國立臺灣大學
National Taiwan University

Neural Tangent Kernel: On Double Descent

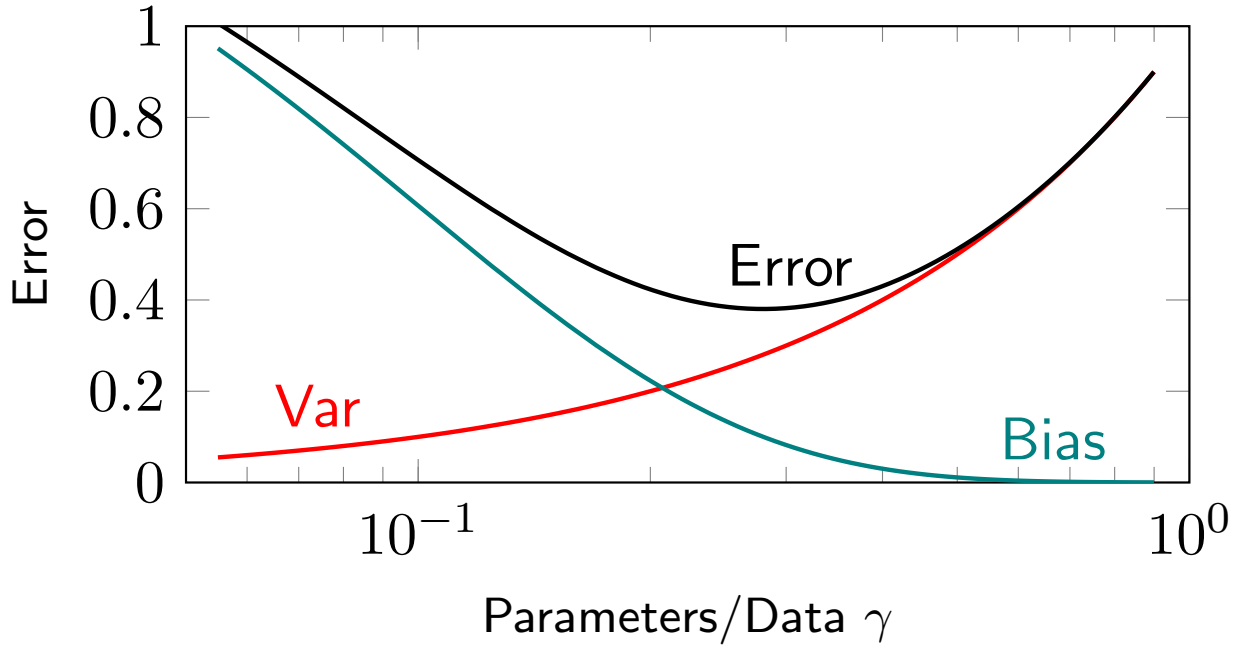
Wen Perng (B10901042)

Physical Theories of (Machine) Learning

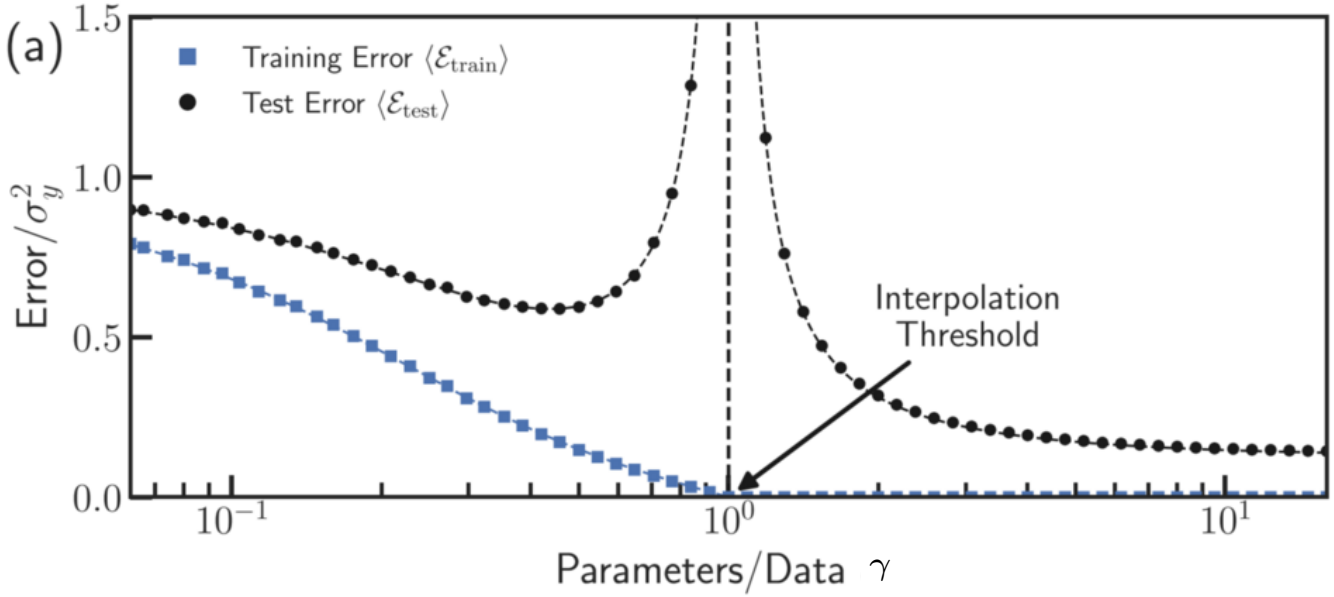
2025 Nov. 12

The generalization error can be separated into bias and variance:

$$\begin{aligned}\mathcal{E}_g &= \mathbb{E}_{\boldsymbol{x},\mathcal{D}}\left[(f_\star(\boldsymbol{x}) - f_{\boldsymbol{\theta}}(\boldsymbol{x}))^2\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\Big(\mathrm{Bias}(f_{\boldsymbol{\theta}}, \boldsymbol{x})\Big)^2 + \mathrm{Var}\left(f_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\Big(f_\star(\boldsymbol{x}) - \mathbb{E}_{\mathcal{D}}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x})\right]\Big)^2 + \mathbb{E}_{\mathcal{D}}\left[\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbb{E}_{\mathcal{D}}\left[f_{\boldsymbol{\theta}}(\boldsymbol{x})\right]\right)^2\right]\right]\end{aligned}$$



Model performances are not trivial:



[Rocks et al. '22]

This is the **double descent** behavior.

- Behavior depends on the **hyperparameters**.
- Explain via neural tangent kernel (NTK), Θ_0.

Using the **replica trick**, we will be able to derive the following double descent curve for ridgeless kernel regression.

