Interpolation can hurt robust generalization even when there is no noise

Konstantin Donhauser*,†, Alexandru Ţifrea*,†, Michael Aerni†, Reinhard Heckel∘,§, Fanny Yang†

ETH Zurich†, Rice University∘, TU Munich§

Phenomenon 1: Double descent

Observed empirically for neural networks and theoretically e.g. for highly overparameterized ($d \gg n$) linear models [1].

- Regularization does not improve generalization, compared to interpolating the training data.
- Overparameterization implicitly controls the variance.

  → Regularization is **redundant**.

Phenomenon 2: Robust overfitting

Observed empirically for neural networks on image data [2].

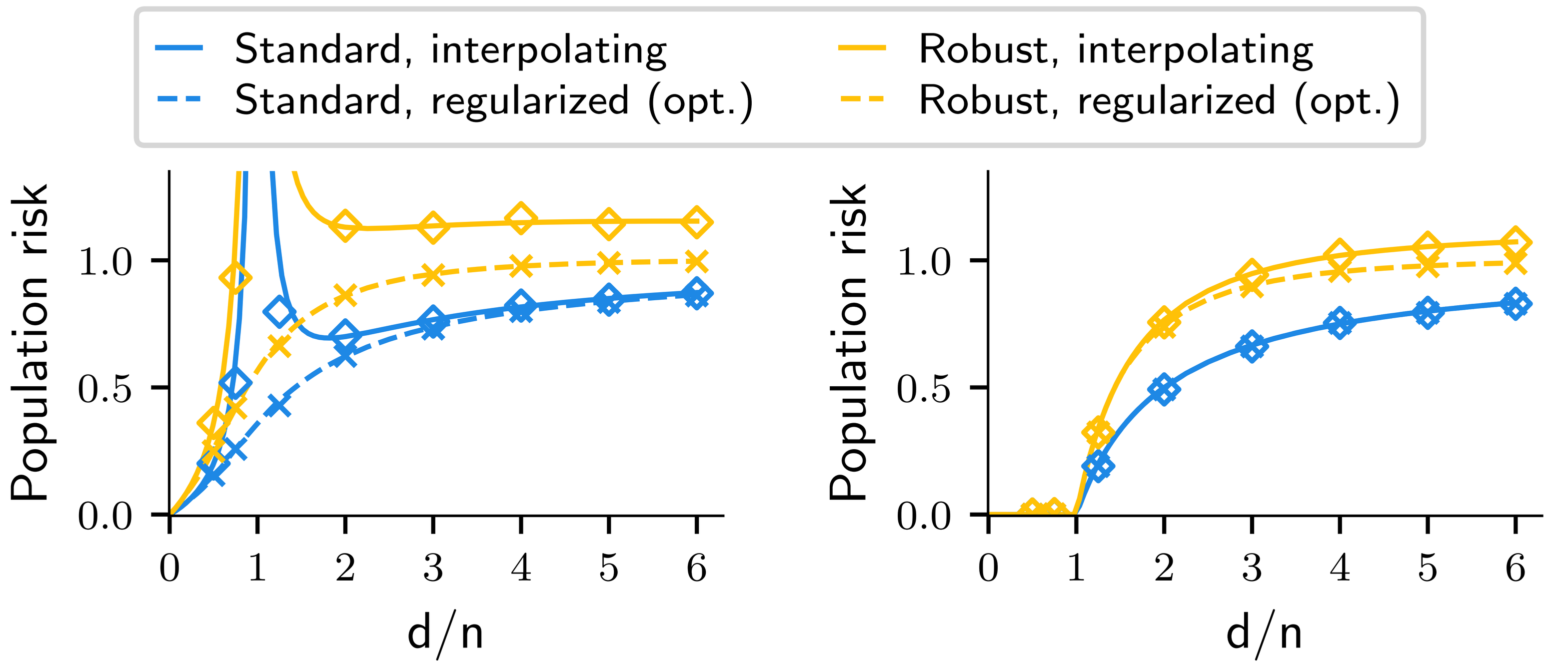
- *Robust* generalization benefits greatly from regularization.
- Prior work has attributed this phenomenon to:
  - noise in the training data
  - non-smooth predictors

Does robust overfitting occur on noiseless data?

