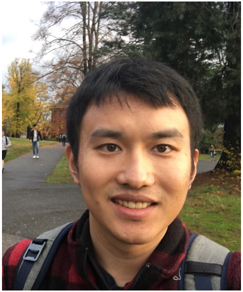
Overparametrized models: Linear theory and its limits

Andrea Montanari

Stanford University

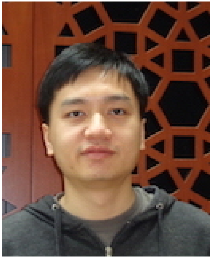
August 12, 2025

Behrooz Ghorbani



Song Mei

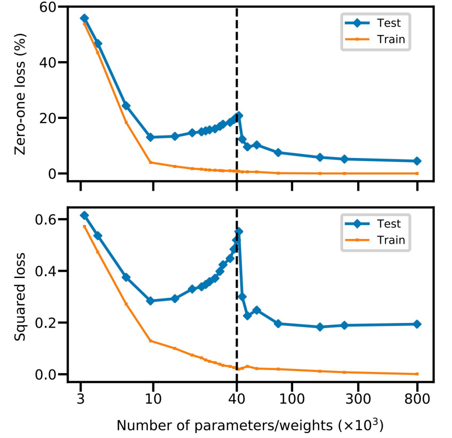
Theodor Misiakiewicz



Yiqiao Zhong

Chen Cheng

Surprise #1: N ... ctice

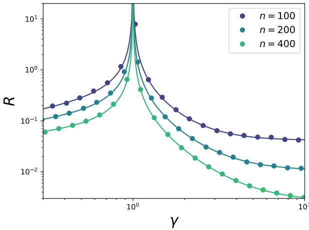


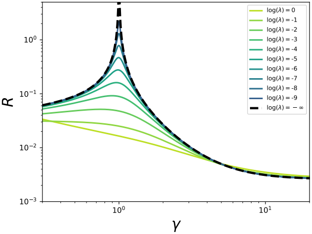
- Model complex enough to 'interpolate' random labels
- Despite this, does well on uncorrupted test samples
- Test error $\gg$ Train error ≈ 0

MNIST: 4,000 images 10 classes; 2-layers. Square loss. Belkin, Hsu, Ma, Mandal, 2018

Surprise #2: Completely general

Test error of ridge(less) regression vs $\gamma = p/n$





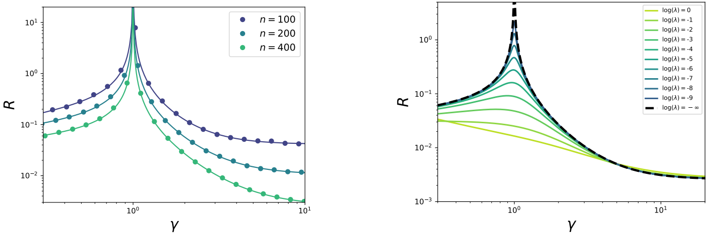
$$y_i = \langle \theta, x_i \rangle + \varepsilon_i\,, \quad x_i \sim N(0, I_d),$$
$$z_i = W^{\mathsf{T}} x_i + g_i\,, \quad W \in \mathbb{R}^{d \times p}\,, \quad g_i \sim N(0, I_d)\,.$$

- x: latent. z: features **Regress y vs z**

Belkin, Hsu, Ma, Mandal, 2019; **Hastie, Montanari, Rosset, Tibshirani, 2021**

Surprise #2: Completely general

Test error of ridge(less) regression vs $\gamma = p/n$



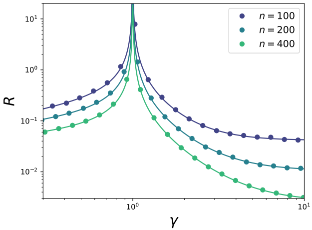
$$y_i = \langle \theta, x_i \rangle + \varepsilon_i\,, \qquad x_i \sim \mathsf{N}(0, \mathbf{I}_d),$$
$$z_i = W^{\mathsf{T}} x_i + g_i\,, \qquad W \in \mathbb{R}^{d\times p}, \quad g_i \sim \mathsf{N}(0, \mathbf{I}_d)\,.$$

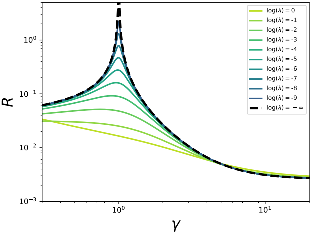
- x: latent. z: features Regress y vs z

Belkin, Hsu, Ma, Mandal, 2019; **Hastie, Montanari, Rosset, Tibshirani, 2021**

Surprise #2: Completely general

Test error of ridge(less) regression vs $\gamma = p/n$



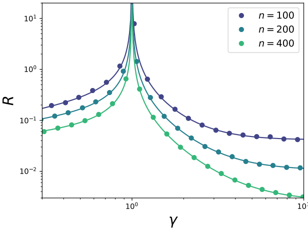


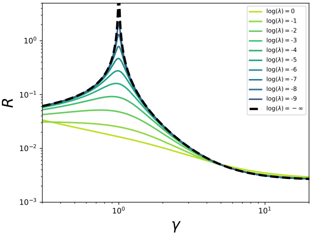
$$y_i = \langle \theta, x_i \rangle + \varepsilon_i\,, \qquad x_i \sim \mathsf{N}(0, \mathbf{I}_d),$$
$$z_i = W^{\mathsf{T}} x_i + g_i\,, \qquad W \in \mathbb{R}^{d \times p}, \quad g_i \sim \mathsf{N}(0, \mathbf{I}_d)\,.$$

- x: latent. z: features **Regress y vs z**

Belkin, Hsu, Ma, Mandal, 2019; **Hastie, Montanari, Rosset, Tibshirani, 2021**

Surprise #2: Completely general





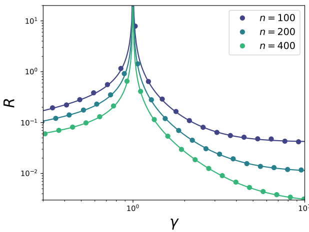
$$y_i = \langle \theta, x_i \rangle + \varepsilon_i\,, \qquad x_i \sim \mathsf{N}(0, \mathbf{I}_d),$$
$$z_i = W^{\mathsf{T}} x_i + g_i\,, \quad W \in \mathbb{R}^{d\times p}\,, \quad g_i \sim \mathsf{N}(0, \mathbf{I}_d),$$

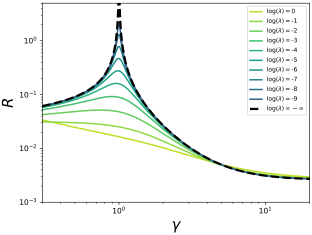
Equivalent description

$$y_i = \langle \beta, z_i \rangle + \tilde{\varepsilon}_i\,, \qquad z_i \sim \mathsf{N}(0, \boldsymbol{\Sigma}_d),$$
$$\boldsymbol{\Sigma} = WW^{\mathsf{T}} + \mathbf{I}_p\,, \quad \beta \in \mathrm{span}(W)\,.$$

General picture? Connection to neural nets?

Surprise #2: Completely general





$$
\begin{aligned}
& y_i = \langle \theta, x_i \rangle + \varepsilon_i\,, \qquad x_i \sim \mathsf{N}(0, \mathbf{I}_d), \\
& z_i = \boldsymbol{W}^{\mathsf{T}} x_i + \mathbf{g}_i\,, \quad \boldsymbol{W} \in \mathbb{R}^{d \times p}\,, \quad \mathbf{g}_i \sim \mathsf{N}(0, \mathbf{I}_d),
\end{aligned}
$$

Equivalent description

$$
\begin{aligned}
& y_i = \langle \boldsymbol{\beta}, z_i \rangle + \tilde{\varepsilon}_i\,, \qquad z_i \sim \mathsf{N}(0, \boldsymbol{\Sigma}_d), \\
& \boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^{\mathsf{T}} + \mathbf{I}_p\,, \quad \boldsymbol{\beta} \in \mathrm{span}(\boldsymbol{W})\,.
\end{aligned}
$$

General picture? Connection to neural nets?

