Surprising benefits of ridge regularization for noiseless regression

Konstantin Donhauser[*,†], Alexandru Țifrea[*,†], Michael Aerni[†], Reinhard Heckel[∘,§], Fanny Yang[†]

ETH Zurich[†], Rice University[∘], TU Munich[§]

[*] Equal contribution

Phenomenon 1: Double Descent

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

- Generalization does not benefit from optimal regularization compared to interpolating the training data.
- Overparameterization implicitly controls the variance

$\rightarrow$ Regularization (e.g. ridge or early stopping) is **redundant**.

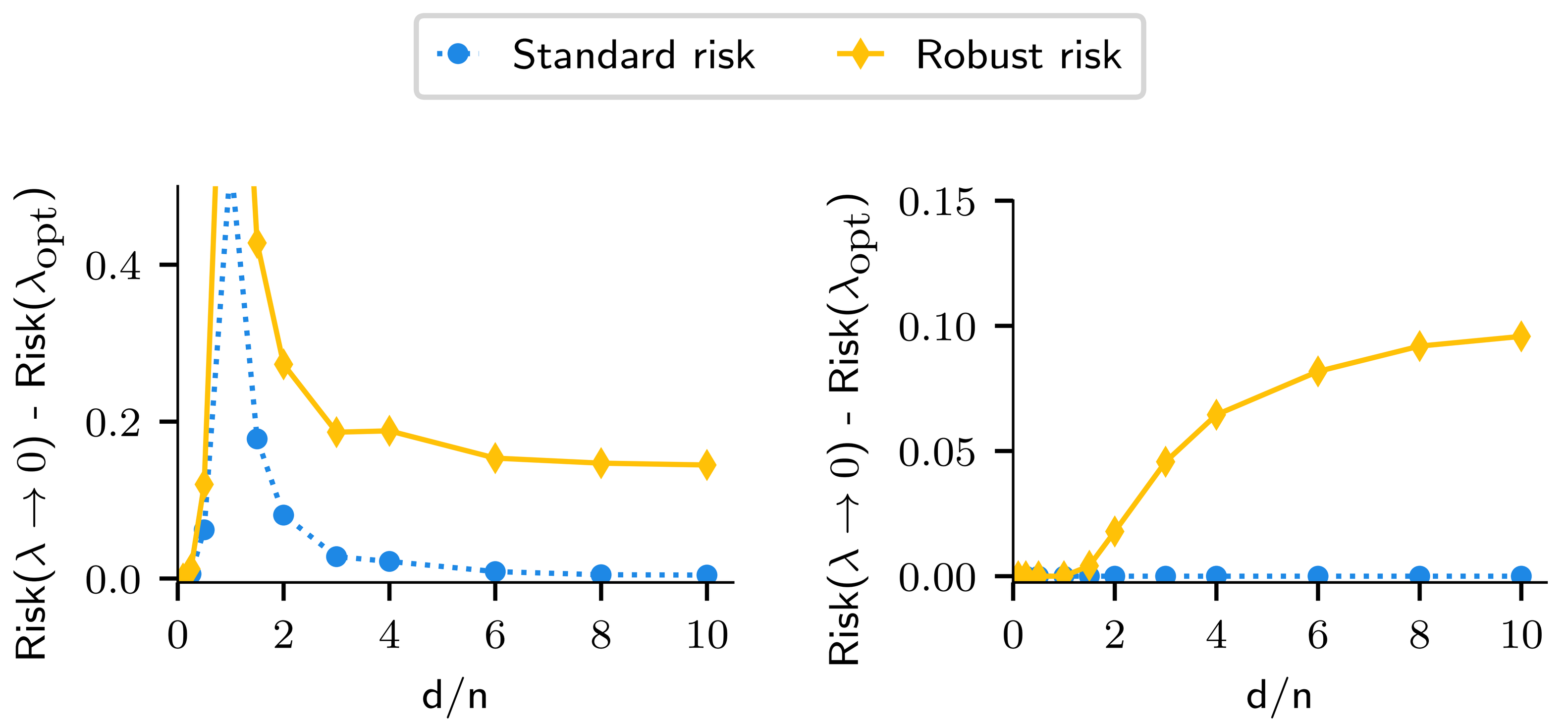
Phenomenon 2: Robust risk overfits

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

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

Does linear regression suffer from robust overfitting?

Yes, even on noiseless training data!



