

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

Song Mei

April 20, 2021

Joint work with Andrea Montanari

Surprises of generalization behavior of neural networks

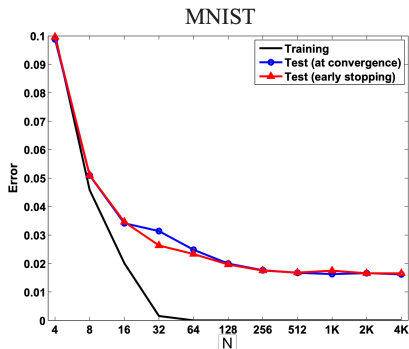
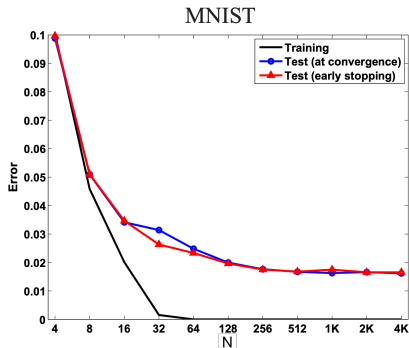


Figure: Experiments on MNIST by [Neyshabur, Tomioka, Srebro, 2014a]

Surprise: why **doesn'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

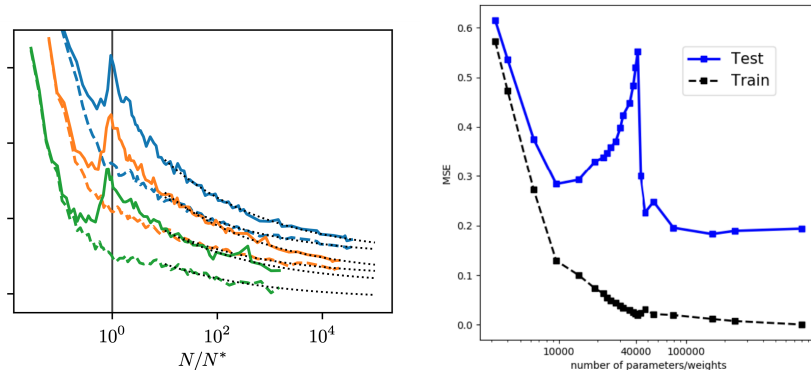


Figure: Experiments on MNIST. Left: [Spigler, Geiger, Ascoli, Sagun, Biroli, Wyart, 2018]. Right: [Belkin, Hsu, Ma, Mandal, 2018].

Double descent

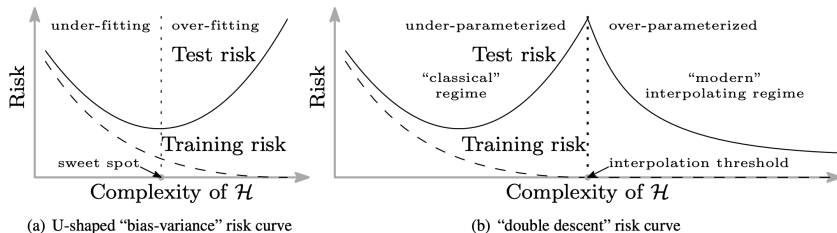


Figure: A cartoon by [Belkin, Hsu, Ma, Mandal, 2018].

