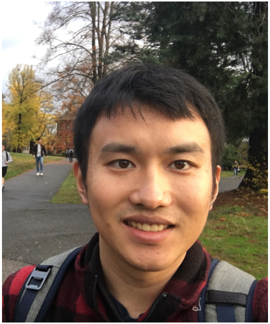
Self-induced regularization:
From linear regression to neural networks

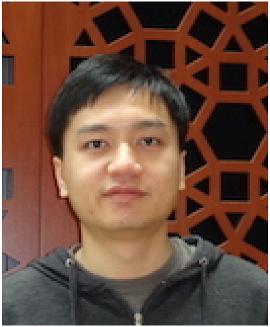
Andrea Montanari

Stanford University

August 10, 2020



Song Mei



Yiqiao Zhong

Supervised learning

- **Data**

$$\{(y_i, \boldsymbol{x}_i)\}_{i\le n} \sim_{iid} \mathbb{P}\,.$$

$\mathbb{P} \in \mathscr{P}(\mathbb{R}\times\mathbb{R}^d)$ unknown.

- **Want**

$$f : \mathbb{R}^d \to \mathbb{R}$$

- **Objective:** Given loss $\boldsymbol{\ell} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}_{\ge 0}$, minimize

$$R(f) := \mathbb{E}\{\boldsymbol{\ell}(y_{\rm new}, f(\boldsymbol{x}_{\rm new}))\}\,, \qquad (y_{\rm new}, \boldsymbol{x}_{\rm new}) \sim \mathbb{P}\,.$$

Classical theory

1. **Empirical Risk Minimization**

$$\min \; \widehat{R}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i))\,, \quad \text{subj. to } f \in \mathcal{F}\,.$$

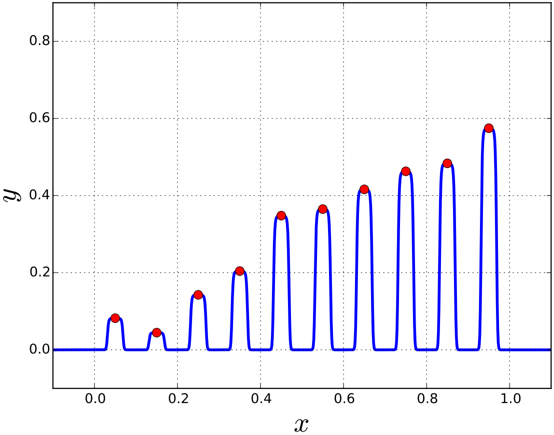
2. **Uniform convergence**

$$\sup_{f\in\mathcal{F}} \left| \widehat{R}_n(f) - R(f) \right| \le \varepsilon(\mathcal{F}, n)\,.$$

3. **Convex optimization**

   Choose ℓ, $\mathcal{F} = \{f(\,\cdot\,;\boldsymbol{\theta}) : \; \boldsymbol{\theta} \in \Theta\}$ so that ERM is convex

Why constrain $f \in \mathcal{F}$? Baby example



Classical viewpoint

- Since ∼2010, none of the three pillars seems to hold anymore[1].

[1] For many applications

Multi-layer (fully connected) neural network

$$\boldsymbol{\theta} = (\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_L)$$

$$\boldsymbol{\theta} \in \boldsymbol{\Theta} := \mathbb{R}^{N_1 \times N_0} \times \cdots \times \mathbb{R}^{N_L \times N_{L-1}}, \quad N_0 = d, N_L = 1,$$

$$f(\cdot; \boldsymbol{\theta}) := \boldsymbol{W}_L \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_{L-1} \circ \cdots \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_1 .$$

where

$$\boldsymbol{W}_\ell(\boldsymbol{x}) := \boldsymbol{W}_\ell \boldsymbol{x},$$

$$\boldsymbol{\sigma}(\boldsymbol{x}) := (\sigma(x_1), \ldots, \sigma(x_N)),$$

Examples: $\sigma(x) = \tanh(x)$, $\sigma(x) = \max(x, 0)$,...

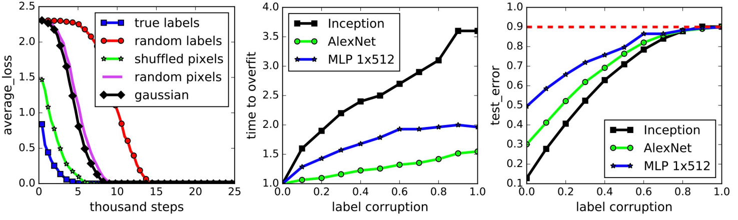
Pillar #3: ~~Convex optimization~~

$$f(\,\cdot\,;\boldsymbol{\theta}) = \boldsymbol{W}_L \circ \boldsymbol{\sigma} \circ \cdots \circ \boldsymbol{\sigma} \circ \boldsymbol{W}_1\,,$$

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(\boldsymbol{x}_i;\boldsymbol{\theta})\big)^2\,.$$

- Highly nonconvex!

Pillar #2: ~~Uniform convergence~~

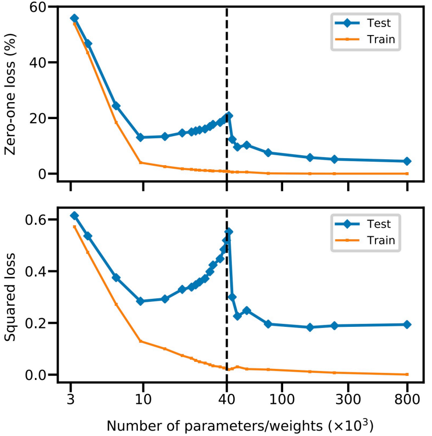


(a) learning curves (b) convergence slowdown (c) generalization error growth

[Zhang, Bengio, Hardt, Recht, Vinyals, 2016]

- $\mathcal{F}$ rich enough to 'interpolate' data points
- Test error $\gg$ Train error ≈ 0

Pillar #2: ~~Uniform convergence~~



- MNIST (subset): 4,000 images in 10 different classes.
- 2-layers Neural Net. Square loss.

