The generalization error of random features model: Precise asymptotics and double descent curve

Song Mei

Stanford University

September 8, 2019

Joint work with Andrea Montanari

Surprises of generalization behavior of neural networks

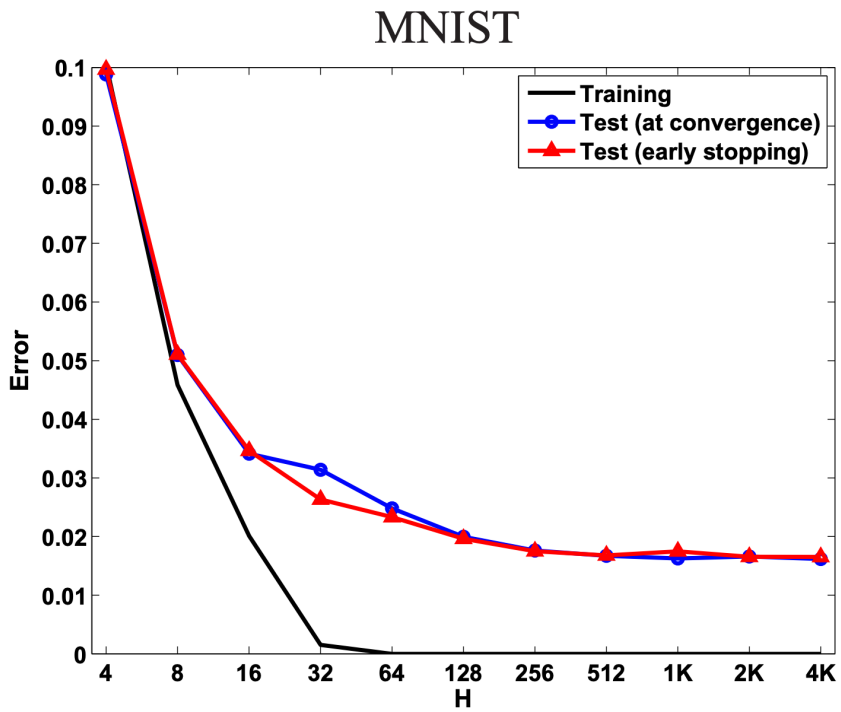
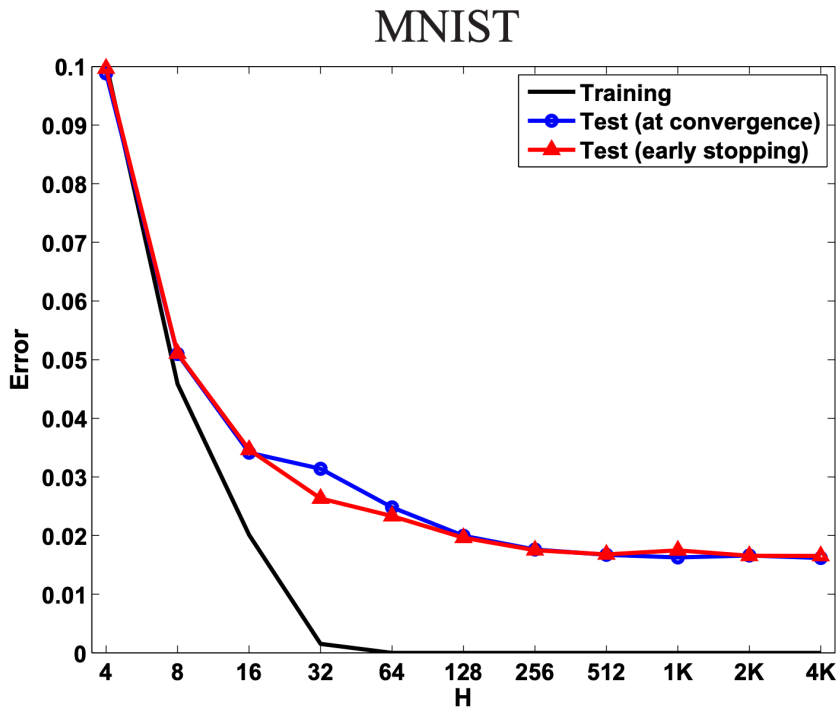


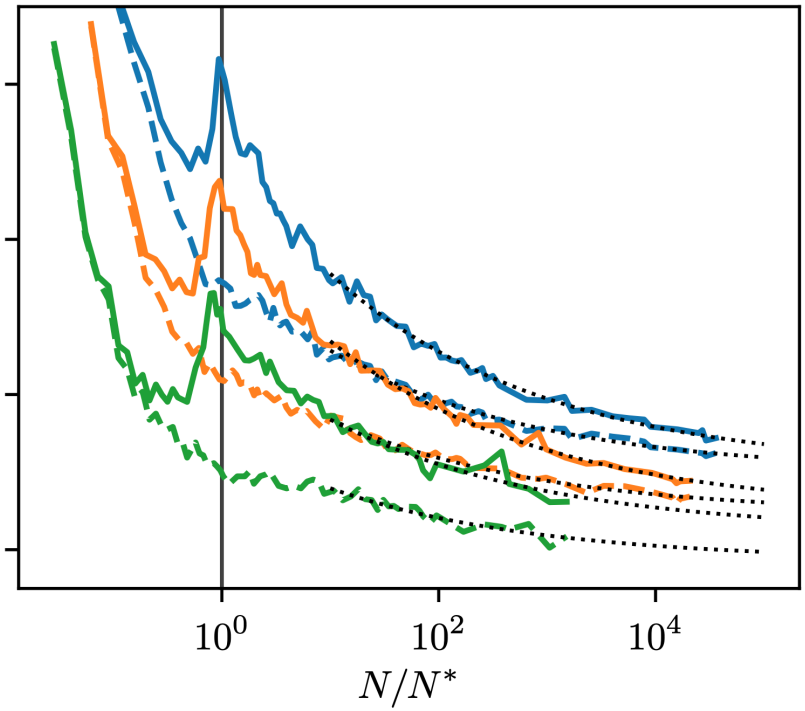
Figure: Experiments on MNIST by [Neyshabur, Tomioka, Srebro, 2014a]