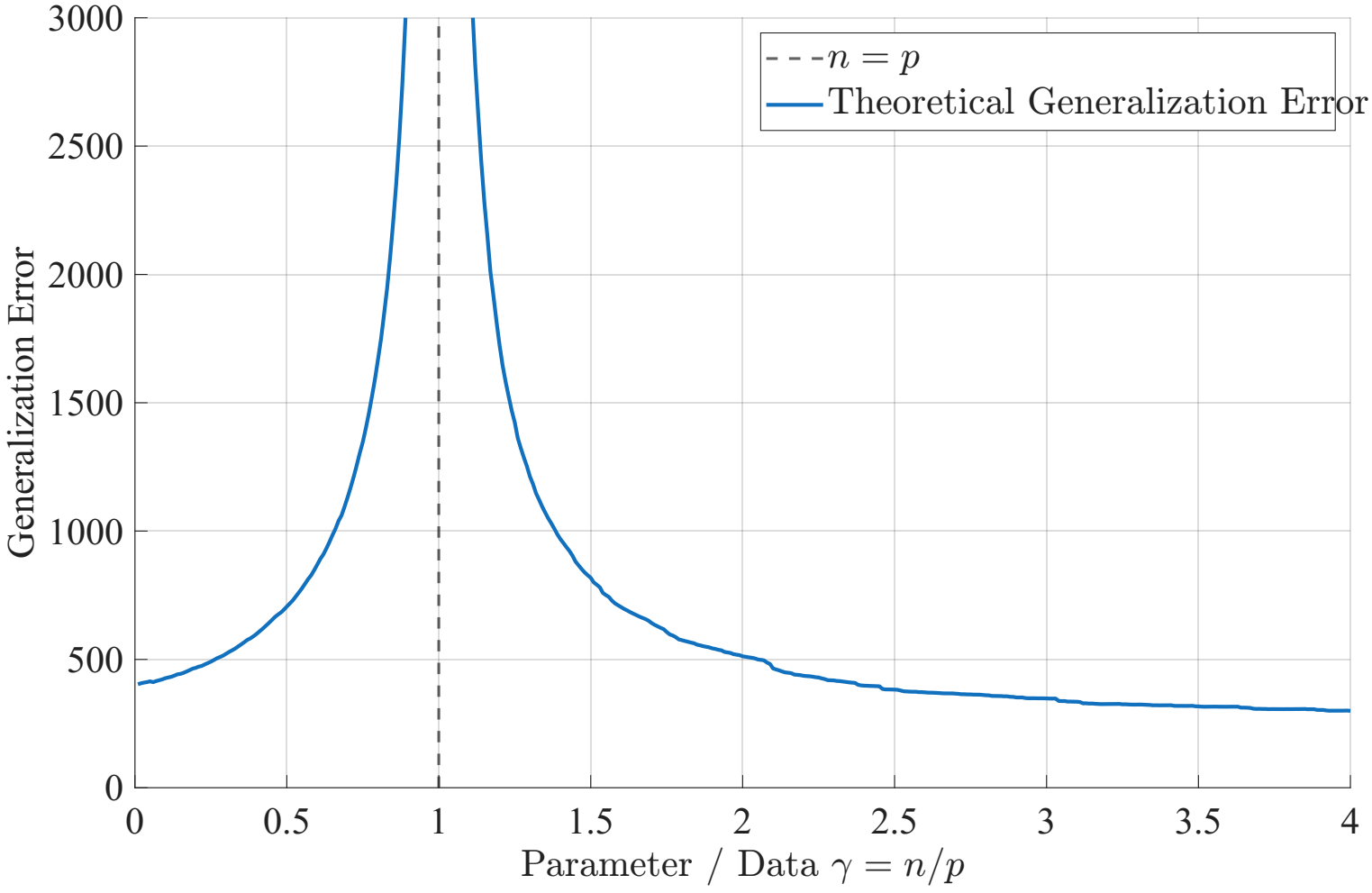
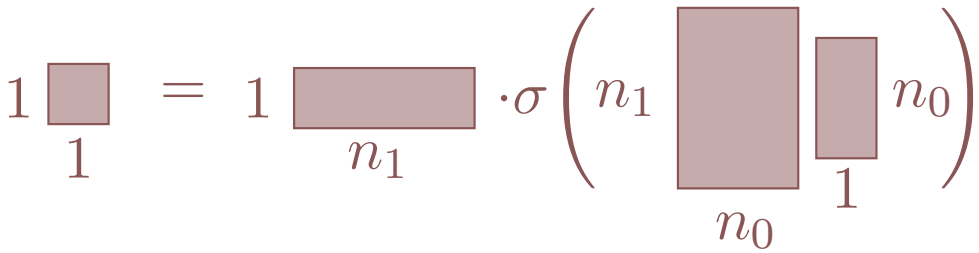


Table of Contents

Table of Contents

Model: with $[W_1]_{ij} \sim \mathcal{N}(0,1)$ and $[W_2]_{ij} \sim \mathcal{N}(0, \sigma_{W_2}^2)$,

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{W_2}{\sqrt{n_1}} \cdot \sigma\left(\frac{W_1}{\sqrt{n_0}}\boldsymbol{x}\right). \tag{1}$$



Trainable parameters:

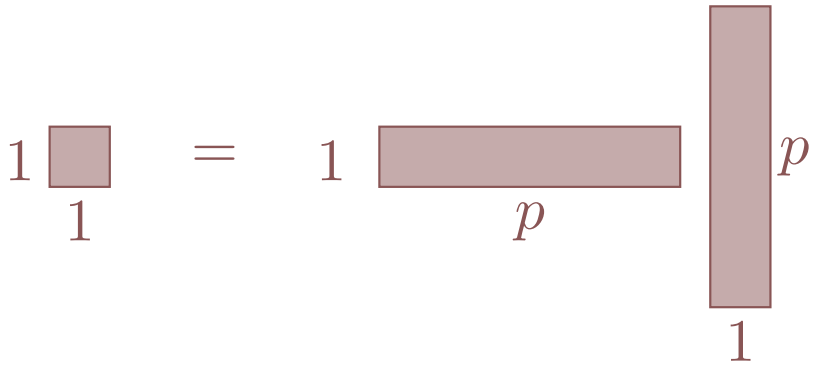
$$\boldsymbol{\theta} = [\mathrm{vec}(W_1); \mathrm{vec}(W_2)] \in \mathbb{R}^{(n_0 n_1 + n_1)\times 1}.$$

Training data: $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{p}$, with

$$X = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_p \end{bmatrix} \in \mathbb{R}^{n_0 \times p},\ Y = \begin{bmatrix} y_1 & \dots & y_p \end{bmatrix} \in \mathbb{R}^{1\times p}.$$

NTK describes the learning dynamics under gradient flow:

$$\frac{\mathrm{d} f_{\boldsymbol{\theta}_t}(\boldsymbol{x})}{\mathrm{d}t} = -\left(f_{\boldsymbol{\theta}_t}(X) - Y\right)\Theta_t(X, \boldsymbol{x}), \tag{2}$$



where

$$\begin{aligned}\Theta_t(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \nabla_{\boldsymbol{\theta}}^{\mathsf{T}} f_{\boldsymbol{\theta}_t}(\boldsymbol{x}_i) \overbrace{\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_t}(\boldsymbol{x}_j)}^{(n_0 n_1 + n_1)\times 1} \\ &= \nabla_{W_1}^{\mathsf{T}} f_{\boldsymbol{\theta}_t}(\boldsymbol{x}_i) \underbrace{\nabla_{W_1} f_{\boldsymbol{\theta}_t}(\boldsymbol{x}_j)}_{n_0 n_1 \times 1} + \nabla_{W_2}^{\mathsf{T}} f_{\boldsymbol{\theta}_t}(\boldsymbol{x}_i) \underbrace{\nabla_{W_2} f_{\boldsymbol{\theta}_t}(\boldsymbol{x}_j)}_{n_1 \times 1}.\end{aligned} \tag{3}$$

As network width goes to infinity: $n_0 \to \infty$ and $n_1/n_0 = \gamma$ fixed, the neural tangent kernel Θ_t becomes **frozen**:

$$\Theta_t \to \Theta_0 \ \forall t \geq 0.$$

The final form is a ridgeless kernel regression: denote $f_{\boldsymbol{\theta}_t}$ by f_t,

$$f_\infty(\boldsymbol{x}) = f_0(\boldsymbol{x}) + \underbrace{(Y - f_0(X))\,\Theta_0^{-1}(X, X)}_{\text{regression coefficients}}\ \underbrace{\Theta_0(X, \boldsymbol{x})}_{\text{random features}}. \qquad (4)$$

As no features are learned, the regression uses the random kernel as random features. This is termed **lazy learning**.