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

The misspecified linear model

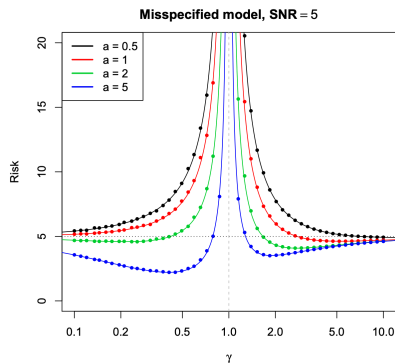
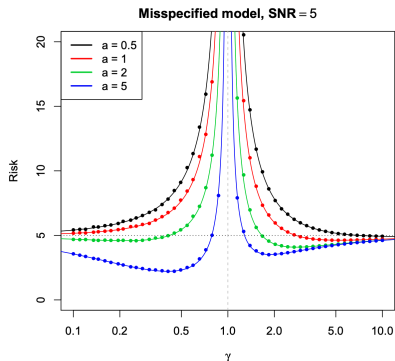
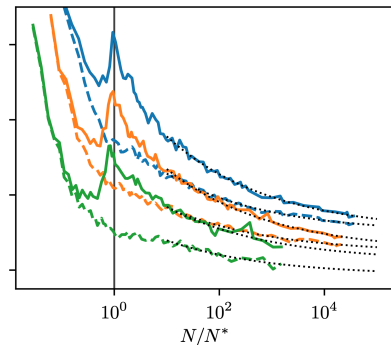


Figure: By [Hastie, Montanari, Rosset, Tibshirani, 2019]. See also [Belkin, Hsu, Xu, 2019].

Model: $y = \langle x_S, \beta_S \rangle + \varepsilon$ for $|S| = k$.

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.

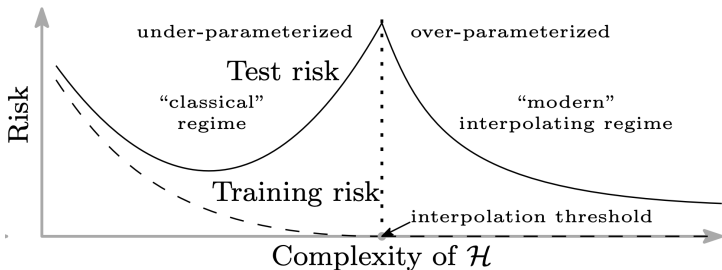


Figure: By [Belkin, Hsu, Ma, Mandal, 2018].

The neural tangent model

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

$$f(x; \theta) = \sigma(W_1 \sigma(W_2 \cdots \sigma(W_L x)))$$

- ▶ NT model: linearization of $f(x; \theta)$ around initialization θ_0 ,

$$f_{\text{NT}}(x; \theta) = \langle \theta, \nabla_{\theta} f(x; \theta_0) \rangle.$$

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

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

Two-layers neural tangent model

The two-layers neural tangent model

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

Random weights

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

An even simpler model

The random features model

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

Random weights

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

Setting

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

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

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

Assumption

- ▶ Proportional regime: $N/d \rightarrow \psi_1$, $n/d \rightarrow \psi_2$, as $d \rightarrow \infty$.
- ▶ Activation: σ sub exponential growth, including ReLU, tanh, etc.
- ▶ Truth function: $f_\star(x) = \langle \beta_1, x \rangle$.

Precise asymptotics

Theorem (M. and Montanari, 2019)

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

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

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

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

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

Explicit formulae

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

$$\begin{aligned}\nu_1 &= \psi_1 \left(-\xi - \nu_2 - \frac{\zeta^2 \nu_2}{1 - \zeta^2 \nu_1 \nu_2} \right)^{-1}, \\ \nu_2 &= \psi_2 \left(-\xi - \nu_1 - \frac{\zeta^2 \nu_1}{1 - \zeta^2 \nu_1 \nu_2} \right)^{-1};\end{aligned}$$

Let

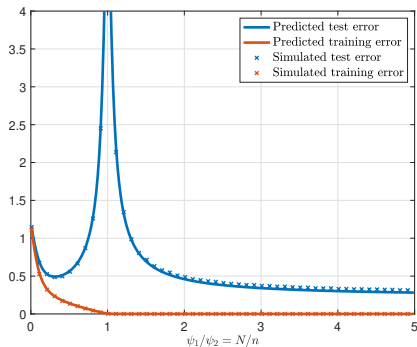
$$\chi \equiv \nu_1(i(\psi_1 \psi_2 \bar{\lambda})^{1/2}) \cdot \nu_2(i(\psi_1 \psi_2 \bar{\lambda})^{1/2}),$$

