Interpolation can hurt robust generalization even when there is no noise

Alexandru Țifrea

joint work with Konstantin Donhauser, Michael Aerni, Reinhard Heckel, Fanny Yang

Role of regularization: Classical narrative

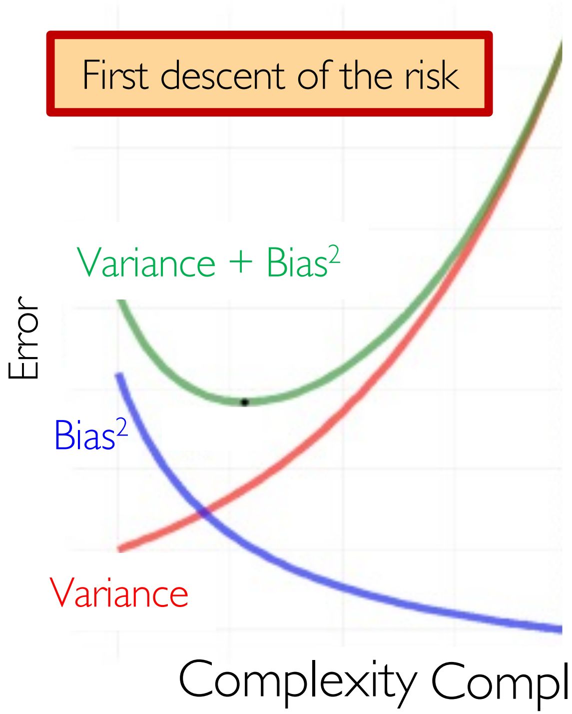
Classical regime (underparameterized)

- Regularization reduces variance ⇒ regularization leads to better generalization

Recent works (overparameterized)

- Variance of the interpolator found by GD vanishes ⇒ regularization is redundant

⇒ always just interpolate!   Not always ☹



Complexity

Second descent of the risk

Empirically: double descent for DNNs (Nakkiran et al)

Theoretically: double descent for linear, random feature models (Hastie et al; Mei et al etc) or kernel methods (e.g. Liang et al)

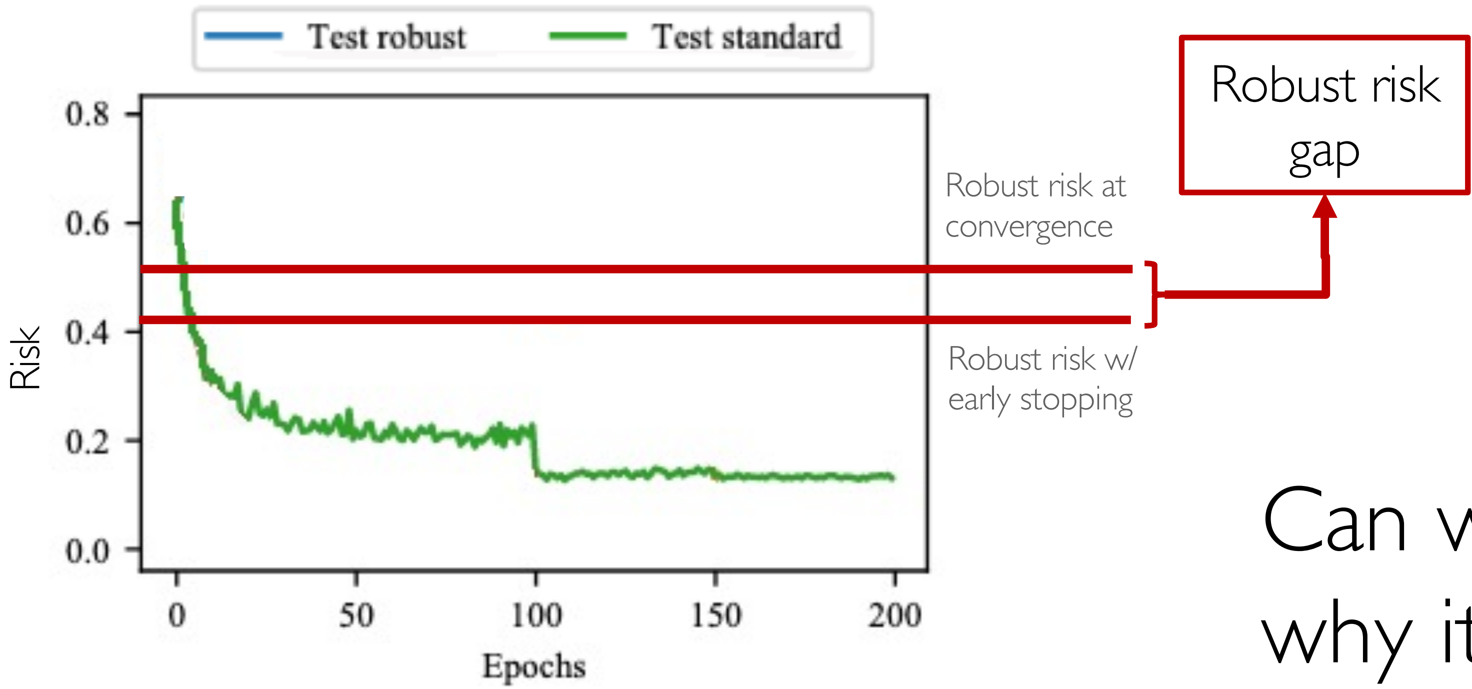
But all these works use the standard test risk for evaluation!