Outline

Examples of linear regression

Setting

- Data

$$(y_1, z_1),\ (y_2, z_2),\ \dots,\ (y_n, z_n)$$
$$y_i \in \mathbb{R}: \qquad \text{'label'}$$
$$z_i \in \mathbb{R}^p: \qquad \text{'features' vector'}.$$

- Distribution

$$y_i = \langle \beta, z_i \rangle + \varepsilon_i\,, \quad \mathbb{E}(\varepsilon_i) = 0,\ \ \mathbb{E}(\varepsilon_i^2) = \tau^2$$

Potentially $p = \infty$

Setting

- Data

$$
\begin{aligned}
&(y_1, \boldsymbol{z}_1),\ (y_2, \boldsymbol{z}_2),\ \dots,\ (y_n, \boldsymbol{z}_n)\\
&y_i \in \mathbb{R}: \qquad \text{`label'}\\
&\boldsymbol{z}_i \in \mathbb{R}^p: \qquad \text{`features' vector'}\,.
\end{aligned}
$$

- Distribution

$$
y_i = \langle \boldsymbol{\beta}, \boldsymbol{z}_i \rangle + \varepsilon_i\,, \quad \mathbb{E}(\varepsilon_i) = 0,\ \ \mathbb{E}(\varepsilon_i^2) = \tau^2
$$

Potentially $p = \infty$

Ridge regression

$$\hat{\boldsymbol{\beta}}(\lambda) := \arg\min_{\mathbf{b}} \left\{ \|\mathbf{y} - \mathbf{Z}\mathbf{b}\|^2 + \lambda \|\mathbf{b}\|^2 \right\}, \quad \mathbf{Z} := \begin{bmatrix} \text{—}\mathbf{z}_1\text{—} \\ \vdots \\ \text{—}\mathbf{z}_n\text{—} \end{bmatrix} \in \mathbb{R}^{p \times p}$$

$$\hat{\beta}(\lambda) := \frac{1}{n} Z^\top \left(K_n + (\lambda/n) I_n \right)^{-1} y\,,$$

$$K_n := \frac{1}{n} Z Z^\top .$$

Ridge regression

$$\hat{\boldsymbol{\beta}}(\lambda) := \arg\min_{\mathbf{b}} \left\{ \|\mathbf{y} - \mathbf{Z}\mathbf{b}\|^2 + \lambda \|\mathbf{b}\|^2 \right\}, \quad \mathbf{Z} := \begin{bmatrix} -\mathbf{z}_1- \\ \vdots \\ -\mathbf{z}_n- \end{bmatrix} \in \mathbb{R}^{p \times p}$$

$$\hat{\boldsymbol{\beta}}(\lambda) := \frac{1}{n}\mathbf{Z}^\mathsf{T}\big(\mathbf{K}_n + (\lambda/n)\mathbf{I}_n\big)^{-1}\mathbf{y}\,,$$
$$\mathbf{K}_n := \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}\,.$$

Test error

$$\widehat{\boldsymbol{\beta}}(\lambda) := \arg\min_{\mathbf{b}} \left\{ \|\mathbf{y} - \mathbf{Z}\mathbf{b}\|^2 + \lambda \|\mathbf{b}\|^2 \right\}.$$

$$\mathscr{R}_{\mathbf{Z}}(\lambda) := \mathbb{E}_{\text{new}}\left\{ \left(y_{\text{new}} - \langle \widehat{\boldsymbol{\beta}}(\lambda), \boldsymbol{z}_{\text{new}} \rangle \right)^2 \right\} - \underbrace{\mathbb{E}_{\text{new}}\left\{ \left(y_{\text{new}} - \langle \boldsymbol{\beta}, \boldsymbol{z}_{\text{new}} \rangle \right)^2 \right\}}_{\text{Bayes}}$$

$$= \|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}^2 \qquad \boldsymbol{\Sigma} := \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}]$$

Test error

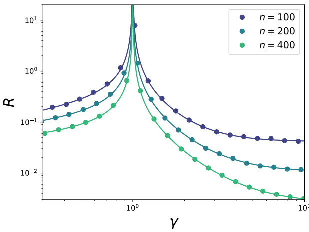
$$\widehat{\boldsymbol{\beta}}(\lambda) := \arg\min_{\mathbf{b}} \left\{ \|\mathbf{y} - \mathbf{Z}\mathbf{b}\|^2 + \lambda \|\mathbf{b}\|^2 \right\}.$$

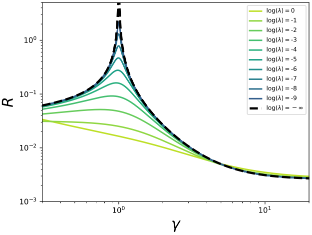
$$\mathscr{R}_{\mathbf{Z}}(\lambda) := \mathbb{E}_{\text{new}}\left\{ \left(y_{\text{new}} - \langle \widehat{\boldsymbol{\beta}}(\lambda), \boldsymbol{z}_{\text{new}} \rangle \right)^2 \right\} - \underbrace{\mathbb{E}_{\text{new}}\left\{ \left(y_{\text{new}} - \langle \boldsymbol{\beta}, \boldsymbol{z}_{\text{new}} \rangle \right)^2 \right\}}_{\text{Bayes}}$$

$$= \|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}\|_{\boldsymbol{\Sigma}}^2 \qquad \boldsymbol{\Sigma} := \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}]$$

Examples

Example #1: Well-concentrated covariates





- $\boldsymbol{z}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_i$, $\mathbb{E}\{\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}\} = \mathbf{I}_p$.
- Concentration properties for $\boldsymbol{x}_i$:

Either: Independent sub-Gaussian coordinates
or: Log-Sobolev
or: ...

Dicker, 2016; Advani, Saxe, 2017; Dobriban, Wager, 2018; Hastie, Montanari, Rosset, Tibshirani, 2021; Cheng, Montanari, 2021; ...

Example #2: Kernel Ridge Regression

Data

$$\boldsymbol{x}_i \sim \mathbb{P} \in \mathscr{P}(\mathbb{R}^d)\,,$$
$$y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i\,, \quad f_* \in \mathcal{H} \subseteq L^2(\mathbb{R}^d;\mathbb{P}),$$

Function space view

$$\hat{f}_\lambda = \mathrm{argmin}_f \Big\{ \sum_{i=1}^{n} \big(y_i - f(\boldsymbol{x}_i)\big)^2 + \lambda \|f\|_{\mathcal{H}}^2 \Big\}\,,$$

$\mathcal{H} =$ Reproducing Kernel Hilbert Space. Kernel K

Example #2: Kernel Ridge Regression (Take 2)

Data

$$\boldsymbol{x}_i \sim \mathbb{P} \in \mathscr{P}(\mathbb{R}^d)\,,$$
$$y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i\,, \quad f_* \in \mathcal{H} \subseteq L^2(\mathbb{P}), \;\; \text{(Hilbert space)}$$

Featurization map

$$\boldsymbol{\phi} : \mathbb{R}^d \to \mathcal{H}_0\,, \quad \boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x})\,.$$

Feature space view $(p = \infty)$

$$\hat{f}_\lambda(\boldsymbol{x}) = \langle \hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\phi}(\boldsymbol{x})\rangle\,, \qquad \boldsymbol{z}_i = \boldsymbol{\phi}(\boldsymbol{x}_i)$$

$$\hat{\boldsymbol{\beta}}_\lambda = \mathrm{argmin}_{\boldsymbol{b}} \Big\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_{\mathcal{H}_0}^2 \Big\}\,,$$

$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{\phi}(\boldsymbol{x}_2)\rangle_{\mathcal{H}_0}\,.$

Example #2: Kernel Ridge Regression (Take 2)

Data

$$\boldsymbol{x}_i \sim \mathbb{P} \in \mathscr{P}(\mathbb{R}^d)\,,$$
$$y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i\,, \quad f_* \in \mathcal{H} \subseteq L^2(\mathbb{P}), \quad \text{(Hilbert space)}$$

Featurization map

$$\boldsymbol{\phi} : \mathbb{R}^d \to \mathcal{H}_0\,, \quad \boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x})\,.$$

Feature space view $(p = \infty)$

$$\hat{f}_\lambda(\boldsymbol{x}) = \langle \hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\phi}(\boldsymbol{x}) \rangle\,, \qquad \boldsymbol{z}_i = \boldsymbol{\phi}(\boldsymbol{x}_i)$$

$$\hat{\boldsymbol{\beta}}_\lambda = \operatorname{argmin}_{\boldsymbol{b}} \Big\{ \big\| \boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b} \big\|_2^2 + \lambda \|\boldsymbol{b}\|_{\mathcal{H}_0}^2 \Big\}\,,$$

$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{\phi}(\boldsymbol{x}_2) \rangle_{\mathcal{H}_0}$.

