Thermodynamic limits for neural networks

Andrea Montanari

Stanford University

November 21, 2019

Setting

- $\{(y_i, \boldsymbol{x}_i)\}_{i \le n}$
- $\boldsymbol{x}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ or $\boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$, $\quad d \gg 1$
- Response (regression)

$$y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i\,, \qquad \varepsilon_i \sim \mathsf{N}(0, \tau^2)$$

Evaluating a model

- Loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ (e.g. $\ell(y_1, y_2) = (y_1 - y_2)^2$)

- Test error

$$R(f) = \mathbb{E}_{\text{new}} \left\{ \ell(y^{\text{new}}, f(x^{\text{new}})) \right\}$$

Evaluating a model

- Loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ (e.g. $\ell(y_1, y_2) = (y_1 - y_2)^2$)

- **Test error**

$$R(f) = \mathbb{E}_{\text{new}} \left\{ \ell(y^{\text{new}}, f(x^{\text{new}})) \right\}$$

Two-layers (fully-connected) neural networks

$$\mathcal{F}_{\mathsf{NN}}^{N} \equiv \left\{ f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \; : \;\; \boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i \leq N}, \;\; a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^d \right\}$$

- $\hat{\theta}_n$ computed on training sample
- **Train/test error**

$$\widehat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x)) \,,$$

$$R(\hat{\theta}_n) = \mathbb{E}_{\text{new}} \{ \ell(y^{\text{new}}, f_{\hat{\theta}_n}(x^{\text{new}})) \}$$

Two-layers (fully-connected) neural networks

$$\mathcal{F}_{\mathsf{NN}}^{N} \equiv \left\{ f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \; : \;\; \boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i \le N}, \;\; a_i \in \mathbb{R}, \boldsymbol{w}_i \in \mathbb{R}^d \right\}$$

- $\hat{\boldsymbol{\theta}}_n$ computed on training sample
- **Train/test error**

$$\widehat{R}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_{\boldsymbol{\theta}}(\boldsymbol{x})) \,,$$

$$R(\hat{\boldsymbol{\theta}}_n) = \mathbb{E}_{\text{new}} \{ \ell(y^{\text{new}}, f_{\hat{\boldsymbol{\theta}}_n}(\boldsymbol{x}^{\text{new}})) \}$$

‘Thermodynamic’ limit?

- Classical limit in Stats/ML: $n \to \infty$ (large sample)
- Here: $N \to \infty$ (large network)

Very rich picture. At least two regimes...

‘Thermodynamic’ limit?

- Classical limit in Stats/ML: $n \to \infty$ (large sample)
- Here: $N \to \infty$ (large network)

Very rich picture. At least two regimes...

‘Thermodynamic’ limit?

- Classical limit in Stats/ML: $n \to \infty$ (large sample)
- Here: $N \to \infty$ (large network)

Very rich picture. At least two regimes...

‘Thermodynamic’ limit?

- Classical limit in Stats/ML: $n \to \infty$ (large sample)
- Here: $N \to \infty$ (large network)

Very rich picture. At least two regimes...

Mean field limit

$$f_\theta(x) = \sum_{i=1}^{N} a_i \, \sigma(\langle w_i, x \rangle)$$

$$f_\theta(x) = \int a \sigma(\langle w, x \rangle) \; \rho(\mathrm{d}a, \mathrm{d}w) \,, \qquad \rho = \frac{1}{N} \sum_{i=1}^{N} \delta_{N a_i, w_i}$$

$$\mathcal{F}_{\mathsf{NN}}^{N} \subseteq \mathcal{F}_{\mathsf{NN}}^{N+1} \subseteq \cdots \subseteq \mathcal{F}_{\mathsf{NN}}^{\infty} \,,$$

$$\mathcal{F}_{\mathsf{NN}}^{\infty} \equiv \left\{ f(x; \rho) = \int a \sigma(\langle w, x \rangle) \; \rho(\mathrm{d}a, \mathrm{d}w) \quad \rho \in \mathscr{P}(\mathbb{R}^{d+1}) \right\} .$$

Mean field limit

$$f_{\boldsymbol{\theta}}(x) = \sum_{i=1}^{N} a_i \, \sigma(\langle w_i, x \rangle)$$

$$f_\theta(x) = \int a \sigma(\langle w, x \rangle) \, \rho(\mathrm{d}a, \mathrm{d}w) \,, \qquad \rho = \frac{1}{N} \sum_{i=1}^{N} \delta_{N a_i, w_i}$$

$$\mathcal{F}_{\mathsf{NN}}^{N} \subseteq \mathcal{F}_{\mathsf{NN}}^{N+1} \subseteq \cdots \subseteq \mathcal{F}_{\mathsf{NN}}^{\infty} \,,$$

$$\mathcal{F}_{\mathsf{NN}}^{\infty} \equiv \left\{ f(x; \rho) = \int a \sigma(\langle w, x \rangle) \, \rho(\mathrm{d}a, \mathrm{d}w) \quad \rho \in \mathscr{P}(\mathbb{R}^{d+1}) \right\} .$$

Mean field limit

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int a \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) \; \rho(\mathrm{d}a, \mathrm{d}\boldsymbol{w}) \,, \quad \rho = \frac{1}{N} \sum_{i=1}^{N} \delta_{N a_i, \boldsymbol{w}_i}$$

$$\mathcal{F}_{\mathsf{NN}}^{N} \subseteq \mathcal{F}_{\mathsf{NN}}^{N+1} \subseteq \cdots \subseteq \mathcal{F}_{\mathsf{NN}}^{\infty} \,,$$

$$\mathcal{F}_{\mathsf{NN}}^{\infty} \equiv \left\{ f(x; \rho) = \int a \sigma(\langle w, x \rangle) \; \rho(\mathrm{d}a, \mathrm{d}w) \quad \rho \in \mathscr{P}(\mathbb{R}^{d+1}) \right\} .$$