Empirically: regularization improves the adversarially robust test risk, even for overparameterized models (Rice et al)

Theoretically: ? ? ?

"Robust overfitting"

Context: Adversarial training ⇒ Low robust risk, i.e. $R_\epsilon(\theta) = \mathbb{E}_{x,y} \max_{\|\delta\|_p \leq \epsilon} \ell(y, f_\theta(x+\delta))$



Can we show why it happens?

Adversarial training w/ *early stopping* for *deep neural networks* on *image data* → low robust risk

Prior explanations for robust overfitting

1) Due to complexity of neural networks (Wu et al)

$\Rightarrow$ *robust overfitting does not occur for linear models*

2) Amplified by noise (Sanyal et al)

$\Rightarrow$ *robust overfitting does not occur for noiseless data*

No! Robust overfitting still occurs!

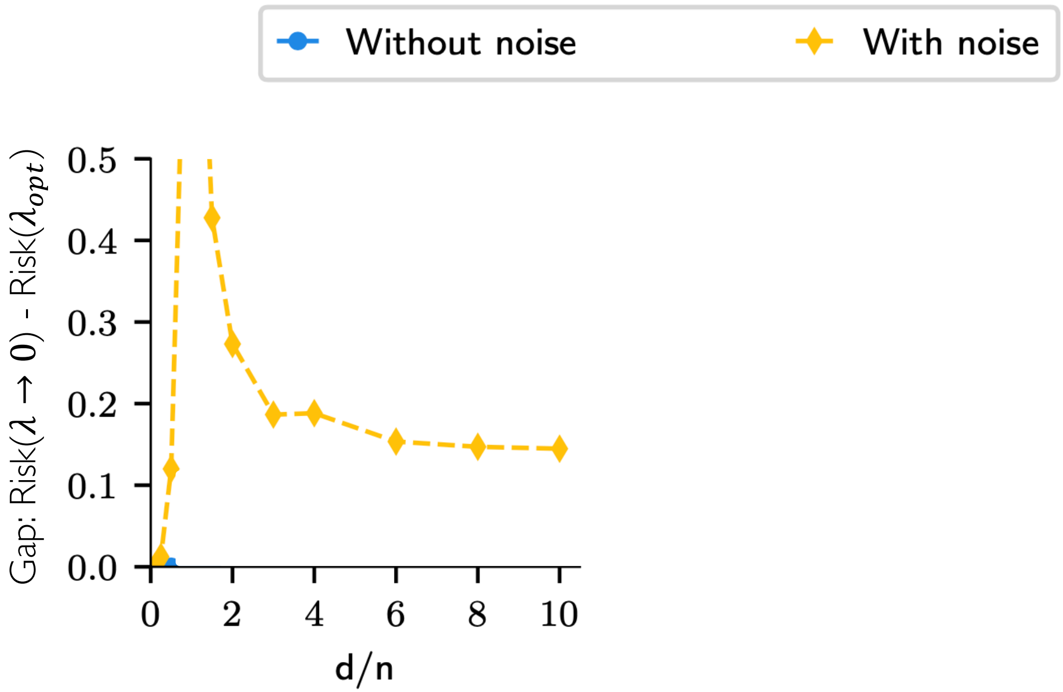
Robust overfitting for linear models and no noise

$$\hat{\theta}_\lambda = \mathrm{argmin}_\theta\, \mathcal{L}_\epsilon(\theta) + \lambda R(\theta)$$