Surprise: why does't higher model complexity ...
... induce larger generalization error?



Partial explanations: The intrinsic model complexity is not the number of parameters, but "some norm" of the weights. This intrinsic model complexity is implicitly controlled by SGD. [Neyshabur, Tomioka, Srebro, 2014b], [Gunasekar, Woodworth, Bhojanapalli, Neyshabur, Srebro, 2017], ...

Train more carefully to better interpolates the data



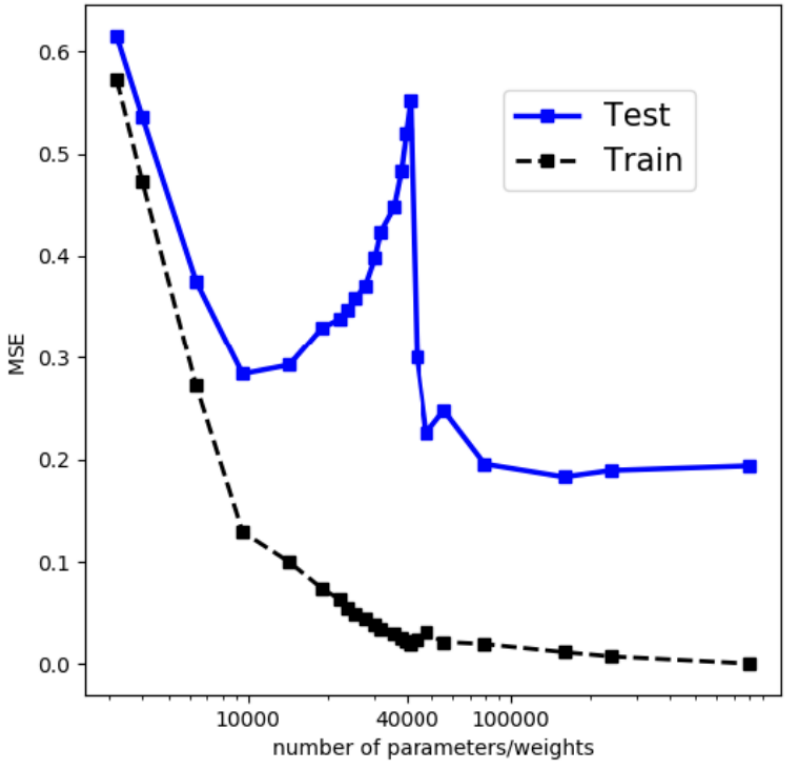
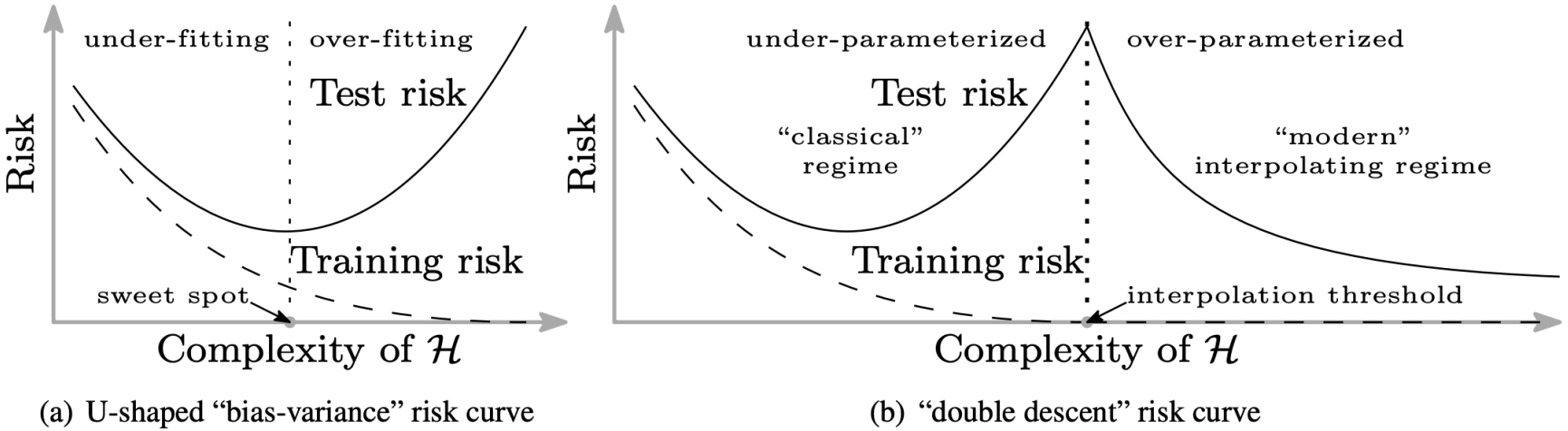


Figure: Experiments on MNIST. Left: [Spigler, Geiger, Ascoli, Sagun, Biroli, Wyart, 2018]. Right: [Belkin, Hsu, Ma, Mandal, 2018].