Mean field limit

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \, \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$$

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int a \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) \; \rho(\mathrm{d}a, \mathrm{d}\boldsymbol{w}) \,, \quad \rho = \frac{1}{N} \sum_{i=1}^{N} \delta_{N a_i, \boldsymbol{w}_i}$$

$$\mathcal{F}_{\mathsf{NN}}^{N} \subseteq \mathcal{F}_{\mathsf{NN}}^{N+1} \subseteq \cdots \subseteq \mathcal{F}_{\mathsf{NN}}^{\infty} \,,$$

$$\mathcal{F}_{\mathsf{NN}}^{\infty} \equiv \left\{ f(\boldsymbol{x}; \rho) = \int a \sigma(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) \; \rho(\mathrm{d}a, \mathrm{d}\boldsymbol{w}) \quad \rho \in \mathscr{P}(\mathbb{R}^{d+1}) \right\} .$$

Train error

$$\widehat{R}_n(\rho) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i; \rho))$$

- **Good news:** Convex
- **Bad news:** Infinite dimensional
- **Question:** Can we optimize it efficiently?
  e.g. approximate ρ by finitely many points

[Nitanda, Suzuki, 2018; Mei, Montanari, Nguyen, 2018; Rotskoff, Vanden-Eijnden, 2018; Chizat, Bach, 2018; Mei, Misiakiewicz, Montanari, 2019; Rotskoff, Jelassi, Bruna,Vanden-Eijnden, 2019; Araújo, Oliveira, Yukimura 2019; Nguyen, 2019; . . .]

Train error

$$\widehat{R}_n(\rho) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i; \rho))$$

- **Good news:** Convex
- **Bad news:** Infinite dimensional
- **Question:** Can we optimize it efficiently?
  e.g. approximate ρ by finitely many points

[Nitanda, Suzuki, 2018; Mei, Montanari, Nguyen, 2018; Rotskoff, Vanden-Eijnden, 2018; Chizat, Bach, 2018; Mei, Misiakiewicz, Montanari, 2019; Rotskoff, Jelassi, Bruna,Vanden-Eijnden, 2019; Araújo, Oliveira, Yukimura 2019; Nguyen, 2019; . . .]

Train error

$$\widehat{R}_n(\rho) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i; \rho))$$

- **Good news:** Convex
- **Bad news:** Infinite dimensional
- **Question:** Can we optimize it efficiently?
  e.g. approximate ρ by finitely many points

[Nitanda, Suzuki, 2018; Mei, Montanari, Nguyen, 2018; Rotskoff, Vanden-Eijnden, 2018; Chizat, Bach, 2018; Mei, Misiakiewicz, Montanari, 2019; Rotskoff, Jelassi, Bruna,Vanden-Eijnden, 2019; Araújo, Oliveira, Yukimura 2019; Nguyen, 2019; . . .]

Train error

$$\widehat{R}_n(\rho) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i; \rho))$$

- **Good news:** Convex
- **Bad news:** Infinite dimensional
- **Question:** Can we optimize it efficiently?
  e.g. approximate ρ by finitely many points

[Nitanda, Suzuki, 2018; Mei, Montanari, Nguyen, 2018; Rotskoff, Vanden-Eijnden, 2018; Chizat, Bach, 2018; Mei, Misiakiewicz, Montanari, 2019; Rotskoff, Jelassi, Bruna,Vanden-Eijnden, 2019; Araújo, Oliveira, Yukimura 2019; Nguyen, 2019; . . .]

Train error

$$\widehat{R}_n(\rho) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i; \rho))$$

- **Good news:** Convex
- **Bad news:** Infinite dimensional
- **Question:** Can we optimize it efficiently?
  e.g. approximate ρ by finitely many points

[Nitanda, Suzuki, 2018; Mei, Montanari, Nguyen, 2018; Rotskoff, Vanden-Eijnden, 2018; Chizat, Bach, 2018; Mei, Misiakiewicz, Montanari, 2019; Rotskoff, Jelassi, Bruna,Vanden-Eijnden, 2019; Araújo, Oliveira, Yukimura 2019; Nguyen, 2019; . . .]

Another approach: lazy limit

- **Idea:** Linearize around a random initialization
- $\Rightarrow$ Linear class $\mathcal{F}_{\mathsf{NT}}$

Neural tangent

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

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

- 'Tangent' to $\mathcal{F}_{\mathsf{NN}}^{N}$ around a random point
- Good approximation: (i) $N \gg n$; (ii) Parameters's scaling

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

Neural tangent

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

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

- 'Tangent' to $\mathcal{F}_{\mathsf{NN}}^{N}$ around a random point
- Good approximation: (i) $N \gg n$; (ii) Parameters's scaling

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

A simpler linear model

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

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

- Random features model

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

This talk

Which phenomena are captured linear models?