Risk($\lambda \to 0$): Robust risk of interpolating GD solution

Risk(λ_{opt}): Robust risk of ridge estimator ($\lambda > 0$)

y-axis: Gap (i.e. positive gap = regularization helps robustness)



Linear regression

1) Robust overfitting for linear models? ✓

2) Robust overfitting for noiseless data? ✓

Can we prove that robust overfitting occurs?

Yes! For linear regression and classification with noiseless data.

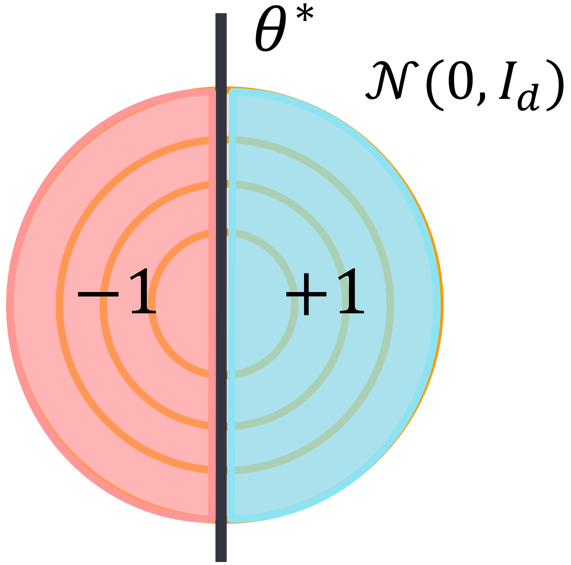
Data model for classification

High-dimensional data ($d > n$) ➡ interpolation is possible

- n i.i.d. covariates $x_i \sim \mathcal{N}(0, I_d)$
- deterministic labels (like e.g. Salehi et al, Sur et al)

$$y_i = \text{sgn}(\langle \theta^*, x_i \rangle) \in \{-1, +1\}$$

$\Rightarrow$ noiseless data



Max-margin interpolator

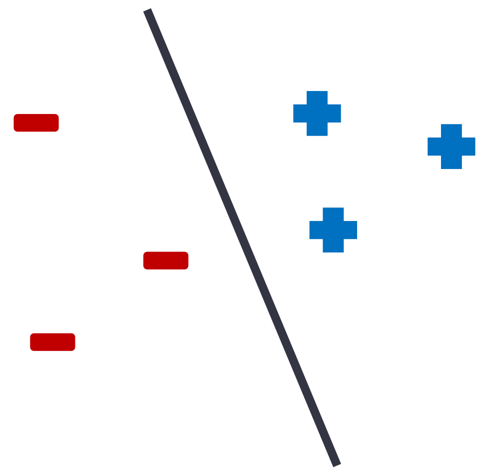
$\hat{\theta}_\lambda = \text{argmin}_\theta \sum_{i=1}^{n} \ell(y_i \langle x_i, \theta \rangle) + \lambda \|\theta\|_2^2$, with ℓ the logistic loss

Standard training (i.e. $\epsilon = 0$)

- unregularized predictor (i.e. $\lambda \to 0$) converges to max-margin estimator

$$\hat{\theta}_0 = \text{argmin}_\theta \|\theta\|_2 \text{ such that } y_i \langle x_i, \theta \rangle \geq 1$$

- the limit of GD on *standard* training loss (Soudry et al)