Belkin, Hsu, Ma, Mandal, 2018; Spigler, Geiger, d'Ascoli, Sagun, Biroli, Wyart, 2018;...

Pillar #1: ~~Empirical Risk Minimization~~

GD/SGD

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \varepsilon_t \nabla_{\boldsymbol{\theta}} \widehat{R}_n(\boldsymbol{\theta}^{t})$$

- Nonconvex optimization
- Many global optima ($\widehat{R}_n(\boldsymbol{\theta}) \approx 0$)
- Output depends on
  - Initialization
  - Step-size schedule ε_t
  - ...

Can we understand all of this mathematically?

1. The big picture
2. A toy model
3. Results: The infinite width limit
4. Results: Random features model
5. Results: Neural tangent model
6. Conclusion and current directions

Ghorbani, Mei, Misiakiewicz, M, arXiv:1904.12191, 1906.08899

Mei, M, arXiv:1908.05355

M, Ruan, Sohn, Yan, arXiv:1911.01544

M, Zhong, arXiv:2007.12826

Related work

- Belkin, Rakhlin, Tsybakov, 2018
- Liang, Rakhlin, 2018
- Hastie, Montanari, Rosset, Tibshirani, 2019
- Belkin, Hsu, Xu,2019
- Bartlett, Long, Lugosi, Tsigler, 2019
- Muthukumar, Vodrahalli, Sahai, 2019
- Many papers in 2020

This work: Sharp asymptotics in the lazy regime of 2-layers nnets
(random features models)

Related work

- Belkin, Rakhlin, Tsybakov, 2018
- Liang, Rakhlin, 2018
- Hastie, Montanari, Rosset, Tibshirani, 2019
- Belkin, Hsu, Xu,2019
- Bartlett, Long, Lugosi, Tsigler, 2019
- Muthukumar, Vodrahalli, Sahai, 2019
- Many papers in 2020

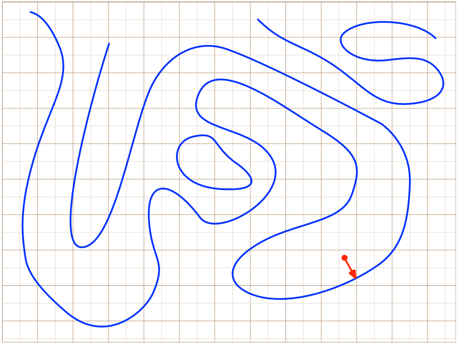
This work: Sharp asymptotics in the lazy regime of 2-layers nnets
(random features models)

The big picture

The big picture: Classical view

Learning	Empirical risk minimization
Tractability	Convexity
Regularization	Penalty in the cost function

The big picture



Learning	~~Empirical risk minim.~~	Algorithmic selection
Tractability	~~Convexity~~	Overparametrization
Regularization	~~Penalty in the cost~~	*'Self-induced regularization'*

Can we understand this rigorously?

Two-layers neural networks

$$\mathcal{F}_{\mathsf{NN}}^{N} \equiv \Big\{ f(\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{W}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \ : \quad a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^d \; \forall i \le N \Big\}.$$

Remark

$$f(\,\cdot\,) \in \mathcal{F}_{\mathsf{NN}}^{N}, \alpha \in \mathbb{R} \;\; \Rightarrow \;\; \alpha f(\,\cdot\,) \in \mathcal{F}_{\mathsf{NN}}^{N}$$

Can we understand this rigorously?

Lazy regime (linearize around a random initialization)

$$\frac{1}{\varepsilon} f(x; a_0 + \varepsilon a, W_0 + \varepsilon W) \approx$$

$$\approx \frac{1}{\varepsilon} f(x; a_0, W_0) + \langle a, \nabla_a f(x; a_0, W_0) \rangle + \langle W, \nabla_W f(x; a_0, W_0) \rangle$$

$$\approx \frac{1}{\varepsilon} f(x; a_0, W_0) + \sum_{i=1}^{N} a_i \sigma(\langle w_{0,i}, x \rangle) + \sum_{i=1}^{N} a_{0,i} \langle w_i, x \rangle \sigma(\langle w_{0,i}, x \rangle)$$

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; ...

Can we understand this rigorously?

Lazy regime (linearize around a random initialization)

$$\frac{1}{\varepsilon} f(x; \boldsymbol{a}_0 + \varepsilon \boldsymbol{a}, \boldsymbol{W}_0 + \varepsilon \boldsymbol{W}) \approx$$

$$\approx \frac{1}{\varepsilon} f(x; \boldsymbol{a}_0, \boldsymbol{W}_0) + \langle \boldsymbol{a}, \nabla_{\boldsymbol{a}} f(x; \boldsymbol{a}_0, \boldsymbol{W}_0)\rangle + \langle \boldsymbol{W}, \nabla_{\boldsymbol{W}} f(x; \boldsymbol{a}_0, \boldsymbol{W}_0)\rangle$$

$$\approx \frac{1}{\varepsilon} f(x; a_0, W_0) + \sum_{i=1}^{N} a_i \sigma(\langle w_{0,i}, x\rangle) + \sum_{i=1}^{N} a_{0,i} \langle w_i, x\rangle \sigma(\langle w_{0,i}, x\rangle)$$

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; ...

Can we understand this rigorously?

Lazy regime (linearize around a random initialization)

$$
\begin{aligned}
&\frac{1}{\varepsilon} f(x; a_0 + \varepsilon a, \boldsymbol{W}_0 + \varepsilon \boldsymbol{W}) \approx \\
&\quad \approx \frac{1}{\varepsilon} f(x; a_0, \boldsymbol{W}_0) + \langle a, \nabla_{a} f(x; a_0, \boldsymbol{W}_0) \rangle + \langle \boldsymbol{W}, \nabla_{\boldsymbol{W}} f(x; a_0, \boldsymbol{W}_0) \rangle \\
&\quad \approx \frac{1}{\varepsilon} f(x; a_0, \boldsymbol{W}_0) + \sum_{i=1}^{N} a_i \boldsymbol{\sigma}(\langle w_{0,i}, x \rangle) + \sum_{i=1}^{N} a_{0,i} \langle w_i, x \rangle \boldsymbol{\sigma}(\langle w_{0,i}, x \rangle)
\end{aligned}
$$

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; . . .