Outline

1. What is captured by linear models?
2. What is not captured by linear models?
3. Conclusion

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

Mei, M, arXiv:1908.05355

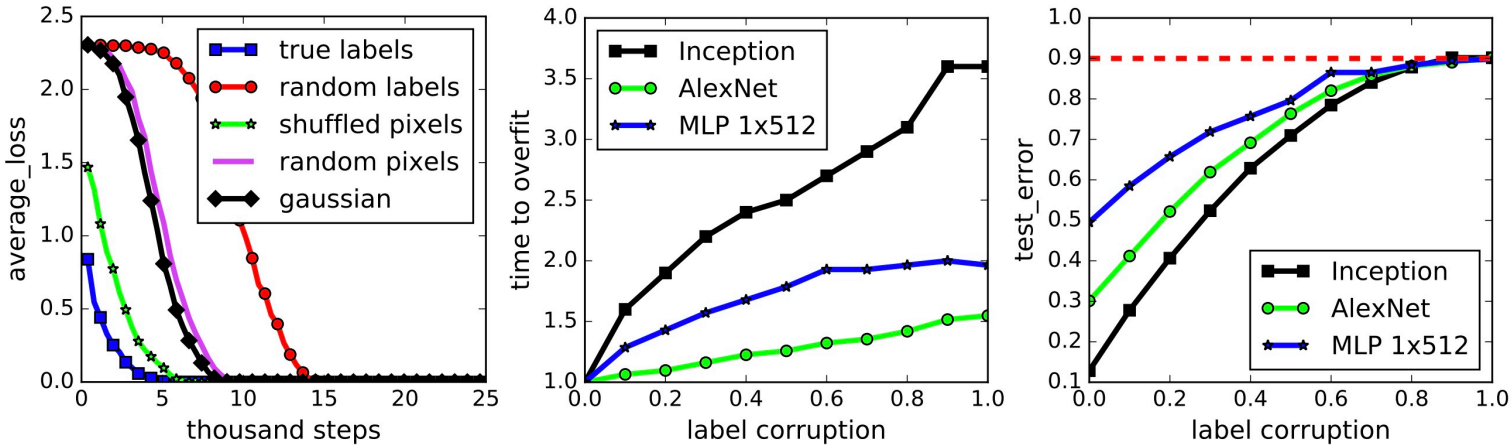
M, Ruan, Sohn, Yan, arXiv:1911.01544

What is captured by linear models?

What is captured by linear models?

