Tight bounds for minimum ℓ_1-norm interpolation of noisy data

Guillaume Wang*, Konstantin Donhauser*, Fanny Yang

ETH Zurich

Motivation

In high dimensions, models that interpolate noisy training data can still generalize well [1]. How come?

"Benign overfitting" explanation [2]: min-ℓ_2-norm interpolation is consistent when covariates are *effectively low-dimensional*, i.e. $d_{\text{eff}} = \text{tr}(\Sigma)/\|\Sigma\|_{\text{op}} \ll n$.

- What about effectively high-dimensional covariates $d_{\text{eff}} = d \gg n$?
- What about other interpolating models?

Can we consistently learn sparse ground truths with minimum-norm interpolators on high-dimensional features?

This work: **YES** for isotropic covariates $x \sim \mathcal{N}(0, I_d)$, sparse ground truth $\|w^*\|_0 \leq \tilde{O}(n)$, and min-$\ell_1$-norm interpolation.

Previous results

Min-ℓ_1-norm interpolation (Basis Pursuit) in our setting was known to

- achieve consistency for zero noise $\sigma = 0$;
- have statistical rate $\|\hat{w} - w^*\|_2^2 \leq O(\sigma^2)$ as $d/n \to \infty$ [3];
- have statistical rate $\|\hat{w} - w^*\|_2^2 \geq \Omega\left(\frac{\sigma^2}{\log(d/n)}\right)$ [4].

We close the gap between upper and lower bound, showing $\|\hat{w} - w^*\|_2^2 \sim \frac{\sigma^2}{\log(d/n)}$. In particular, Basis Pursuit is consistent even in the presence of noise.

Remark. *In practice, ℓ_1-norm penalization (LASSO) is preferable to interpolation when noise is present.*

Main result

Problem setting:

- Data model: covariates $x \sim \mathcal{N}(0, I_d)$, noisy observations $y = \langle w^*, x\rangle + \xi$ where $\xi \sim \mathcal{N}(0, \sigma^2)$.
- Prediction error $\mathbb{E}_{x,y}(\langle \hat{w}, x\rangle - y)^2 = \|\hat{w} - w^*\|_2^2 + \sigma^2$.
- We study the min-ℓ_1-norm interpolator defined by

$$\hat{w} = \operatorname*{argmin}_w \|w\|_1 \quad \text{such that} \quad \forall i,\ \langle x_i, w\rangle = y_i.$$

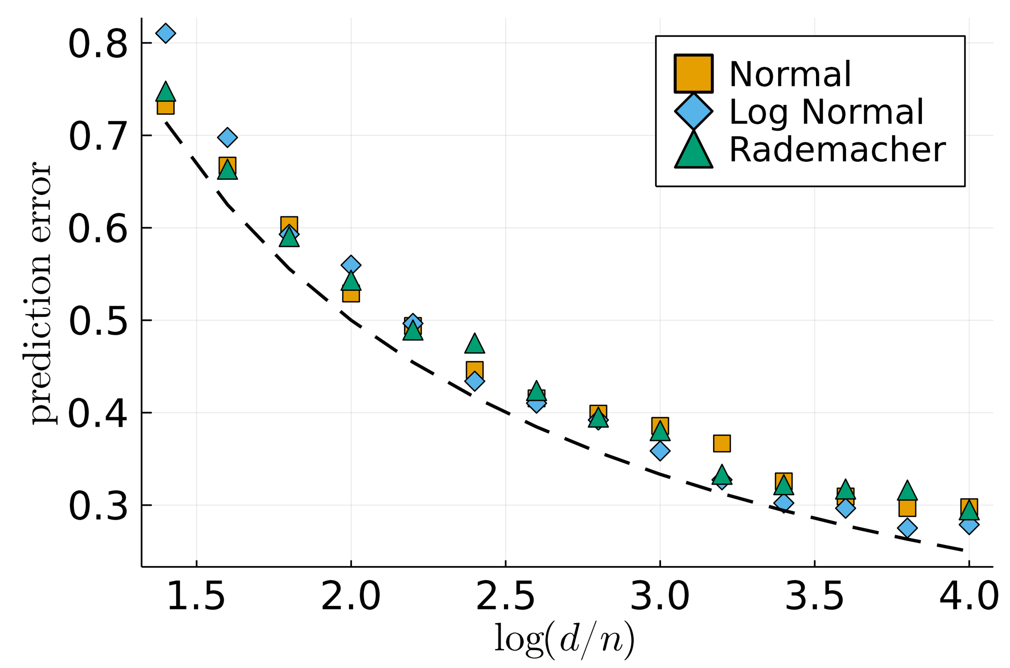
Main result: Non-asymptotic matching upper and lower bounds for prediction error of min-ℓ_1-norm interpolator.