<u>Robust</u> max-margin interpolator

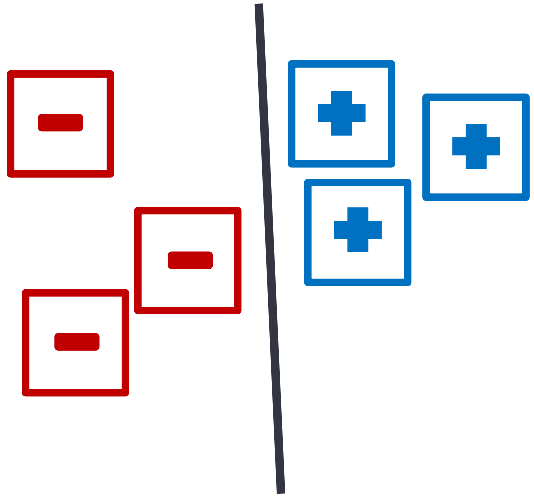
$$\hat{\theta}_\lambda = \text{argmin}_\theta \sum_{i=1}^{n} \max_{\|\delta\|_\infty \le \epsilon} \ell(y_i\langle x_i + \delta, \theta\rangle) + \lambda\|\theta\|_2^2$$ with ℓ the logistic loss

ℓ_∞-*adversarial* training (i.e. $\epsilon > 0$)

- unregularized predictor (i.e. $\lambda \to 0$) converges to *robust* max-margin estimator wrt ℓ_∞-perturbations

$$\hat{\theta}_0 = \text{argmin}_\theta \|\theta\|_2 \text{ such that } y_i\langle x_i, \theta\rangle - \epsilon\|\theta\|_1 \ge 1$$

- the limit of GD on *adversarial* training loss



Main result for linear classification

$$\hat{\theta}_\lambda = \text{argmin}_\theta \underbrace{\mathcal{L}_\epsilon(\theta)}_{\epsilon\text{-adv. loss}} + \lambda \|\theta\|_2$$

Theorem *DTAHY'21 (informal)* – better robustness with ridge regularization

For a sparse ground truth, we derive the limit of the robust risk as $d, n \to \infty$ and $d/n \to \gamma$:

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

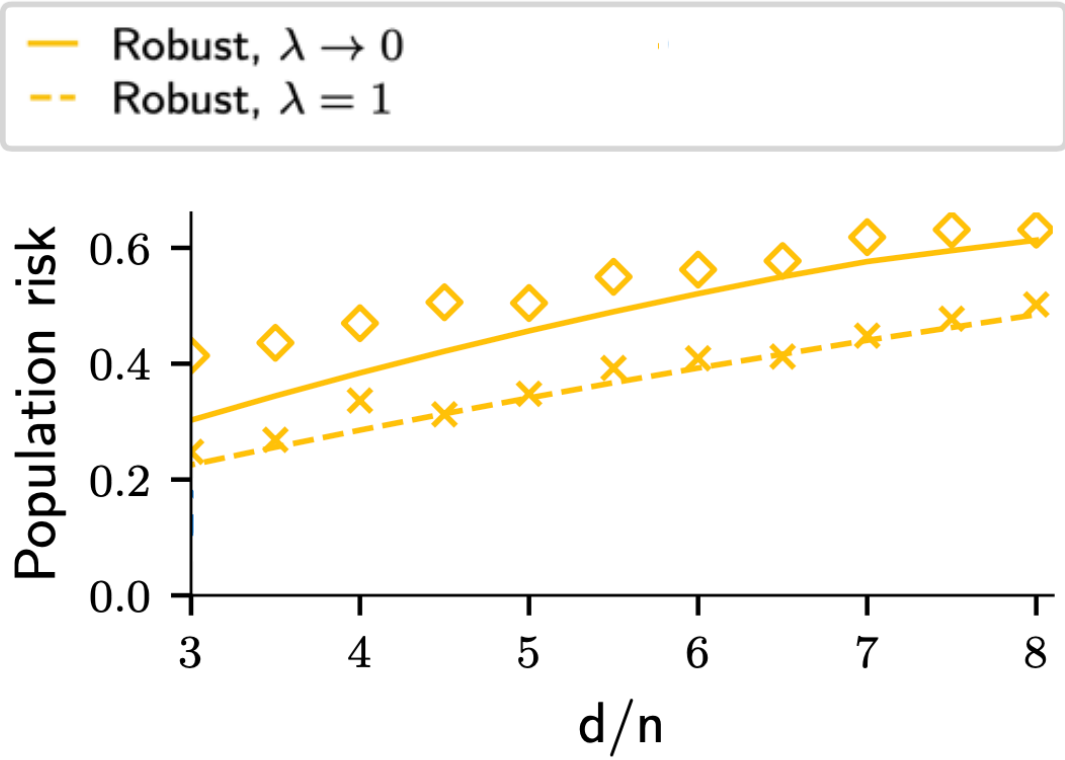
In particular, for some $\lambda_{opt} > 0$:

$$\underbrace{\mathcal{R}_{\lambda_{opt}}(\epsilon, \gamma)}_{\text{regularized}} < \underbrace{\lim_{\lambda \to 0} \mathcal{R}_\lambda(\epsilon, \gamma)}_{\text{interpolating}}$$

Proof: Uses the *Convex Gaussian Minimax Theorem* and Gaussian concentration.

scalar optimization problem → original optimization problem (i.e. minimize training loss)

Main result for linear classification



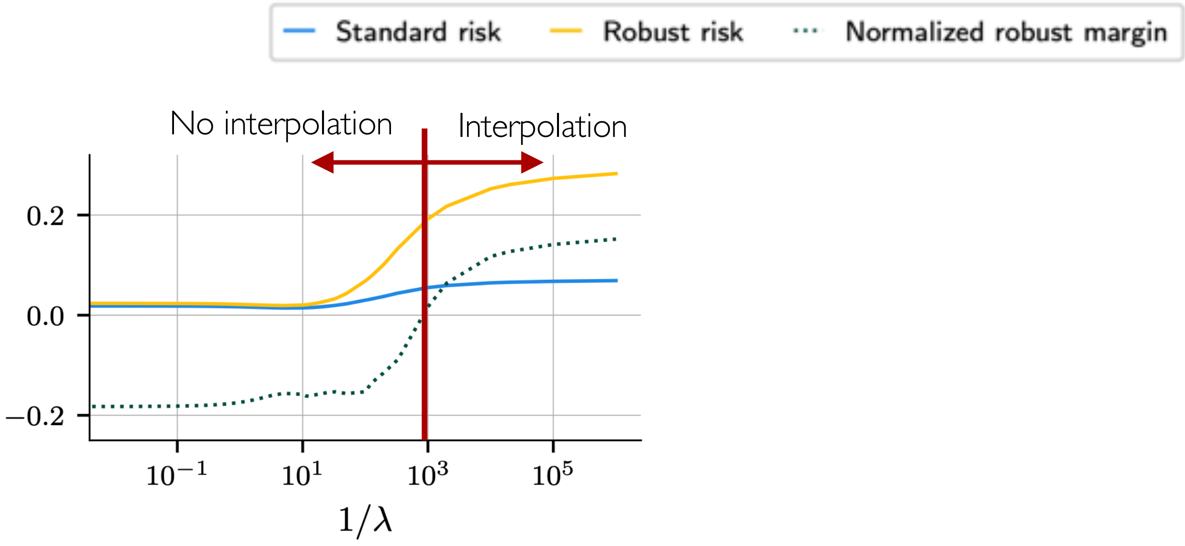
⇒ regularization reduces the robust risk even for d > n

⇒ trend persists also for finite d, n simulations

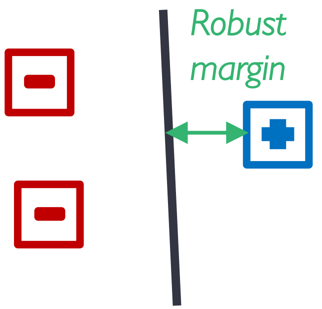
Lines: asymptotic risks (theory)

Markers: risks for finite d, n (simulations)