- Interpolation phase transition
- Double descent

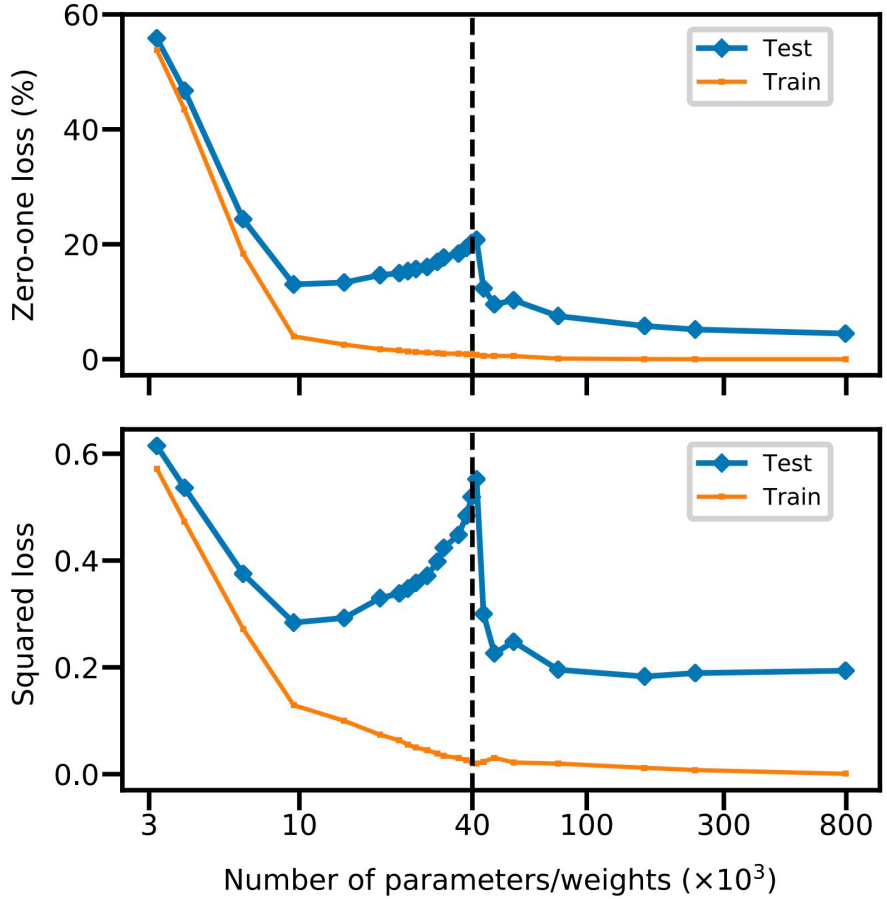
Using deep nets to fit noise



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

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

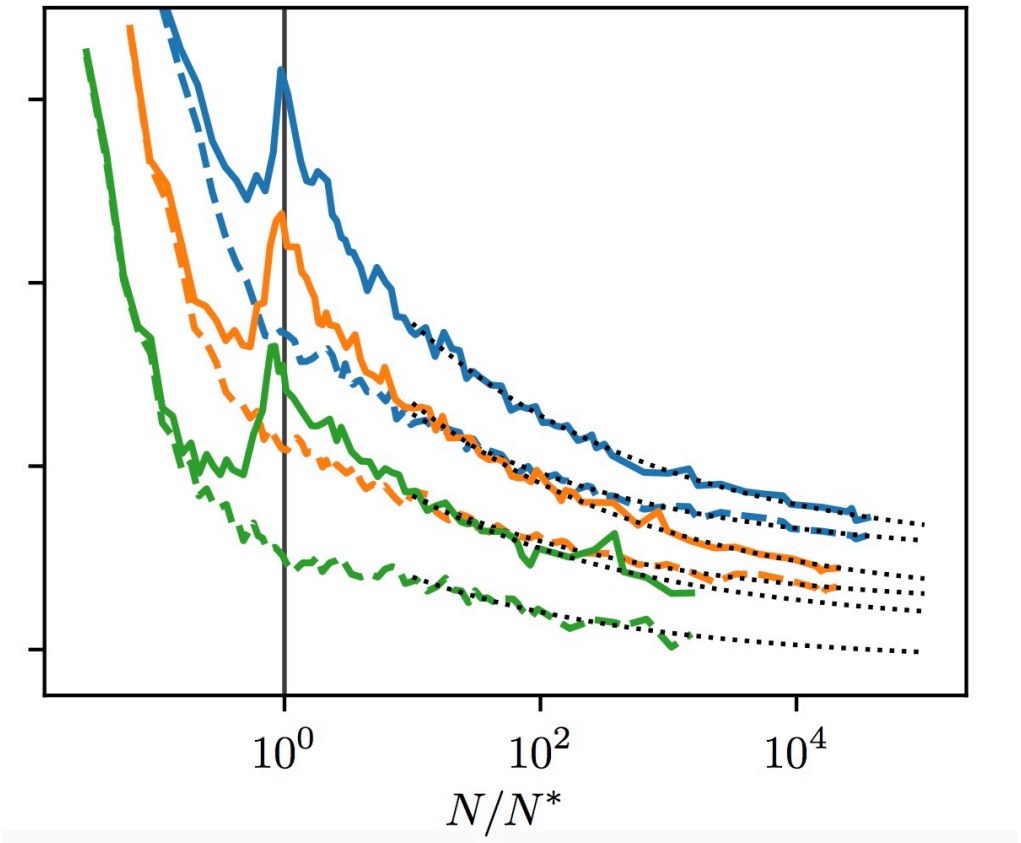
Interpolating, yet not overfitting



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

[Belkin, Hsu, Ma, Mandal, 2018]

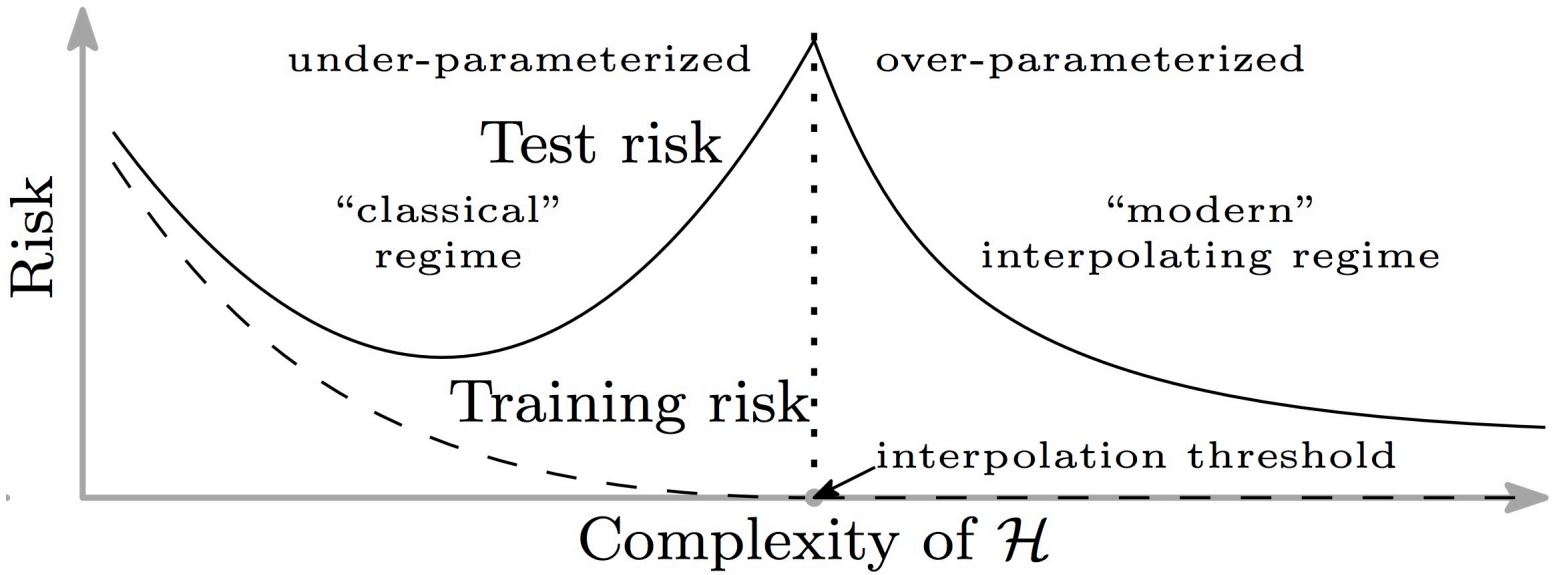
Interpolating, yet not overfitting



- MNIST: 50,000 images in 2 different classes.
- 5-layers Neural Net. Quadratic hinge loss.

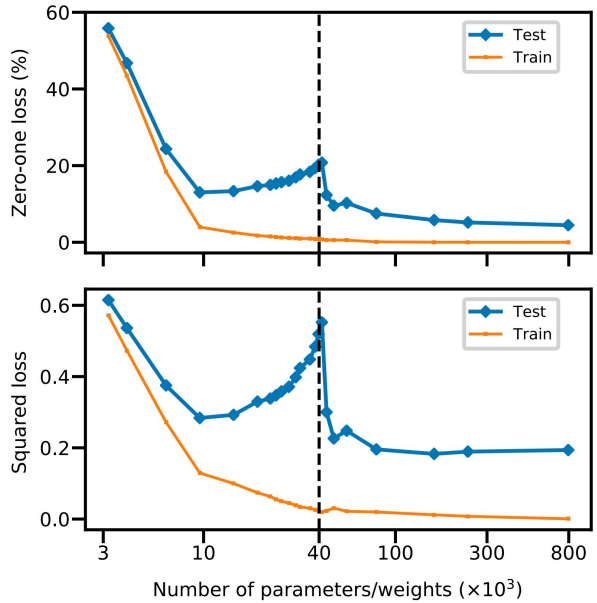
[Spigler, Geiger, d'Ascoli, Sagun, Biroli, Wyart, 2018]