Example #3: Random Features Regression

Data

$$\begin{aligned} \boldsymbol{x}_i &\sim \mathbb{P} \in \mathscr{P}(\mathbb{R}^d)\,, \\ y_i &= f_*(\boldsymbol{x}_i) + \varepsilon_i\,, \quad f_* \in \mathcal{H} \subseteq L^2(\mathbb{P}), \end{aligned}$$

Two-layer network with random first layer ($p = N$)

$$\hat{f}(\boldsymbol{x}; \mathbf{b}) = \sum_{i=1}^{N} b_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)\,, \qquad \boldsymbol{w}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}),$$

$$\hat{\mathbf{b}}_\lambda = \mathrm{argmin}_{\mathbf{b}} \Big\{ \sum_{i=1}^{N} \big(y_i - f(\boldsymbol{x}_i; \mathbf{b})\big)^2 + \lambda \|\mathbf{b}\|_2^2 \Big\}\,.$$

Neal, 1996; Balcan, Blum, Vempala, 2006; Rahimi, Recht, 2008

Example #3: Random Features Regression (Take 2)

Data

$$x_i \sim \mathbb{P} \in \mathscr{P}(\mathbb{R}^d)\,,$$
$$y_i = f_*(x_i) + \varepsilon_i\,, \quad f_* \in \mathcal{H} \subseteq L^2(\mathbb{P}),$$

Two-layer network with random first layer $(p = N)$

$$\hat{f}(x; \mathbf{b}) = \sum_{i=1}^{N} b_i\, \sigma(\langle \boldsymbol{w}_i, x\rangle)\,, \qquad \boldsymbol{w}_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}),$$

$$\boldsymbol{z}_i = \boldsymbol{\phi}_{\boldsymbol{W}}(x_i) := \Big(\sigma(\langle \boldsymbol{w}_1, x_i\rangle), \sigma(\langle \boldsymbol{w}_2, x_i\rangle), \dots, \sigma(\langle \boldsymbol{w}_N, x_i\rangle)\Big)^{\mathsf{T}},$$

$$\hat{\mathbf{b}}_\lambda = \mathrm{argmin}_{\mathbf{b}} \Big\{ \|\mathbf{y} - \mathbf{Z}\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_2^2 \Big\}\,.$$

Neal, 1996; Balcan, Blum, Vempala, 2006; Rahimi, Recht, 2008

Example #4: (Neural) Tangent Regression

- Parametric model $\alpha f(\,\cdot\,;\theta) : \mathbb{R}^d \to \mathbb{R}$ (parameters: (α, θ))
- Linearize around SGD initialization θ^0

$$\alpha f(x;\theta^0 + \alpha^{-1}\mathbf{b}) = \alpha\, f(x;\theta^0) + \langle \mathbf{b}, \nabla_\theta f(x;\theta^0)\rangle + O(\alpha^{-1})$$

$$= \text{const.} + \underbrace{\langle \mathbf{b}, \nabla_\theta f(x;\theta^0)\rangle}_{f_{NT}(x;\mathbf{b})} + O(\alpha^{-1})$$

Jacot, Gabriel, Hongler, 2018; Du, Zhai, Poczos, Singh 2018; Allen-Zhu, Li, Song 2018; Chizat, Bach, 2019; Ghorbani, Mei, Misiakiewicz, M, 2019; Arora, Du, Hu, Li, Salakhutdinov, Wang, 2019; Oymak, Soltanolkotabi, 2019;...

Example #4: (Neural) Tangent Regression

Two-layer neural net

$$f(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{W}) = \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle)$$

$$f_{\mathsf{NT}}(\boldsymbol{x}; \overline{\boldsymbol{b}}, \boldsymbol{b}) = \sum_{j=1}^{N} \langle \boldsymbol{b}_j, \boldsymbol{x} \rangle \sigma'(\langle \boldsymbol{w}_j^0, \boldsymbol{x} \rangle) + \sum_{j=1}^{N} \overline{b}_j \sigma(\langle \theta_j^0, \boldsymbol{x} \rangle) \,.$$

Example #4: (Neural) Tangent Regression

Two-layer neural net

$$f(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{W}) = \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle)$$

$$f_{\rm NT}(\boldsymbol{x}; \overline{\boldsymbol{b}}, \boldsymbol{b}) = \sum_{j=1}^{N} \langle \boldsymbol{b}_j, \boldsymbol{x} \rangle \sigma'(\langle \boldsymbol{w}_j^0, \boldsymbol{x} \rangle) + \cancel{\sum_{j=1}^{N} \overline{b}_j \sigma(\langle \boldsymbol{w}_j^0, \boldsymbol{x} \rangle)}\,.$$

Two-layer network with random first layer ($p = Nd$)

$$z_i = \phi_W(\boldsymbol{x}_i) := \left(\boldsymbol{x}_i^{\mathsf{T}} \sigma'(\langle \boldsymbol{w}_1^0, \boldsymbol{x}_i \rangle), \boldsymbol{x}_i^{\mathsf{T}} \sigma'(\langle \boldsymbol{w}_2^0, \boldsymbol{x}_i \rangle), \dots, \boldsymbol{x}_i^{\mathsf{T}} \sigma'(\langle \boldsymbol{w}_N^0, \boldsymbol{x}_i \rangle)\right)^{\mathsf{T}},$$

$$\hat{\boldsymbol{b}}_\lambda = \operatorname{argmin}_{\boldsymbol{b}} \left\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_2^2 \right\}.$$

Example #4: (Neural) Tangent Regression

Two-layer neural net

$$f(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{W}) = \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle)$$

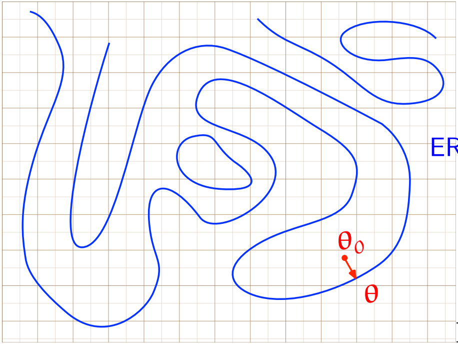
$$f_{\mathrm{NT}}(\boldsymbol{x}; \overline{\boldsymbol{b}}, \boldsymbol{b}) = \sum_{j=1}^{N} \langle \boldsymbol{b}_j, \boldsymbol{x} \rangle \sigma'(\langle \boldsymbol{w}_j^0, \boldsymbol{x} \rangle) + \cancel{\sum_{j=1}^{N} \overline{b}_j \sigma(\langle \boldsymbol{w}_j^0, \boldsymbol{x} \rangle)}\,.$$

Two-layer network with random first layer $(p = Nd)$

$$\boldsymbol{z}_i = \boldsymbol{\phi}_{\boldsymbol{W}}(\boldsymbol{x}_i) := \left(\boldsymbol{x}_i^{\mathsf{T}} \sigma'(\langle \boldsymbol{w}_1^0, \boldsymbol{x}_i \rangle), \boldsymbol{x}_i^{\mathsf{T}} \sigma'(\langle \boldsymbol{w}_2^0, \boldsymbol{x}_i \rangle), \dots, \boldsymbol{x}_i^{\mathsf{T}} \sigma'(\langle \boldsymbol{w}_N^0, \boldsymbol{x}_i \rangle) \right)^{\mathsf{T}},$$

$$\hat{\boldsymbol{b}}_\lambda = \mathrm{argmin}_{\boldsymbol{b}} \left\{ \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{b}\|_2^2 \right\}.$$

Heuristic connection with neural nets



$\mathsf{ERM}_{n,d} := \{\theta : \widehat{\mathscr{R}}_n(\theta) = 0\}$

$\mathbb{R}^p$ (parameter space)

Connection with 'lazy training' of neural nets

'Zero' initialization

$$f(\boldsymbol{x};\boldsymbol{w}) = \frac{\gamma}{\sqrt{N}} \sum_{j=1}^{N} b_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle)\,,$$

$$b_1 = \cdots = b_{N/2} = +1\,, \;\; b_{N/2+1} = \cdots = b_N = -1\,, \quad \text{not evolving}\,,$$

$$(\boldsymbol{w}_j : j \le N/2) \sim \mathrm{Unif}(\mathbb{S}^{d-1})\,, \quad \boldsymbol{w}_{N/2+j} = \boldsymbol{w}_j\,.$$

$f(x; w) = \frac{\gamma}{\sqrt{N}} \sum_{j=1}^{N} b_j \sigma(\langle w_j, x \rangle)$

Theorem (Chizat, Bach, 2018; Oymak, Soltanolkotabi, 2020)

Assume σ *is generic,* $x_i \sim N(0, \mathbf{I}_d)$*,* y_i *subgaussian. Then for any* $C_0 > 0$ *there exists* $C > 0$ *such that the following happens. If*

$$Nd \geq C\, d \log d\,, \quad \gamma \geq C\sqrt{\frac{n^2}{Nd}}\,,$$

then, with probability at least $1 - 2\exp(-n/C)$*:*

1. *Exponential convergence of gradient flow. For all* $t \geq 0$*,*

$$\widehat{\mathscr{R}}_n(\theta_t) \leq \widehat{\mathscr{R}}_n(\theta_0)\, e^{-\lambda_* t}\,, \qquad \lambda_* := \gamma^2 (d/n)/C$$