Double descent



(a) U-shaped "bias-variance" risk curve

(b) "double descent" risk curve

Figure: A cartoon by [Belkin, Hsu, Ma, Mandal, 2018].

- ✓ Peak at the interpolation thresholds.
- ✓ Monotonic decreasing in the overparameterized regime.
- ✓ Global minimum when the number of parameters is infinity.

The misspecified linear model

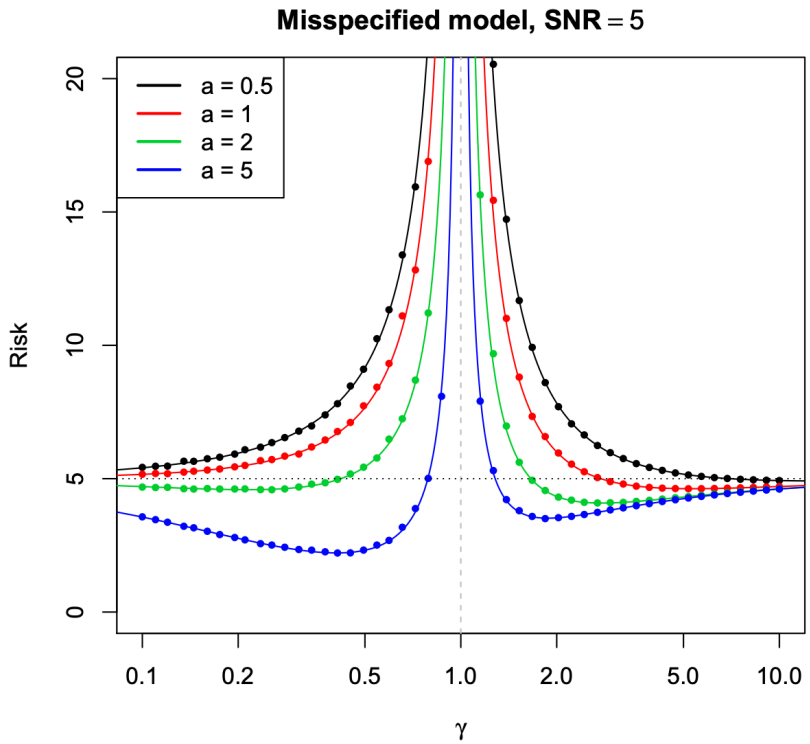
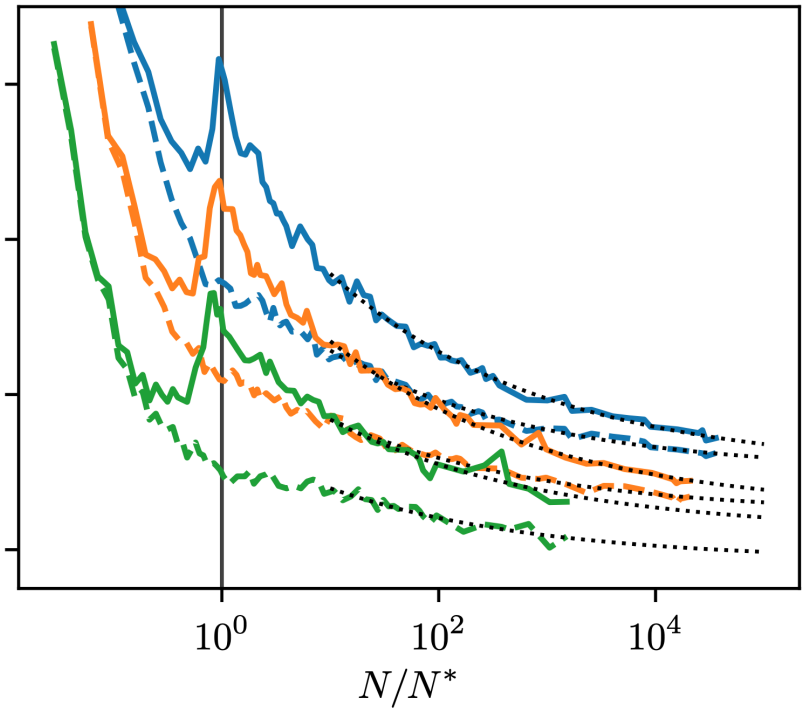
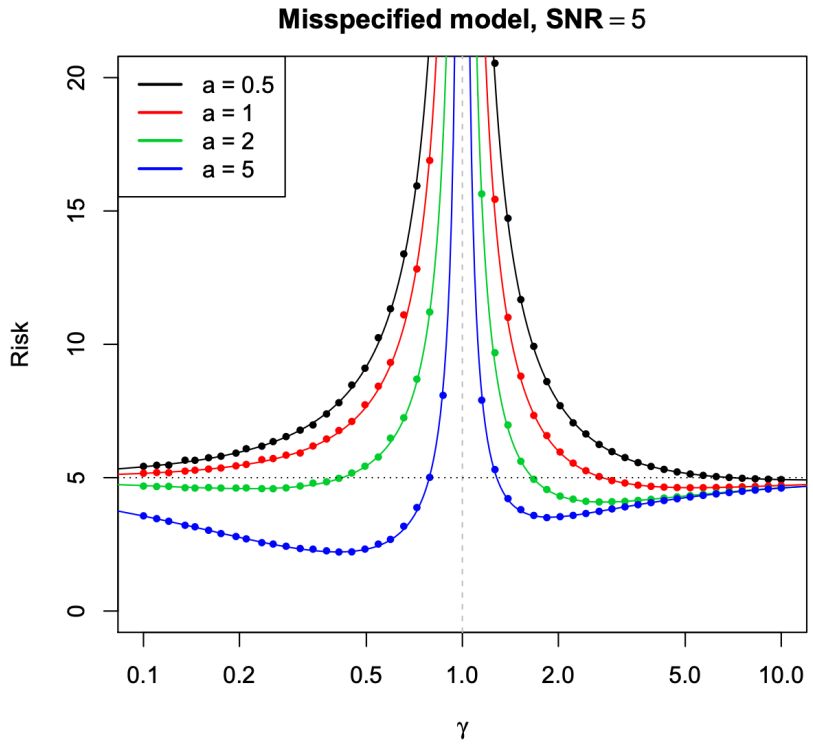


