Theory of Generalization Reading Group

Reconciling Modern Machine Learning Practice and the Bias–Variance Tradeoff

Belkin et al. (2019) arXiv:1812.11118

Outline

1. Motivation: The Interpolation Puzzle
2. Interpolation and Capacity
3. The Double-Descent Risk Curve
4. Evidence Across Model Families
5. Why Might Overparameterization Help?
6. What Changed?

Motivation: The Interpolation Puzzle

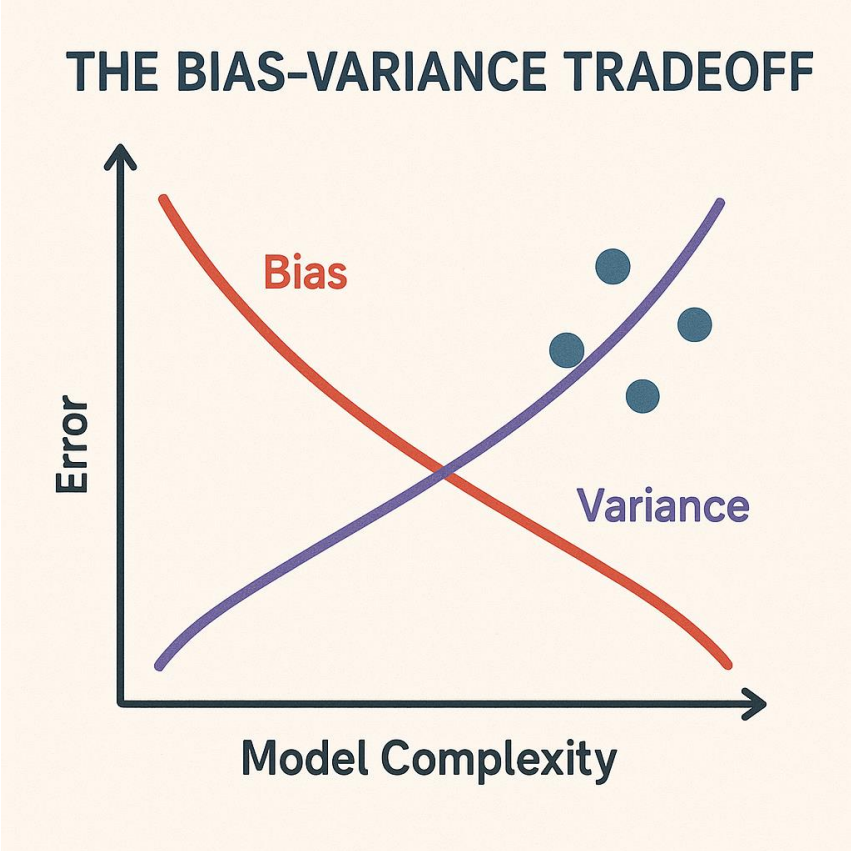
$$\mathbb{E}_{D,\varepsilon}\left[(y-\hat{f}_D(x))^2\right] = \underbrace{\left(\mathbb{E}_D[\hat{f}_D(x)]-f(x)\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}_D\left[\left(\hat{f}_D(x)-\mathbb{E}_D[\hat{f}_D(x)]\right)^2\right]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Noise}}$$

Increasing model complexity should:

- **reduce bias**
- **increase variance**

Test error should follow a U-shaped curve.

Yet modern machine learning practice appears to contradict this prediction. Can interpolation and good generalization coexist?



Capacity and the Interpolation Threshold

Capacity

A parameter C indexing a nested family of hypothesis classes

$$\mathcal{H}_{C_1} \subseteq \mathcal{H}_{C_2} \subseteq \cdots \qquad (C_1 < C_2),$$

so that increasing C increases the expressive power of the model.

Importantly: C is an *operational* notion of capacity used to traverse a model family, not necessarily t he quantity governing generalization.

Interpolation

Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a training set

and $\mathcal{H}_C$ a hypothesis class of capacity C

The Empirical Risk:

$$R_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i), \qquad h \in \mathcal{H}_C.$$

We call an interpolation

$$R_n(h) = 0$$

or $\qquad h(x_i) = y_i, \qquad i = 1, \ldots, n.$

We call the **interpolation threshold**

$$C^\star = \inf \{C : \exists h \in \mathcal{H}_C \text{ such that } R_n(h) = 0\}.$$

The smallest capacity for which the training data become perfectly fit.