Can we understand this rigorously?

Lazy regime (linearize around a random initialization)

$$\frac{1}{\varepsilon} f(x; a_0 + \varepsilon a, W_0 + \varepsilon W) \approx$$
$$\approx \frac{1}{\varepsilon} f(x; a_0, W_0) + \langle a, \nabla_a f(x; a_0, W_0) \rangle + \langle W, \nabla_W f(x; a_0, W_0) \rangle$$
$$\approx \frac{1}{\varepsilon} f(x; a_0, W_0) + \sum_{i=1}^{N} a_i \sigma(\langle w_{0,i}, x \rangle) + \sum_{i=1}^{N} a_{0,i} \langle w_i, x \rangle \sigma(\langle w_{0,i}, x \rangle)$$

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; . . .

$$\frac{1}{\varepsilon} f(x; \boldsymbol{a}_0 + \varepsilon \boldsymbol{a},\, \boldsymbol{W}_0 + \varepsilon \boldsymbol{W}) \approx \underbrace{\sum_{i=1}^{N} a_i \boldsymbol{\sigma}(\langle \boldsymbol{w}_{0,i}, \boldsymbol{x}\rangle)}_{\mathcal{F}_{\mathsf{RF}}^{N}(\boldsymbol{W}_0)} + \underbrace{\sum_{i=1}^{N} \langle \boldsymbol{b}_i, \boldsymbol{x}\rangle \boldsymbol{\sigma}(\langle \boldsymbol{w}_{0,i}, \boldsymbol{x}\rangle)}_{\mathcal{F}_{\mathsf{NT}}^{N}(\boldsymbol{W}_0)}$$

$$\mathcal{F}_{\mathsf{RF}}^{N}(\boldsymbol{W}) := \Big\{ f(\boldsymbol{x}; \boldsymbol{a}) = \sum_{i=1}^{N} a_i\, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) \;:\; \quad a_i \in \mathbb{R}\; \forall i \le N \Big\},$$

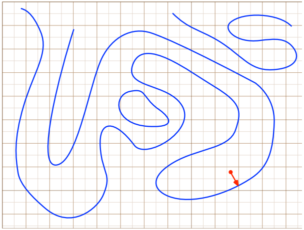
$$\mathcal{F}_{\mathsf{NT}}^{N}(\boldsymbol{W}) := \Big\{ f(\boldsymbol{x}; \boldsymbol{a}) = \sum_{i=1}^{N} \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle\, \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle) \;:\; \quad \boldsymbol{a}_i \in \mathbb{R}^d\; \forall i \le N \Big\},$$

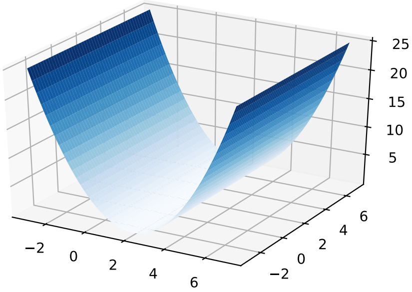
$$\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_N] \quad \boldsymbol{w}_i \sim_{iid} \mathsf{Unif}(\mathbb{S}^{d-1}(1))$$

Training: Ridge regression

$$\widehat{\boldsymbol{a}}_{\rm RR}(\lambda) := \arg\min_{\boldsymbol{a}\in\mathbb{R}^N} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\sf RF/NT}(\boldsymbol{x}_i, \boldsymbol{a}))^2 + \lambda \|\boldsymbol{a}\|_2^2 \right\} .$$

Why (ridgeless) ridge regression?





Why (ridgeless) ridge regression?

Gradient descent

$$\widehat{\boldsymbol{a}}_{k+1} = \widehat{\boldsymbol{a}}_k - t_k \nabla_{\boldsymbol{a}} \widehat{R}_n(\widehat{\boldsymbol{a}}_k)\,,$$
$$\widehat{R}_n(\boldsymbol{a}) := \text{Empirical square loss}\,.$$

Remark: In the overparametrized regime

$$\lim_{k\to\infty} \widehat{\boldsymbol{a}}_k = \lim_{\lambda\to 0} \widehat{\boldsymbol{a}}_{\rm RR}(\lambda)\,.$$

Self-induced regularization: A toy model

Linear regression

- Data $(\boldsymbol{y}, \boldsymbol{X}) = \{(y_i, \boldsymbol{x}_i)\}_{i \le n}$, $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$

- Ridge regularization

$$\widehat{\boldsymbol{\beta}}(\gamma) := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\{ \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \gamma \|\boldsymbol{\beta}\|_2^2 \right\},$$
$$= \frac{1}{d} (\gamma \boldsymbol{I}_d + \hat{\boldsymbol{\Sigma}}_X)^{-1} \boldsymbol{X}^\mathsf{T} \boldsymbol{y}$$

- $\hat{\boldsymbol{\Sigma}}_X := \boldsymbol{X}^\mathsf{T} \boldsymbol{X} / n$

Linear regression with a twist

- Data $(\boldsymbol{y}, \boldsymbol{X}) = \{(y_i, \boldsymbol{x}_i)\}_{i \le n}$
- Add noise to the covariates $\boldsymbol{z}_i = \boldsymbol{x}_i + \alpha \boldsymbol{g}_i$, $\boldsymbol{g}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$; $\boldsymbol{Z} = (\boldsymbol{z}_i)_{i \le n}$
- Ridge regularization

$$\widehat{\boldsymbol{\beta}}(\gamma; \alpha) := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\{ \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \gamma \|\boldsymbol{\beta}\|_2^2 \right\},$$
$$= \frac{1}{d} (\gamma \boldsymbol{I}_d + \hat{\boldsymbol{\Sigma}}_Z)^{-1} \boldsymbol{Z}^\mathsf{T} \boldsymbol{y}$$