Noisy observations *Noiseless observations*

Problem setting

- We study the **linear ridge regression** estimator:

$$\hat{\theta}_\lambda := \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}(\langle\theta, x_i\rangle - y_i)^2 + \lambda\|\theta\|_2^2.$$

- If $d/n > 1$, $\lambda \to 0$ yields the **minimum ℓ_2-norm interpolator**:

$$\hat{\theta}_0 := \arg\min_\theta \|\theta\|_2 \text{ such that for all } i, \langle\theta, x_i\rangle = y_i.$$

- Evaluation with respect to the **consistent robust risk** with ℓ_2 perturbations:

$$\mathbf{R}_\epsilon(\theta) := \mathbb{E}_{X\sim\mathbb{P}} \max_{\|\delta\|_2\le\epsilon, \langle\theta^\star,\delta\rangle=0} (\langle\theta-\theta^\star, X+\delta\rangle)^2$$

Theoretical result

High-dimensional data model:

- n i.i.d. covariates $x_i \sim \mathcal{N}(0, I_d)$.
- observations $y_i = \langle\theta^\star, x_i\rangle + \xi_i$ with noise $\xi_i \sim \mathcal{N}(0, \sigma^2 I_d)$.
- $d, n \to \infty$, $d/n \to \gamma$.

Theorem. *Define* $m(z) = \frac{1-\gamma-z-\sqrt{(1-\gamma-z)^2-4\gamma z}}{2\gamma z}$ *and let* m' *be its derivative. Let* $\mathcal{P} = \mathcal{B} + \mathcal{V} - \lambda^2(m(-\lambda))^2$ *and* $\mathcal{B} = \lambda^2 m'(-\lambda)$, $\mathcal{V} = \sigma^2\gamma(m(-\lambda) - \lambda m'(-\lambda))$ *be the asymptotic bias and variance. Then,*

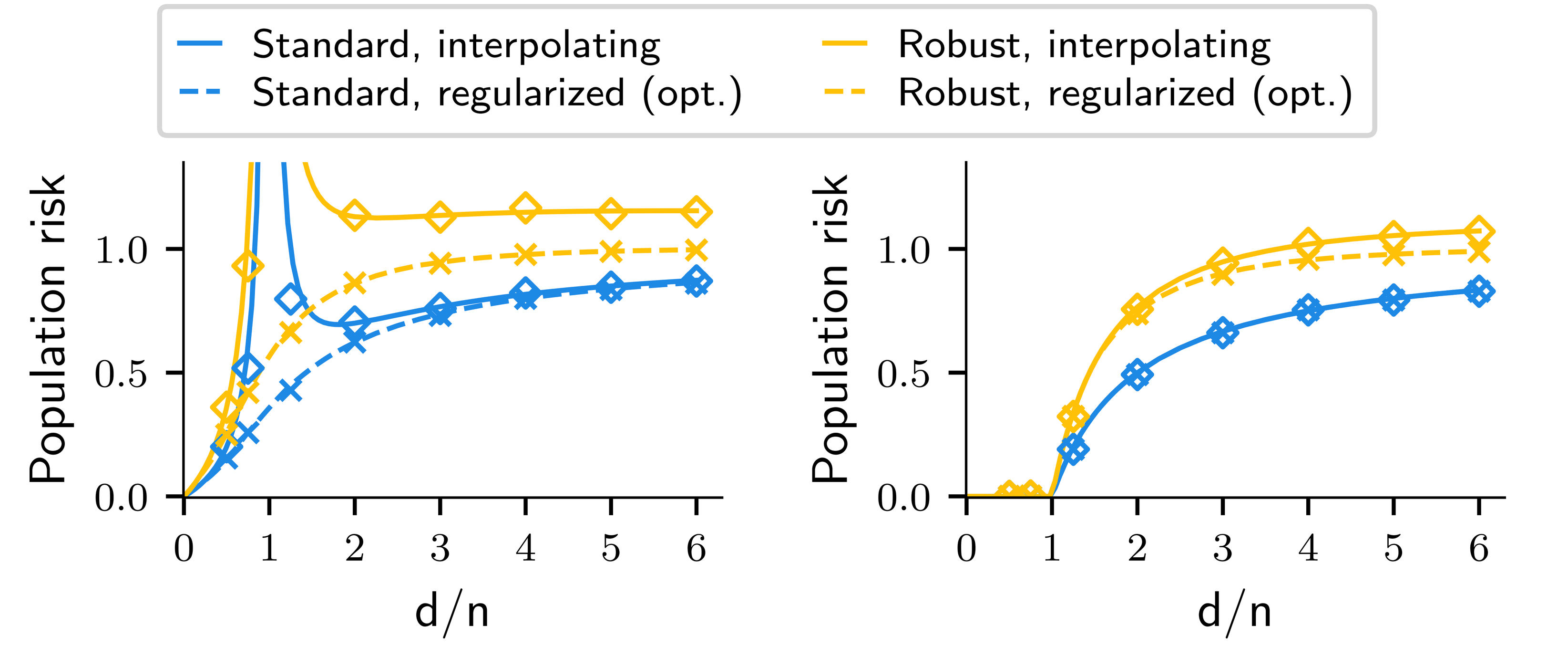
$$\mathbf{R}_\epsilon(\hat{\theta}_\lambda) \xrightarrow{a.s.} \mathcal{B} + \mathcal{V} + \epsilon^2\mathcal{P} + \sqrt{\frac{8\epsilon^2}{\pi}\mathcal{P}(\mathcal{B}+\mathcal{V})}$$

Furthermore, the standard risk $\mathbf{R}(\hat{\theta}_\lambda) \rightarrow \mathcal{B} + \mathcal{V}$ *a.s.*

$\rightarrow$ We can compute the asymptotic standard and robust risks.

