Maximizing the robust margin provably overfits on noiseless data

Konstantin Donhauser*,†, Alexandru Țifrea*,†, Michael Aerni†, Reinhard Heckel°,§, Fanny Yang†
ETH Zurich†, Rice University°, TU Munich§

Robust overfitting

- Adversarial training with regularization → more robust than unregularized estimator.
- First observed for neural networks and image data sets [1].
- Prior work has attributed this phenomenon to: (1) noise in the training data; (2) non-smooth predictors.

Does robust overfitting occur on noiseless data?

Can we prove that this happens?

Robust linear classification

- Evaluation with the **robust risk** with 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:

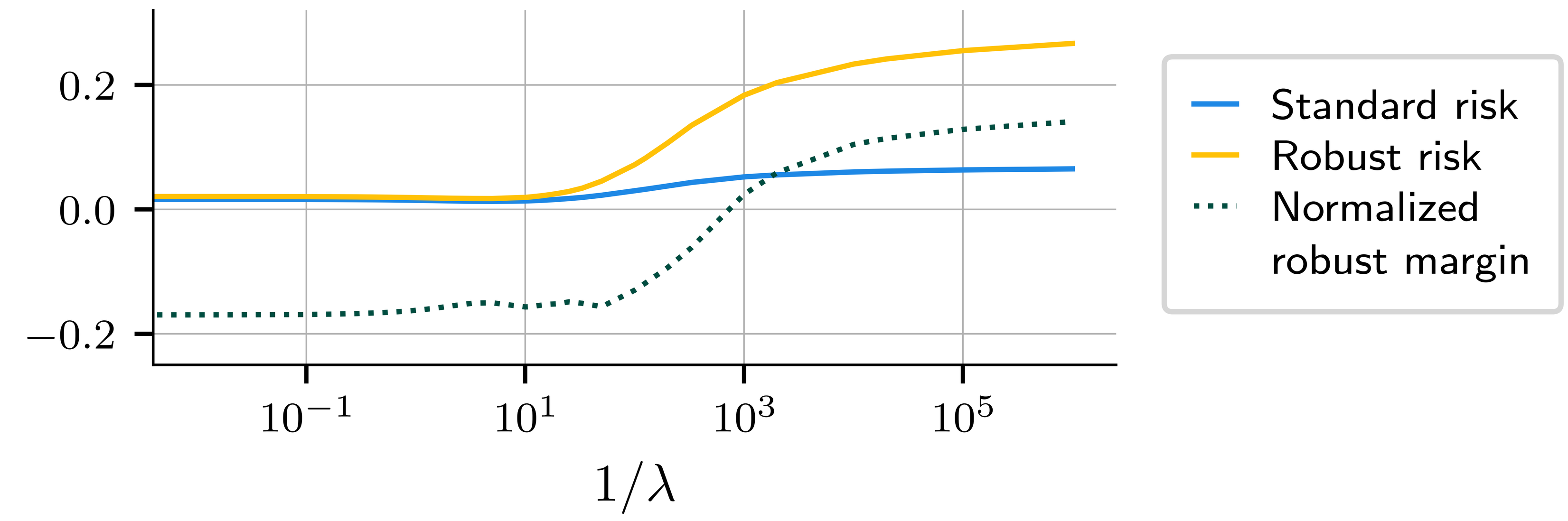
$$\hat{\theta}_\lambda := \arg\min_\theta \frac{1}{n}\sum_{i=1}^n \max_{\delta\in\mathcal{U}(\epsilon)} \ell(\langle\theta, x_i+\delta\rangle y_i) + \lambda\|\theta\|_2^2.$$

- For $\lambda \to 0 \Rightarrow$ maximizes the robust margin of the data.

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

Avoiding $\hat{\theta}_0$ via ridge regularization

Ridge regularization ($\lambda > 0$) yields a negative robust margin → **avoids the max-margin estimator**.