Figure: By [Hastie, Montanari, Rosset, Tibshirani, 2019]. See also [Belkin, Hsu, Xu, 2019].

$$\text{Model: } y = \langle x_S, \beta_S \rangle + \varepsilon \text{ for } |S| = k.$$
$$\text{Fitting: } L(\beta) = \hat{\mathbb{E}}[(y - \langle x, \beta \rangle)^2]$$

The misspecified linear model





- ✓ Peak at the interpolation thresholds.
- ? Monotonic decreasing in the overparameterized regime.
- ? Global minimum when the number of parameters is infinity.

Goal: find a tractable model that exhibits all the features of the double descent curve.

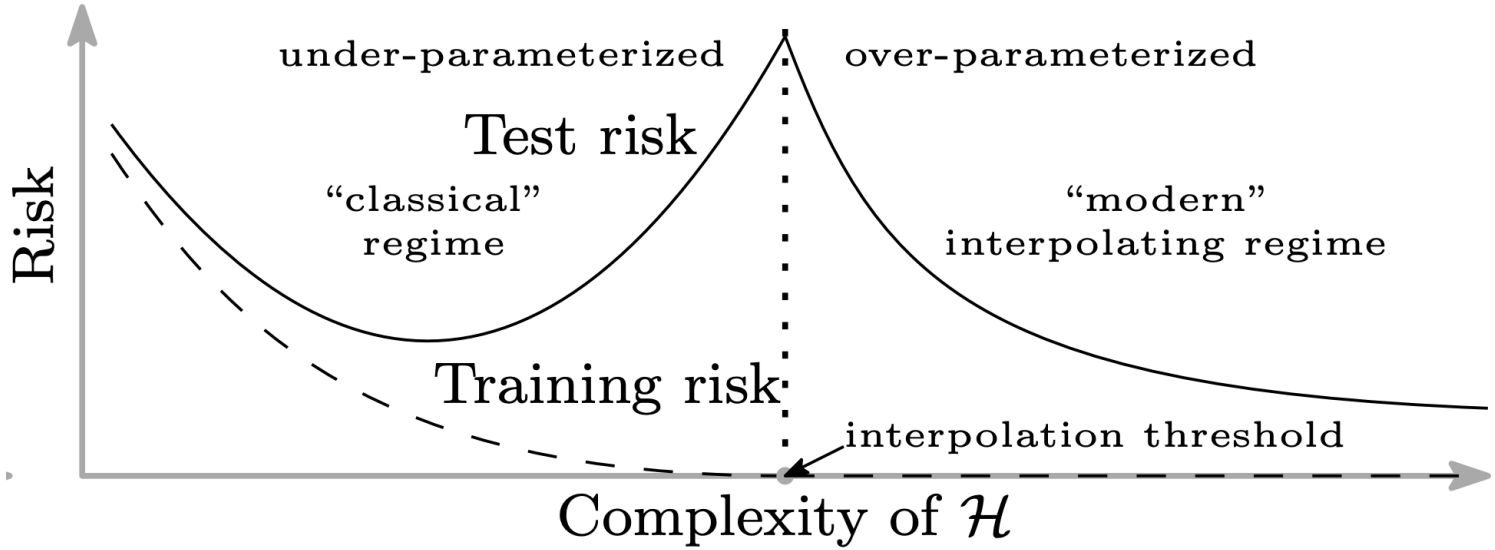


Figure: By [Belkin, Hsu, Ma, Mandal, 2018].

The neural tangent model

- Let $f(\boldsymbol{x};\boldsymbol{\theta})$ be a multi-layers neural network

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sigma(\boldsymbol{W}_1\sigma(\boldsymbol{W}_2\cdots\sigma(\boldsymbol{W}_L\boldsymbol{x})))$$

- NT model: linearization of $f(\boldsymbol{x};\boldsymbol{\theta})$ around initialization $\boldsymbol{\theta}_0$,

$$f_{\mathsf{NT}}(\boldsymbol{x};\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x};\boldsymbol{\theta}_0)\rangle.$$

  [Jacot, Gabriel, Hongler, 2018], [Du, Zhai, Poczos, Singh, 2018], [Chizat, Bach, 2018b].

- Under some conditions of initialization and learning rate, the trajactory of neural tangent model and neural network is uniformly close.

