Fast rates for noisy interpolation require rethinking the effects of inductive bias

International Conference on Machine Learning

Konstantin Donhauser, joint work with N. Ruggeri, S. Stojanovic and F. Yang

Statistical Machine Learning group, CS department, ETH Zurich

Motivation

- Large overparameterized models: regularization is not necessary for good generalization
- Even when training data is noisy ⇨ Counter-intuitive from a classical statistical viewpoint
- Inspired a new line of research studying *simple* high-dimensional interpolators

Motivation

- Large overparameterized models: regularization is not necessary for good generalization
- Even when training data is noisy ⇨ Counter-intuitive from a classical statistical viewpoint
- Inspired a new line of research studying *simple* high-dimensional interpolators

Setting: High-dimensional linear models $f(x) = w^\top x$

Regression:

Classification:

Interpolators:

for $p \in [1,2]$

Regression: $\hat{w} = \mathrm{argmin}_w \|w\|_p \; s.t. \; y = Xw$

Classification: $\hat{w} = \mathrm{argmin}_w \|w\|_p \; s.t. \; y_i \langle x_i, w \rangle \geq 1 \; \forall i$

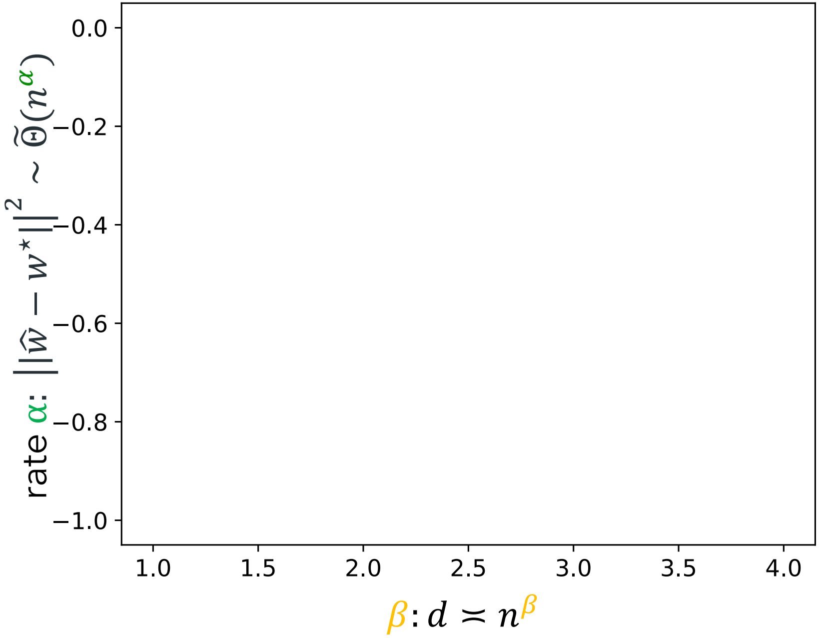
Data distribution for regression

Data model: n samples (x_i, y_i) with

- $y_i = \langle w^\star, x_i \rangle + \xi_i$ with $x_i \sim N(0, I_d)$
- structured (sparse) ground truth

  $w^\star = (1,0, \dots, 0)$
- $d \asymp n^\beta$ with $\beta > 1$ and n large enough