and

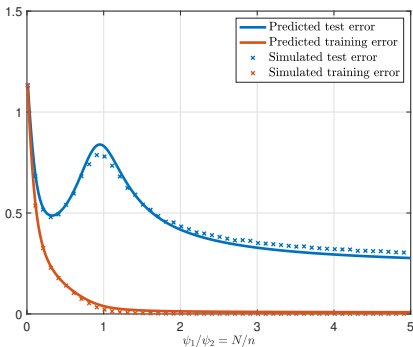
$$\begin{aligned}\mathcal{E}_0(\zeta, \psi_1, \psi_2, \bar{\lambda}) &\equiv -\chi^5 \zeta^6 + 3\chi^4 \zeta^4 + (\psi_1 \psi_2 - \psi_2 - \psi_1 + 1)\chi^3 \zeta^6 - 2\chi^3 \zeta^4 - 3\chi^3 \zeta^2 \\ &\quad + (\psi_1 + \psi_2 - 3\psi_1 \psi_2 + 1)\chi^2 \zeta^4 + 2\chi^2 \zeta^2 + \chi^2 + 3\psi_1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

We then have

$$\mathcal{B}(\zeta, \psi_1, \psi_2, \bar{\lambda}) \equiv \frac{\mathcal{E}_1(\zeta, \psi_1, \psi_2, \bar{\lambda})}{\mathcal{E}_0(\zeta, \psi_1, \psi_2, \bar{\lambda})}, \quad \mathcal{V}(\zeta, \psi_1, \psi_2, \bar{\lambda}) \equiv \frac{\mathcal{E}_2(\zeta, \psi_1, \psi_2, \bar{\lambda})}{\mathcal{E}_0(\zeta, \psi_1, \psi_2, \bar{\lambda})}.$$