Two-layers neural tangent model

The two-layers neural tangent model

$$f_{\mathsf{NT}}(x; \{a_j\}, \{t_j\}) = \underbrace{\sum_{j=1}^{N} a_j \sigma(\langle w_j, x \rangle)}_{\text{Second layer linearization}} + \underbrace{\sum_{j=1}^{N} \langle t_j, x \rangle \sigma'(\langle w_j, x \rangle)}_{\text{First layer linearization}}.$$

Random weights

$$w_j \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}).$$

An even simpler model

The random features model

$$f_{\mathsf{RF}}(\boldsymbol{x}; \boldsymbol{a}) = \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle).$$

Random weights

$$\boldsymbol{w}_j \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1}).$$

Setting

- n data points, N features, in dimension d, proportional as $d \to \infty$.
- Data $(\boldsymbol{x}_i)_{i\in n} \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$, $\mathbb{E}[\varepsilon_i^2] = \tau^2$.
- Features $(\boldsymbol{w}_j)_{j\in[N]} \sim_{iid} \mathrm{Unif}(\mathbb{S}^{d-1})$.
- Random feature regression: $\hat{\boldsymbol{a}}_\lambda = \arg\min_{\boldsymbol{a}} L_\lambda(\boldsymbol{a})$,

$$L_\lambda(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\Big[\Big(y_i - \frac{1}{N}\sum_{j=1}^{N} a_j\sigma(\langle \boldsymbol{x}_i, \boldsymbol{w}_j\rangle)\Big)^2\Big] + \frac{\lambda N}{d}\|\boldsymbol{a}\|_2^2,$$

$$R(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{x},y}\Big[\Big(f_\star(\boldsymbol{x}) - \frac{1}{N}\sum_{j=1}^{N} a_j\sigma(\langle \boldsymbol{x}, \boldsymbol{w}_j\rangle)\Big)^2\Big].$$

Assumption

- Proportional regime: $N/d \to \psi_1$, $n/d \to \psi_2$, as $d \to \infty$.
- Activation: σ sub exponential growth, including ReLU, tanh, etc.
- Truth function: $f_\star(\boldsymbol{x}) = \langle \boldsymbol{\beta}_1, \boldsymbol{x} \rangle$.

Precise asymptotics

Theorem (M. and Montanari, 2019)

Assume $f_\star(x) = \langle \beta_1, x \rangle$ *and define (for* $G \sim \mathcal{N}(0,1)$*)*

$$\mu_1 = \mathbb{E}[\sigma(G)G], \quad \mu_\star^2 = \mathbb{E}[\sigma(G)^2] - \mathbb{E}[\sigma(G)]^2 - \mathbb{E}[\sigma(G)G]^2, \quad \zeta = \mu_1/\mu_\star.$$

Let $N/d \to \psi_1$, $n/d \to \psi_2$, *as* $d \to \infty$. *Then for any* $\lambda > 0$, *we have*

$$R_{\mathsf{RF}}(\hat{a}_\lambda; f_\star) = \|\beta_1\|_2^2 \cdot \mathscr{B}(\zeta, \psi_1, \psi_2, \lambda/\mu_\star^2) + \tau^2 \cdot \mathscr{V}(\zeta, \psi_1, \psi_2, \lambda/\mu_\star^2) + o_{d,\mathbb{P}}(1),$$

where functions $\mathscr{B}$ *and* $\mathscr{V}$ *are given explicitly below.*

Explicit formulae

Let the functions $\nu_1, \nu_2 : \mathbb{C}_+ \to \mathbb{C}_+$ be the unique solution of

$$\nu_1 = \psi_1 \Big(-\xi - \nu_2 - \frac{\zeta^2 \nu_2}{1 - \zeta^2 \nu_1 \nu_2} \Big)^{-1},$$
$$\nu_2 = \psi_2 \Big(-\xi - \nu_1 - \frac{\zeta^2 \nu_1}{1 - \zeta^2 \nu_1 \nu_2} \Big)^{-1};$$

Let

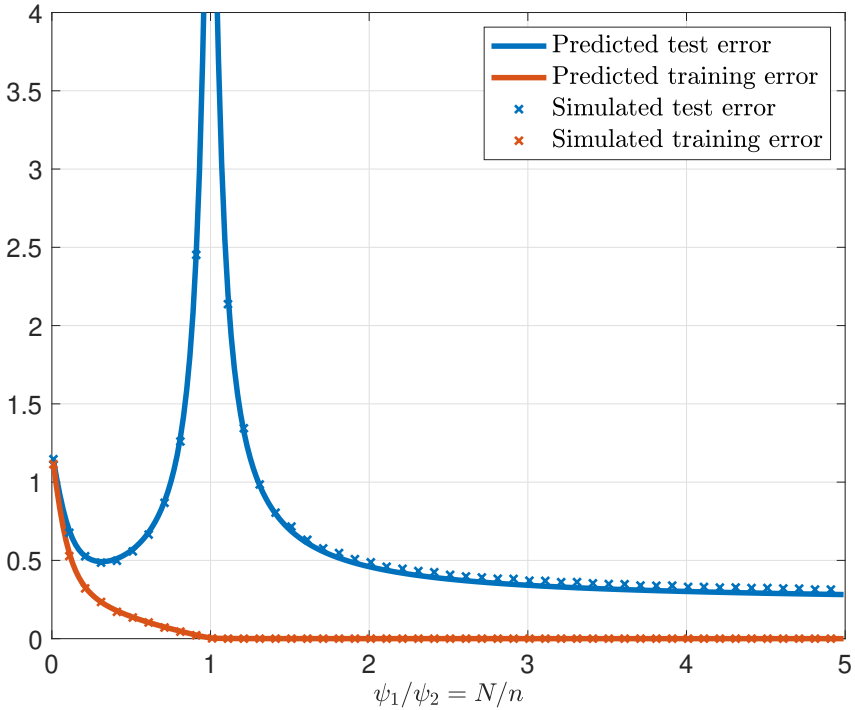
$$\chi \equiv \nu_1(i(\psi_1\psi_2\overline{\lambda})^{1/2}) \cdot \nu_2(i(\psi_1\psi_2\overline{\lambda})^{1/2}),$$