2. *Neural tangent model is a good approximation*

$$\left\| f(\,\cdot\,;\theta_t) - f_{NT}(\,\cdot\,;\theta_t^{NT}) \right\|_{L^2(\mathbb{P})} \leq C\left\{ \frac{1}{\gamma}\sqrt{\frac{n^2}{Nd}} + \frac{1}{\gamma^2}\sqrt{\frac{n^5}{Nd^4}} \right\}.$$

$f(\boldsymbol{x};\boldsymbol{w}) = \frac{\gamma}{\sqrt{N}} \sum_{j=1}^{N} b_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle)$

Theorem (Chizat, Bach, 2018; Oymak, Soltanolkotabi, 2020)

Assume σ *is generic,* $x_i \sim N(0, \mathbf{I}_d)$*,* y_i *subgaussian. Then for any* $C_0 > 0$ *there exists* $C > 0$ *such that the following happens. If*

$$Nd \geq C\, d \log d\,, \quad \gamma \geq C\sqrt{\frac{n^2}{Nd}}\,,$$

then, with probability at least $1 - 2\exp(-n/C)$*:*

1. *Exponential convergence of gradient flow. For all* $t \geq 0$*,*

$$\widehat{\mathscr{R}}_n(\theta_t) \leq \widehat{\mathscr{R}}_n(\theta_0)\, e^{-\lambda_* t}\,, \qquad \lambda_* := \gamma^2 (d/n)/C$$

2. *Neural tangent model is a good approximation*

$$\left\| f(\,\cdot\,;\theta_t) - f_{\mathrm{NT}}(\,\cdot\,;\theta_t^{\mathrm{NT}}) \right\|_{L^2(\mathbb{P})} \leq C\left\{ \frac{1}{\gamma}\sqrt{\frac{n^2}{Nd}} + \frac{1}{\gamma^2}\sqrt{\frac{n^5}{Nd^4}} \right\}.$$

$f(\boldsymbol{x};\boldsymbol{w}) = \frac{\gamma}{\sqrt{N}}\sum_{j=1}^{N} b_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle)$

Theorem (Chizat, Bach, 2018; Oymak, Soltanolkotabi, 2020)

Assume σ is generic, $x_i \sim N(0, \mathbf{I}_d)$, y_i subgaussian. Then for any $C_0 > 0$ there exists $C > 0$ such that the following happens. If

$$Nd \geq C\, d \log d\,, \quad \gamma \geq C\sqrt{\frac{n^2}{Nd}}\,,$$

then, with probability at least $1 - 2\exp(-n/C)$:

1. *Exponential convergence of gradient flow. For all $t \geq 0$,*

$$\widehat{\mathscr{R}}_n(\theta_t) \leq \widehat{\mathscr{R}}_n(\theta_0)\, e^{-\lambda_* t}\,, \qquad \lambda_* := \gamma^2 (d/n)/C$$

2. *Neural tangent model is a good approximation*

$$\left\| f(\,\cdot\,;\theta_t) - f_{\rm NT}(\,\cdot\,;\theta_t^{\rm NT})\right\|_{L^2(\mathbb{P})} \leq C\left\{\frac{1}{\gamma}\sqrt{\frac{n^2}{Nd}} + \frac{1}{\gamma^2}\sqrt{\frac{n^5}{Nd^4}}\right\}.$$

A general formula

Assumptions: $p = \infty$ $\quad (\boldsymbol{\beta}, \boldsymbol{z}_i \in \text{Hilbert})$

$$y_i = \langle \boldsymbol{\beta}, \boldsymbol{z}_i \rangle + \varepsilon_i\,, \quad \mathbb{E}(\varepsilon_i) = \mathbb{E}(\varepsilon_i \boldsymbol{z}_i) = 0, \;\; \mathbb{E}(\varepsilon_i^2) = \tau^2$$

1. $\mathrm{Tr}(\boldsymbol{\Sigma}) < \infty$ and (wlog) $\|\boldsymbol{\Sigma}\| = 1$.
2. $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\beta}\| < \infty$.
3. $(\sigma_i)_{i \geq 1}$: ordered eigenvalues of $\boldsymbol{\Sigma}$. For all $1 \leq k \leq n$:

$$\sum_{l=k}^{\infty} \sigma_l \leq d_{\boldsymbol{\Sigma}} \sigma_k\,.$$

4. $\boldsymbol{u}_i := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}_i$ satisfies a Hanson-Wright inequality.
   Implied by any of the following:
   (a) Independent sub-Gaussian coordinates.
   (b) Concentration $1-$Lipschitz convex function

General result

Theorem (Cheng, M, 2022)

- *$\mathscr{R}_{\mathsf{Z}}(\lambda)$ be the test error of ridge regression (conditional on Z)*
- *$\mathscr{R}^{s}_{n}(\lambda)$ (non-random) test error in the equivalent sequence model.*

Then

$$\mathscr{R}_{\mathsf{Z}}(\lambda) = (1 + \mathrm{err}_n)\mathscr{R}^{s}_{n}(\lambda)\,,$$

where err_n *is small with high probability, provided ...*

- Multiplicative error!
- All that follows will be 'special cases'
  ['Historically,' we proved them independently/earlier]
- $\mathscr{R}^{s}_{n}(\lambda)$ = Risk in a sequence model equivalent

General result

Theorem (Cheng, M, 2022)

- $\mathscr{R}_{\mathsf{Z}}(\lambda)$ *be the test error of ridge regression (conditional on* Z*)*
- $\mathscr{R}^{s}_{n}(\lambda)$ *(non-random) test error in the equivalent sequence model.*

Then

$$\mathscr{R}_{\mathsf{Z}}(\lambda) = (1 + \mathsf{err}_n)\mathscr{R}^{s}_{n}(\lambda)\,,$$

where err_n *is small with high probability, provided ...*

- Multiplicative error!
- All that follows will be 'special cases'
  ['Historically,' we proved them independently/earlier]
- $\mathscr{R}^{s}_{n}(\lambda)$ = Risk in a sequence model equivalent

General result

Theorem (Cheng, M, 2022)

- $\mathscr{R}_{\mathsf{Z}}(\lambda)$ *be the test error of ridge regression (conditional on* Z*)*
- $\mathscr{R}^{s}_{n}(\lambda)$ *(non-random) test error in the equivalent sequence model.*

Then

$$\mathscr{R}_{\mathsf{Z}}(\lambda) = (1 + \mathsf{err}_n)\mathscr{R}^{s}_{n}(\lambda)\,,$$

where err_n *is small with high probability, provided ...*

- Multiplicative error!
- All that follows will be 'special cases'
  ['Historically,' we proved them independently/earlier]
- $\mathscr{R}^{s}_{n}(\lambda)$ = Risk in a sequence model equivalent

General result

Theorem (Cheng, M, 2022)

- *$\mathscr{R}_{\mathsf{Z}}(\lambda)$ be the test error of ridge regression (conditional on Z)*
- *$\mathscr{R}^{\mathsf{s}}_{\mathsf{n}}(\lambda)$ (non-random) test error in the equivalent sequence model.*

Then

$$\mathscr{R}_{\mathsf{Z}}(\lambda) = (1 + \mathsf{err}_{\mathsf{n}})\mathscr{R}^{\mathsf{s}}_{\mathsf{n}}(\lambda)\,,$$

where $\mathsf{err}_{\mathsf{n}}$ is small with high probability, provided …

- Multiplicative error!
- All that follows will be 'special cases'
  ['Historically,' we proved them independently/earlier]
- $\mathscr{R}^{\mathsf{s}}_{\mathsf{n}}(\lambda)$ = Risk in a sequence model equivalent

Key quantity in the sequence model equivalent

Effective regularization $\lambda_\star(\lambda)$:

$$n - \frac{\lambda}{\lambda_\star} = \sum_{i \geq 1} \frac{\sigma_i}{\sigma_i + \lambda_\star}$$

'Self-induced' regularization

$$\lim_{\lambda \to 0+} \lambda_\star(\lambda) = \lambda_\star(0) > 0\,.$$

One key quantity controlling err_n $\quad (\lambda = 0+)$

$$\chi_n := \frac{\sigma_{\lfloor \eta n \rfloor} d_{\Sigma} \log^2(d_{\Sigma})}{\kappa n \lambda_\star(0)}, \qquad d_{\Sigma} := \max_{k \le n} \sum_{l \ge k} \frac{\sigma_l}{\sigma_k}.$$

- $\lambda_\star \asymp \sigma_{\lfloor cn \rfloor}$.
- Need $\chi_n \le n^{1/3-\varepsilon}$
- Need $d_{\Sigma} \le n^{4/3-\varepsilon}$

The actual theorem

1. The ratio between effective dimension and regularization parameter:

$$\chi_n(\lambda) := 1 + \frac{\sigma_{\lfloor \eta n \rfloor} \mathsf{d}_{\boldsymbol{\Sigma}} \log^2(\mathsf{d}_{\boldsymbol{\Sigma}})}{\lambda}. \tag{20}$$

   Here η is a constant that only depends on $\mathsf{C}_{\boldsymbol{x}}$, and hence we will leave it implicit.