Experimental Setup: Random Fourier Features (RFF)

Our **model family $\mathcal{H}_N$** is: $h(x) = \sum_{k=1}^{N} a_k \phi(x; v_k), \qquad \phi(x; v) = e^{i\langle v, x\rangle}$ Where: $v_1, \ldots, v_N \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d).$

- Randomly sampled first layer
- Trainable coefficients $a_1, a_2 \ldots, a_n$
- Similar to a 2-layer neural network with fixed first-layer weights

Learning Procedure

Given $(x_1, y_1), \ldots, (x_n, y_n),$

We solve the empirical risk minimization

$$\min_{h \in \mathcal{H}_N} \frac{1}{n} \sum_{i=1}^{n} \big(h(x_i) - y_i\big)^2.$$

In the overparametrized regime ($N > n$):

Infinitely many interpolating solutions satisfy

$$h(x_i) = y_i.$$

We choose the interpolator with: $\boxed{\min \|a\|_2}$

Among all solutions

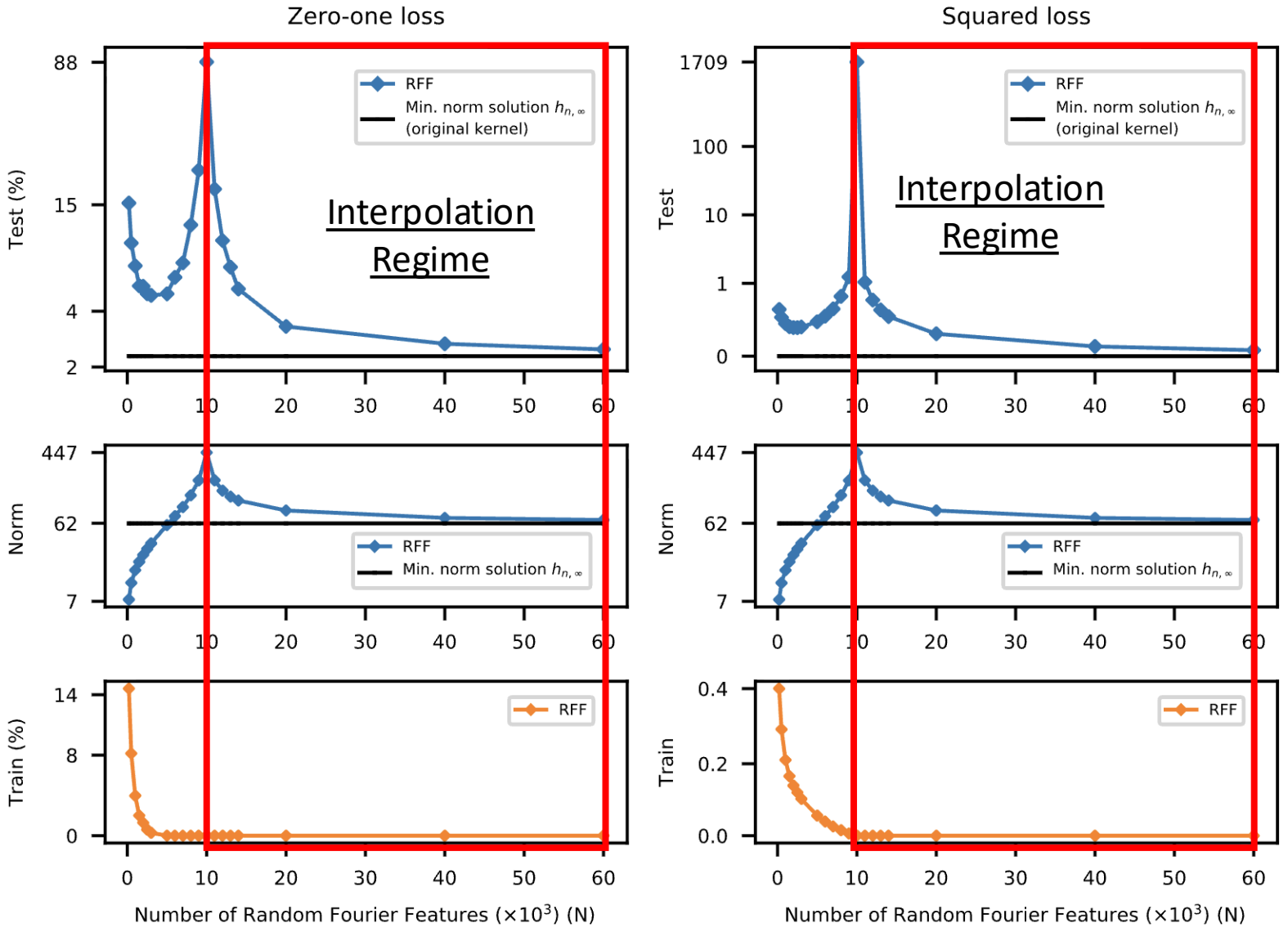
Capacity for this class of models is defined as: $\boxed{C = N}$

The Double Descent Risk for RFF models

On **MNIST**:

- Test risks (log scale),
- coefficient ℓ_2 norms (log scale),
- Training risks of the RFF model predictors $h_{n,N}$ learned on a subset of MNIST (n = 10^4, 10 classes).
- The interpolation threshold is achieved at N = 10^4.