and

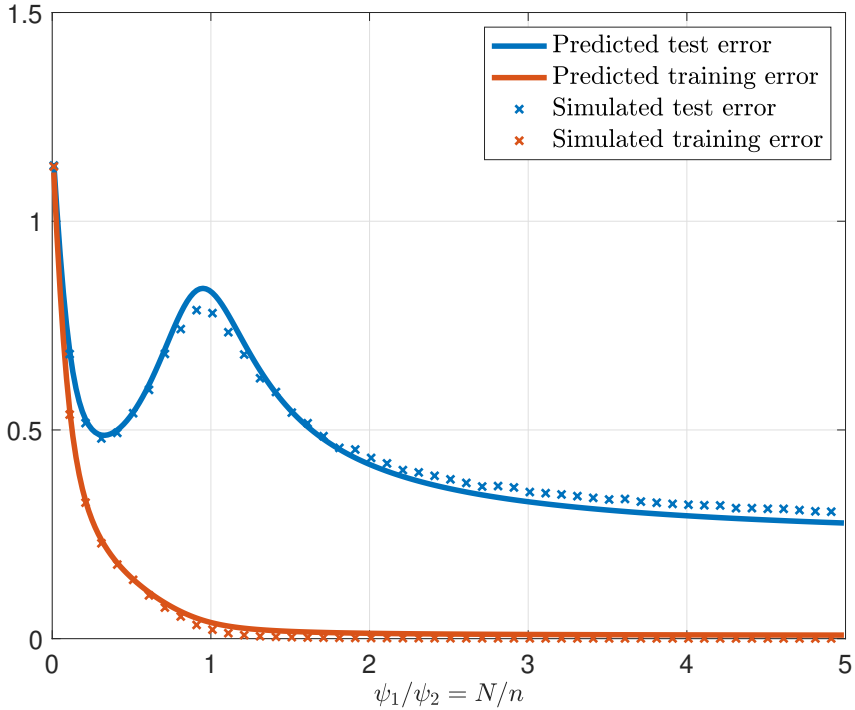
$$\begin{aligned}
\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv & -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2 \\
& + (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2, \\
\mathscr{E}_1(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv & \ \psi_2\chi^3\zeta^4 - \psi_2\chi^2\zeta^2 + \psi_1\psi_2\chi\zeta^2 - \psi_1\psi_2, \\
\mathscr{E}_2(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv & \ \chi^5\zeta^6 - 3\chi^4\zeta^4 + (\psi_1 - 1)\chi^3\zeta^6 + 2\chi^3\zeta^4 + 3\chi^3\zeta^2 + (-\psi_1 - 1)\chi^2\zeta^4 - 2\chi^2\zeta^2 - \chi^2.
\end{aligned}$$

We then have

$$\mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv \frac{\mathscr{E}_1(\zeta, \psi_1, \psi_2, \overline{\lambda})}{\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})}, \qquad \mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv \frac{\mathscr{E}_2(\zeta, \psi_1, \psi_2, \overline{\lambda})}{\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})}.$$