Kobak, Lomond, Sanchez, 2018

Linear regression with a twist

$$
\begin{aligned}
\hat{\boldsymbol{\Sigma}}_Z &= \frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{Z} \\
&= \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} + \frac{\alpha}{n}\boldsymbol{X}^\top\boldsymbol{G} + \frac{\alpha}{n}\boldsymbol{G}^\top\boldsymbol{X} + \frac{\alpha^2}{n}\boldsymbol{G}^\top\boldsymbol{G} \\
&\approx \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} + \alpha^2\boldsymbol{I}_d\,.
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\beta}(\gamma;\alpha) &= \frac{1}{d}(\gamma\boldsymbol{I}_d + \hat{\boldsymbol{\Sigma}}_Z)^{-1}\boldsymbol{Z}^\top\boldsymbol{y} \\
&\approx \frac{1}{d}((\gamma+\alpha^2)\boldsymbol{I}_d + \hat{\boldsymbol{\Sigma}}_X)^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\
&\approx \widehat{\beta}(\gamma+\alpha^2;0)
\end{aligned}
$$

Covariates noise $\approx$ Ridge regularization

Linear regression with a twist

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_Z &= \frac{1}{n}\boldsymbol{Z}^\top \boldsymbol{Z} \\ &= \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} + \frac{\alpha}{n}\boldsymbol{X}^\top \boldsymbol{G} + \frac{\alpha}{n}\boldsymbol{G}^\top \boldsymbol{X} + \frac{\alpha^2}{n}\boldsymbol{G}^\top \boldsymbol{G} \\ &\approx \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} + \alpha^2 \boldsymbol{I}_d\,.\end{aligned}$$

$$\begin{aligned}\widehat{\beta}(\gamma;\alpha) &= \frac{1}{d}(\gamma \boldsymbol{I}_d + \hat{\boldsymbol{\Sigma}}_Z)^{-1}\boldsymbol{Z}^\top \boldsymbol{y} \\ &\approx \frac{1}{d}((\gamma+\alpha^2)\boldsymbol{I}_d + \hat{\boldsymbol{\Sigma}}_X)^{-1}\boldsymbol{X}^\top \boldsymbol{y} \\ &\approx \widehat{\beta}(\gamma+\alpha^2;0)\end{aligned}$$

Covariates noise $\approx$ Ridge regularization

Linear regression with a twist

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_Z &= \frac{1}{n}\boldsymbol{Z}^\top \boldsymbol{Z} \\ &= \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} + \frac{\alpha}{n}\boldsymbol{X}^\top \boldsymbol{G} + \frac{\alpha}{n}\boldsymbol{G}^\top \boldsymbol{X} + \frac{\alpha^2}{n}\boldsymbol{G}^\top \boldsymbol{G} \\ &\approx \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} + \alpha^2 \boldsymbol{I}_d\,.\end{aligned}$$

$$\begin{aligned}\widehat{\beta}(\gamma;\alpha) &= \frac{1}{d}(\gamma \boldsymbol{I}_d + \hat{\boldsymbol{\Sigma}}_Z)^{-1}\boldsymbol{Z}^\top \boldsymbol{y} \\ &\approx \frac{1}{d}((\gamma+\alpha^2)\boldsymbol{I}_d + \hat{\boldsymbol{\Sigma}}_X)^{-1}\boldsymbol{X}^\top \boldsymbol{y} \\ &\approx \widehat{\beta}(\gamma+\alpha^2;0)\end{aligned}$$

Covariates noise $\approx$ Ridge regularization

Self-induced regularization

Nonlinear ridgeless high-dimensional model

$$\Downarrow$$

Simpler model with positive ridge regularization

The role of covariate noise is played by nonlinearity

Self-induced regularization

Nonlinear ridgeless high-dimensional model

$$\Downarrow$$

Simpler model with positive ridge regularization

The role of covariate noise is played by nonlinearity

Results: The infinite width limit

Connection with kernels

$$\widehat{\boldsymbol{a}}_{\rm RR}(\lambda) := \arg\min_{\boldsymbol{a}\in\mathbb{R}^N} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\sf RF/NT}(\boldsymbol{x}_i, \boldsymbol{a}))^2 + \frac{N\lambda}{d}\|\boldsymbol{a}\|_2^2 \right\},$$

$$f_{\sf RF}(\boldsymbol{x}; \boldsymbol{a}) := \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle),$$

$$f_{\sf NT}(\boldsymbol{x}; \boldsymbol{a}) := \sum_{i=1}^{N} \langle \boldsymbol{a}_i, \boldsymbol{x}\rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle).$$

Function space formulation

$$\hat{f}_{\mathrm{RR},\lambda} := \arg \min_{\hat{f}:\mathbb{R}^d \to \mathbb{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(\boldsymbol{x}_i))^2 + \frac{\lambda}{d} \|\hat{f}\|_{K_N}^2 \right\},$$

$$K_{\mathsf{RF},N}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \frac{1}{N} \sum_{i=1}^{N} \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}_1 \rangle) \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x}_2 \rangle),$$

$$K_{\mathsf{NT},N}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \frac{1}{N} \sum_{i=1}^{N} \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}_1 \rangle) \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x}_2 \rangle).$$

(random kernel!)

Connection with kernels

Very wide limit

$$K_{\mathsf{RF},N}(\boldsymbol{x}_1, \boldsymbol{x}_2) \to K_{\mathsf{RF}}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \mathsf{E}_{\boldsymbol{w}}\{\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_1\rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_2\rangle)\}$$

$$K_{\mathsf{NT},N}(\boldsymbol{x}_1, \boldsymbol{x}_2) \to K_{\mathsf{NT}}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \mathsf{E}_{\boldsymbol{w}}\{\sigma'(\langle \boldsymbol{w}, \boldsymbol{x}_1\rangle)\sigma'(\langle \boldsymbol{w}, \boldsymbol{x}_2\rangle)\}$$

Rahimi, Recht; 2008; Bach, 2016; Daniely, Frostig, Gupta, Singer, 2017; *Jacot, Gabriel, Hongler, 2018*;...