In the wide network limit, the training result is equivalent to a ridgeless ($\lambda \to 0$) kernel regression with NTK Θ_0, with solution of the form[1]

$$f_\infty(\boldsymbol{x}) = \boldsymbol{v}^\mathsf{T} \Theta_0(X, \boldsymbol{x}).$$

By **Mercer's theorem**, the following decomposition exists:

$$\Theta_0(\boldsymbol{x}, \boldsymbol{x}') = \sum_\rho \psi_\rho(\boldsymbol{x}) \psi_\rho(\boldsymbol{x}'), \tag{5}$$

$$\int p(\boldsymbol{x}) \, \Theta_0(\boldsymbol{x}, \boldsymbol{x}') \psi_\rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \eta_\rho \psi_\rho(\boldsymbol{x}'), \tag{6}$$

$$\int p(\boldsymbol{x}) \, \psi_\rho(\boldsymbol{x}) \psi_{\rho'}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \eta_\rho \delta_{\rho\rho'}. \tag{7}$$

The integrations are over $\boldsymbol{x} \in \mathbb{R}^n$.

[1] For simplicity, assume centering $f_0(X) = 0$.

By **Moore–Aronszajn theorem**, the NTK forms an RKHS:

$$\mathcal{H} = \left\{ \sum_{\rho} w_\rho \psi_\rho(\cdot) \right\} = \left\{ \boldsymbol{w}^\mathsf{T} \boldsymbol{\psi}(\cdot) \right\} \tag{8}$$

with

$$\Lambda := \mathbb{E}_{\boldsymbol{x}} \left[\boldsymbol{\psi}(\boldsymbol{x}) \boldsymbol{\psi}^\mathsf{T}(\boldsymbol{x}) \right] = \mathrm{diag}(\eta_1, \eta_2, \ldots). \tag{9}$$

The teacher (target) and the student (learned) functions both lies in $\mathcal{H}$:[2]

$$f_\star(\cdot) = \boldsymbol{w}_\star^\mathsf{T} \boldsymbol{\psi}(\cdot), \qquad f_{\boldsymbol{\theta}}(\cdot) = \boldsymbol{w}^\mathsf{T} \boldsymbol{\psi}(\cdot).$$

depends on training

[2] If $f_\star \notin \mathcal{H}$, it is equivalent to having noise in the sampling of $f_\star$.

Table of Contents

Let $\boldsymbol{x}$ follow the same data distribution as the training dataset $\mathcal{D}$,

$$
\begin{aligned}
\mathcal{E}_g &= \mathbb{E}_{\boldsymbol{x}}\left[\|f_\star(\boldsymbol{x}) - f_\infty(\boldsymbol{x})\|^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[(f_\star(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2\right] \\
&\quad - 2\mathbb{E}_{\boldsymbol{x}}\left[(f_\star(\boldsymbol{x}) - f_0(\boldsymbol{x}))\,(Y - f_0(X))\,\Theta_0^{-1}(X,X)\Theta(X,\boldsymbol{x})\right] \\
&\quad + \mathbb{E}_{\boldsymbol{x}}\left[\left((Y - f_0(X))\,\Theta_0^{-1}(X,X)\Theta_0(X,\boldsymbol{x})\right)^2\right] \\
&= a + \mathrm{Tr}\left\{B\,\Theta_0^{-1}(X,X)\right\} \\
&\quad + \sum c_i \mathrm{Tr}\left\{C_i\,\Theta_0^{-1}(X,X)\,D_i\,\Theta_0^{-1}(X,X)\right\}
\end{aligned}
$$

The generalization depends on the **spectrum** of the **empirical** NTK.

For the two-layer network:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{W_2}{\sqrt{n_1}} \cdot \sigma\left(\frac{W_1}{\sqrt{n_0}}\boldsymbol{x}\right) \quad (1)$$

$$\nabla_{W_1} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \mathbb{R}^{n_0 n_1 \times 1},\ \nabla_{W_2} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \mathbb{R}^{n_1 \times 1}.$$

Thus, the empirical NTK is

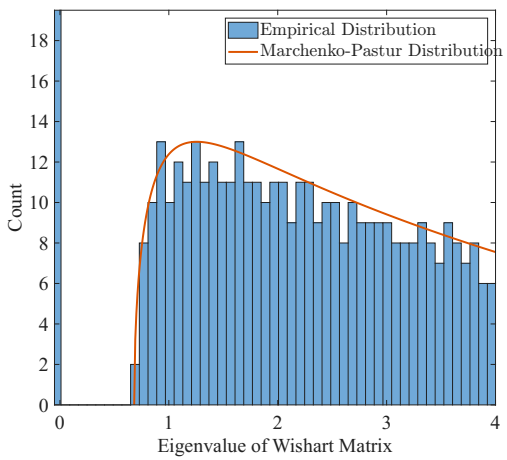
$$\Theta_0(X, X) = p \underbrace{\begin{bmatrix} \vdots \\ \nabla_{W_1}^{\mathsf{T}} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \\ \vdots \end{bmatrix}}_{n_0 n_1} \underbrace{\begin{bmatrix} \cdots \nabla_{W_1} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) \cdots \end{bmatrix}}_{p} n_0 n_1$$

$$+ p \underbrace{\begin{bmatrix} \vdots \\ \nabla_{W_2}^{\mathsf{T}} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \\ \vdots \end{bmatrix}}_{n_1} \underbrace{\begin{bmatrix} \cdots \nabla_{W_2} f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) \cdots \end{bmatrix}}_{p} n_1.$$