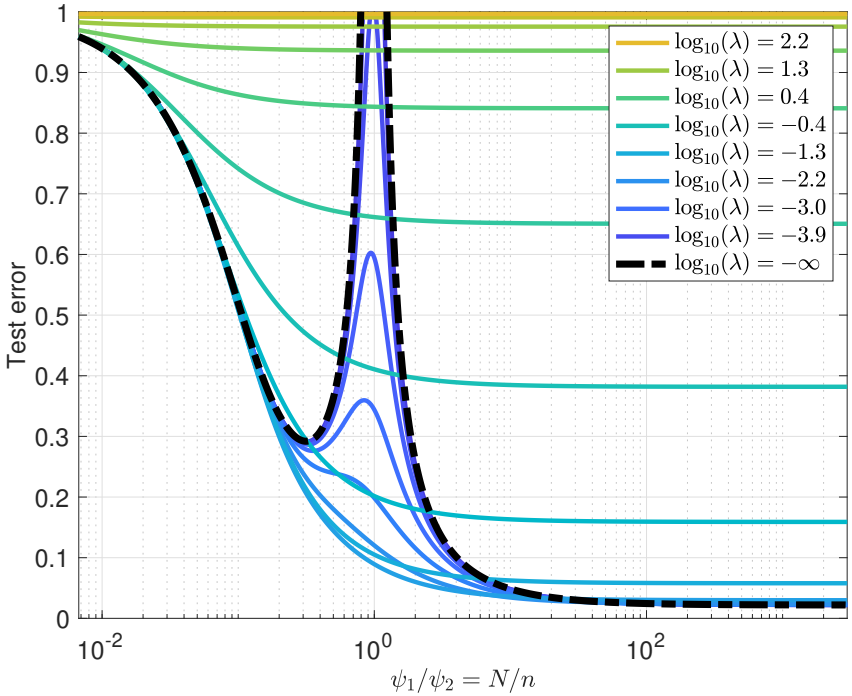
$\lambda = 0+$



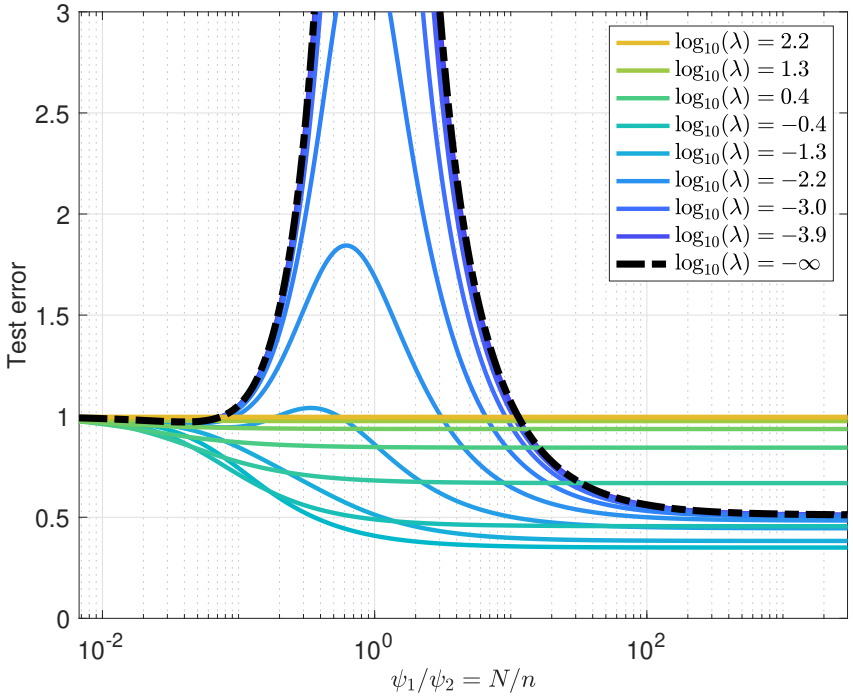
$\lambda = 3 \times 10^{-4}$

- ✓ Peak at the interpolation thresholds.
- ✓ Monotonic decreasing in the overparameterized regime.
- ✓ Global minimum when the number of parameters is infinity.