Setting

- $\{(y_i, \boldsymbol{x}_i)\}_{i \le n}$ iid
- $\boldsymbol{x}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ or $\boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$, $\quad d \gg 1$
- $$y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i \,, \qquad \varepsilon_i \sim \mathsf{N}(0, \tau^2)\,.$$

Prediction error of Kernel Ridge Regression

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

Assume σ *continuous,* $|\sigma(x)| \le c_0 \exp(c_1|x|)$. *Let* $\ell \in \mathbb{Z}$, *and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$, $\varepsilon > 0$. *Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$,

$$R_{\mathrm{KRR}}(f_*; \lambda) = \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2 + o_d(1)(\|f_*\|_{L^2}^2 + \tau^2)\,,$$

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

Further, no kernel method can do better.

- Optimal error → interpolants ($\lambda = 0$)

Generalizes El Karoui, 2010

Prediction error of Kernel Ridge Regression

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

Assume σ *continuous,* $|\sigma(x)| \le c_0 \exp(c_1|x|)$. *Let* $\ell \in \mathbb{Z}$, *and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$, $\varepsilon > 0$. *Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$,

$$R_{\mathsf{KRR}}(f_*; \lambda) = \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2 + o_d(1)(\|f_*\|_{L^2}^2 + \tau^2)\,,$$

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

Further, no kernel method can do better.

- Optimal error → interpolants ($\lambda = 0$)

Generalizes El Karoui, 2010

Prediction error of Kernel Ridge Regression

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

Assume σ *continuous,* $|\sigma(x)| \le c_0 \exp(c_1|x|)$. *Let* $\ell \in \mathbb{Z}$, *and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$, $\varepsilon > 0$. *Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$,

$$R_{\mathsf{KRR}}(f_*; \lambda) = \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2 + o_d(1)(\|f_*\|_{L^2}^2 + \tau^2)\,,$$

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

Further, no kernel method can do better.

- Optimal error $\rightarrow$ interpolants ($\lambda = 0$)

Generalizes El Karoui, 2010

Prediction error of Kernel Ridge Regression

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

Assume σ *continuous,* $|\sigma(x)| \le c_0 \exp(c_1|x|)$*. Let* $\ell \in \mathbb{Z}$*, and assume* $d^{\ell+\varepsilon} \le n \le d^{\ell+1-\varepsilon}$*,* $\varepsilon > 0$*. Then, for any* $\lambda \in [0, \lambda_*(\sigma)]$*,*

$$R_{\mathsf{KRR}}(f_*; \lambda) = \|\mathsf{P}_{>\ell} f_*\|_{L^2}^2 + o_d(1)(\|f_*\|_{L^2}^2 + \tau^2)\,,$$

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

Further, no kernel method can do better.

- Optimal error $\rightarrow$ interpolants ($\lambda = 0$)

Generalizes El Karoui, 2010

Results: Random features model

Random features model

$$\mathcal{F}_{\mathsf{RF}}^{N}(\boldsymbol{W}) := \Big\{ f(\boldsymbol{x};\boldsymbol{a}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \ : \quad a_i \in \mathbb{R} \ \forall i \le N \Big\},$$

$$\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_N] \quad \boldsymbol{w}_i \sim_{iid} \mathsf{Unif}(\mathbb{S}^{d-1}(1))$$

- Number of parameters: N
- Number of samples: n
- Degrees of freedom in the target (polynomial of degree ℓ): d^{ℓ}.

Neal, 1996; Balcan, Blum, Vempala 2006; Rahimi, Recht; 2008; Bach, 2016

Proportional asymptotics

- $n \asymp d$
- $N \asymp d$

Focus on linear targets (d degrees of freedom)

- Target function

$$f_*(\boldsymbol{x}) = \langle \boldsymbol{\beta}_0, \boldsymbol{x} \rangle + f_*^{\rm NL}(\boldsymbol{x})$$

$f_*^{\rm NL}$ non-linear isotropic.

- $\|\boldsymbol{\beta}_0\|_2 = F_1$, $\|f_*^{\rm NL}\|_{L^2} = F_*$

- $n, N, d \to \infty$: $N/d \to \psi_1$, $n/d \to \psi_2$.

- $R(\hat{f}_\lambda) \equiv$ prediction error

Precise asymptotics

Theorem (Mei, M. 2019)

Decompose $\sigma(x) = \sigma_0 + \sigma_1 x + \sigma^{\mathrm{NL}}(x)$ *where (for* $G \sim \mathsf{N}(0,1)$*)*

$$\mathbb{E}[G\sigma^{\mathrm{NL}}(G)] = \mathbb{E}[\sigma^{\mathrm{NL}}(G)] = 0, \qquad \zeta^2 := \frac{\sigma_1^2}{\mathbb{E}[\sigma^{\mathrm{NL}}(G)^2]}\,.$$

Then, for any $\overline{\lambda} = \lambda/\overline{b}_*^2 > 0$

$$R(\hat{f}_\lambda) = F_1^2 \mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) + (\tau^2 + F_*^2)\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) + F_*^2 + o_d(1)\,,$$

where $\mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda})$, $\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda})$ *are explicitly given below.*

Variance computed in [Hastie, M, Rosset, Tibshirani, 2019]

Precise asymptotics

Theorem (Mei, M. 2019)

Decompose $\sigma(x) = \sigma_0 + \sigma_1 x + \sigma^{\rm NL}(x)$ *where (for* $G \sim \mathsf{N}(0,1)$*)*

$$\mathbb{E}[G\sigma^{\rm NL}(G)] = \mathbb{E}[\sigma^{\rm NL}(G)] = 0, \qquad \zeta^2 := \frac{\sigma_1^2}{\mathbb{E}[\sigma^{\rm NL}(G)^2]}\,.$$

Then, for any $\overline{\lambda} = \lambda/\overline{b}_*^2 > 0$

$$R(\hat{f}_\lambda) = F_1^2 \mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) + (\tau^2 + F_*^2)\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) + F_*^2 + o_d(1)\,,$$

where $\mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda})$, $\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda})$ *are explicitly given below.*

Variance computed in [Hastie, M, Rosset, Tibshirani, 2019]

Explicit formulae

Let $(\nu_1(\xi), \nu_2(\xi))$ 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

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

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

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

- *Kernel inner product random matrices*

Cheng, Singer, 2016; Do, Vu, 2017; Fan, M, 2017; Pennington Wohra, 2018;...