The classical U-curve becomes a double descent curve



Why Does the Interpolation Peak Occur?

Linear - random-feature models case

Assume a linear model in feature space: $y = X\beta^\star + \varepsilon, \qquad X \in \mathbb{R}^{n\times p}.$

The minimum-norm interpolator is $\hat{\beta} = X^\dagger y.$ where $\boxed{X^\dagger = X^\top (XX^\top)^{-1}}$

Using the SVD, $X = U\Sigma V^\top,\ X^\dagger = V\Sigma^{-1}U^\top.$

Substitute the noisy labels: $\hat{\beta} = X^\dagger X\beta^\star + X^\dagger\varepsilon.$

The second term is the problem: $X^\dagger\varepsilon = V\Sigma^{-1}U^\top\varepsilon.$

Near interpolation, $p \approx n \implies \sigma_{\min}(X) \approx 0.$ (Marchenko–Pastur law)

Hence $\dfrac{1}{\sigma_{\min}(X)} \gg 1.$ So small noise components are strongly amplified.

Near interpolation ⟹ noise amplification ⟹ test error peak

Hastie, Trevor, et al. "Surprises in high-dimensional ridgeless least squares interpolation." Annals of statistics 50.2 (2022): 949.

Why Does Error Decrease Again?

Beyond interpolation $(N > n)$

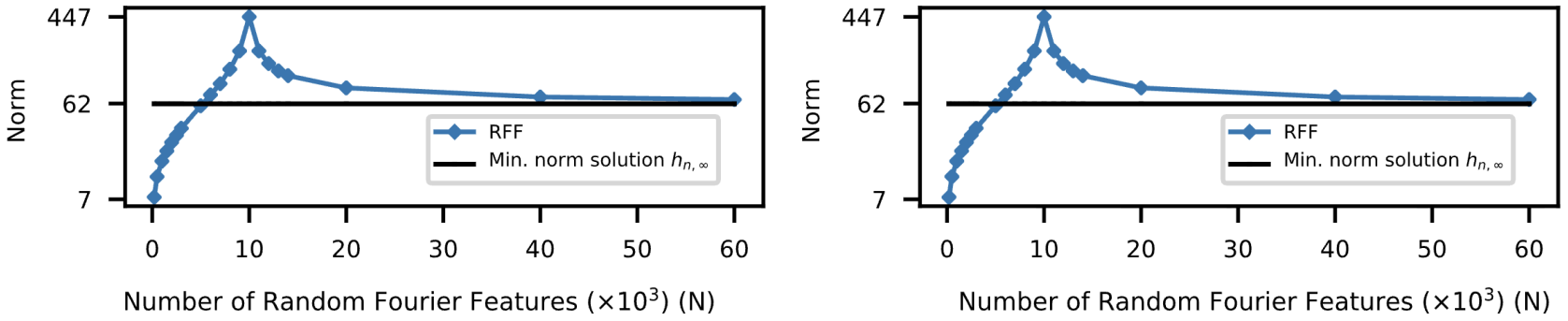
The system: $X_\Phi a = y$

has infinitely many interpolating solutions. $\mathcal{S}_N = \{a : X_\Phi a = y\}.$

Minimum-norm interpolation

Belkin selects $a^\star = \arg\min_{X_\Phi a = y} \|a\|_2.$

As N increases, the hypotheses class becomes larger and the learner can choose interpolators with smaller norm



And this is due to the definition of the optimal RFF solution before (and to the simplicity bias of SGD later on)