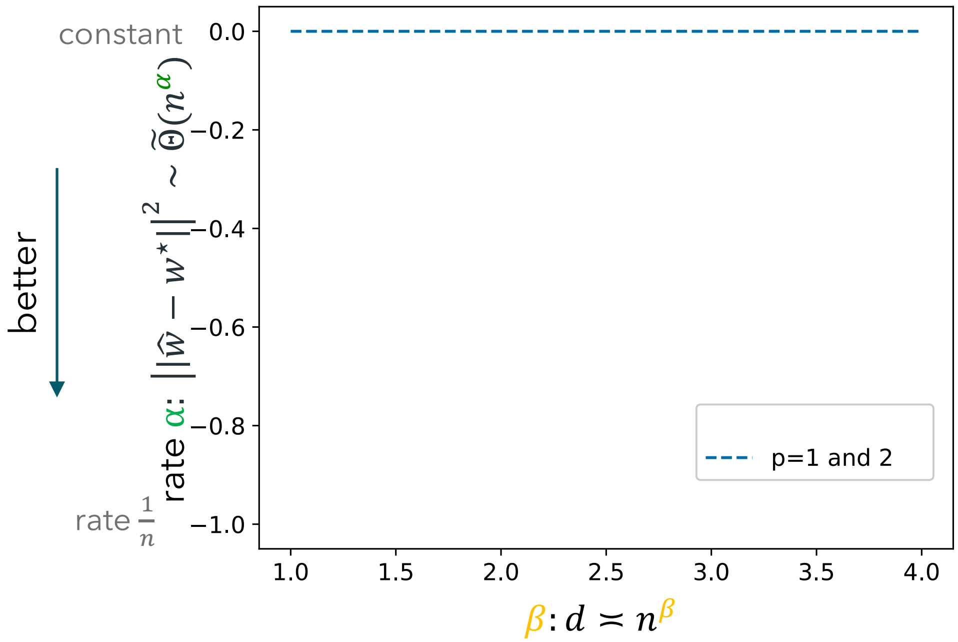
Data distribution for regression



Data model: n samples (x_i, y_i) with

- $y_i = \langle w^\star, x_i \rangle + \xi_i$ with $x_i \sim N(0, I_d)$
- structured (sparse) ground truth $w^\star = (1,0,\dots,0)$
- $d \asymp n^\beta$ with $\beta > 1$ and n large enough

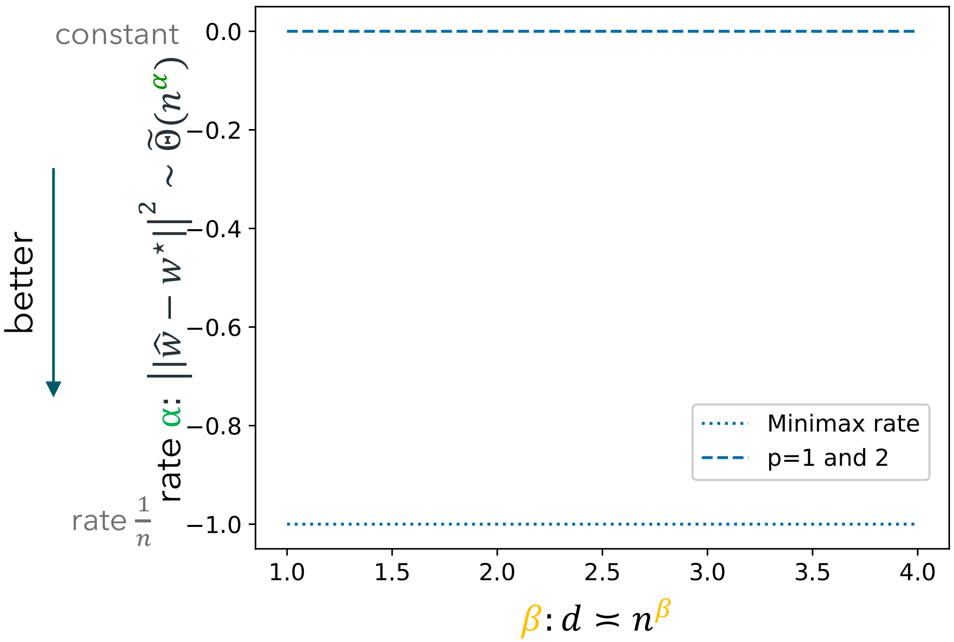
Prior work: Slow rates for $p = 1$ or $p = 2$



- Interpolators with $p = 1, 2$:

  $||\hat{w} - w^\star||^2 = \tilde{\Theta}(1) \rightarrow \alpha = 0$

Prior work: Slow rates for $p = 1$ or $p = 2$



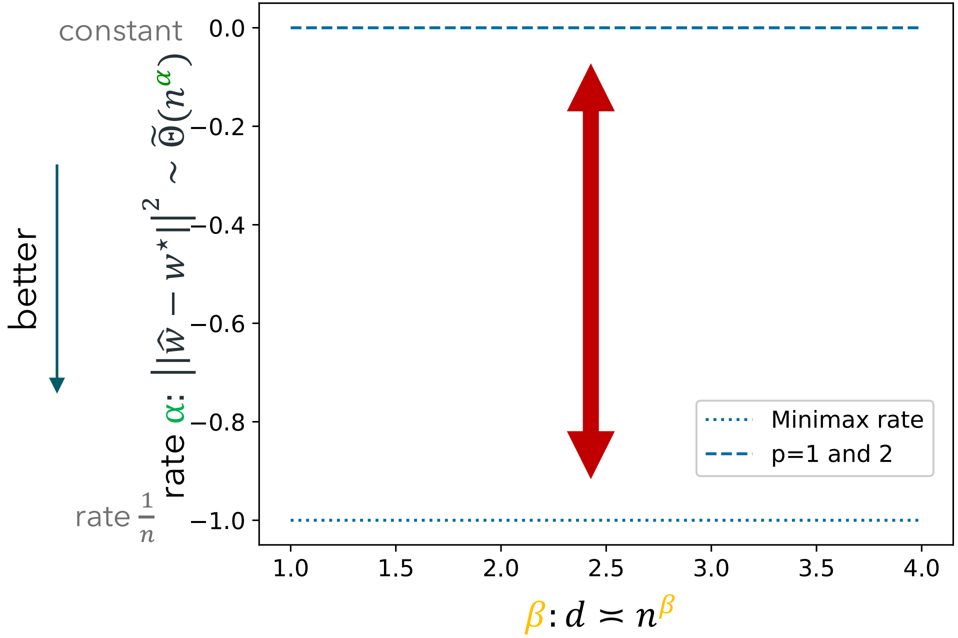
- Interpolators with $p = 1, 2$:

  $||\hat{w} - w^\star||^2 = \tilde{\Theta}(1) \rightarrow \alpha = 0$

- minimax optimal rate (e.g., ℓ_1-norm regularized estimator LASSO)

  $||\hat{w} - w^\star||^2 = \tilde{\Theta}(n^{-1}) \rightarrow \alpha = -1$

Prior work: Slow rates for $p = 1$ or $p = 2$



- Interpolators with $p = 1, 2$:

  $||\hat{w} - w^\star||^2 = \widetilde{\Theta}(1) \to \alpha = 0$

- minimax optimal rate (e.g., ℓ_1-norm regularized estimator LASSO)

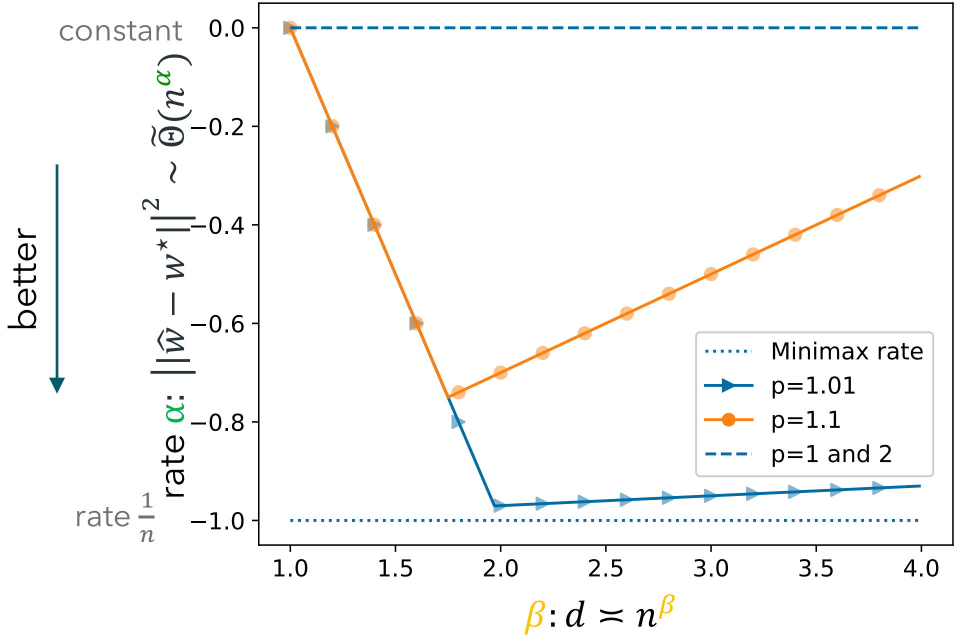
  $||\hat{w} - w^\star||^2 = \widetilde{\Theta}(n^{-1}) \to \alpha = 1$

Interpolators have far from optimal prediction performance?