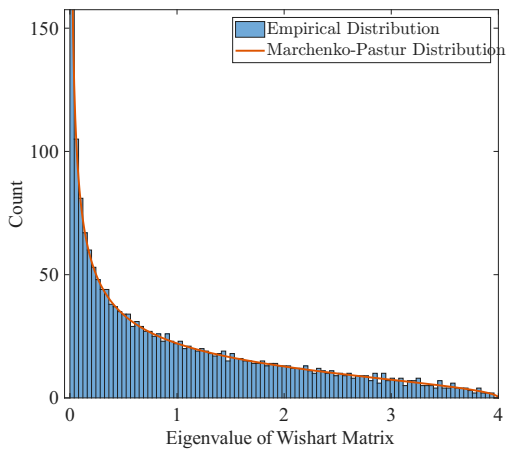
The empirical NTK contains terms of the form

$$\frac{1}{n} F^{\mathsf{T}} F,$$

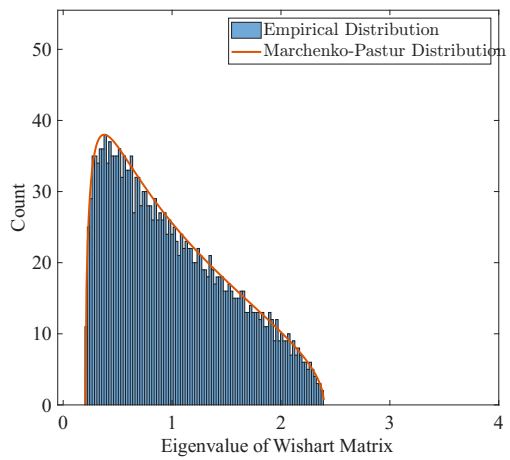
where F is of size $n_0 n_1 \times p$ or $n_1 \times p$ with **stochastic** entries. The limiting spectrum is of **Marchenko–Pastur**-like form[3].



(a) $n/p = 0.67$



(b) $n/p = 1.00$

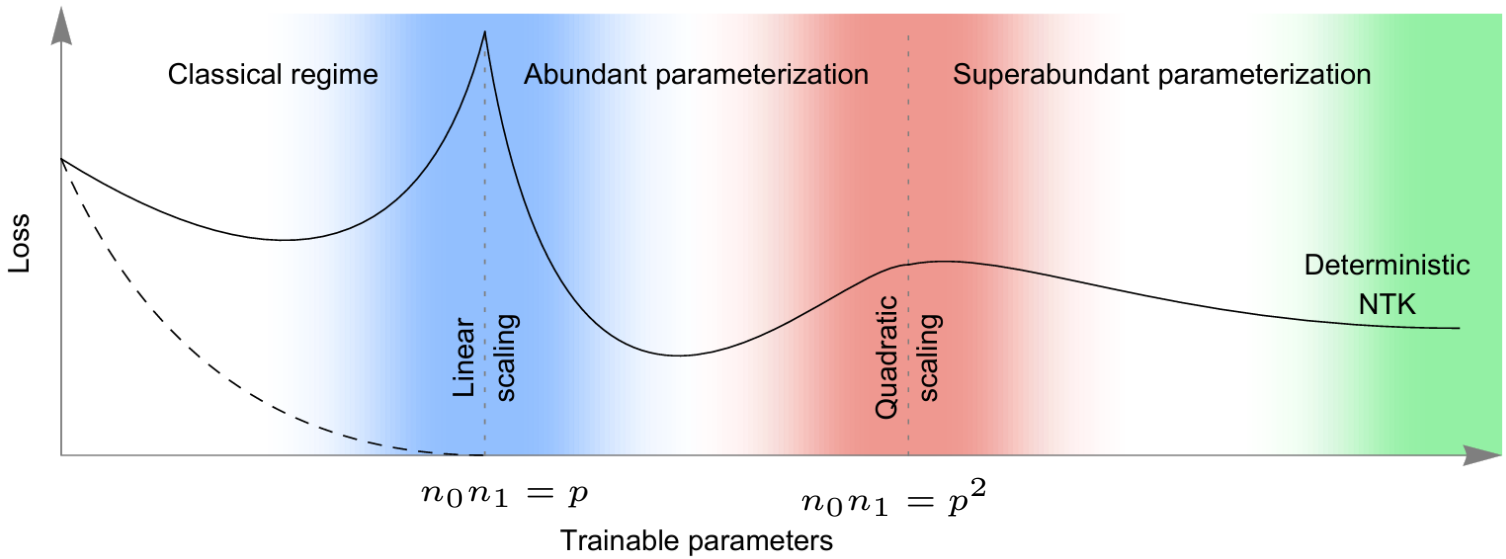


(c) $n/p = 3.33$

[3] The figure above is only an illustration if there is no nonlinearity.

The rank of NTK undergoes **transition** at $n_0 n_1 = p$ and $n_1 = p$.

The figure below shows an illustration of $n_0 = n_1 \to \infty$.



[Adlam et al. '20]

A quantitative analysis requires knowledge in **nonlinear random matrix theory** [AP20; S23].

Table of Contents

Since NTK training is similar to **kernel ridge regression**, let us study the generalization behavior of the latter first!

There is label noise: $y_i = f_\star(\boldsymbol{x}_i) + \varepsilon_i$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The dataset is identified as $\mathcal{D} = \{(\boldsymbol{x}_i, \varepsilon_i)\}_{i=1}^p$.

With these, the training loss function for kernel ridge regression is

$$\begin{aligned}
\mathcal{E}_\lambda(\boldsymbol{w}) &= \sum_{i\in\mathcal{D}} \left(y_i - f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right)^2 + \lambda \left\|f_{\boldsymbol{\theta}}\right\|_{\mathcal{H}}^2 \\
&= \sum_{i\in\mathcal{D}} \Big((\boldsymbol{w}_\star - \boldsymbol{w})^\mathsf{T} \boldsymbol{\psi}(\boldsymbol{x}_i) + \varepsilon_i\Big)^2 + \lambda \left\|\boldsymbol{w}\right\|^2 \\
&= \sum_{i\in\mathcal{D}} \Big(\overline{\boldsymbol{w}}^\mathsf{T} \boldsymbol{\psi}(\boldsymbol{x}_i) + \varepsilon_i\Big)^2 + \lambda \left\|\boldsymbol{w}_\star - \overline{\boldsymbol{w}}\right\|^2,
\end{aligned}$$

where $\overline{\boldsymbol{w}} = \boldsymbol{w}_\star - \boldsymbol{w}$. Denote $\overline{\boldsymbol{w}}^* := \arg\min \mathcal{E}_\lambda(\overline{\boldsymbol{w}})$.

Similarly, the generalization error will be:

$$\begin{aligned}\mathcal{E}_g(\overline{\boldsymbol{w}}^*) &= \mathbb{E}_{\boldsymbol{x}}\left[(f_\star(\boldsymbol{x}) - f_{\boldsymbol{\theta}}(\boldsymbol{x}))^2\right] \\ &= (\boldsymbol{w}_\star - \boldsymbol{w}^*)^\mathsf{T}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{\psi}(\boldsymbol{x})\boldsymbol{\psi}^\mathsf{T}(\boldsymbol{x})\right](\boldsymbol{w}_\star - \boldsymbol{w}^*) \\ &= \overline{\boldsymbol{w}}^{*\mathsf{T}}\Lambda\overline{\boldsymbol{w}}^*.\end{aligned}$$

The typical generalization error will be

$$\begin{aligned}\mathcal{E}_g &= \mathbb{E}_{\mathcal{D}}\left[\mathcal{E}_g(\overline{\boldsymbol{w}}^*)\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[\int \mathcal{E}_g(\overline{\boldsymbol{w}})\ \delta(\overline{\boldsymbol{w}} - \overline{\boldsymbol{w}}^*)\ \mathrm{d}\overline{\boldsymbol{w}}\right].\end{aligned}$$

$$\boxed{\mathcal{E}_g = \mathbb{E}_{\mathcal{D}}\left[\int \mathcal{E}_g(\overline{\boldsymbol{w}})\ \delta(\overline{\boldsymbol{w}} - \overline{\boldsymbol{w}}^*)\ \mathrm{d}\overline{\boldsymbol{w}}\right]}$$

Consider

$$Z = \int \mathrm{d}\overline{\boldsymbol{w}}\ \exp\left\{\beta\left(-\mathcal{E}_\lambda(\overline{\boldsymbol{w}})\right)\right\},$$

then

$$\delta(\overline{\boldsymbol{w}} - \overline{\boldsymbol{w}}^*) = \lim_{\beta\to\infty} \frac{1}{Z} \mathrm{e}^{-\beta\mathcal{E}_\lambda(\overline{\boldsymbol{w}})}.$$

How do we incorporate the $\mathcal{E}_g$ term? By including a **source term**.