The double-descent curve



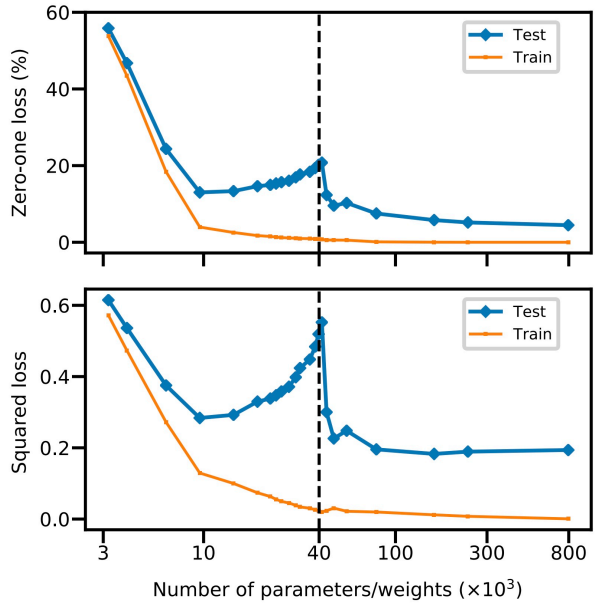
[Belkin, Hsu, Ma, Mandal, 2018]

Desiderata



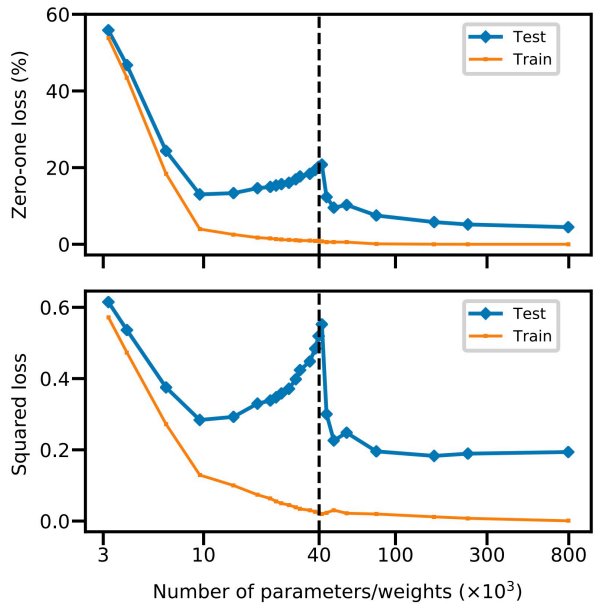
- ✓ Peak at the interpolation threshold
- ✓ Global minimum in the overparametrized regime
- ✓ Monotone decreasing in the overparametrized regime
- ✓ Optimal regularization $\to 0$.

Desiderata



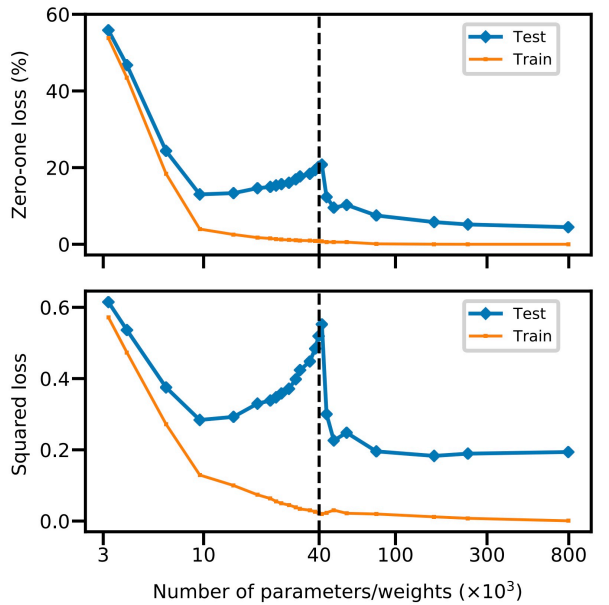
- ✓ Peak at the interpolation threshold
- ✓ Global minimum in the overparametrized regime
- ✓ Monotone decreasing in the overparametrized regime
- ✓ Optimal regularization $\to 0$.

Desiderata



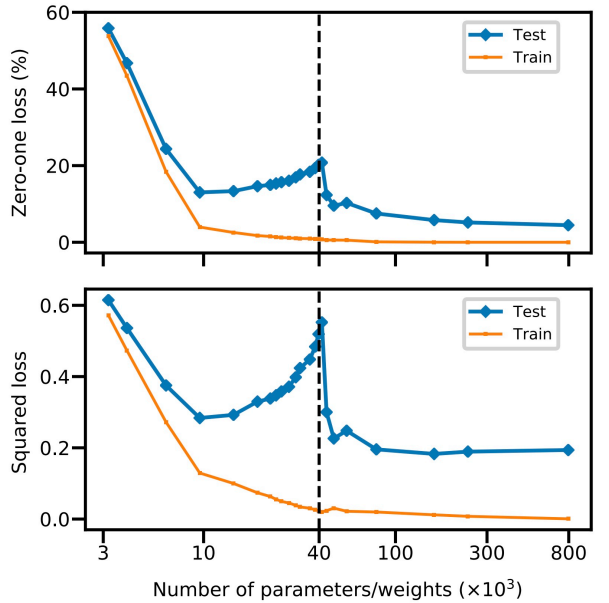
- ✓ Peak at the interpolation threshold
- ✓ Global minimum in the overparametrized regime
- ✓ Monotone decreasing in the overparametrized regime
- ✓ Optimal regularization $\to 0$.

