Theoretical Analysis of Double Descent

Jikai Jin

Stanford ICME

February 12, 2026

This Presentation

- **The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve. Song Mei & Andrea Montanari** (arXiv:1908.05355v5)
- Goal: Explains **double descent** in a random feature ridge regression model.

Mathematical Setting: Random Features Ridge Regression

Data: $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$.
Teacher:

$$y_i = f_d(x_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i] = 0,\ \mathbb{E}[\varepsilon_i^2] = \tau^2.$$

Random features: draw $\theta_1, \dots, \theta_N \overset{\text{iid}}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and set

$$Z_{ia} = \frac{1}{\sqrt{d}}\,\sigma\left(\frac{\langle x_i, \theta_a\rangle}{\sqrt{d}}\right), \qquad i \le n,\ a \le N.$$

Ridge regression (second layer):

$$\hat{a}_\lambda \in \arg\min_{a \in \mathbb{R}^N} \frac{1}{n}\|y - Za\|_2^2 + \lambda\|a\|_2^2, \qquad \hat{f}_\lambda(x) = \sum_{a=1}^{N} \hat{a}_{\lambda,a}\,\sigma\left(\frac{\langle x, \theta_a\rangle}{\sqrt{d}}\right).$$

What Makes This “Non-Standard”?

Proportional asymptotics:

$$\psi_1 \equiv N/d \to \text{const}, \qquad \psi_2 \equiv n/d \to \text{const}.$$

- **Overparameterization:** N is tunable.
- **Interpolation:** $N \gtrsim n$ allows (near) zero training error even with noise.
- **Nonlinear feature map:** the feature map is nonlinear in x.
- **Nonlinear ground-truth model:**

$$f_d(x) = \beta_{d,0} + \langle \beta_{d,1}, x \rangle + f_d^{\mathrm{NL}}(x),$$

with f_d^{NL} an isotropic Gaussian process on the d-dimensional sphere. Signal strength: $\|\beta_{d,1}\|_2^2 \to F_1^2$. Nonlinear power: $\mathbb{E}_x[f_d^{\mathrm{NL}}(x)^2] \to F_*^2$.

Empirical Observations

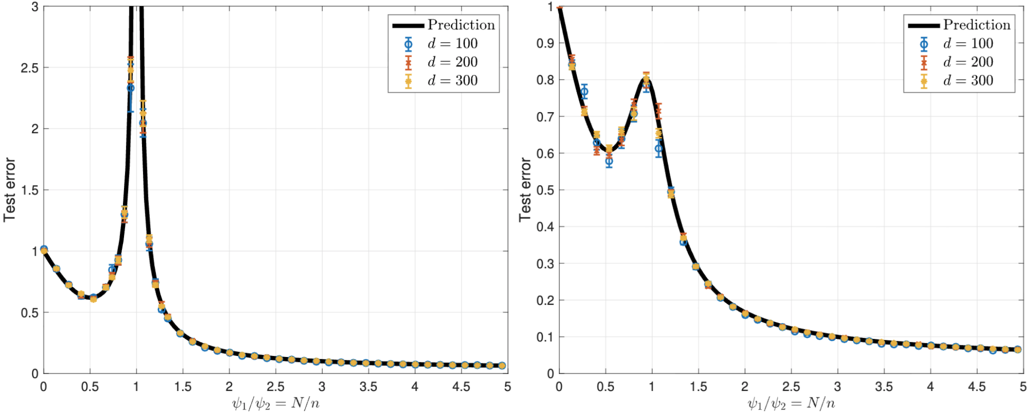

Figure 1: Random features ridge regression with ReLU activation ($\sigma = \max\{x, 0\}$). Data are generated via $y_i = \langle \boldsymbol{\beta}_1, \boldsymbol{x}_i \rangle$ (zero noise) with $\|\boldsymbol{\beta}_1\|_2^2 = 1$, and $\psi_2 = n/d = 3$. Left frame: regularization $\lambda = 10^{-8}$ (we didn't set $\lambda = 0$ exactly for numerical stability). Right frame: $\lambda = 10^{-3}$. The continuous black line is our theoretical prediction, and the colored symbols are numerical results for several dimensions d. Symbols are averages over 20 instances and the error bars report the standard error of the means over these 20 instances.

Main result: Decomposition of Test Error

$\mathcal{E}_{\text{test}}(\lambda) \to$ (squared) bias on *linear* signal $+$ variance / noise amplification $+$ irreducible nonlinear error.

Ridge on random features. Via smoothing matrix $P_\lambda := Z(Z^\top Z + \lambda I)^{-1} Z^\top$.

What is actually learned? Write

$$y = s + \eta, \qquad s_i := \langle \beta_{d,1}, x_i \rangle \ \text{(learnable linear signal)}, \qquad \eta_i := y_i - s_i.$$

Two consequences of $\hat{y} = P_\lambda y$.

- **Signal shrinkage (bias):** $s - \hat{y}_{\text{signal}} = (I - P_\lambda)s \implies \frac{1}{n}\|(I - P_\lambda)s\|^2$. Assuming that $\|\beta_{d,1}\|_2^2 \to F_1^2$, $\frac{1}{n}\|(I - P_\lambda)s\|^2 \to F_1^2\, B(\cdot)$.
- **Noise amplification (variance):** if $\mathbb{E}[\eta\eta^\top \mid Z] \approx (\tau^2 + F_*^2) I_n$, then $\mathbb{E}\left[\frac{1}{n}\|P_\lambda \eta\|^2 \,\middle|\, Z\right] = \frac{\tau^2 + F_*^2}{n} \operatorname{Tr}(P_\lambda^2) \to (\tau^2 + F_*^2)\, V(\cdot)$.