By utilizing the partition function [CBP21]:

$$Z = \int \mathrm{d}\overline{\boldsymbol{w}} \, \exp\left\{\beta\left(-\mathcal{E}_\lambda(\overline{\boldsymbol{w}}) + t \cdot \mathcal{E}_g(\overline{\boldsymbol{w}})\right)\right\}, \tag{10}$$

the test error under kernel ridge regression is

$$\mathcal{E}_g(\overline{\boldsymbol{w}}^*) = \lim_{\beta\to\infty} \frac{1}{\beta} \frac{\partial}{\partial t} \log Z\Big|_{t=0}. \tag{11}$$

$\overline{\boldsymbol{w}}^* = \arg\min \mathcal{E}_\lambda(\overline{\boldsymbol{w}})$

Then, the generalization error is

$$\mathcal{E}_g = \mathbb{E}_{\mathcal{D}}\left[\mathcal{E}_g(\overline{\boldsymbol{w}}^*)\right] = \lim_{\beta\to\infty} \frac{1}{\beta} \frac{\partial}{\partial t} \mathbb{E}_{\mathcal{D}}\left[\log Z\right]\Big|_{t=0}. \tag{12}$$

Thus, it requires the **replica trick**[4]!

[4] I apologize for using a different notation as in the paper.

Recap: what is the **replica trick**?

$$\mathbb{E}_{\mathcal{D}}[\log Z] = \mathbb{E}_{\mathcal{D}}\left[\lim_{x\to\infty} \frac{Z^x - 1}{x}\right] = \lim_{x\to 0} \frac{\mathbb{E}_{\mathcal{D}}[Z^x] - 1}{x}.$$

Steps:

1. For integer n, compute Z^n with replicated state parameter $\boldsymbol{w}^a$ and identical disorder parameter $\mathcal{D}$.
2. Apply $\mathbb{E}_{\mathcal{D}}[\cdot]$.
3. Along the way, apply
   - Hubbard–Stratonovich transformation, and
   - Fourier transform relation of delta distribution.
4. Compute saddle point equation.

$$\mathbb{E}_{\mathcal{D}}[Z^n] = \int \underbrace{\left(\prod_{a=1}^{n} \mathrm{d}\overline{\boldsymbol{w}}^a\right)}_{\mathscr{D}[\overline{\boldsymbol{w}}]} \cdot \mathrm{e}^{\beta t \sum_{a=1}^n \mathcal{E}_g(\overline{\boldsymbol{w}}^a)} \cdot \mathbb{E}_{\mathcal{D}}\left[\mathrm{e}^{-\beta \sum_{a=1}^n \mathcal{E}_\lambda(\overline{\boldsymbol{w}}^a)}\right],$$

where

$$= \mathbb{E}_{\mathcal{D}}\left[\exp\left(-\beta \sum_{a=1}^{n} \sum_{i \in \mathcal{D}} \left(\overline{\boldsymbol{w}}^{a\mathsf{T}} \boldsymbol{\psi}(\boldsymbol{x}_i) + \varepsilon_i\right)^2\right)\right] \cdot \mathrm{e}^{-\beta\lambda \sum_{a=1}^n \|\boldsymbol{w}^a\|^2}$$

$$= \left(\mathbb{E}_{(\boldsymbol{x},\varepsilon)}\left[\exp\left(-\beta \sum_{a=1}^{n} \underbrace{\left(\overline{\boldsymbol{w}}^{a\mathsf{T}} \boldsymbol{\psi}(\boldsymbol{x}) + \varepsilon\right)}_{:= q^a \sim \mathcal{N}}\right)^2\right]\right)^p \cdot \mathrm{e}^{-\beta\lambda \sum_{a=1}^n \|\boldsymbol{w}^a\|^2}$$

$$\rightsquigarrow \begin{cases} \mathbb{E}_{(\boldsymbol{x},\varepsilon)}[q^a] = \overline{\boldsymbol{w}}^{a\mathsf{T}} \mathbb{E}_{(\boldsymbol{x},\varepsilon)}\left[\boldsymbol{\psi}(\boldsymbol{x})\right] \overset{!}{=} 0, \\ \mathbb{E}_{(\boldsymbol{x},\varepsilon)}[q^a q^b] = \overline{\boldsymbol{w}}^{a\mathsf{T}} \Lambda \overline{\boldsymbol{w}}^b + \sigma^2 =: C^{ab}. \end{cases}$$

$$= \left(\det(\mathbb{1} + 2\beta C)\right)^{p/2} \cdot \mathrm{e}^{-\beta\lambda \sum_{a=1}^n \|\boldsymbol{w}_\star - \overline{\boldsymbol{w}}^a\|^2}.$$

$$\mathbb{E}_{\mathcal{D}}[Z^n] = \int \mathscr{D}[\overline{\boldsymbol{w}}] \cdot \mathrm{e}^{-\beta \sum_{a=1}^{n} \left(\lambda \|\boldsymbol{w}_\star - \overline{\boldsymbol{w}}^a\|^2 - t\mathcal{E}_g(\overline{\boldsymbol{w}}^a)\right)} \, \mathrm{e}^{-\frac{p}{2}\log\det(\mathbb{1}+2\beta C)},$$

where

$$= \int \mathscr{D}[C] \, \mathrm{e}^{-\frac{p}{2}\log\det(\mathbb{1}+2\beta C)} \prod_{a \geq a'} \delta\left(C^{ab} - \overline{\boldsymbol{w}}^{a\mathsf{T}} \Lambda \overline{\boldsymbol{w}}^b - \sigma^2\right)$$

$$\int \frac{\mathrm{d}\widehat{C}^{ab}}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}\widehat{C}^{ab}(C^{ab} - \overline{\boldsymbol{w}}^{a\mathsf{T}} \Lambda \overline{\boldsymbol{w}^b} - \sigma^2)}$$

$$= \int \frac{\mathscr{D}[C]\,\mathscr{D}[\widehat{C}]}{(2\pi)^{\frac{n(n+1)}{2}}} \mathrm{e}^{-\frac{p}{2}\log\det(\mathbb{1}+2\beta C) + \mathrm{i}\sum_{a \geq b} \widehat{C}^{ab}(C^{ab} - \overline{\boldsymbol{w}}^{a\mathsf{T}} \Lambda \overline{\boldsymbol{w}^b} - \sigma^2)},$$

with $\mathscr{D}[C] = \prod_{a \geq b} \mathrm{d}C^{ab}$ and $\mathscr{D}[\widehat{C}] = \prod_{a \geq b} \mathrm{d}\widehat{C}^{ab}$.