What do these formulae mean?

‘Noisy linear features model’

Nonlinear features

$$\hat{f}(\boldsymbol{x}_i; \boldsymbol{a}) = \langle \boldsymbol{a}, \boldsymbol{x}_i \rangle \,,$$
$$u_{ij} = \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle) = \sigma_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + \sigma^{\mathrm{NL}}(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle)$$

Noisy linear features

$$\hat{f}_{\boldsymbol{a}}(\boldsymbol{x}_i) = \langle \boldsymbol{a}, \tilde{\boldsymbol{u}} \rangle \,,$$
$$\tilde{u}_{ij} = \sigma_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + \sigma_* \, z_{ij} \,, \qquad (z_{ij}) \sim_{iid} \mathsf{N}(0,1)$$
$$\sigma_* := \|\sigma^{\mathrm{NL}}\|_{L^2}$$

Gaussian, correlated

'Noisy linear features model'

Nonlinear features

$$\hat{f}(\boldsymbol{x}_i; \boldsymbol{a}) = \langle \boldsymbol{a}, \boldsymbol{x}_i \rangle \,,$$
$$u_{ij} = \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle) = \sigma_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + \sigma^{\rm NL}(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle)$$

Noisy linear features

$$\hat{f}_{\boldsymbol{a}}(\boldsymbol{x}_i) = \langle \boldsymbol{a}, \tilde{\boldsymbol{u}} \rangle \,,$$
$$\tilde{u}_{ij} = \sigma_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + \sigma_* \, z_{ij} \,, \qquad (z_{ij}) \sim_{iid} \mathsf{N}(0,1)$$
$$\sigma_* := \|\sigma^{\rm NL}\|_{L^2}$$

Gaussian, correlated

Conceptual version of our theorem

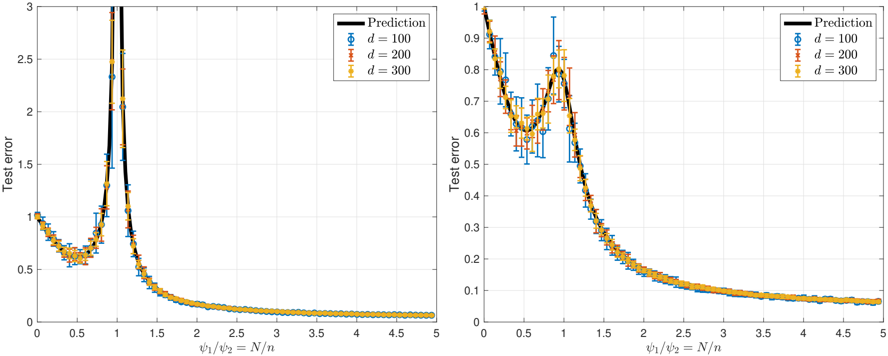
Theorem (Mei, M, 2019)

Consider random-features ridge regression in the proportional asymptotics

$$d \to \infty, \quad N/d \to \psi_1, \quad n/d \to \psi_2.$$

Then the nonlinear features model and noisy linear features model are 'asymptotically equivalent.'