2. The ratio between regularization and effective regularization

$$\kappa := \min\left(\frac{\lambda}{n\lambda_\star}; 1 - \frac{\lambda}{n\lambda_\star}\right) > 0. \tag{21}$$

3. For a positive semi-definite operator $\boldsymbol{Q}$, define the modified population resolvent:

$$\mathscr{R}_0(\mu_0, \mu; \boldsymbol{Q}) := \operatorname{Tr}\left(\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{Q} \boldsymbol{\Sigma}^{\frac{1}{2}} (\mu_0 \boldsymbol{I} + \mu \boldsymbol{\Sigma})^{-1}\right). \tag{22}$$

   Letting $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}$, $\|\boldsymbol{\theta}\| < \infty$, we consider the ratio

$$\rho(\lambda) := \frac{\mathscr{R}_0(\lambda_\star, 1; \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} / \|\boldsymbol{\theta}\|^2)}{\mathscr{R}_0(\lambda_\star, 1; \boldsymbol{I})} \in (0, 1]. \tag{23}$$

We next present our master theorem for ridge regression: its proof is postponed to Section 6.

Theorem 1 (Ridge regression)**.** *Under Assumption 1, for any positive integers k and D, there exist constants $\eta = \eta(\mathsf{C}_{\boldsymbol{x}}) \in (0, 1/2)$ and $\mathsf{C} = \mathsf{C}(\mathsf{C}_{\boldsymbol{x}}, D) > 0$ such that the following hold. Define $\chi_n(\lambda), \kappa, \rho(\lambda)$ as above (with $\eta = \eta(\mathsf{C}_{\boldsymbol{x}})$ in Eq. (20)).*

If it holds that

$$\chi_n(\lambda)^3 \log^2 n \le \mathsf{C} n \kappa^{4.5}, \qquad n^{-2D+1} = \mathcal{O}\left(\sqrt{\frac{\kappa^3 \log^2 n}{\max\{n, \lambda\}}}\right),$$

then for all $n = \Omega_{k,D}(1)$, with probability $1 - \mathcal{O}_k(n^{-D+1})$ we have:

1. ***Variance approximation.***

$$|\mathscr{V}_{\boldsymbol{X}}(\lambda) - \mathsf{V}_n(\lambda)| = \mathcal{O}_{k, \mathsf{C}_{\boldsymbol{x}}, D}\left(\frac{\chi_n(\lambda)^3 \log^2 n}{n^{1-\frac{1}{k}} \kappa^{9.5}}\right) \cdot \mathsf{V}_n(\lambda).$$

2. ***Bias approximation.*** *If we additionally have $\chi_n(\lambda)^3 \log^2 n \le \mathsf{C} n \kappa^{4.5} \sqrt{\rho(\lambda)}$ and $\lambda k n^{-\frac{1}{k}} \le n\kappa/2$, for all $n = \Omega_{k,D}(1)$, we have*

$$|\mathscr{B}_{\boldsymbol{X}}(\lambda) - \mathsf{B}_n(\lambda)| = \mathcal{O}_{k, \mathsf{C}_{\boldsymbol{x}}, D}\left(\frac{\lambda_\star(\lambda)^{k+1}}{n\kappa^3} + \frac{\chi_n(\lambda)^3 \log^2 n}{\sqrt{\rho(\lambda)} n^{1-\frac{1}{k}} \kappa^{8.5}}\right) \cdot \mathsf{B}_n(\lambda).$$

Remark 3.1 The condition $\|\beta\|_{...} < \infty$ in Assumption 1 amounts to requiring that the coef

Equivalent sequence model

$$\theta_i := \langle \boldsymbol{\beta}, \boldsymbol{v}_i \rangle\,, \qquad \boldsymbol{v}_i := i\text{-th eigenvectors of } \boldsymbol{\Sigma}$$

$$y_i^s = \sigma_i^{1/2}\theta_i + \frac{\omega}{\sqrt{n}} g_i\,, \quad (g_i)_{i\geq 1} \sim_{\text{iid}} \mathsf{N}(0,1)\,,$$

$$\hat{\theta}_i^s := \text{argmin}_{t\in\mathbb{R}}\left\{(y_i^s - \sigma_i^{1/2}t)^2 + \lambda_\star t^2\right\} = \frac{\sigma_i^{1/2}}{\sigma_i + \lambda_\star}\cdot y_i^s\,.$$

Effective noise level and regularization $\omega, \lambda_\star$

$$\omega^2 = \tau^2 + \underbrace{\mathbb{E}_{\mathbf{g}}\left\{\sum_{i\geq 1}\sigma_i(\hat{\theta}_i^s - \theta_i)^2\right\}}_{\mathscr{R}_n^s}, \qquad n - \frac{\lambda}{\lambda_\star} = \sum_{i\geq 1}\frac{\sigma_i}{\sigma_i + \lambda_\star}$$

Equivalent sequence model

$$\theta_i := \langle \boldsymbol{\beta}, \boldsymbol{v}_i \rangle\,, \qquad \boldsymbol{v}_i := i\text{-th eigenvectors of } \boldsymbol{\Sigma}$$

$$y_i^s = \sigma_i^{1/2}\theta_i + \frac{\omega}{\sqrt{n}} g_i\,, \quad (g_i)_{i\ge 1} \sim_{iid} \mathsf{N}(0,1)\,,$$

$$\hat{\theta}_i^s := \mathrm{argmin}_{t\in\mathbb{R}}\big\{(y_i^s - \sigma_i^{1/2}t)^2 + \lambda_\star t^2\big\} = \frac{\sigma_i^{1/2}}{\sigma_i + \lambda_\star}\cdot y_i^s\,.$$

Effective noise level and regularization $\omega, \lambda_\star$

$$\omega^2 = \tau^2 + \underbrace{\mathbb{E}_{\mathbf{g}}\Big\{\sum_{i\ge 1}\sigma_i(\hat{\theta}_i^s - \theta_i)^2\Big\}}_{\mathscr{R}_n^s}\,, \qquad n - \frac{\lambda}{\lambda_\star} = \sum_{i\ge 1}\frac{\sigma_i}{\sigma_i + \lambda_\star}$$

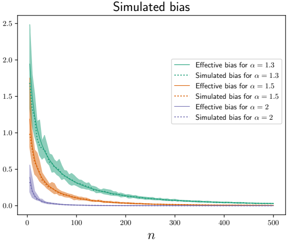
Specific eigenvalue structures

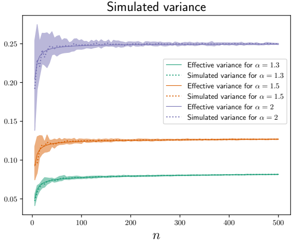
Power law decay: $\sigma_i = i^{-\alpha}$, $\alpha > 1$

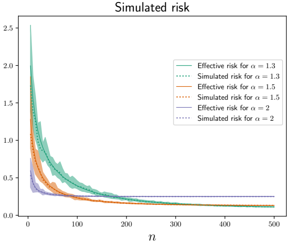
Critical decay: $\sigma_i = i^{-1}(1 + \log i)^{-\alpha'}$.

See paper for many other examples

Power law decay; $\lambda = 0+$

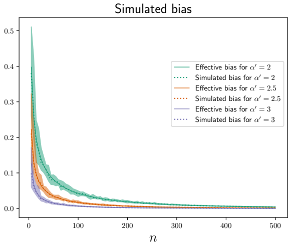


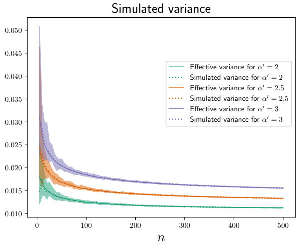


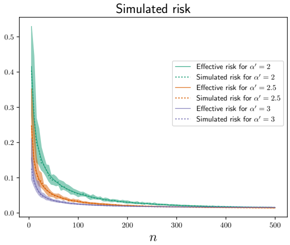


- Variance does not decrease with n.
- Need to use larger λ.