$$\mathbb{E}_{\mathcal{D}}[Z^n] = \int \frac{\mathscr{D}[C]\,\mathscr{D}[\widehat{C}]\,\mathscr{D}[\overline{\boldsymbol{w}}]}{(2\pi)^{\frac{n(n+1)}{2}}} \mathrm{e}^{-\beta\sum_a \left(\lambda\|\boldsymbol{w}_\star - \overline{\boldsymbol{w}}^a\|^2 - t\overline{\boldsymbol{w}}^{a\,\mathsf{T}}\Lambda\overline{\boldsymbol{w}}^a\right)}$$
$$\cdot\, \mathrm{e}^{-\frac{p}{2}\log\det(\mathbb{1}+2\beta C)+\mathrm{i}\sum_{a\geq b}\widehat{C}^{ab}(C^{ab}-\overline{\boldsymbol{w}}^{a\,\mathsf{T}}\Lambda\overline{\boldsymbol{w}}^b-\sigma^2)}$$
$$= \mathrm{e}^{-\frac{n(n+1)}{2}\log(2\pi)-\beta\lambda n\boldsymbol{w}_\star^{\mathsf{T}}\boldsymbol{w}_\star}$$
$$\times \int \mathscr{D}[C]\,\mathscr{D}[\widehat{C}]\ \mathrm{e}^{-pG_E}\ \mathrm{e}^{-G_S}\ \mathrm{e}^{\mathrm{i}\sum_{a\geq b}\widehat{C}^{ab}(C^{ab}-\sigma^2)}$$

Note that e^{-G_S} is a Gaussian integral in $\overline{\boldsymbol{w}}$:

$$\mathrm{e}^{-G_S} = \int \mathrm{d}\overline{\boldsymbol{w}}\cdot\exp\left\{-\beta\overline{\boldsymbol{w}}^{\mathsf{T}}\,X\,\overline{\boldsymbol{w}} + \beta\,\boldsymbol{\xi}^{\mathsf{T}}\,\overline{\boldsymbol{w}}\right\}.$$

$X^{ab} = (\lambda\mathbb{1} - t\Lambda + \frac{\mathrm{i}}{2\beta}\widehat{C}^{aa}\Lambda)\delta_{ab} + \frac{\mathrm{i}}{2\beta}\widehat{C}^{ab}\Lambda$

$\xi^a = 2\lambda\boldsymbol{w}_\star$

$[\overline{\boldsymbol{w}}^1;\ldots;\overline{\boldsymbol{w}}^n]$

We thus have

$$\mathrm{e}^{-G_S} = (\det(\beta X/\pi))^{-1/2}\, \mathrm{e}^{\frac{1}{4\beta}\boldsymbol{\xi}^{\mathsf{T}} X^{-1}\boldsymbol{\xi}}.$$

Combining all the above, we obtain the integral

$$\boxed{\mathbb{E}_{\mathcal{D}}\left[Z^n\right] = \mathrm{e}^{O(n)} \int \mathscr{D}[C]\, \mathscr{D}[\widehat{C}]\, \mathrm{e}^{-n\beta\, S[C,\widehat{C}]}},$$

$$S[C,\widehat{C}] = \frac{p}{2n\beta}\log\det(\mathbb{1}+2\beta C) + \frac{1}{2n\beta}\log\det X + \frac{1}{4n\beta^2}\boldsymbol{\xi}^{\mathsf{T}} X^{-1}\boldsymbol{\xi} - \frac{\mathrm{i}}{n\beta}\sum_{a\geq b}\widehat{C}^{ab}(C^{ab}-\sigma^2).$$

This requires **saddle-point method** to find the typical C and $\widehat{C}$!

By the **convexity** of the problem, we can assume the symmetric replica ansatz:

$$C_0 = C^{aa}, \qquad \widehat{C}_0 = \widehat{C}^{aa}, \qquad C_1 = C^{a\neq b}, \qquad \widehat{C}_1 = \widehat{C}^{a\neq b}.$$

Knowing that $n \to 0$ and $p \to \infty$,

$$\begin{aligned} S[C,\widehat{C}] = \Bigg[& \frac{p}{2\beta}\left(\log\left(1+2\beta(C_0-C_1)\right) + \frac{2\beta C_1}{1+2\beta(C_0-C_1)}\right) \\ & + \frac{1}{2\beta}\left(\log\det\left(\lambda\mathbb{1} + \left(\frac{\mathrm{i}}{\beta}\left(\widehat{C}_0 - \frac{1}{2}\widehat{C}_1\right) - t\right)\Lambda\right) + \mathrm{Tr}\left\{\frac{\mathrm{i}}{2\beta}\widehat{C}_1\Lambda\left(\lambda\mathbb{1} + \left(\frac{\mathrm{i}}{\beta}\left(\widehat{C}_0 - \frac{1}{2}\widehat{C}_1\right) - t\right)\Lambda\right)^{-1}\right\}\right) \\ & + \frac{1}{4\beta^2}\boldsymbol{w}_\star^{\mathsf{T}}\left(\lambda\mathbb{1} + \left(\frac{\mathrm{i}}{\beta}\left(\widehat{C}_0 - \frac{1}{2}\widehat{C}_1\right) - t\right)\Lambda\right)^{-1}\boldsymbol{w}_\star \\ & - \frac{\mathrm{i}}{\beta}\left(\widehat{C}_0(C_0-\sigma^2) - \frac{1}{2}\widehat{C}_1(C_1-\sigma^2)\right)\Bigg] + O(n). \end{aligned}$$

The saddle point $(C^*, \widehat{C}^*)$ then satisfies

$$0 = \frac{\partial S}{\partial C_0} = \frac{\partial S}{\partial C_1} = \frac{\partial S}{\partial \widehat{C}_0} = \frac{\partial S}{\partial \widehat{C}_1}.$$

By defining

$$\kappa(t) := \lambda\left(1 + 2\beta(C_0^* - C_1^*)\right),$$

one obtains

$$\mathbb{E}_{\mathcal{D}}[Z^n] \approx \mathrm{e}^{O(n)}\mathrm{e}^{-n\beta S[C^*,\widehat{C}^*]}$$

$$\Rightarrow \mathbb{E}_{\mathcal{D}}[\log Z] = \lim_{n\to 0}\frac{1}{n}\left(O(n) - n\beta S[C^*,\widehat{C}^*]\right)$$

$$\Rightarrow \mathcal{E}_g = \lim_{\beta\to\infty}\lim_{n\to 0}\frac{\partial}{\partial t}S[C^*,\widehat{C}^*]\Big|_{t=0}$$

$$\stackrel{!}{=} \frac{\zeta}{1-\zeta}\left(\sigma^2 + \kappa^2\sum_{\rho}\frac{\eta_\rho w_{\star,\rho}^2}{(p\eta_\rho+\kappa)^2}\right) + \kappa^2\sum_{\rho}\frac{\eta_\rho w_{\star,\rho}^2}{(p\eta_\rho+\kappa)^2},$$

where $\kappa := \kappa(0)$ and $\kappa'(0) = \kappa^2\zeta/(1-\zeta)$. [5]