Desiderata



- ✓ Peak at the interpolation threshold
- ✓ Global minimum in the overparametrized regime
- ✓ Monotone decreasing in the overparametrized regime
- ✓ Optimal regularization $\to 0$.

Desiderata



- ✓ Peak at the interpolation threshold
- ✓ Global minimum in the overparametrized regime
- ✓ Monotone decreasing in the overparametrized regime
- ✓ Optimal regularization $\to 0$.

Random features model

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

- $\boldsymbol{W} = [\boldsymbol{w}_1, \dots, \boldsymbol{w}_N]$ random; $\quad \boldsymbol{w}_i \sim \mathsf{Unif}(\mathbb{S}^{d-1}(1))$, i.i.d.

Ridge regression

$$\widehat{\boldsymbol{a}}(\lambda) = \operatorname{argmin}_{\boldsymbol{a}\in\mathbb{R}^N} \left\{ \widehat{\mathbb{E}}_n \left[\left(y - \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle \right)^2 \right] + \frac{N\lambda}{d} \|\boldsymbol{a}\|_2^2 \right\}$$

Setting

- Random features ridge regression (as above.)
- Proportional regime $n, N, d \to \infty$,

$$\frac{N}{d} \to \psi_1, \quad \frac{n}{d} \to \psi_2 \, .$$

- $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$
- $f_*(\boldsymbol{x}) = \langle \boldsymbol{\beta}_0, \boldsymbol{x} \rangle$ (Higher order terms $\to$ See paper)

Setting

- Random features ridge regression (as above.)

- Proportional regime $n, N, d \to \infty$,

$$\frac{N}{d} \to \psi_1, \quad \frac{n}{d} \to \psi_2 \, .$$

- $y_i = f_*(x_i) + \varepsilon_i$
- $f_*(x) = \langle \beta_0, x \rangle$ (Higher order terms $\to$ See paper)

Setting

- Random features ridge regression (as above.)
- Proportional regime $n, N, d \to \infty$,

$$\frac{N}{d} \to \psi_1, \quad \frac{n}{d} \to \psi_2 \, .$$

- $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$
- $f_*(\boldsymbol{x}) = \langle \boldsymbol{\beta}_0, \boldsymbol{x} \rangle$ (Higher order terms $\to$ See paper)

Setting

- Random features ridge regression (as above.)
- Proportional regime $n, N, d \to \infty$,

$$\frac{N}{d} \to \psi_1, \quad \frac{n}{d} \to \psi_2 \, .$$

- $y_i = f_*(\boldsymbol{x}_i) + \varepsilon_i$
- $f_*(\boldsymbol{x}) = \langle \boldsymbol{\beta}_0, \boldsymbol{x} \rangle$ (Higher order terms $\to$ See paper)

Precise asymptotics

Theorem (Mei, M. 2019)

Assume $f_*(x) = \langle \beta_0, x \rangle$ *and define (for* $G \sim \mathsf{N}(0,1)$*)*

$$b_*^2 = \mathbb{E}[\sigma(G)^2] - \mathbb{E}[\sigma(G)]^2 - b_1^2, \quad b_1 = \mathbb{E}[G\sigma(G)], \qquad \zeta \equiv \frac{b_1^2}{b_*^2}\,.$$

Then, for any $\lambda > 0$*, we have*

$$R_{\mathsf{RF}}(\hat{f}_\lambda) = \|\beta_0\|_2^2 \mathscr{B}(\zeta, \psi_1, \psi_2, \lambda/\overline{b}_*^2) + \tau^2 \mathscr{V}(\zeta, \psi_1, \psi_2, \lambda/\overline{b}_*^2) + o_d(1)\,,$$

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)

Assume $f_*(x) = \langle \beta_0, x \rangle$ *and define (for* $G \sim \mathsf{N}(0,1)$*)*

$$b_*^2 = \mathbb{E}[\sigma(G)^2] - \mathbb{E}[\sigma(G)]^2 - b_1^2, \quad b_1 = \mathbb{E}[G\sigma(G)], \qquad \zeta \equiv \frac{b_1^2}{b_*^2}\,.$$

Then, for any $\lambda > 0$*, we have*

$$R_{\mathsf{RF}}(\hat{f}_\lambda) = \|\beta_0\|_2^2 \mathscr{B}(\zeta, \psi_1, \psi_2, \lambda/\overline{b}_*^2) + \tau^2 \mathscr{V}(\zeta, \psi_1, \psi_2, \lambda/\overline{b}_*^2) + o_d(1)\,,$$

