Interpolation can hurt robust generalization even when there is no noise
Konstantin Donhauser*,†, Alexandru Tifrea*‡, 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 CLASSIFICATION**
Evaluation with the **robust risk** wrt $\ell_\infty$-perturbations:
$$R_\epsilon(\theta) := \mathbb{E}_{x \sim P} \max_{\delta \in \mathbb{U}(\epsilon)} \|\text{sgn}(\langle \theta, x + \delta \rangle) - \text{sgn}(\langle \theta, x \rangle)\|_\infty.$$  
- We use adversarial training to obtain a robust estimator:
  $$\hat{\theta}_\lambda := \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{\delta \in \mathbb{U}(\epsilon)} \ell(\langle \theta, x_i + \delta \rangle y_i) + \lambda \|\theta\|_2^2.$$  
- For $\lambda \to 0$, it maximizes the robust margin of the data:
  $$\hat{\theta}_0 := \arg \min_{\theta} \|\theta\|_2 \text{ such that for all } i, \max_{\delta \in \mathbb{U}(\epsilon)} \ell(\langle \theta, x_i + \delta \rangle) \geq 1.$$  

**THEORETICAL RESULT FOR CLASSIFICATION**
**Problem setting:**
- Data model: covariates $x \sim N(0, I_d)$, deterministic labels given by $y = \text{sgn}(\theta^* x) \in \{-1, +1\}$ → **Noiseless data!**
- We consider **linear classifiers** trained with the logistic loss.

**Theorem.** For a sparse ground truth, we derive the limit $R_\lambda(\epsilon, \gamma)$ of the robust risk as $d, n \to \infty$ and $d/n \to \gamma$:
$$R_\epsilon(\hat{\theta}_\lambda) \xrightarrow{\text{prob}} R_\lambda(\epsilon, \gamma).$$
In particular, for some $\lambda_{\text{opt}} > 0$: $R_\lambda(\epsilon, \gamma) < \lim_{\gamma \to 0} R_\lambda(\epsilon, \gamma)$. **Remark:** Regularization still leads to smaller robust risk, even in the presence of noise.

**REFERENCES**