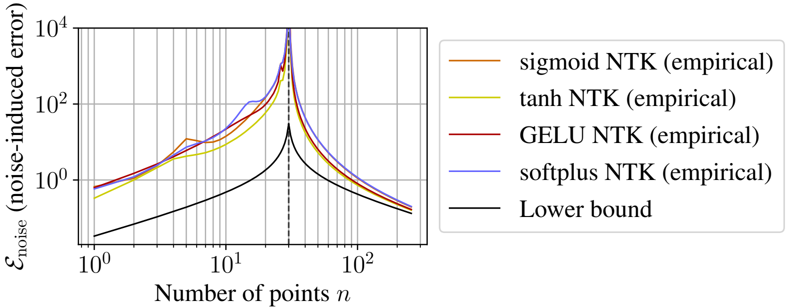
- Random Fourier features for kernel with continuous spectrum
- Input x with non-atomic distribution (every point has probability zero)

When are the assumptions satisfied? (3)



- Polynomial kernel $k(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = (\langle \boldsymbol{x}, \tilde{\boldsymbol{x}} \rangle + c)^m$, $c > 0$, with $p := \binom{m+d}{m}$
- Input x with (Lebesgue) density

When are the assumptions satisfied? (4)



- Simple computational verification for analytic random feature maps
- Here: Finite-width NTK, input x with (Lebesgue) density

Remarks

- Lower bound is asymptotically sharp for $n, p \to \infty$, $p/n \to \gamma \in (0, \infty)$
- **Key takeaway**: Cannot prevent Double Descent by engineering the feature map