Does this provably happen even for linear models?

Robust linear regression

Ridge regularization **avoids the min-norm interpolator**.



Noisy observations *Noiseless observations*

- The lowest robust risks are not obtained by the min-norm interpolators, but by the regularized estimators.
  - holds true even for **noiseless data!**

Robust linear classification

- Evaluation with the **robust risk** wrt ℓ_∞-perturbations:

$$\mathbf{R}_\epsilon(\theta) := \mathbb{E}_{X\sim\mathbb{P}} \max_{\delta\in\mathcal{U}_c(\epsilon)} \mathbb{1}_{\mathrm{sgn}(\langle\theta, X+\delta\rangle)\neq\mathrm{sgn}(\langle\theta^\star, X\rangle)}$$

- We use adversarial training to obtain a robust estimator:

$$\hat{\theta}_\lambda := \arg\min_\theta \frac{1}{n}\sum_{i=1}^n \max_{\delta\in\mathcal{U}_c(\epsilon)} \ell(\langle\theta, x_i+\delta\rangle y_i) + \lambda\|\theta\|_2^2.$$

- For $\lambda \to 0$, it maximizes the robust margin of the data:

$$\hat{\theta}_0 := \arg\min_\theta \|\theta\|_2 \text{ such that for all } i, \max_{\delta\in\mathcal{U}_c(\epsilon)} y_i\langle\theta, x_i+\delta\rangle \geq 1.$$

Theoretical result for classification

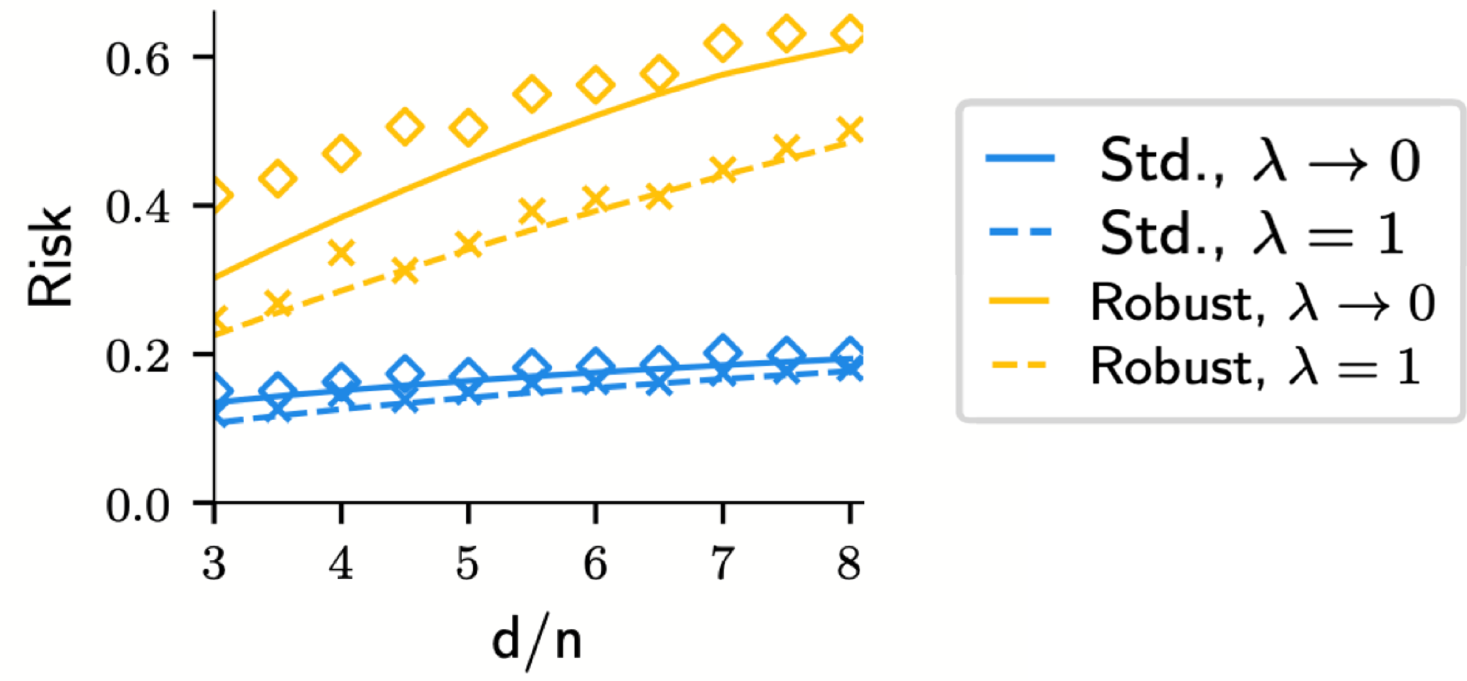
Problem setting:

- Data model: covariates $x \sim \mathcal{N}(0, I_d)$, deterministic labels given by $y = \mathrm{sgn}\langle\theta^\star, x\rangle \in \{-1, +1\}$. → **Noiseless data!**
- We consider **linear classifiers** trained with the logistic loss.

Theorem. *For a sparse ground truth, we derive the limit $\mathcal{R}_\lambda(\epsilon,\gamma)$ of the robust risk as $d, n \to \infty$ and $d/n \to \gamma$:*

$$\mathbf{R}_\epsilon(\hat{\theta}_\lambda) \xrightarrow{prob} \mathcal{R}_\lambda(\epsilon,\gamma)$$

In particular, for some $\lambda_{opt} > 0$: $\underbrace{\mathcal{R}_{\lambda_{opt}}(\epsilon,\gamma)}_{\text{regularized}} < \underbrace{\lim_{\lambda\to 0}\mathcal{R}_\lambda(\epsilon,\gamma)}_{\text{interpolating}}$.

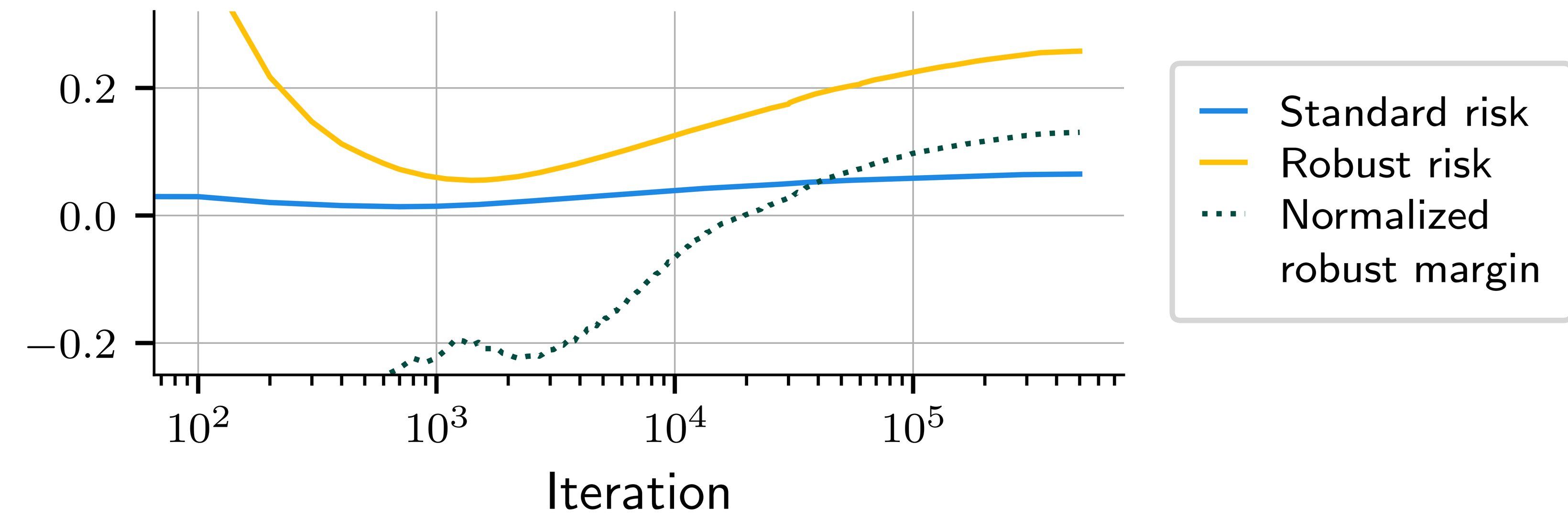


Lines: asymptotic risks (theory)
Markers: risks for finite d, n (simulations)

- Ridge regularization **avoids the max-margin estimator**.