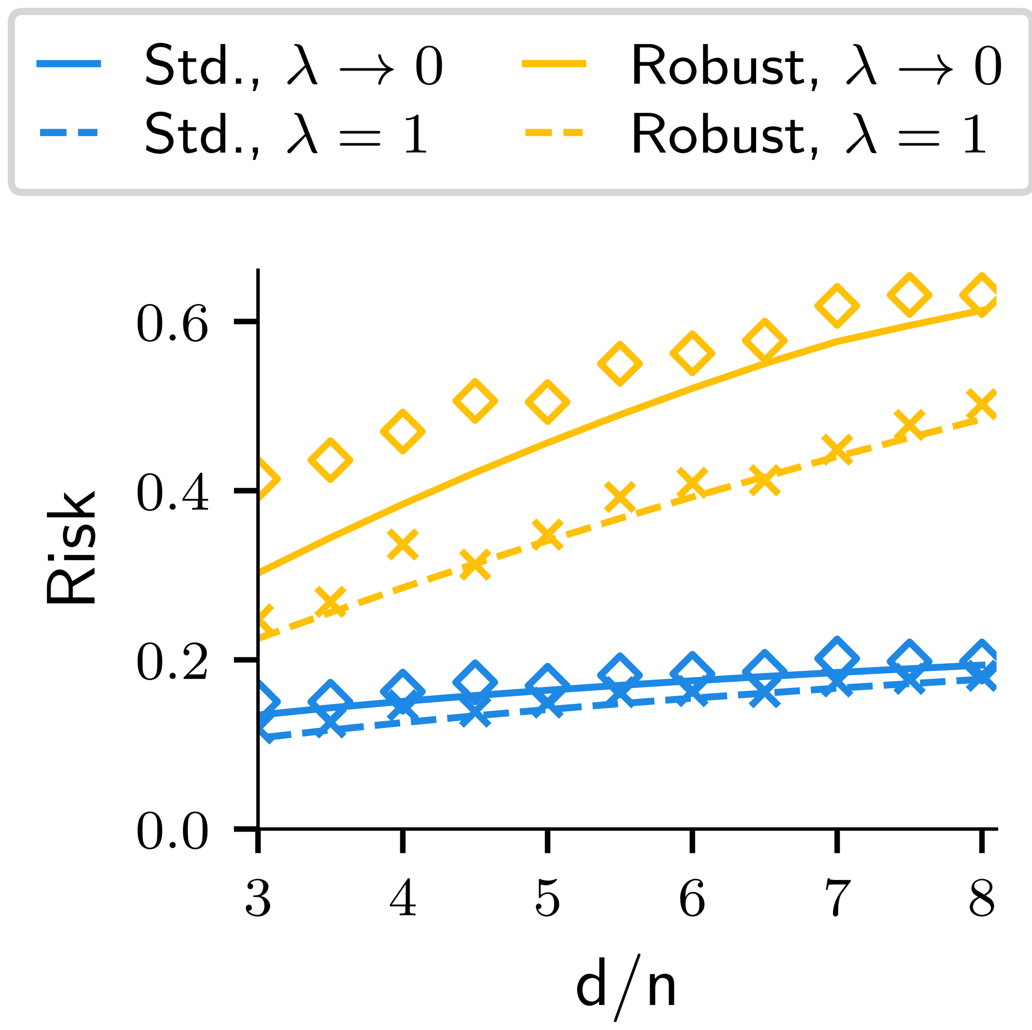
→ the lowest standard and robust risks are not obtained by the max-margin classifier, but by the regularized ones.

Theoretical result

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!**
- Sparse ground truth $\theta^\star = (1, 0, \ldots, 0)^\top$.
- We consider **linear classifiers** trained with the logistic loss.

Main result: We derive expressions for standard and robust risks in the asymptotic regime as $d, n \to \infty$ and $d/n \to \gamma$.



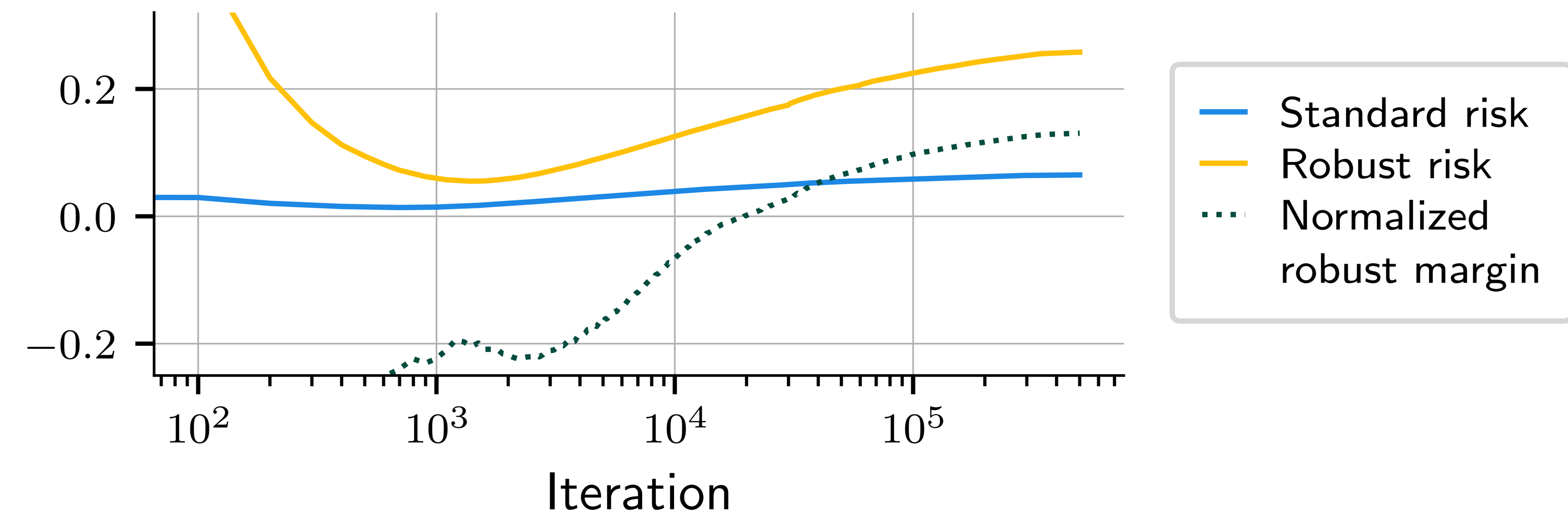
Lines: asymptotic risks (theory).
Markers: the risks for finite d, n (simulations).