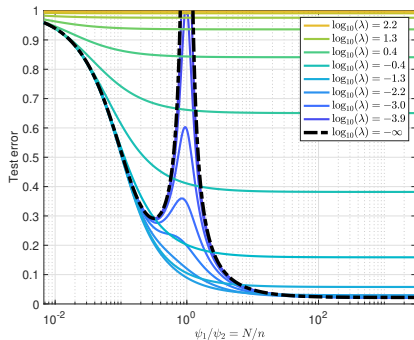
$\lambda = 0+$



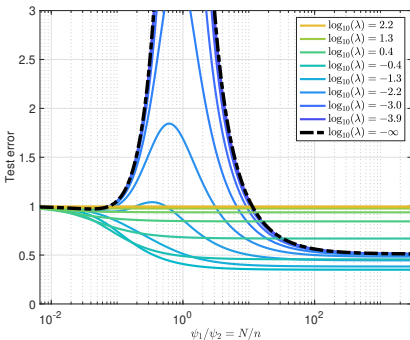
$\lambda = 3 \times 10^{-4}$

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

Further insights

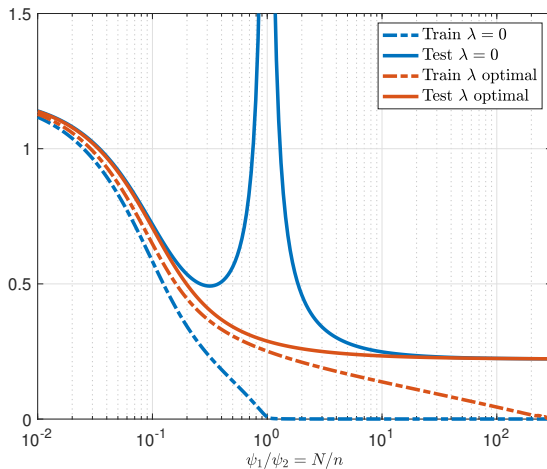


SNR = 5

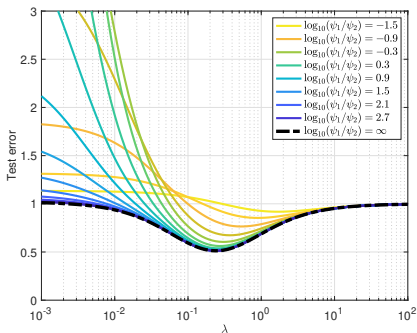
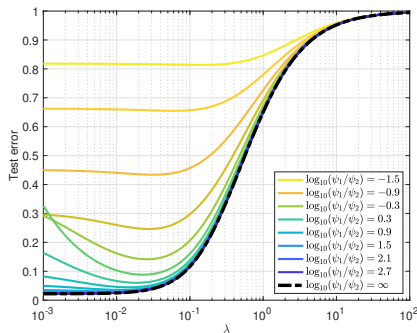


SNR = 1/5

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



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



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

Heuristic derivation and proof strategy

- ▶ Use linear pencil method (**free energy trick**) to reduce the task to computing the **log determinant of a random matrix**.
- ▶ Heuristic derivation: Gaussian equivalence principle + replica trick.
- ▶ Rigorous proof: Stieltjes transform + leave one out.

Linear pencil method

Expected test error can be decomposed as summation of product of dependent random matrices $\mathbf{H}, \mathbf{Q}, \mathbf{Z}, (\mathbf{Z}\mathbf{Z}^\top + \lambda\mathbf{I})$.

- ▶ Want to calculate $\mathbb{E}[\text{tr}(\mathbf{H}(\mathbf{Z}\mathbf{Z}^\top + \lambda\mathbf{I})^{-1}\mathbf{Z}\mathbf{Q}\mathbf{Z}^\top(\mathbf{Z}\mathbf{Z}^\top + \lambda\mathbf{I})^{-1})]$.
- ▶ Define a block matrix

$$\mathbf{A}(s, t) = \begin{bmatrix} s\mathbf{H} + \lambda\mathbf{I} & \mathbf{Z} \\ \mathbf{Z}^\top & t\mathbf{Q} + \lambda\mathbf{I} \end{bmatrix}.$$

- ▶ We have

$$\begin{aligned} & \mathbb{E}[\text{tr}(\mathbf{H}(\mathbf{Z}\mathbf{Z}^\top + \lambda\mathbf{I})^{-1}\mathbf{Z}\mathbf{Q}\mathbf{Z}^\top(\mathbf{Z}\mathbf{Z}^\top + \lambda\mathbf{I})^{-1})] \\ &= \partial_s \partial_t \mathbb{E}[\log(\det(\mathbf{A}(s, t)))]. \end{aligned}$$

Intuition: $\log(\det(\mathbf{A}(s, t)))$ is like the free energy. Need to be clever to construct $\mathbf{A}$.

Log determinant and Gaussian equivalent principle

- ▶ If all the element of $\mathbf{A}$ are related to Gaussians, then the log determinant can be calculated using the "replica method":

$$\begin{aligned}\mathbb{E}[\log \det(\mathbf{A})] &\approx -2 \log \mathbb{E}[\det(\mathbf{A})^{-1/2}] \\ &= -2 \log \mathbb{E}\left[\int_{\mathbb{R}^d} \exp\{-\langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle / 2\} d\mathbf{x}\right].\end{aligned}$$

- ▶ However, there is $\mathbf{Z}$ matrix

$$\mathbf{Z}_{ij} = \sigma(\langle \mathbf{x}_i, \boldsymbol{\theta}_j \rangle / \sqrt{d}).$$

- ▶ The Gaussian equivalence principle:

$$\overline{\mathbf{Z}}_{ij} = \mu_0 + \mu_1 \langle \mathbf{x}_i, \boldsymbol{\theta}_j \rangle / \sqrt{d} + \mu_* z_{ij},$$

where z_{ij} are iid Gaussian independent of $\mathbf{x}_i, \boldsymbol{\theta}_j$.

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.