[5] I have checked till the last step.

$$\mathcal{E}_g = \underbrace{\frac{\zeta}{1-\zeta}\left(\sigma^2 + \kappa^2 \sum_\rho \frac{\eta_\rho w_{\star,\rho}^2}{(p\eta_\rho+\kappa)^2}\right)}_{\text{variance}} + \underbrace{\kappa^2 \sum_\rho \frac{\eta_\rho w_{\star,\rho}^2}{(p\eta_\rho+\kappa)^2}}_{\text{bias}}, \tag{13}$$

where[6]

$$\kappa := \kappa(0) = \lambda + \sum_\rho \frac{\kappa\eta_\rho}{p\eta_\rho + \kappa}, \ \zeta = \sum_\rho \frac{p\eta_\rho^2}{(\kappa + p\eta_\rho)^2}. \tag{14}$$

Note: The function $\kappa(t)$ actually has close relation with the spectrum of the empirical NTK.

[6]Note that if $f_\star \notin \mathcal{H}$, some additional terms need to be included: $\sigma^2 \mapsto \sigma^2 + \|\boldsymbol{a}_\star\|^2$ and $\mathcal{E}_g \mapsto \mathcal{E}_g + \|\boldsymbol{a}_\star\|^2$, where $\boldsymbol{a}_\star$ is the out-of-RKHS part of $f_\star$. See the paper [CBP21] for the full picture.

Consider a linear regression with $\psi_\rho : \mathbb{R}^D \to \mathbb{R}$, $\boldsymbol{x} \mapsto x_\rho$, $\eta_\rho = 1$, $D = 400$, $p = 100$, $\sigma = 0.5$, and $\boldsymbol{w}_\star$, $\boldsymbol{x}$'s are all normally distributed. Setting $\rho = 1, \ldots, n$, we have:

$$\kappa = \cancel{\lambda}^{0} + \sum_{\rho=1}^{n} \frac{\kappa \cancel{\eta_\rho}^{1}}{p \cancel{\eta_\rho}^{1} + \kappa} = \frac{n\kappa}{p + \kappa},$$

$$\kappa^2 - (n-p)\kappa = 0 \rightsquigarrow \kappa = \begin{cases} n - p, & n \geq p; \\ 0, & n < p. \end{cases}$$

Similarly,

$$\zeta = \begin{cases} p/n, & n \geq p; \\ n/p, & n < p. \end{cases}$$

Denote $\gamma = n/p =$ parameter / data,

$$\|\boldsymbol{w}_{\star,n}\|^2 = \sum_{\rho=1}^{n} w_{\star,\rho}^2 \text{ and } \|\boldsymbol{a}_{\star,n}\|^2 = \sum_{\rho=n+1}^{\infty} w_{\star,\rho}^2.$$

This example has closed form solution for the generalization error:

$$\mathcal{E}_g = \begin{cases} \dfrac{\gamma-1}{\gamma}\|\boldsymbol{w}_{\star,n}\|^2 + \dfrac{\sigma^2+\|\boldsymbol{a}_{\star,n}\|^2}{\gamma-1} + \|\boldsymbol{a}_{\star,n}\|^2, & \gamma \geq 1; \\ \dfrac{\gamma}{1-\gamma}\left(\sigma^2+\|\boldsymbol{a}_{\star,n}\|^2\right) + \|\boldsymbol{a}_{\star,n}\|^2, & 1 > \gamma \leq 0. \end{cases}$$

Example – Double Descent (3)

Run the code yourself! [link]

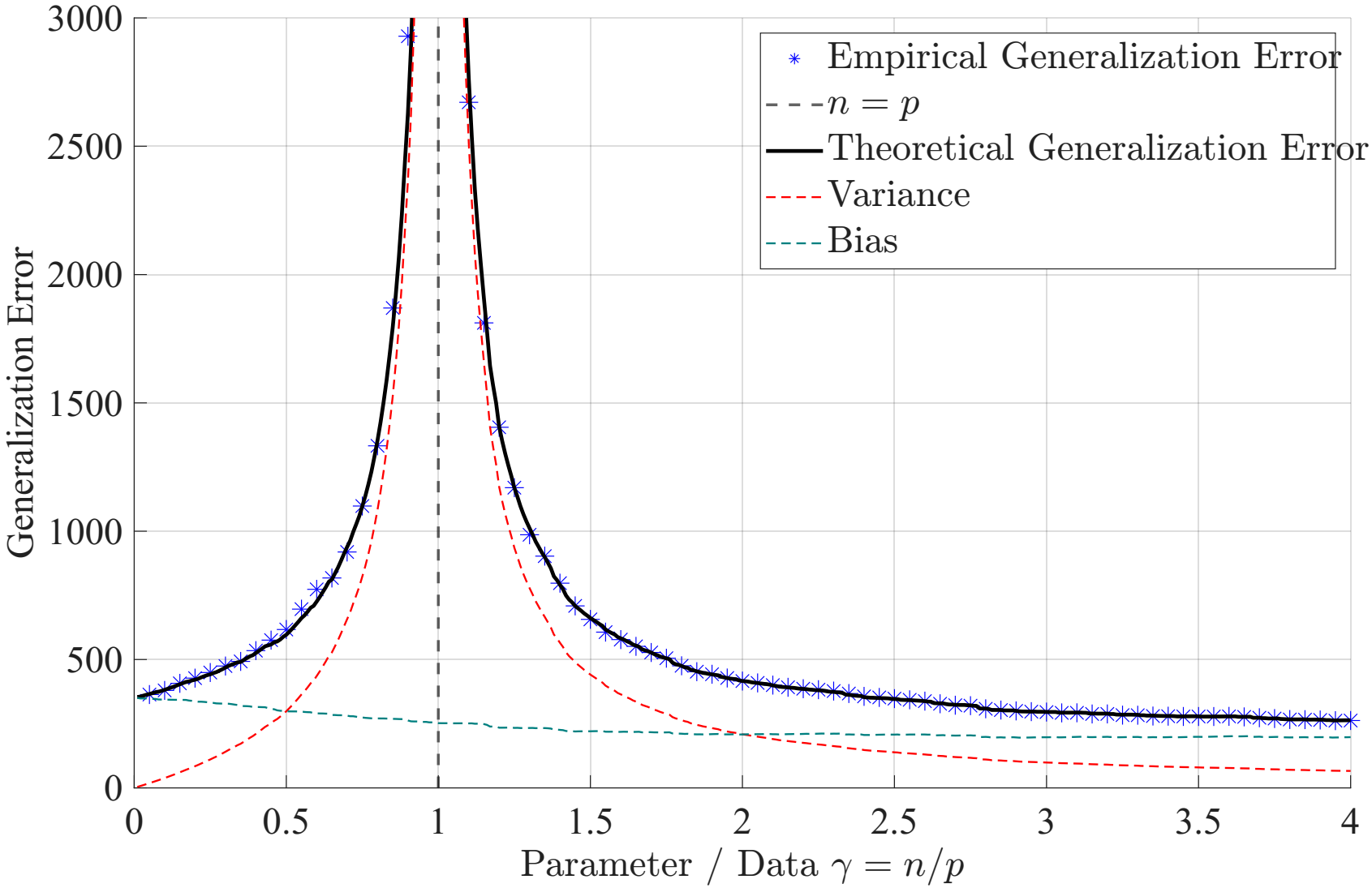
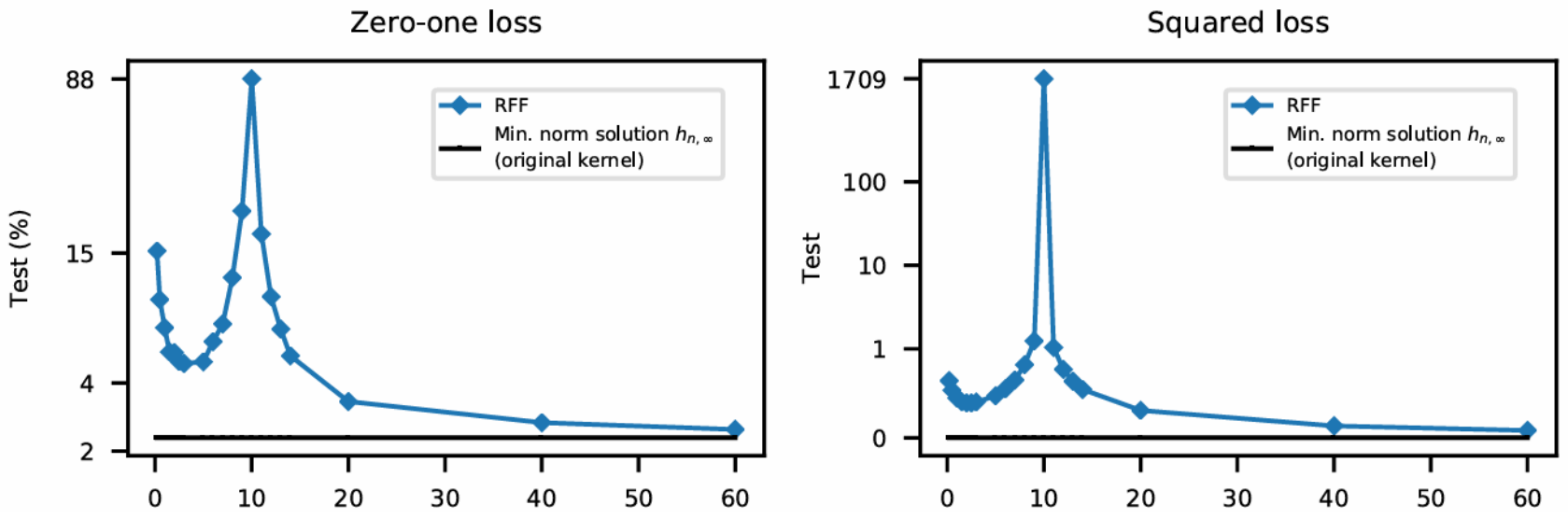