What Drives Bias B?

$$\mathcal{E}_{\text{test}}(\lambda) \approx F_1^2\, B(\zeta, \psi_1, \psi_2, \bar{\lambda}) + (\tau^2 + F_*^2)\, V(\zeta, \psi_1, \psi_2, \bar{\lambda}) + F_*^2.$$

where $\bar{\lambda} := \lambda/\mu_*^2$.

Problem size:

$$\psi_1 := \frac{N}{d}, \quad \psi_2 := \frac{n}{d}, \quad \gamma := \frac{\psi_1}{\psi_2} = \frac{N}{n}.$$

- **Underparameterized ($\gamma < 1$):** limited feature dimension $\Rightarrow$ more signal is missed $\Rightarrow$ B tends to be large.
- **Increase ψ_1 at fixed ψ_2 (wider model):** typically improves signal transfer $\Rightarrow B \downarrow$ (away from $\gamma \approx 1$).
- **Increase ψ_2 at fixed ψ_1 (more data):** moves away from interpolation ($\gamma \downarrow$) and stabilizes estimation $\Rightarrow$ B improves smoothly.

Feature quality & regularization:

$$\zeta := \frac{\mu_1}{\mu_*} \quad \text{(signal/residual; shape).}$$

Activation moments ($G \sim \mathcal{N}(0,1)$): $\mu_1 = \mathbb{E}[G\sigma(G)]$, $\mu_*^2 = \mathbb{E}[\sigma(G)^2] - \mathbb{E}[\sigma(G)]^2 - \mu_1^2$.

- **Feature informativeness ($\zeta \uparrow$):** more predictable/linear component per residual $\Rightarrow$ less shrinkage on the learnable signal $\Rightarrow B \downarrow$.
- **Regularization ($\bar{\lambda} \uparrow$):** stronger ridge filtering $\Rightarrow$ more shrinkage $\Rightarrow B \uparrow$.

Takeaway: bias decreases with width/data ($\psi_1 \uparrow$, $\psi_2 \uparrow$) and feature informativeness ($\zeta \uparrow$), but increases with ridge strength ($\bar{\lambda} \uparrow$).

What Drives Variance V?

$$\mathcal{E}_{\text{test}}(\lambda) \approx F_1^2\, B(\zeta, \psi_1, \psi_2, \bar{\lambda}) + (\tau^2 + F_*^2)\, V(\zeta, \psi_1, \psi_2, \bar{\lambda}) + F_*^2.$$

where $\bar{\lambda} := \lambda/\mu_*^2$.

Problem size:

$$\psi_1 := \frac{N}{d}, \quad \psi_2 := \frac{n}{d}, \quad \gamma := \frac{\psi_1}{\psi_2} = \frac{N}{n}.$$

Feature quality & regularization:

$$\zeta := \frac{\mu_1}{\mu_*} \quad \text{(signal/residual; shape)}.$$

- **Interpolation boundary:** V peaks near $\gamma \approx 1$ (Gram ill-conditioning), especially when $\bar{\lambda}$ is small.
- **Move away from $\gamma \approx 1$:** either $\gamma \ll 1$ or $\gamma \gg 1$ $\Rightarrow$ better-conditioned spectrum $\Rightarrow$ V is moderate.
- **Far overparameterized ($\gamma \gg 1$):** after the spike, amplification typically decreases.

- **Regularization ($\bar{\lambda} \uparrow$):** suppresses small-eigenvalue amplification $\Rightarrow V \downarrow$ (kills the spike).
- **Residual nonlinearity ($\mu_* \uparrow$):** at fixed λ, $\bar{\lambda} = \lambda/\mu_*^2 \downarrow \Rightarrow$ effectively weaker ridge $\Rightarrow V \uparrow$ near $\gamma \approx 1$.
- **Feature informativeness ($\zeta \uparrow$):** less noise-dominated features $\Rightarrow$ typically smaller amplification $\Rightarrow V \downarrow$.

Takeaway: variance spikes near interpolation ($\gamma \approx 1$) unless ridge is strong; $\bar{\lambda}$ largely controls the spike.

Summary

- **Model:** random features ridge regression is a simple 2-layer NN where only the last layer is trained.
- **Phenomenon:** test error can *increase* near $N \approx n$ (interpolation), but *decrease again* for $N \gg n \quad \Rightarrow$ **double descent**.
- **Mechanism:** near $N \approx n$, the feature Gram matrix is ill-conditioned, so noise gets amplified (large variance).
- **Decomposition to remember:**

  $$\underbrace{F_1^2\, B}_{\text{bias}} + \underbrace{(\tau^2 + F_*^2)\, V}_{\text{variance}} + \underbrace{F_*^2}_{\text{irreducible}} .$$

  Regularization (or early stopping) mainly reduces V; width/data mainly reduce B.
- **Practical takeaway:** if possible, **use a wider model and tune regularization**—it can beat the "just-right" size near $N \approx n$.