Preventing interpolation ⇒ lower robust risk



(a) Benefit of ridge regularization



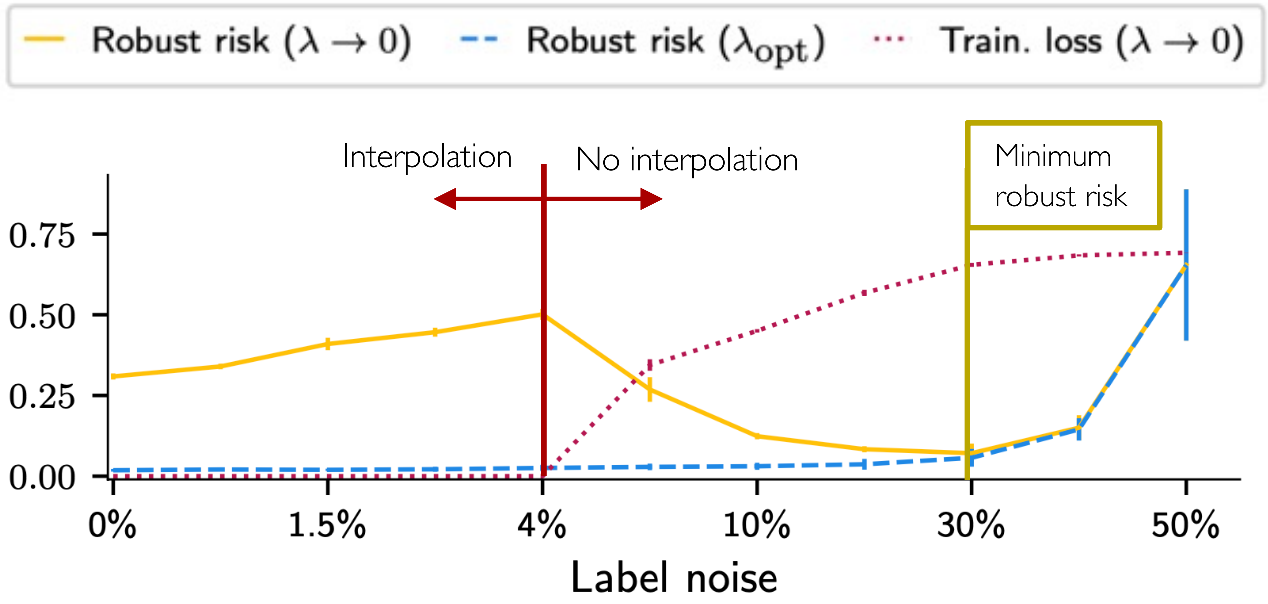
Regularize enough to prevent interpolation ⇒ lower robust risk

- negative robust margin ~ no interpolation ⇒ minimum robust risk
- What if we use other means to prevent interpolation?

An unorthodox way to prevent interpolation

Introduce a small amount of artificial label noise in the training data

$\longrightarrow$ avoids the robust max-margin estimator!



Remark: not advocating for label noise as a method to improve robustness

- regularization still leads to smaller robust risk

Conclusion & Future work

Summary: We show that avoiding the GD interpolating solution can be beneficial in the high-dimensional regime even for noiseless data and linear function classes.

- first formal proof of robust overfitting

Future work:

- extend proof to early stopping regularization for logistic regression
- extend our theoretical analysis to more complex model classes (e.g. random feature regression, shallow NNs etc)

Thank you!

References

- Hastie et al, Surprises in high-dimensional ridgeless least squares interpolation, 2020.
- Liang et al, Just interpolate: Kernel ''ridgeless" regression can generalize, Ann. Statist. 2020.
- Mei et al, The generalization error of random features regression: Precise asymptotics and double descent curve, CPAM 2021.
- Nakkiran et al, Deep double descent: Where bigger models and more data hurt, ICLR 2020.
- Rice et al, Overfitting in adversarially robust deep learning, ICML 2020.
- Salehi et al, The impact of regularization on high-dimensional logistic regression, NeurIPS 2019.
- Sanyal et al, How benign is benign overfitting?, ICLR 2021.
- Soudry et al, The implicit bias of gradient descent on separable data, JMLR 2018.
- Sur et al, A modern maximum-likelihood theory for high-dimensional logistic regression, PNAS, 2019.
- Wu et al, Do Wider Neural Networks Really Help Adversarial Robustness?, arXiv, 2020.