Theorem. *Suppose $\|w^*\|_0 \leq \kappa_1 \frac{n}{\log(d/n)^5}$ for some constant $\kappa_1 > 0$. There exist constants $\kappa_2, \kappa_3, \kappa_4, c_1, c_2, c_3 > 0$ such that, if $n \geq \kappa_2$ and $\kappa_3 n \log(n)^2 \leq d \leq \exp(\kappa_4 n^{1/5})$,*

$$\left| \|\hat{w} - w^*\|_2^2 - \frac{\sigma^2}{\log(d/n)} \right| \leq c_1 \frac{\sigma^2}{\log(d/n)^{3/2}}$$

with probability $\geq 1 - c_2 \exp\left(-\frac{n}{\log(d/n)^5}\right) - d\exp(-c_3 n)$.

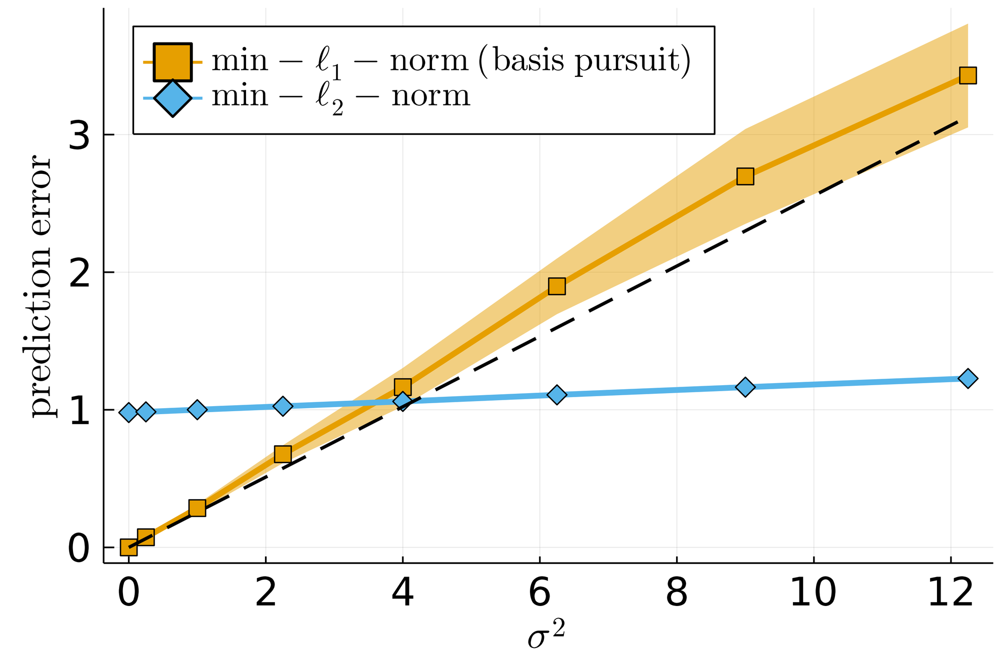
Experimental validation



Dashed curve: theoretical rate
Orange squares: experimental rate for Normal-distributed features (our setting)

Conjecture: min-ℓ_1-norm interpolation also has statistical rate $\sim \frac{\sigma^2}{\log(d/n)}$ for certain heavy-tailed feature distributions.

Comparison to min-ℓ_2 interpolation



Min-ℓ_1-norm interpolation is sensitive to the noise level σ^2; min-ℓ_2-norm interpolation has similar (non-vanishing) prediction error across all values of σ^2.

Trade-off between structural bias vs. sensitivity to noise:

- Min-ℓ_1-norm interpolation (squares):
  - ✓ strong structural bias,
  - ✓ efficient noiseless recovery of sparse signals,
  - ✗ but poor rate in the presence of noise.
- Min-ℓ_2-norm interpolation (diamonds):
  - ✗ no structural bias (except towards zero),
  - ✗ fails to recover any non-zero signal even in the absence of noise,
  - ✓ but does not suffer from overfitting of the noise.

References

[1] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, 2021.

[2] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *PNAS*, 2020.

[3] G. Chinot, M. Löffler, and S. van de Geer, "On the robustness of minimum-norm interpolators," *arXiv:2012.00807*, 2021.

[4] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, "Harmless interpolation of noisy data in regression," *IEEE Journal on Selected Areas in Information Theory*, 2020.