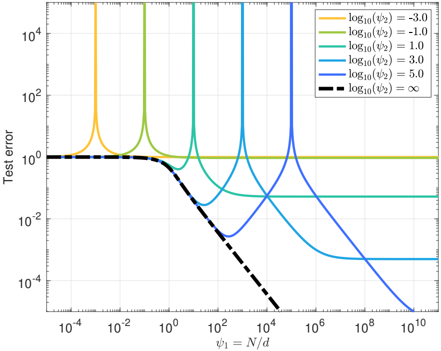
Simulations vs theory



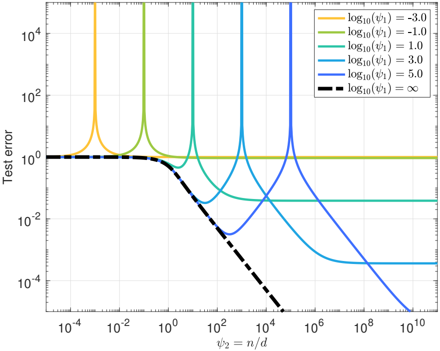
$\lambda = 0+$ $\lambda > 0$

Insigths

Insight #1: Optimum at $N/n \to \infty$

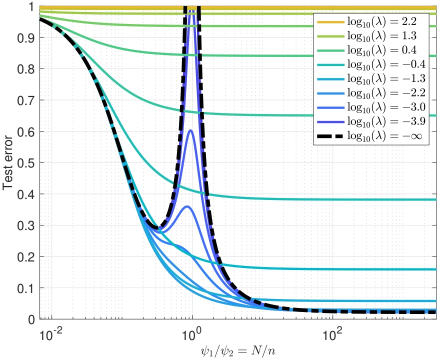


$\lambda = 0+$

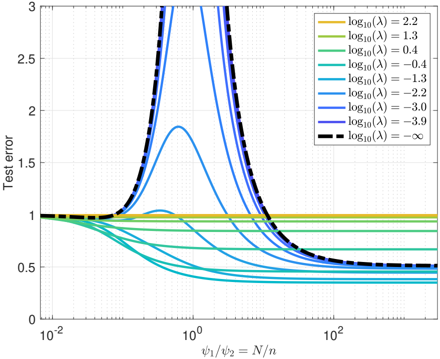


$\lambda = 0+$

Insight #2: No double descent for optimal λ

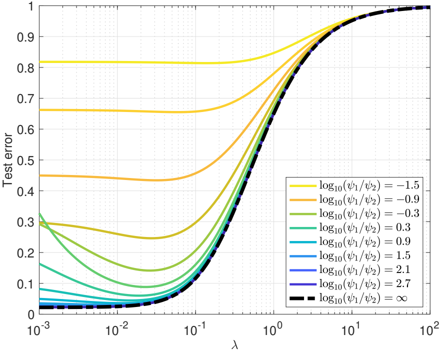


SNR= 5

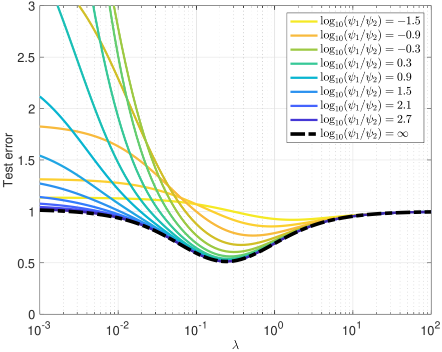


SNR= 1/5

Insight #3: $\lambda = 0+$ optimal at high SNR



SNR= 5



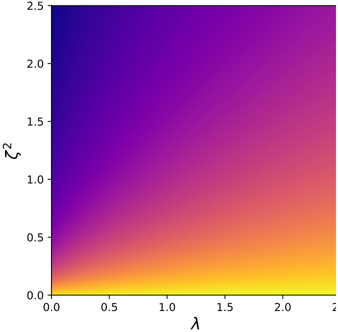
SNR= 1/10

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

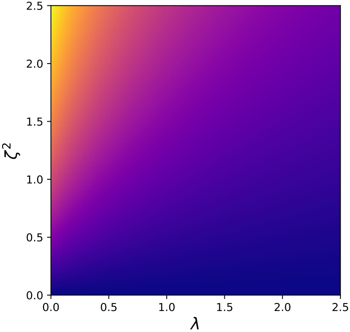
Insight #4: Self-induced regularization

- **Wide limit** $\psi_1 = N/d \to \infty$, $\psi_2 = n/d < \infty$

Insight #4: Self-induced regularization

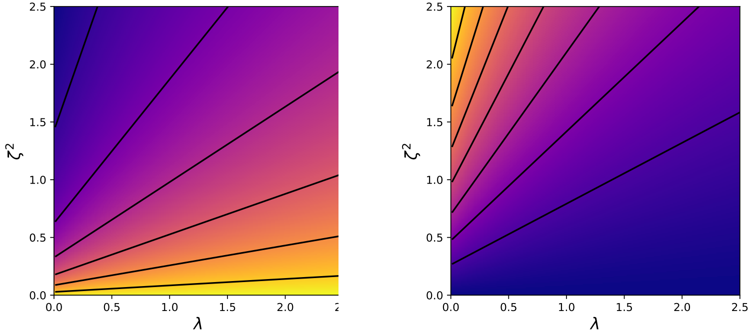


Bias



Variance

Insight #4: Self-induced regularization



Bias

Variance

$$\text{Decreasing } \zeta^2 := \frac{\mathbb{E}\{\sigma(G)G\}^2}{\mathbb{E}[\sigma^{\mathrm{NL}}(G)^2]} \Leftrightarrow \text{Increasing } \lambda$$

Insight #4: Self-induced regularization

- **Normalized regularization**

$$r = \frac{\overline{\lambda} + \psi_2^{-1}}{\zeta^2}, \quad \zeta^2 := \frac{\mathbb{E}\{\sigma(G)G\}^2}{\mathbb{E}[\sigma^{\mathrm{NL}}(G)^2]}$$

- **Bias, Variance**

$$\mathscr{B}(\zeta, \infty, \psi_2, \overline{\lambda}) = \frac{1}{(1+\omega)^2 - \omega^2/\psi_2},$$

$$\mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) = \frac{\omega^2/\psi_2}{(1+\omega)^2 - \omega^2/\psi_2},$$

$\omega = \overline{\omega}(1/r)$ increasing in $1/r$.

Results: Neural tangent model

Neural tangent model

$$\mathcal{F}_{\mathsf{NT}}^{N}(\boldsymbol{W}) := \left\{ f(\boldsymbol{x};\boldsymbol{a}) = \sum_{i=1}^{N} \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle \, \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \ : \quad a_i \in \mathbb{R} \ \forall i \le N \right\},$$

$$\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_N] \quad \boldsymbol{w}_i \sim_{iid} \mathsf{Unif}(\mathbb{S}^{d-1}(1))$$

- Number of parameters: Nd
- Number of samples: n
- Degrees of freedom in the target (polynomial of degree ℓ): d^ℓ.

Jacot, Gabriel, Hongler, 2018

Polynomial asymptotics

- $N, n, d \to \infty$
- $d^{\varepsilon} \leq N$, for some $\varepsilon > 0$
- $N \leq Cd$ for some $C > 0$

The interpolation phase transition

Theorem (M, Zhong, 2020)

There exists $C < \infty$ such that, if $Nd/(\log d)^C \geq n$, then an NT interpolator exists with high probability for any choice of the responses $(y_i)_{i \leq n}$.

- Interpoaltion requires $Nd \geq n$
- Recent related results:
  Daniely 2020; Bubeck, Eldan, Lee, Mikulincer, 2020

Key technical result

Empirical kernel

$$\boldsymbol{K} = \left(\frac{1}{Nd} \sum_{\ell=1}^{N} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \sigma'(\langle \boldsymbol{w}_\ell, \boldsymbol{x}_i \rangle) \sigma'(\langle \boldsymbol{w}_\ell, \boldsymbol{x}_j \rangle) \right)_{i,j \le N}$$

Theorem (M, Zhong, 2020)

There exists a matrix $\boldsymbol{E} \succeq 0$, $\mathrm{rank}(\boldsymbol{E}) \le N$, *such that*

$$\boldsymbol{K} \gtrsim (v(\sigma) - o(1)) \boldsymbol{I}_n + \boldsymbol{E}, \qquad v(\sigma) := \mathsf{Var}_{G \sim \mathsf{N}(0,1)}(\sigma'(G)).$$

Effective linear model

$$\widehat{\beta}(\gamma) := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\{ \frac{1}{d} \sum_{i=1}^{n} \big(y_i - \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle\big)^2 + \gamma \|\boldsymbol{\beta}\|_2^2 \right\},$$

$$R_{\rm lin}(\gamma) := \mathbb{E}_{\boldsymbol{x}_{\rm new}} \big[(\langle \boldsymbol{\beta}_0, \boldsymbol{x}_{\rm new} \rangle - \langle \widehat{\boldsymbol{\beta}}(\gamma), \boldsymbol{x}_{\rm new} \rangle)^2 \big] .$$

Very well understood!