Critical decay; $\lambda = 0+$







- Variance does not decreases with n!
- Benign overfitting [Bartlett, Long, Lugosi, Tsigler, 2020]

Benign overfitting

Simplifying formulas in the seq. model

Determine

$$n - \frac{\lambda}{\lambda_\star} = \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-1}\right)$$

Then $\mathscr{R}_n^s(\lambda) = \mathrm{B}_n^s(\lambda) + \mathrm{V}_n^s(\lambda)$:

$$\mathrm{B}_n^s(\lambda) = \frac{\lambda_\star^2 \langle \boldsymbol{\beta}, (\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2}\, \boldsymbol{\Sigma}\boldsymbol{\beta}\rangle}{1 - n^{-1}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2}\right)},$$

$$\mathrm{V}_n^s(\lambda) = \frac{\tau^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2}\right)}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2}\right)}.$$

Eigenvalue decay assumption

$$\mathrm{Tr}\left(\mathbf{\Sigma}^2(\mathbf{\Sigma}+\lambda_\star\mathbf{I})^{-2}\right) \le \mathrm{Tr}\left(\mathbf{\Sigma}(\mathbf{\Sigma}+\lambda_\star\mathbf{I})^{-1}\right) = n - \frac{\lambda}{\lambda_\star}$$

Assume: $\quad \mathrm{Tr}\left(\mathbf{\Sigma}^2(\mathbf{\Sigma}+\lambda_\star\mathbf{I})^{-2}\right) \le n(1-c_\star^{-1})$

Then

$$\begin{aligned} V_n^s(\lambda) &= \frac{\tau^2 \mathrm{Tr}\left(\mathbf{\Sigma}^2(\mathbf{\Sigma}+\lambda_\star\mathbf{I})^{-2}\right)}{n - \mathrm{Tr}\left(\mathbf{\Sigma}^2(\mathbf{\Sigma}+\lambda_\star\mathbf{I})^{-2}\right)} \\ &\le \frac{c_\star \tau^2}{n} \mathrm{Tr}\left(\mathbf{\Sigma}^2(\mathbf{\Sigma}+\lambda_\star\mathbf{I})^{-2}\right) \end{aligned}$$

Eigenvalue decay assumption

$$\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \mathbf{I})^{-2}\right) \le \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star \mathbf{I})^{-1}\right) = n - \frac{\lambda}{\lambda_\star}$$

$$\text{Assume:} \quad \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \mathbf{I})^{-2}\right) \le n\left(1 - c_\star^{-1}\right)$$

Then

$$\begin{aligned} V_n^s(\lambda) &= \frac{\tau^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \mathbf{I})^{-2}\right)}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \mathbf{I})^{-2}\right)} \\ &\le \frac{c_\star \tau^2}{n} \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star \mathbf{I})^{-2}\right) \end{aligned}$$

Eigenvalue decay assumption

$$\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\mathbf{I})^{-2}\right) \le \mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_\star\mathbf{I})^{-1}\right) = n - \frac{\lambda}{\lambda_\star}$$

$$\text{Assume:}\quad \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\mathbf{I})^{-2}\right) \le n\left(1 - c_\star^{-1}\right)$$

Then

$$\begin{aligned} V_n^s(\lambda) &= \frac{\tau^2 \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\mathbf{I})^{-2}\right)}{n - \mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\mathbf{I})^{-2}\right)} \\ &\le \frac{c_\star \tau^2}{n}\mathrm{Tr}\left(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_\star\mathbf{I})^{-2}\right) \end{aligned}$$

Continuing...

$$
\begin{aligned}
V_n^s(\lambda) &\le \frac{c_\star \tau^2}{n} \operatorname{Tr}\left(\boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2}\right) \\
&\le \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell = k_\star + 1}^{\infty} \frac{\sigma_\ell^2}{\lambda_\star^2} \right\} \\
&\le \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell = k_\star + 1}^{\infty} \frac{\sigma_\ell^2}{\lambda_\star^2} \right\} \\
&\le \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell = k_\star + 1}^{\infty} \frac{\sigma_\ell^2}{\sigma_{k_\star + 1}^2} \right\}
\end{aligned}
$$

For $k_\star := \max\{k : \ k \ge \lambda_\star\}$

Continuing...

$$V_n^s(\lambda) \le \frac{c_\star \tau^2}{n} \mathrm{Tr}\left(\mathbf{\Sigma}^2 (\mathbf{\Sigma} + \lambda_\star \mathbf{I})^{-2}\right)$$

$$\le \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell=k_\star+1}^{\infty} \frac{\sigma_\ell^2}{\lambda_\star^2} \right\}$$

$$\le \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell=k_\star+1}^{\infty} \frac{\sigma_\ell^2}{\lambda_\star^2} \right\}$$

$$\le \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell=k_\star+1}^{\infty} \frac{\sigma_\ell^2}{\sigma_{k_\star+1}^2} \right\}$$

For $k_\star := \max\{k : \; k \ge \lambda_\star\}$

Continuing...

$$
\begin{aligned}
V_n^s(\lambda) &\leq \frac{c_\star \tau^2}{n} \mathrm{Tr}\left(\boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2}\right) \\
&\leq \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell = k_\star + 1}^{\infty} \frac{\sigma_\ell^2}{\lambda_\star^2} \right\} \\
&\leq \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell = k_\star + 1}^{\infty} \frac{\sigma_\ell^2}{\lambda_\star^2} \right\} \\
&\leq \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell = k_\star + 1}^{\infty} \frac{\sigma_\ell^2}{\sigma_{k_\star + 1}^2} \right\}
\end{aligned}
$$

For $k_\star := \max\{k : \; k \geq \lambda_\star\}$

Continuing. . .

$$
\begin{aligned}
V_n^s(\lambda) &\leq \frac{c_\star \tau^2}{n} \operatorname{Tr}\left(\boldsymbol{\Sigma}^2 (\boldsymbol{\Sigma} + \lambda_\star \mathbf{I})^{-2} \right) \\
&\leq \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell = k_\star + 1}^{\infty} \frac{\sigma_\ell^2}{\lambda_\star^2} \right\} \\
&\leq \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell = k_\star + 1}^{\infty} \frac{\sigma_\ell^2}{\lambda_\star^2} \right\} \\
&\leq \frac{c_\star \tau^2}{n} \left\{ k_\star + \sum_{\ell = k_\star + 1}^{\infty} \frac{\sigma_\ell^2}{\sigma_{k_\star + 1}^2} \right\}
\end{aligned}
$$

For $k_\star := \max\{k : \ k \geq \lambda_\star\}$

Benign overfitting

Proposition (Cheng, M, 2022)

Let $k_\star := \max\{k : \sigma_k \geq \lambda_\star\}$, $b_k := \sigma_k / \sigma_{k+1}$ *and*

$$r_q(k) := \sum_{\ell > k} \left(\frac{\sigma_\ell}{\sigma_{k+1}} \right)^q, \quad \bar{r}(k) := \frac{r_1(k)^2}{r_2(k)}.$$

Then,

$$V_n(\lambda) \leq c_\star \tau^2 \left(\frac{k_\star}{n} + \frac{r_2(k_\star)}{n} \right) \leq c_\star \tau^2 \left(\frac{k_\star}{n} + \frac{4 b_{k_\star}^2 n}{\bar{r}(k_\star)} \right),$$

$$B_n(\lambda) \leq c_\star \left(\sigma_{k_\star}^2 \|\boldsymbol{\beta}_{\leq k_\star}\|_{\boldsymbol{\Sigma}^{-1}}^2 + \|\boldsymbol{\beta}_{> k_\star}\|_{\boldsymbol{\Sigma}}^2 \right).$$

- Consistent if:
  - $1 \ll k_\star \ll n$.
  - $\bar{r}(k_\star) \to \infty$
  - $\|\boldsymbol{\beta}_{> k_\star}\|_{\boldsymbol{\Sigma}}^2 \to 0$
- cf. Bartlett, Long, Lugosi, Tsigler, 2020; Bartlett, Tsigler, 2021

Kernel ridge regression

High-dimensional setting

- $\boldsymbol{x}_i \sim \mathrm{Unif}(\sqrt{d}\mathbb{S}^{d-1})$
- $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i, \qquad f_* \in L^2(\mathbb{R}^d; \mathbb{P})$
- $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = h(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle / d), \quad \mathbb{E}[h(G)\mathrm{He}_k(G)] \neq 0$ for all k.

$$\hat{f}_\lambda = \mathrm{argmin}_f \Big\{ \sum_{i=1}^n \big(y_i - f(\boldsymbol{x}_i)\big)^2 + \lambda \|f\|_{\mathcal{H}}^2 \Big\}.$$

General abstract assumptions in [Mei, Misiakiewicz, Montanari, 2021]

Staircase phenomenon

Theorem (Ghorbani, Mei, Misiakiewicz, M. 2019)

Let $\ell \in \mathbb{Z}$*, and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$, $\varepsilon > 0$*. Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$,

$$R_\infty(f_*;\lambda) = \|P_{>\ell} f_*\|_{L^2}^2 + \|f_*\|_{L^2}^2 o_d(1)\,,$$

$$P_{>\ell} f_* = \textit{Projection of } f_* \textit{ onto deg. } > \ell \textit{ polynomials}$$

Further, no inner product kernel method can do better.