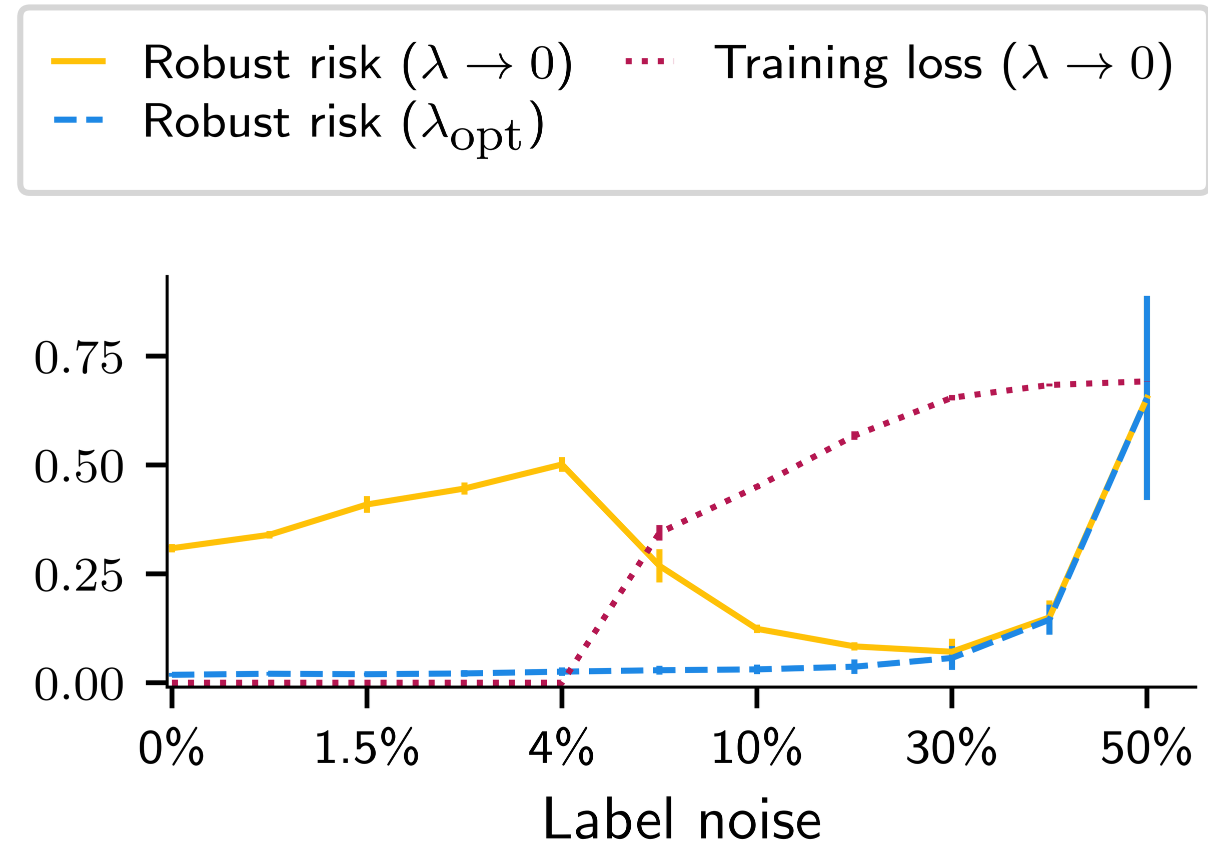
Other ways to avoid $\hat{\theta}_0$

1. Early stopping **avoids the max-margin estimator** and achieves a lower robust risk.



2. Adding artificial label noise prevents a vanishing training loss and **avoids the max-margin estimator**.

   Surprising consequence: Smaller robust risk, compared to the max-margin interpolator of the original clean data.



Remark: Regularization still leads to smaller robust risk, even in the presence of noise.

References

[1] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *arXiv:1903.08560*, 2019.

[2] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *ICML*, 2020.

[3] C. Thrampoulidis, S. Oymak, and B. Hassibi, "Regularized linear regression: A precise analysis of the estimation error," in *COLT*, 2015.