Very well understood!

$$R_{\mathrm{lin}}(\delta,\gamma) = \|\beta^*\|_2^2 \mathscr{B}_{\mathrm{lin}}(\delta,\gamma) + \sigma_\varepsilon^2 \mathscr{V}_{\mathrm{lin}}(\delta,\gamma), \ ,$$

When $n, d \to \infty$, $n/d \to \delta \in (0,\infty)$:

$$\mathscr{B}_{\mathrm{lin}}(\delta,\gamma) = \frac{1}{2}\left\{1 - \delta + \sqrt{(\delta-1+\gamma)^2+4\gamma} - \frac{\gamma(1+\delta+\gamma)}{\sqrt{(\delta-1+\gamma)^2+4\gamma}}\right\} + o(1)\,,$$

$$\mathscr{V}_{\mathrm{lin}}(\delta,\gamma) = \frac{1}{2}\left\{-1 + \frac{\delta+\gamma+1}{\sqrt{(\delta-1+\gamma)^2+4\gamma}}\right\} + o(1)\,.$$

Hastie, M, Rosset, Tibshirani, 2019

Generalization

Theorem (M, Zhong, 2020)

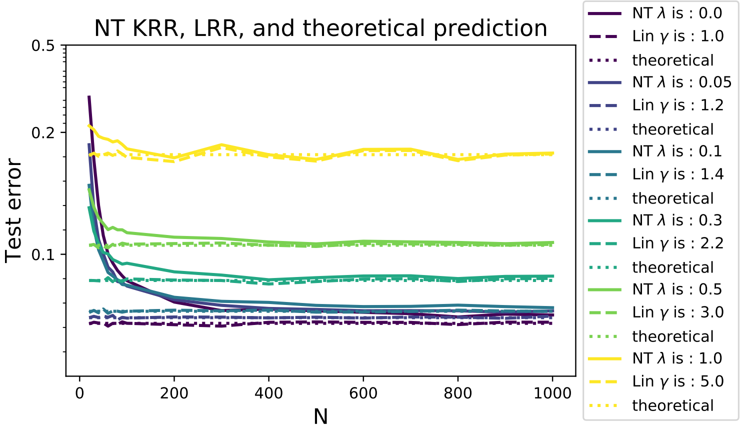
Assume that $n \geq \varepsilon_0 d$ *for a constant* $\varepsilon_0 > 0$, *and that* $\mathbb{E}[\sigma'(G)] \neq 0$. *Recall* $v(\sigma) = \mathsf{Var}(\sigma'(G))$.
If $Nd/(\log^C(Nd)) \geq n$, *then for any* $\lambda \geq 0$,

$$R_{\mathsf{NT}}(\lambda) = R_{\text{lin}}(\gamma_{\text{eff}}(\lambda, \sigma)) + O_{d,\mathbb{P}}\left(\sqrt{\frac{n(\log d)^C}{Nd}}\right), \qquad \textit{where}$$

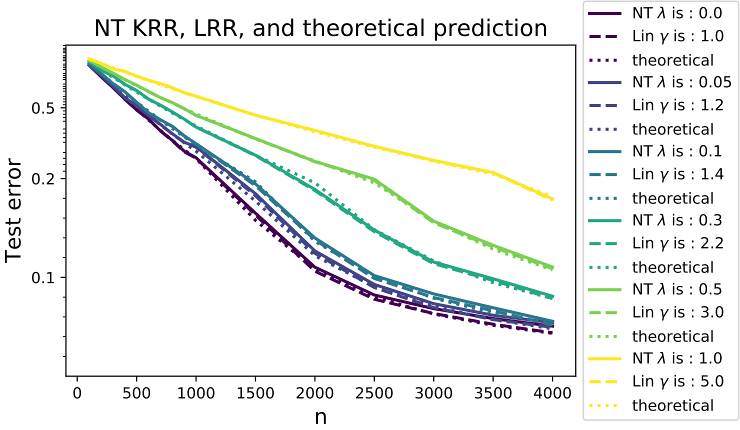
$$\gamma_{\text{eff}}(\lambda, \sigma) := \frac{\lambda + v(\sigma)}{\{\mathbb{E}[\sigma'(G)]\}^2}.$$

In particular, ridgeless NT $\approx$ *linear regression with* $\gamma = v(\sigma)/\{\mathbb{E}[\sigma'(G)]\}^2$.

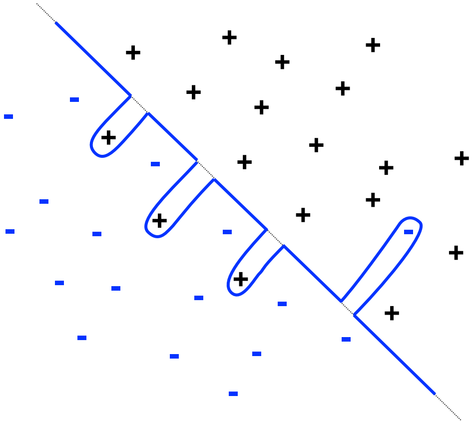
NT vs linear regression



NT vs linear regression vs Random matrix theory



Intuitive picture



Conclusion and current directions

Conclusion

- Do not need to carefully trade model complexity vs sample size.
- Optimal generalization with no/minimal regularization
- Self-induced regularization

Conclusion

- Do not need to carefully trade model complexity vs sample size.
- Optimal generalization with no/minimal regularization
- Self-induced regularization

Conclusion

- Do not need to carefully trade model complexity vs sample size.
- Optimal generalization with no/minimal regularization
- Self-induced regularization

Open problems

- Other losses (classification)

  [M, Ruan, Sohn, Yan, 2019]

- Anisotropic data distributions

  [Ghorbani, Mei, Misiakiewicz, M, 2020]

- Sharp results for NT models

- Sharp results under polynomial scalings

Thanks!