This paper: Fast rates for $p \in (1,2)$

Main Theorem: non-asymptotic upper and lower bounds for $p \in (1,2)$

This paper: Fast rates for $p \in (1,2)$

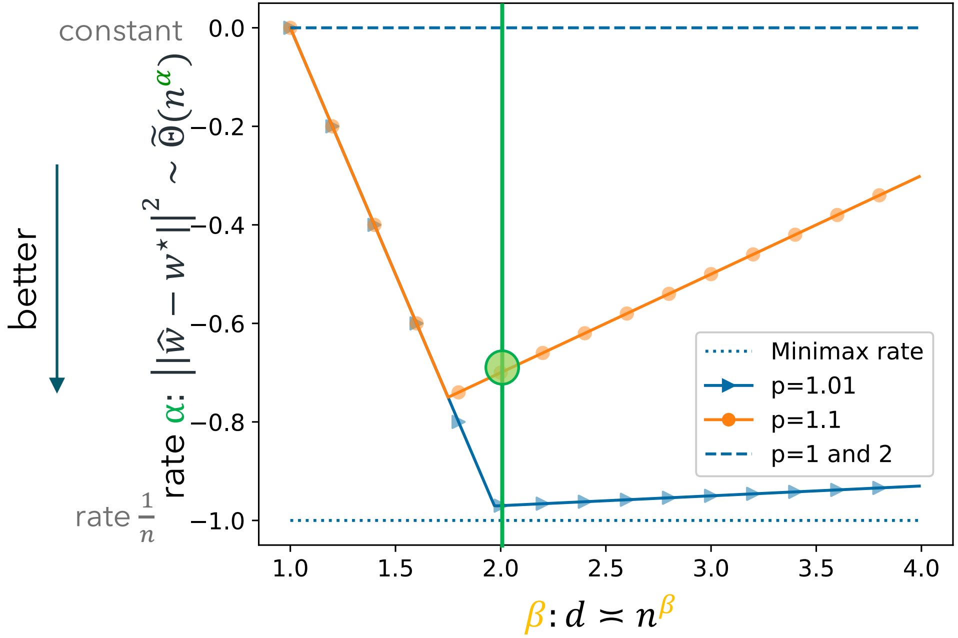


Main Theorem: non-asymptotic upper and lower bounds for $p \in (1,2)$

For example, for $\beta = 2$, i.e. $d \asymp n^2$

- $p = 1.1 \rightarrow \|\hat{w} - w^\star\|^2 \approx n^{-0.7}$

This paper: Fast rates with $p \in (1,2)$

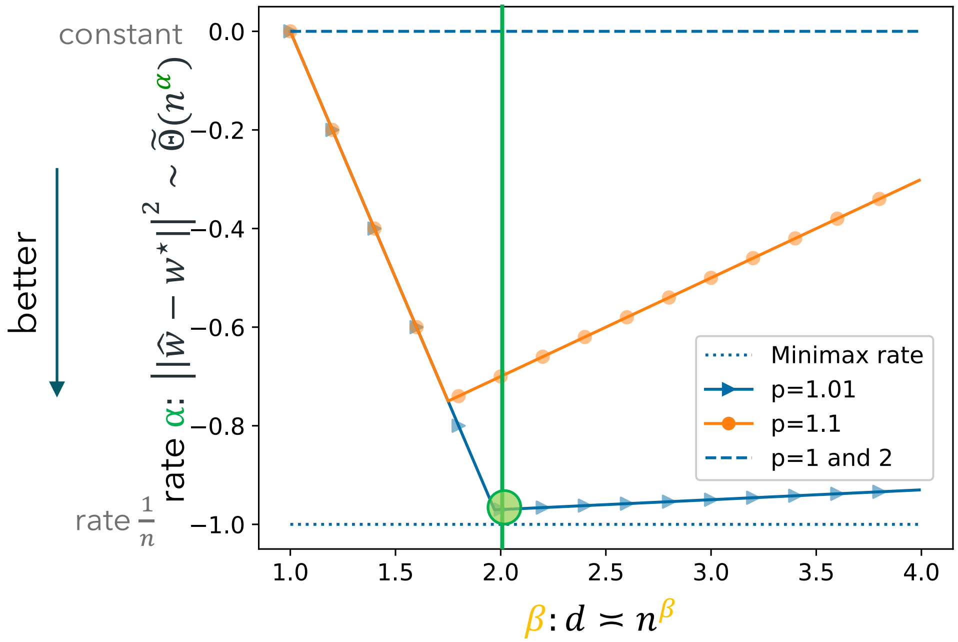


Main Theorem: non-asymptotic upper and lower bounds for $p \in (1,2)$

For example, for $\beta = 2$, i.e. $d \asymp n^2$

- $p = 1.1 \rightarrow ||\hat{w} - w^\star||^2 \approx n^{-0.7}$

This paper: Fast rates with $p \in (1,2)$



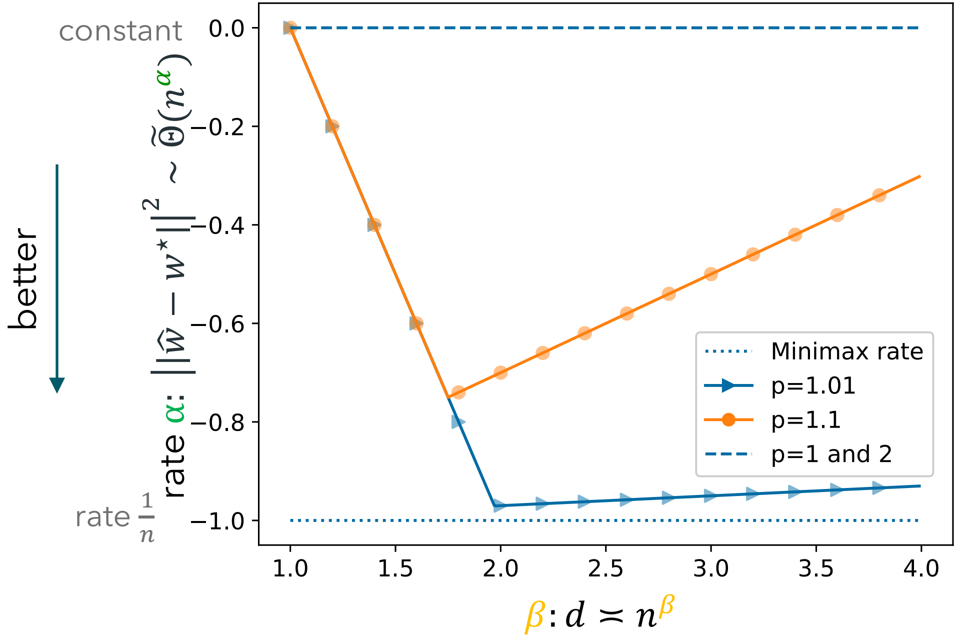
Main Theorem: non-asymptotic upper and lower bounds for $p \in (1,2)$

For example, for $\beta = 2$, i.e. $d \asymp n^2$

- $p = 1.1 \rightarrow \|\hat{w} - w^\star\|^2 \approx n^{-0.7}$
- $p = 1.01 \rightarrow \|\hat{w} - w^\star\|^2 \approx n^{-1}$

Can achieve rates even close to the minimax lower bound $\tilde{O}(n^{-1})$

This paper: Fast rates with $p \in (1,2)$



Main Theorem: non-asymptotic upper and lower bounds for $p \in (1,2)$

For example, for $\beta = 2$, i.e. $d \asymp n^2$

- $p = 1.1 \ \rightarrow ||\hat{w} - w^\star||^2 \approx n^{-0.7}$
- $p = 1.01 \ \rightarrow ||\hat{w} - w^\star||^2 \approx n^{-1}$

Similar results also hold for classification

Thanks for listening!

We are looking forward to seeing you at
the poster #1109