- The proof uses the *Convex Gaussian Minimax Theorem* [2].
- regularization leads to estimators with smaller robust risks

  → even in high-dimensional settings (i.e. $d > n$), where overfitting is most unexpected.

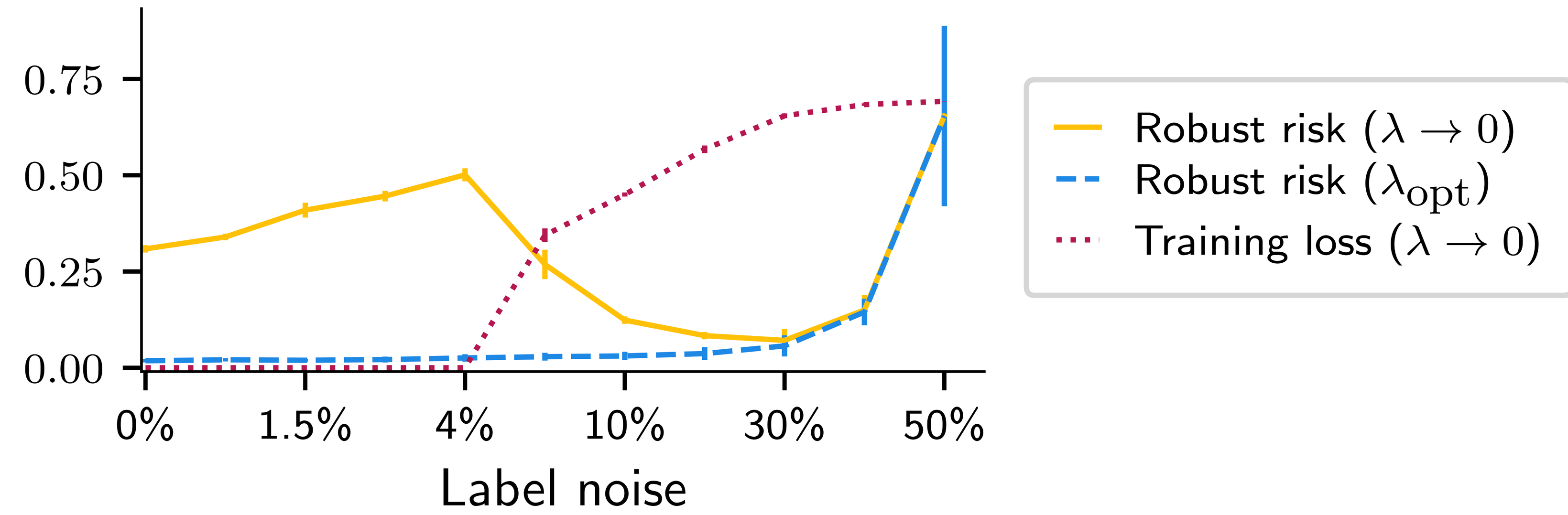
Other ways to avoid $\hat{\theta}_0$

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



2. Adding artificial label noise prevents a vanishing training loss → **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.

Conclusion

Regularization is crucial in order to achieve low robust risk.
→ even for **high-dimensional** and **noiseless** data

References

[1] L. Rice, E. Wong, and Z. Kolter, “Overfitting in adversarially robust deep learning,” in *ICML*, 2020, pp. 8093–8104.
[2] C. Thrampoulidis, S. Oymak, and B. Hassibi, “Regularized linear regression: A precise analysis of the estimation error,” in *COLT*, 2015, pp. 1683–1709.