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

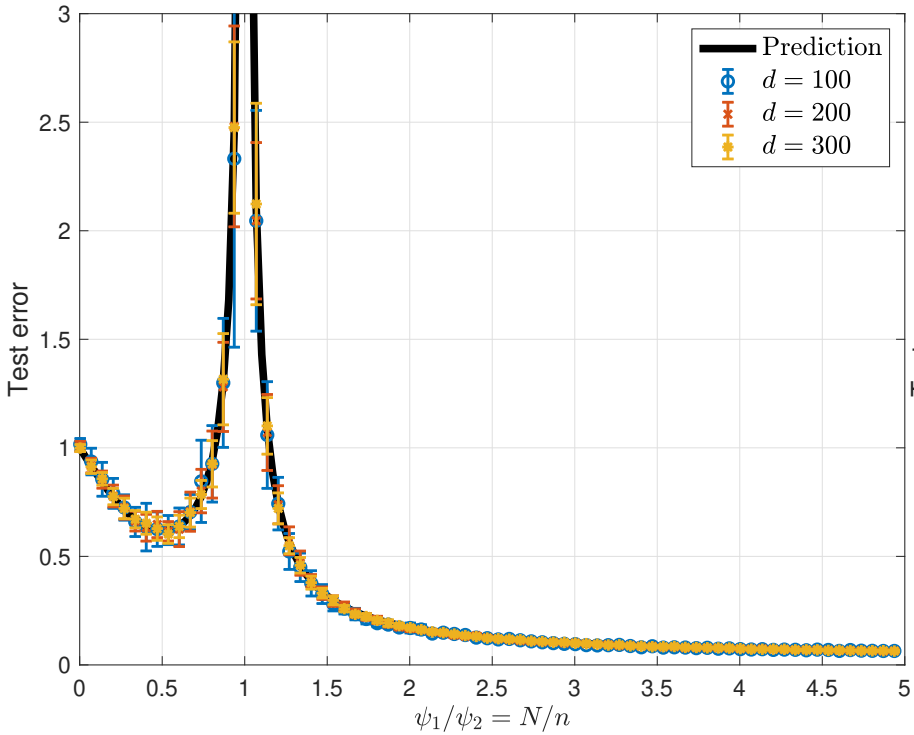
$$\begin{aligned}
\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv\ & -\chi^5\zeta^6 + 3\chi^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\chi^3\zeta^6 - 2\chi^3\zeta^4 - 3\chi^3\zeta^2 \\
& + (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\chi^2\zeta^4 + 2\chi^2\zeta^2 + \chi^2 + 3\psi_1\psi_2\chi\zeta^2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

We then have

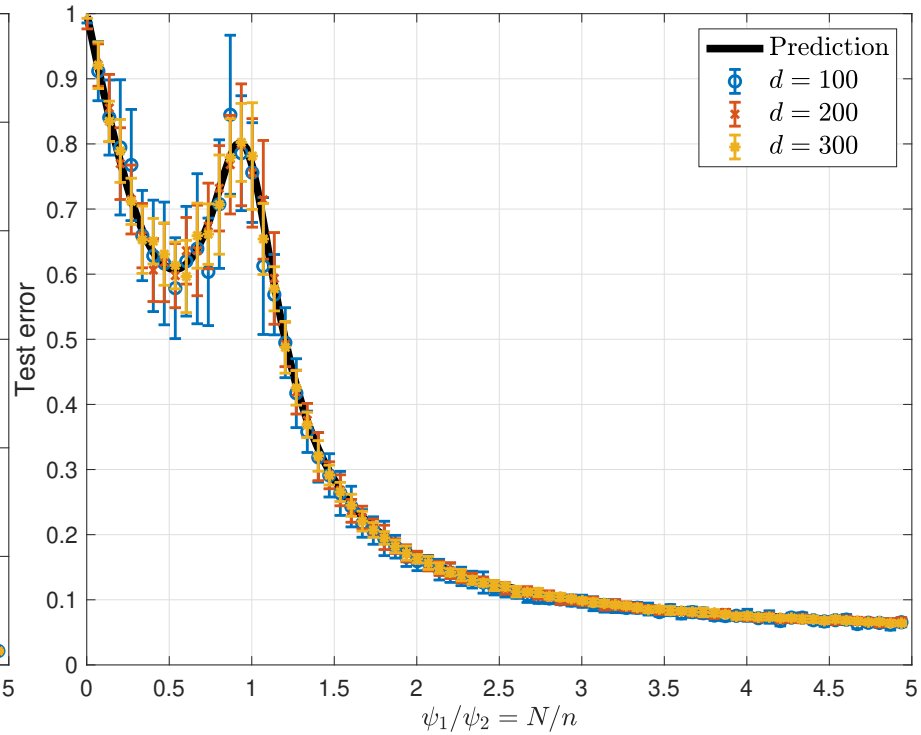
$$\mathscr{B}(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv \frac{\mathscr{E}_1(\zeta, \psi_1, \psi_2, \overline{\lambda})}{\mathscr{E}_0(\zeta, \psi_1, \psi_2, \overline{\lambda})}, \qquad \mathscr{V}(\zeta, \psi_1, \psi_2, \overline{\lambda}) \equiv \frac{\mathscr{E}_2(\zeta, \psi_1, \psi_2, \overline{\lambda})}{\mathscr{E}_0(\zeta,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

Random matrix theory for kernel inner product random matrices

[Cheng, Singer, 2013; Do, Vu 2013; Fan, M. 2019; Penington, Wohra, 2017;. . .]