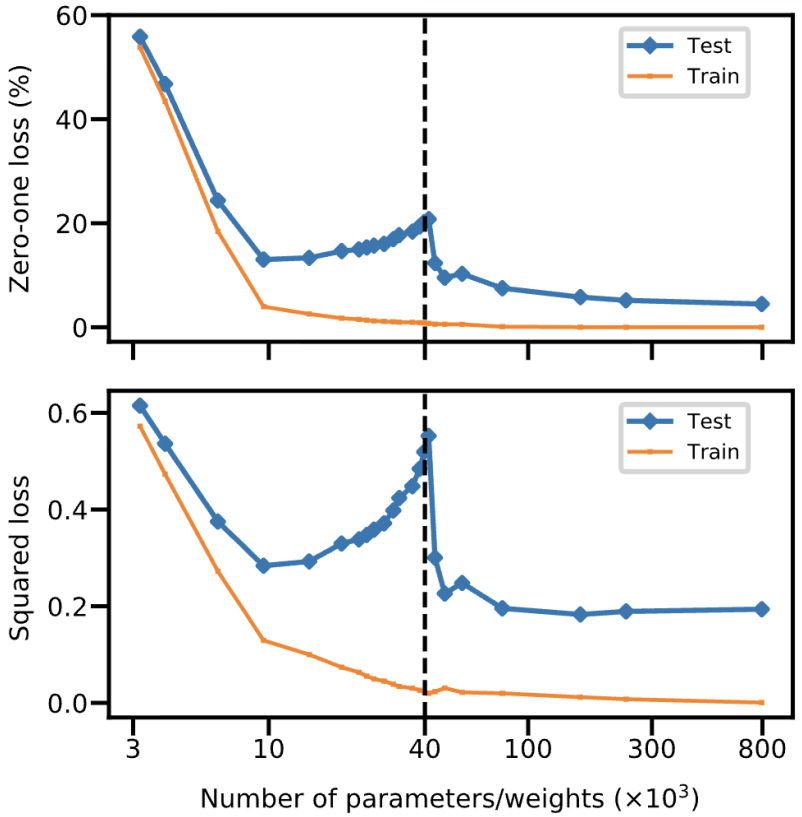
Double Descent Persist Beyond Random Features

Neural Networks

One-hidden-layer fully connected networks.

$$f(x) = \sum_{k=1}^{h} a_k \, \sigma(w_k^\top x + b_k),$$

Capacity: number of hidden units

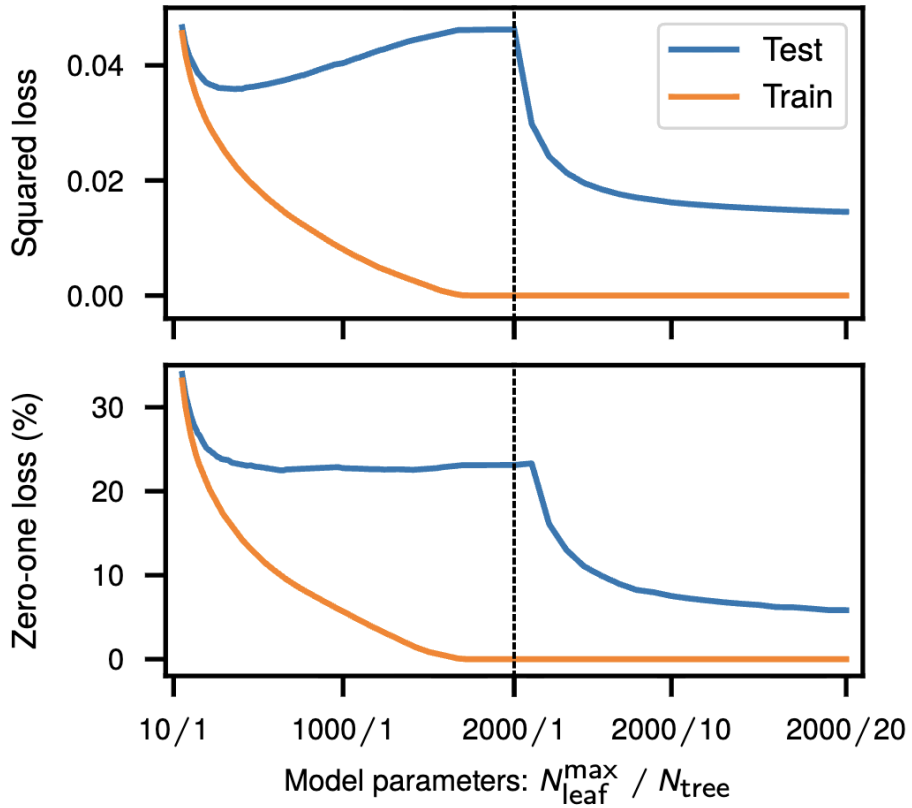


Tree-Based Methods

- Decision trees
- Random forests
- Boosting

Capacity:

- leaves
- ensemble size



Reconciling Modern Practice and the Bias–Variance Trade-off

Classical picture

Increasing capacity $\Longrightarrow$ decreasing bias + increasing variance

Best generalization occurs before interpolation

Belkin et al. (2019)