Theoretical predictions

Theoretical predictions (lines) for $d, n \to \infty$ and experimental results (markers) for finite d, n.



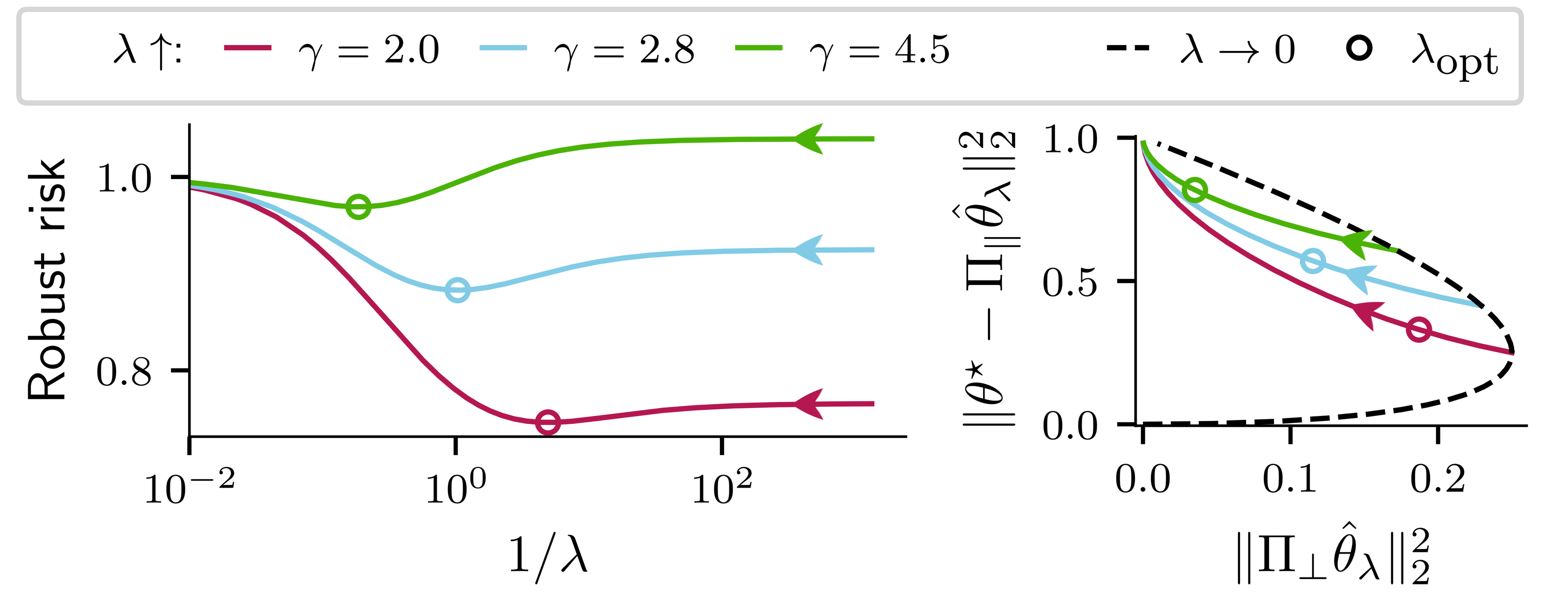
Noisy observations *Noiseless observations*

- Theoretical predictions match simulations for finite d, n.
- **Standard risk:** No overfitting thanks to implicit regularization for large d/n.
- **Robust risk:** Overfitting even for noiseless data and large d/n.

Intuitive explanation

For noiseless observations, both risks depend only on:

- Fit in the direction of the ground truth: $\|\theta^\star - \Pi_\parallel\hat{\theta}_\lambda\|_2^2$.
- Orthogonal misfit: $\|(I - \Pi_\parallel)\hat{\theta}_\lambda\|_2^2$.



$\rightarrow$ Robust risk **punishes orthogonal misfit stronger** than standard risk, leading to $\lambda_{\text{opt}} > 0$.

References

[1] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *arXiv preprint arXiv:1903.08560*, 2019.
[2] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *ICML*, 2020, pp. 8093–8104.
[3] E. Dobriban and S. Wager, "High-dimensional asymptotics of prediction: Ridge regression and classification," *The Annals of Statistics*, pp. 247 – 279, 2018.