Further insights

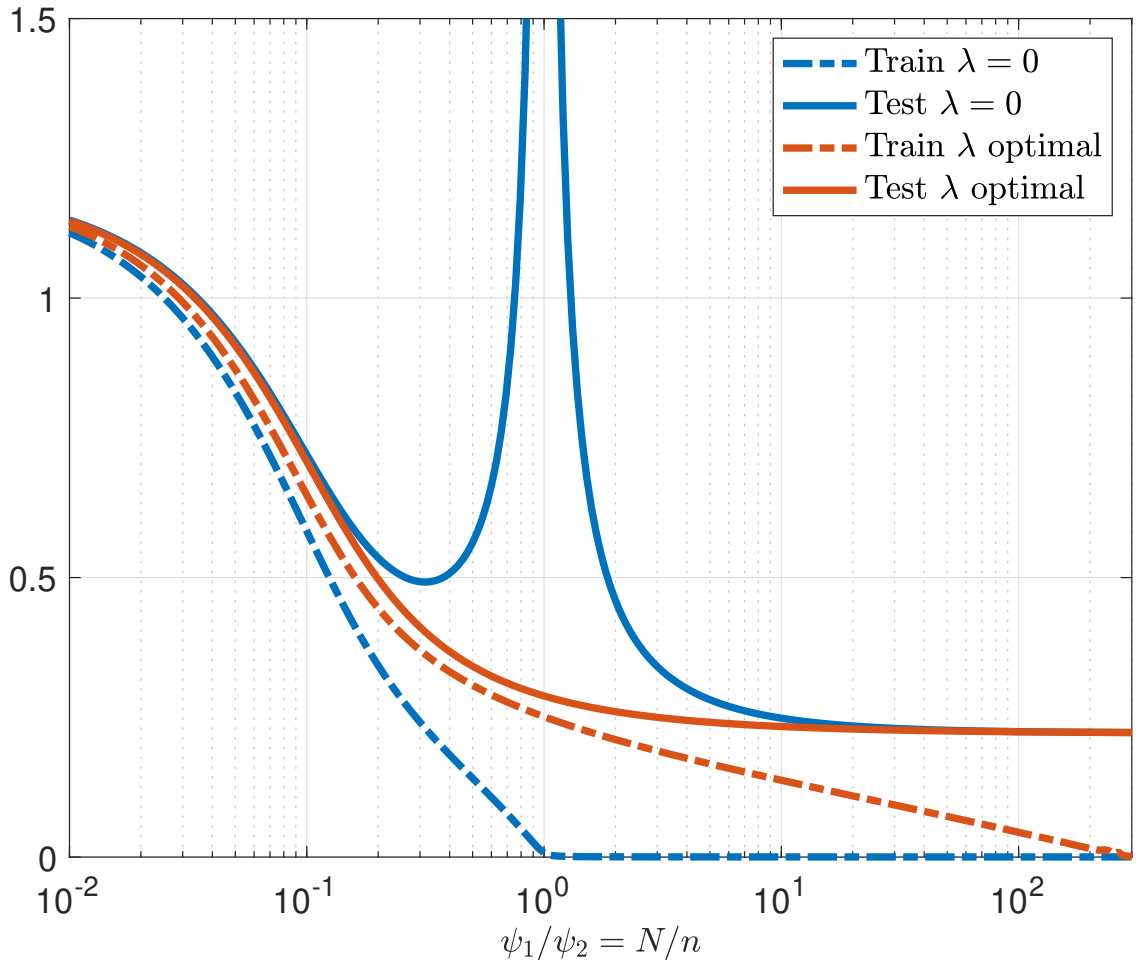


SNR $= 5$

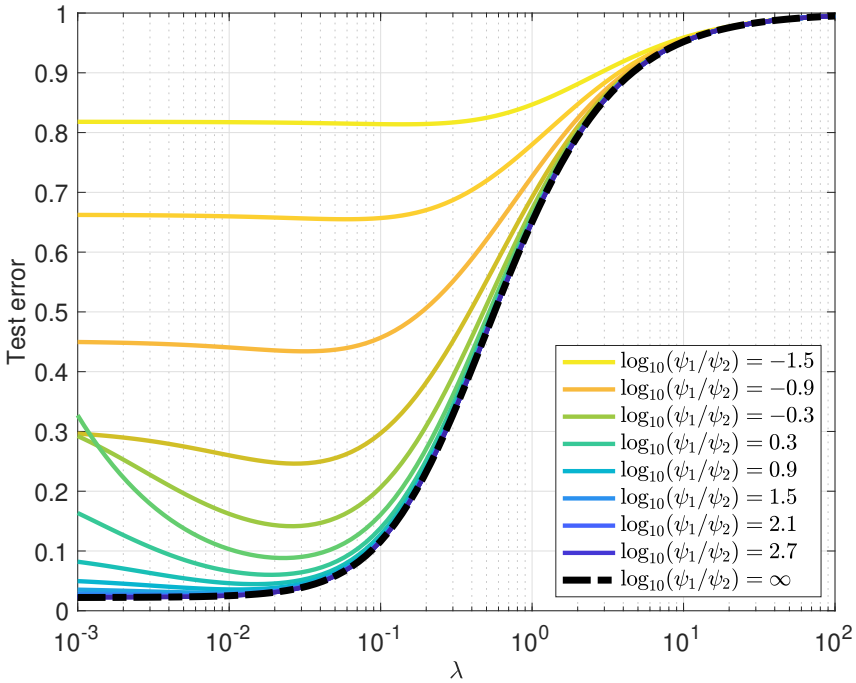


SNR $= 1/5$

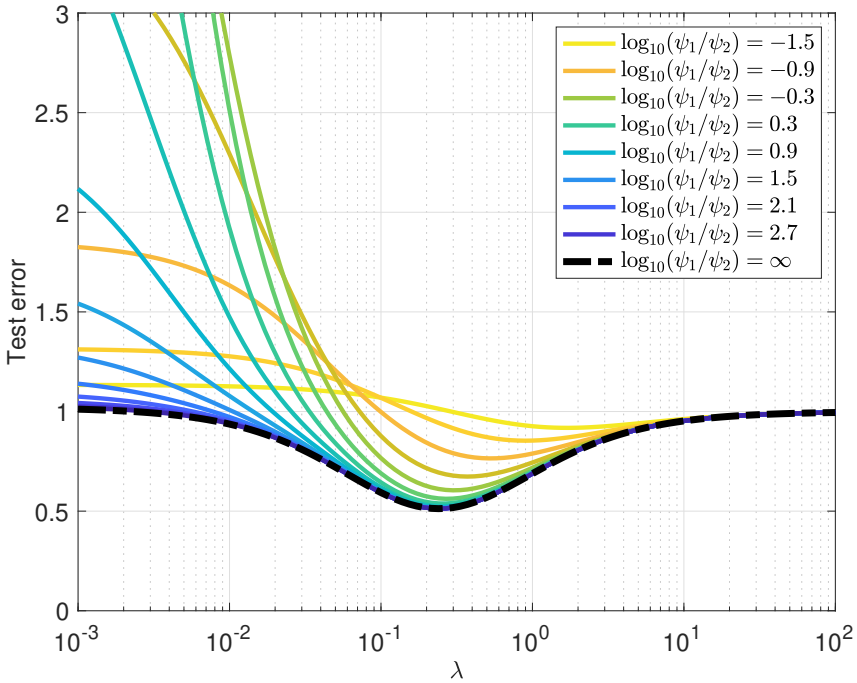
For any λ, the minimum generalization error is achieved at $N/n \to \infty$.



For optimal λ, the generalization error is monotonically decreasing in N/n.



$\mathrm{SNR} = 5$



$\mathrm{SNR} = 1/10$

- High SNR: minimum at $\lambda = 0+$;
- Low SNR: minimum at $\lambda > 0$.

Proof strategy

Random matrix theory for the random kernel inner product matrices

Conclusion

- Number of parameters is not the right model complexity to control the generalization error (we already know this).
- The double descent phenomenon also appears in linearized neural networks.
- When SNR is high, without regularization could be better than with regularization.