- Pointwise result (valid any for fixed f_*)
- $\lambda = 0$: optimal (interpolation)

Liang, Rakhlin, Zhai, 2019: $\|f_*\|_K \le C$, upper bounds on the rates. Bartlett, Rakhlin, M., 2021: Sharp results for $n \asymp d$.

Staircase phenomenon

Theorem (Ghorbani, Mei, Misiakiewicz, M. 2019)

Let $\ell \in \mathbb{Z}$, *and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$, $\varepsilon > 0$. *Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$,

$$R_\infty(f_*; \lambda) = \|P_{>\ell} f_*\|_{L^2}^2 + \|f_*\|_{L^2}^2 o_d(1)\,,$$

$$P_{>\ell} f_* = \textit{Projection of } f_* \textit{ onto deg. } > \ell \textit{ polynomials}$$

Further, no inner product kernel method can do better.

- Pointwise result (valid any for fixed f_*)
- $\lambda = 0$: optimal (interpolation)

Liang, Rakhlin, Zhai, 2019: $\|f_*\|_K \le C$, upper bounds on the rates. Bartlett, Rakhlin, M., 2021: Sharp results for $n \asymp d$.

Staircase phenomenon

Theorem (Ghorbani, Mei, Misiakiewicz, M. 2019)

Let $\ell \in \mathbb{Z}$, *and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$, $\varepsilon > 0$. *Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$,

$$R_\infty(f_*;\lambda) = \|P_{>\ell} f_*\|_{L^2}^2 + \|f_*\|_{L^2}^2 o_d(1)\,,$$

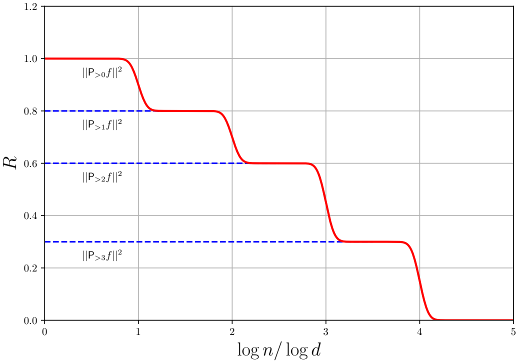
$$P_{>\ell} f_* = \textit{Projection of } f_* \textit{ onto deg. } > \ell \textit{ polynomials}$$

Further, no inner product kernel method can do better.

- Pointwise result (valid any for fixed f_*)
- $\lambda = 0$: optimal (interpolation)

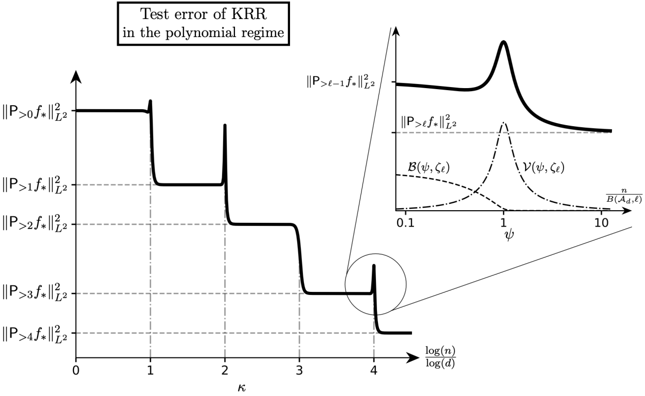
Liang, Rakhlin, Zhai, 2019: $\|f_*\|_K \le C$, upper bounds on the rates. Bartlett, Rakhlin, M., 2021: Sharp results for $n \asymp d$.

Sketch



- If $n \leq d^{1.99}$ can fit only linear functions.
- Valid for any inner product kernel
- Includes fully connected multi-layer nets

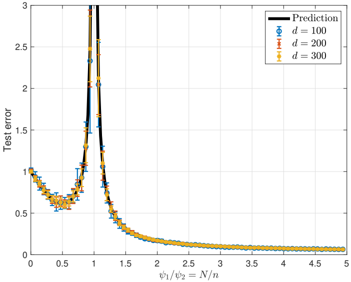
Developments



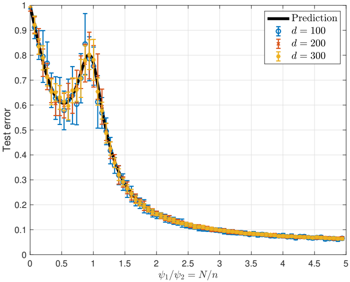
Misiakiewicz, 2022; Xiao, Pennington, 2022; Hu, Lu, 2022; Misiakiewicz, Saeed, 2024

Random features

Precise proportional asymptotics: $N/d \to \psi_1$, $n/d \to \psi_2$.



$\lambda = 0+$



$\lambda > 0$

- Solid line: Theoretical prediction

Neural tangent

Neural Tangent: Parameters view

$$f_{NT}(\boldsymbol{x};\mathbf{b}) = \sum_{j=1}^{N} \langle \mathbf{b}_j, \boldsymbol{x}\rangle \sigma'(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle)\,.$$

$$\boldsymbol{z}_i =:= \big(\boldsymbol{x}_i^{\mathsf{T}}\sigma'(\langle \boldsymbol{w}_1, \boldsymbol{x}_i\rangle), \boldsymbol{x}_i^{\mathsf{T}}\sigma'(\langle \boldsymbol{w}_2, \boldsymbol{x}_i\rangle), \dots, \boldsymbol{x}_i^{\mathsf{T}}\sigma'(\langle \boldsymbol{w}_N, \boldsymbol{x}_i\rangle)\big)^{\mathsf{T}}\,,$$

$$\hat{\mathbf{b}}_\lambda = \mathrm{argmin}_{\mathbf{b}}\Big\{\|\mathbf{y} - \mathbf{Z}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2\Big\}\,.$$

Neural Tangent: Function view

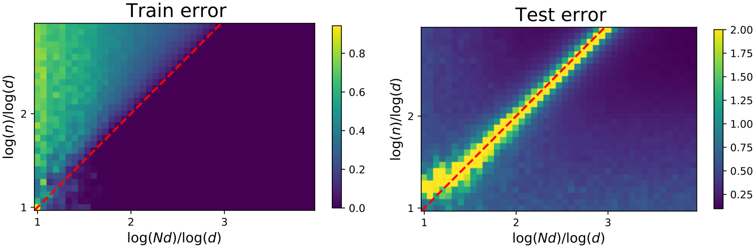
$$\hat{f}_{NT} = \mathrm{argmin}_f \Big\{ \sum_{i=1}^{n} \big(y_i - f(\boldsymbol{x}_i)\big)^2 + \lambda \|f\|_{K_N}^2 \Big\}.$$

$\|\cdot\|_{K_N}$: RKHS norm

$$K_N(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{Nd} \sum_{i=1}^{N} \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}_1 \rangle) \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}_2 \rangle)$$

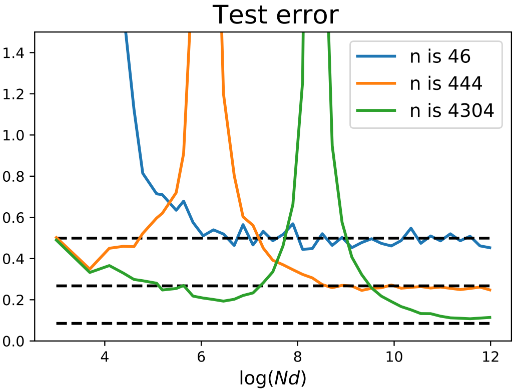
- Can we approximate it by $K(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathbb{E} K_N(\boldsymbol{x}_1, \boldsymbol{x}_2)$?

