Surprising benefits of ridge regularization for noiseless regression

Konstantin Donhauser*,†, Alexandru Tifrea*†, 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
  → 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

PROBLEM SETTING
- We study the linear ridge regression estimator:
  \[
  \hat{\theta}_\lambda := \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} ((\theta, x_i) - y_i)^2 + \lambda \|\theta\|_2.
  \]
- If \(d/n > 1\), \(\lambda \rightarrow 0\) yields the minimum \(\ell_2\)-norm interpolator:
  \[
  \theta_0 := \arg\min_\theta \|\theta\|_2 \text{ such that for all } i, (\theta, x_i) = y_i.
  \]
- Evaluation with respect to the consistent robust risk with \(\ell_2\) perturbations:
  \[
  R_\epsilon(\theta) := \mathbb{E}_{X \sim P} \max_{|\|\theta\|_2 \leq \epsilon} ((\theta - \theta^*, X + \delta)^2)
  \]

THEORETICAL RESULT
High-dimensional data model:
- \(n\) i.i.d. covariates \(x_i \sim \mathcal{N}(0, I_d)\).
- observations \(y_i = (\theta^*, x_i) + \xi_i\) with noise \(\xi_i \sim \mathcal{N}(0, \sigma^2 I_d)\).
- \(d, n \rightarrow \infty, d/n \rightarrow \gamma\).

Theorem. Define \(m(\gamma) = \frac{1-\gamma^2+(1-\gamma^2)\sqrt{(1-\gamma^2)^2-4\gamma^2}}{2\gamma^2}\) and let \(m'\) be its derivative. Let \(P = B + \mathbb{V} - \lambda^2(m(-\lambda))\) and \(B = \lambda^2 m'(-\lambda)\), \(\mathbb{V} = \sigma^2\gamma(m(-\lambda) - \lambda m'(-\lambda))\) be the asymptotic bias and variance. Then,
\[
R_\epsilon(\hat{\theta}_\lambda) \overset{a.s.}{\rightarrow} B + \mathbb{V} + \epsilon^2 P + \sqrt{8\sigma^4 \pi}\mathbb{V}(P + \mathbb{V})
\]
Furthermore, the standard risk \(R(\hat{\theta}_\lambda) \rightarrow B + \mathbb{V}\) a.s.

→ We can compute the asymptotic standard and robust risks.

THEORETICAL PREDICTIONS
Theoretical predictions (lines) for \(d, n \rightarrow \infty\) and experimental results (markers) for finite \(d, n\).
- 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: \(\|\theta^* - \Pi_{\theta} \|_2^2\).
- Orthogonal misfit: \(\| (I - \Pi_{\theta}) \theta^* \|_2^2\).

\[
\lambda \uparrow; \quad \gamma = 2.0 \quad \gamma = 2.8 \quad \gamma = 4.5 \quad \lambda \rightarrow 0 \quad \lambda_{opt}
\]

→ Robust risk punishes orthogonal misfit stronger than standard risk, leading to \(\lambda_{opt} > 0\).

REFERENCES

Does linear regression suffer from robust overfitting?
Yes, even on noiseless training data!

--

PHENOMENON SETTING

Problem setting:

- We study the linear ridge regression estimator:
  \[
  \hat{\theta}_\lambda := \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} ((\theta, x_i) - y_i)^2 + \lambda \|\theta\|_2.
  \]
- If \(d/n > 1\), \(\lambda \rightarrow 0\) yields the minimum \(\ell_2\)-norm interpolator:
  \[
  \theta_0 := \arg\min_\theta \|\theta\|_2 \text{ such that for all } i, (\theta, x_i) = y_i.
  \]
- Evaluation with respect to the consistent robust risk with \(\ell_2\) perturbations:
  \[
  R_\epsilon(\theta) := \mathbb{E}_{X \sim P} \max_{|\|\theta\|_2 \leq \epsilon} ((\theta - \theta^*, X + \delta)^2)
  \]

Theoretical result:
High-dimensional data model:
- \(n\) i.i.d. covariates \(x_i \sim \mathcal{N}(0, I_d)\).
- observations \(y_i = (\theta^*, x_i) + \xi_i\) with noise \(\xi_i \sim \mathcal{N}(0, \sigma^2 I_d)\).
- \(d, n \rightarrow \infty, d/n \rightarrow \gamma\).

Theorem. Define \(m(\gamma) = \frac{1-\gamma^2+(1-\gamma^2)\sqrt{(1-\gamma^2)^2-4\gamma^2}}{2\gamma^2}\) and let \(m'\) be its derivative. Let \(P = B + \mathbb{V} - \lambda^2(m(-\lambda))\) and \(B = \lambda^2 m'(-\lambda)\), \(\mathbb{V} = \sigma^2\gamma(m(-\lambda) - \lambda m'(-\lambda))\) be the asymptotic bias and variance. Then,
\[
R_\epsilon(\hat{\theta}_\lambda) \overset{a.s.}{\rightarrow} B + \mathbb{V} + \epsilon^2 P + \sqrt{8\sigma^4 \pi}\mathbb{V}(P + \mathbb{V})
\]
Furthermore, the standard risk \(R(\hat{\theta}_\lambda) \rightarrow B + \mathbb{V}\) a.s.

→ We can compute the asymptotic standard and robust risks.