Table of Contents

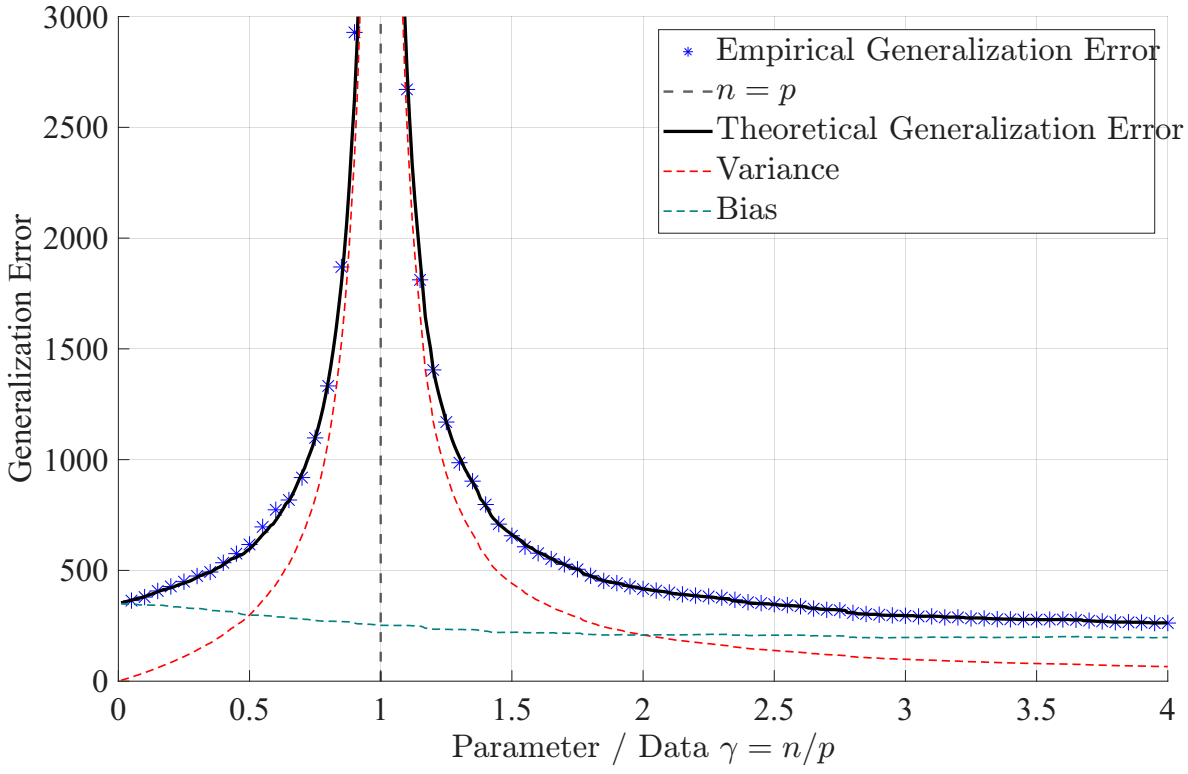
The double descent phenomena have been seen not only in regression or toy model, but also in real dataset classification (e.g., MNIST).



[Belkin et al. '19]

It is known that for neural networks (MLP) with large width, its behavior is similar to ridgeless kernel regression. This talk then linked it with the double descent in kernel regression.

The blowup is caused by variance of small-eigenvalue features from the empirical NTK.



Using the replica trick, we saw the non-monotonic behaviors come from the variance.

[AP20] Ben Adlam and Jeffrey Pennington. *The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization*. 2020. arXiv: 2008.06786 [stat.ML]. URL: https://arxiv.org/abs/2008.06786.

[S23] Roland Speicher. *High Dimensional Analysis: Random Matrices and Machine Learning*. Lecture at Saarland University. 2023. URL: https://rolandspeicher.com/lectures/course-on-high-dimensional-analysis-random-matrices-and-machine-learning-summer-term-2023/.

[CBP21] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. "Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks". In: *Nature Communications* 12.1 (May 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-23103-1. URL: http://dx.doi.org/10.1038/s41467-021-23103-1.