A small experiment $(d = 20)$



- $f_*(x) = g(\langle \beta_*, x \rangle)$, $g =$ deg-4 polynomial

A small experiment $(d = 20)$



- $f_*(\boldsymbol{x}) = g(\langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle)$, g = deg-4 polynomial
- NT ridge regression vs Kernel Ridge(-less) Regression

$$\hat{f} := \arg \min_{f:\mathbb{R}^d \to \mathbb{R}} \left\{ \|f\|_K \text{ subj. to } f(\boldsymbol{x}_i) = y_i \ \forall i \le n \right\},$$

Rigorous confirmation

$$\mathscr{R}_N(f_*;\lambda) := \text{Risk of linearized network}\,.$$
$$\mathscr{R}_\infty(f_*;\lambda) := \text{Risk of KRR}\,.$$

Theorem (M, Zhong, 2020, 2021)

Assume $d^\ell \ll n \ll d^{\ell+1}$ *for some integer* ℓ. *Then*

$$\mathscr{R}_N(f_*;\lambda) = \mathscr{R}_\infty(f_*;\lambda) + O\left(\|f_*\|_{L^2}^2 \sqrt{\frac{n(\log n)^C}{Nd}}\right)$$

Insights

$$R_N(f_*;\lambda) = R_\infty(f_*;\lambda) + O\left(\sqrt{\frac{n(\log n)^C}{Nd}}\right)$$

Insight 1: Risk constant for $Nd \gtrsim n(\log n)^C$

Insight 2: Overparametrization does not hurt

Insight 3: Interpolation ($\lambda = 0$) nearly optimal (see KRR result)

Intuition for 3?

Insights

$$R_N(f_*;\lambda) = R_\infty(f_*;\lambda) + O\left(\sqrt{\frac{n(\log n)^C}{Nd}}\right)$$

Insight 1: Risk constant for $Nd \gtrsim n(\log n)^C$

Insight 2: Overparametrization does not hurt

Insight 3: Interpolation ($\lambda = 0$) nearly optimal (see KRR result)

Intuition for 3?

Limitations of the linear regime

A simple example

Data

$$y_i = \sigma(\langle \boldsymbol{w}_*, \boldsymbol{x}_i \rangle) + \varepsilon_i\,, \quad \boldsymbol{w}_* \in \mathbb{S}^{d-1}\,, \;\; \varepsilon_i \sim \mathsf{N}(0, \tau^2)\,.$$

Machine learning model

$$f(\boldsymbol{x}; \boldsymbol{W}) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x} \rangle)\,,$$

$$(a_j^0)_{j \le N} \sim_{\text{iid}} \mathrm{P}_A\,, \;\; (\boldsymbol{w}_j^0)_{j \le N} \sim_{\text{iid}} \mathrm{Unif}(\mathbb{S}^{d-1})$$

What do we know? $\quad N = 1$

$$f(\boldsymbol{x};\boldsymbol{w}) = a\sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)\,,$$

$$\widehat{\mathscr{R}}_n(\theta) = \frac{1}{2n}\sum_{i=1}^{N}\big(y_i - \gamma\, a\sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)\big)^2\,, \qquad \theta = (a, \boldsymbol{w})\,,$$

$$\dot{\theta}_t = -\nabla \widehat{\mathscr{R}}_n(\theta)\,.$$

What do we know? $N = 1$

$$f(x; w) = a\sigma(\langle w, x\rangle)\,.$$

Proposition (Bai, Mei, M. 2018)

Assume $\gamma = a = 1$ *fixed,* $\sigma \in C^3(\mathbb{R})$*, strictly increasing with bounded derivatives. If* $n \ge C d \log d$*, then*

1. *Gradient flow converges exponentially fast:*

$$\|w_t - \hat{w}\|_2 \le C_0\, e^{-c_* t}\,.$$

2. *Small generalization error:*

$$\mathscr{R}(\hat{w}) - \min_{w \in \mathbb{R}^d} \mathscr{R}(w) \le C_1 \frac{d \log n}{n}\,.$$

Similar results for gradient descent.

What do we know? $\quad N \gg 1$

$$f(x; W) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} a_j \sigma(\langle w_j, x \rangle)$$

Proposition

If σ *generic,* $N \geq C(\varepsilon)\,(d \log d \vee n^2/d)$ *then:*

1. *Gradient flow converges to zero risk*

$$\widehat{\mathscr{R}}_n(f(\,\cdot\,;\theta_t)) \leq \widehat{\mathscr{R}}_n(f(\,\cdot,;\theta_0))\, e^{-c_* t}$$

2. *The neural tangent model is accurate* $|\mathscr{R}(f(\,\cdot\,;\theta_t)) - \mathscr{R}(f_{\mathrm{NT}}(\,\cdot\,;\theta_t))| \leq \varepsilon$.
3. *The generalization error is large. For* $d^\ell \ll n \leq d^{\ell+1}$

$$\mathscr{R}(f(\,\cdot\,;\hat{\theta})) - \min_{\theta} \mathscr{R}(f(\,\cdot\,;\theta)) = \left\|P_{>\ell}\sigma\right\|^2_{L^2(N(0,1))} + O(\varepsilon)\,.$$

What do we know? $N \gg 1$

$$f(x; W) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} a_j \sigma(\langle w_j, x \rangle)$$

Proposition

If σ *generic,* $N \geq C(\varepsilon)\,(d \log d \vee n^2/d)$ *then:*

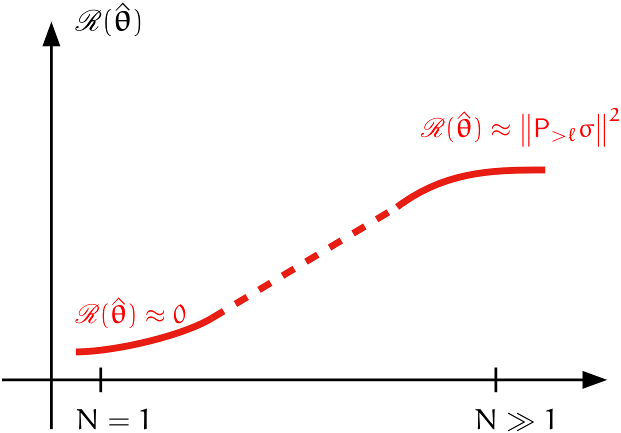
1. *Gradient flow converges to zero risk*

$$\widehat{\mathscr{R}}_n(f(\,\cdot\,;\theta_t)) \leq \widehat{\mathscr{R}}_n(f(\,\cdot,;\theta_0))\, e^{-c_* t}$$

2. *The neural tangent model is accurate* $|\mathscr{R}(f(\,\cdot\,;\theta_t)) - \mathscr{R}(f_{\rm NT}(\,\cdot\,;\theta_t))| \leq \varepsilon$.
3. *The generalization error is large. For* $d^\ell \ll n \leq d^{\ell+1}$

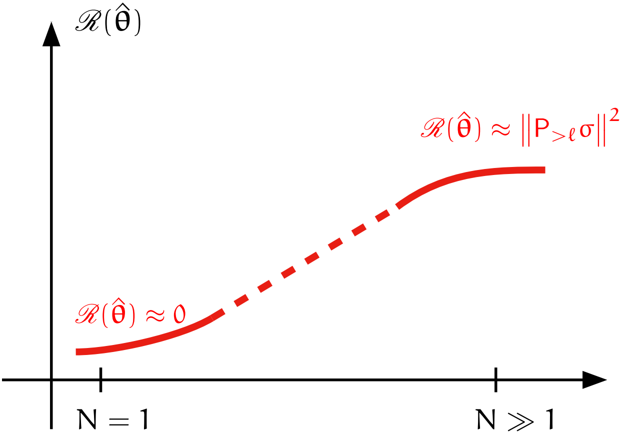
$$\mathscr{R}(f(\,\cdot\,;\hat{\theta})) - \min_{\theta} \mathscr{R}(f(\,\cdot\,;\theta)) = \left\|P_{>\ell}\sigma\right\|^2_{L^2(N(0,1))} + O(\varepsilon)\,.$$

A cartoon



A paradox?

A cartoon



A paradox?

Conclusion

Conclusion

- Linear models elucidate generalization puzzle in deep learning!
- Neural tangent model: Interpolating is optimal at high SNR
- Towards a unified theory

Thanks!

Conclusion

- Linear models elucidate generalization puzzle in deep learning!
- Neural tangent model: Interpolating is optimal at high SNR
- Towards a unified theory

Thanks!

Conclusion

- Linear models elucidate generalization puzzle in deep learning!
- Neural tangent model: Interpolating is optimal at high SNR
- Towards a unified theory

Thanks!

Conclusion

- Linear models elucidate generalization puzzle in deep learning!
- Neural tangent model: Interpolating is optimal at high SNR
- Towards a unified theory

Thanks!