Increasing capacity beyond interpolation reveals a second regime:

Interpolation does not imply overfitting

Overparametrization can improve generalization

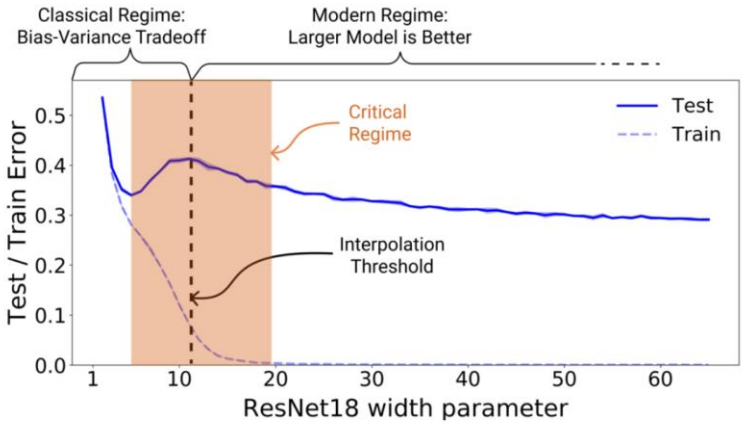
Belkin's experiments suggest that the complexity controlling generalization is not simply the number of parameters. The norm of the selected interpolating solution appears to be a more relevant quantity.

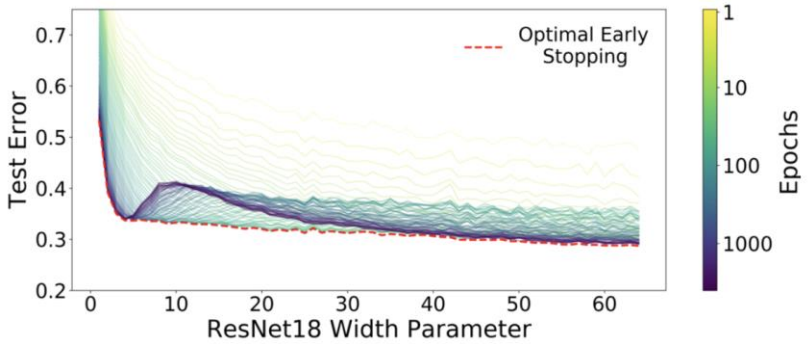
Other Work

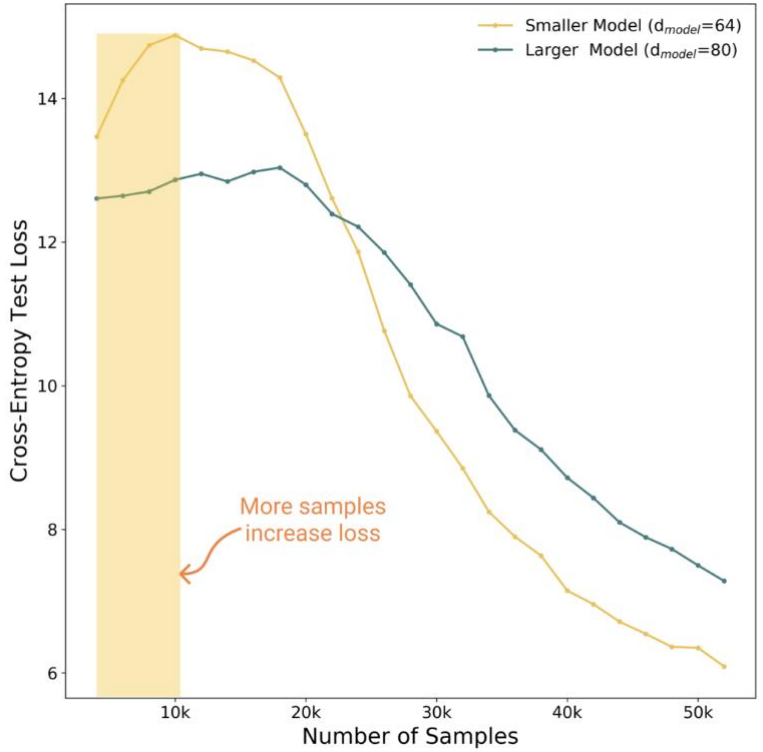
Deep Double Descent: Where Bigger Models and More Data Hurt

https://arxiv.org/abs/1912.02292

Model-wise, sample-wise, epoch-wise double descent...







Open question: Why do increasing model size and training time often seem to have qualitatively similar effects on generalization?