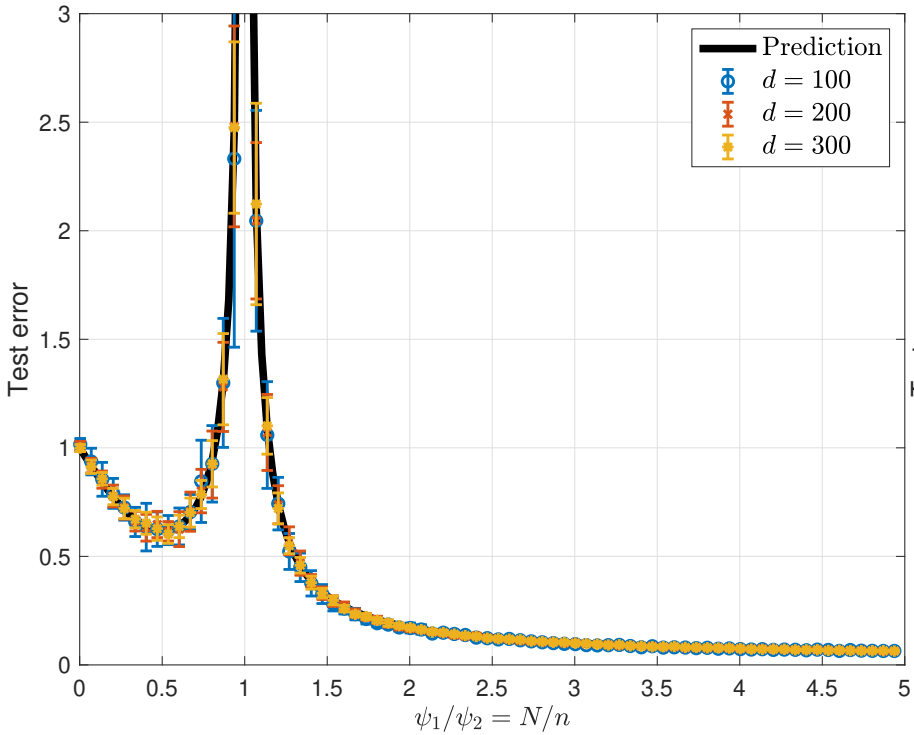


$\lambda = 0+$

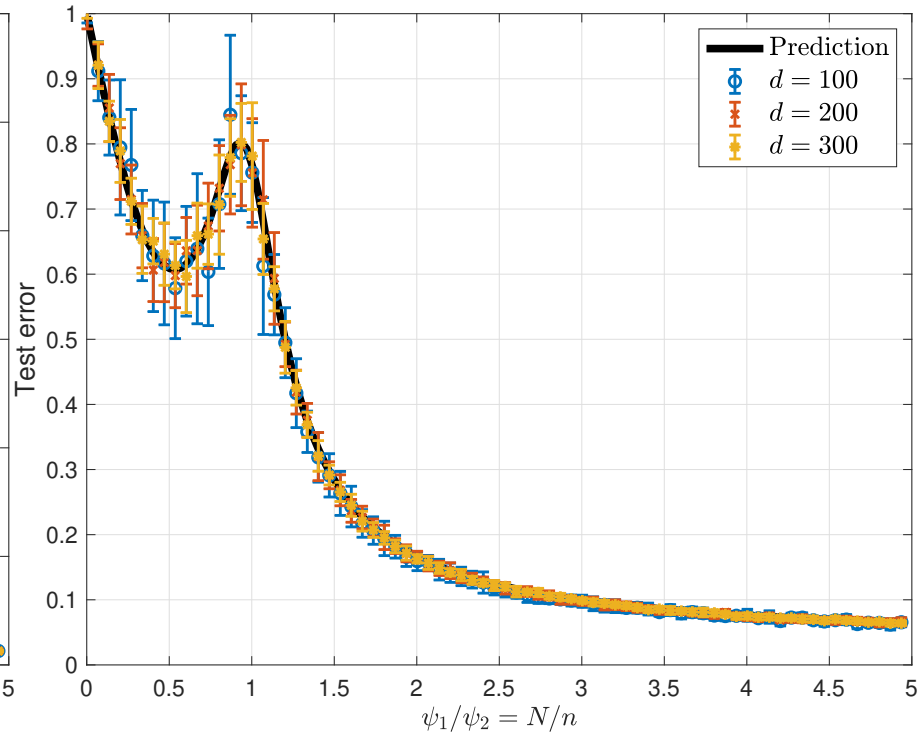


$\lambda > 0$

- ✓ Peak at the interpolation threshold
- ✓ Global minimum in the overparametrized regime
- ✓ Monotone decreasing in the overparametrized regime



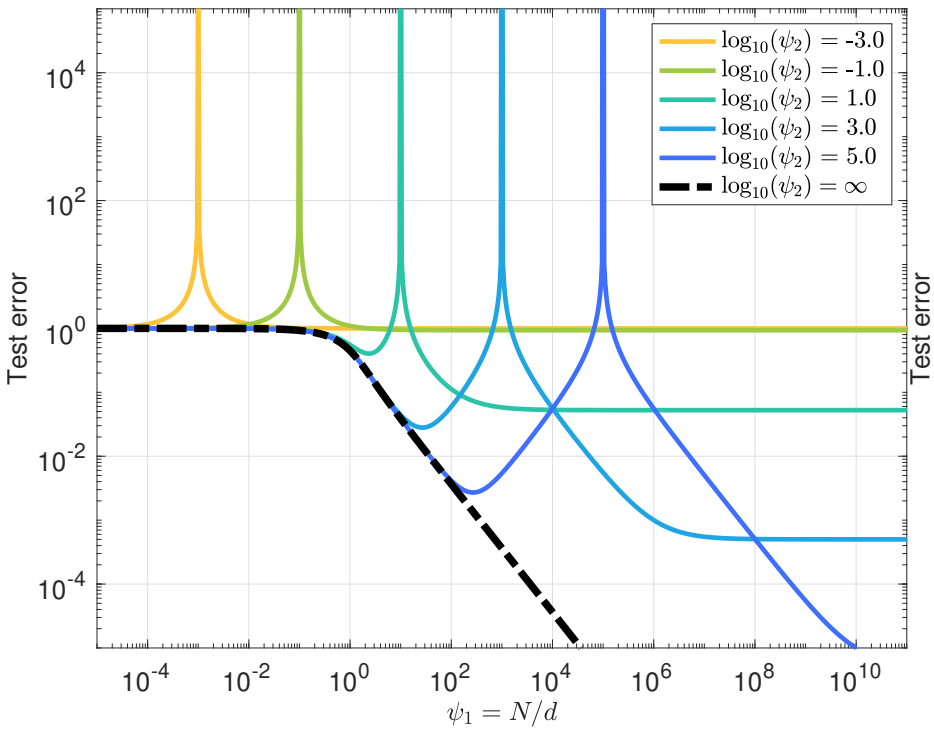
$\lambda = 0+$



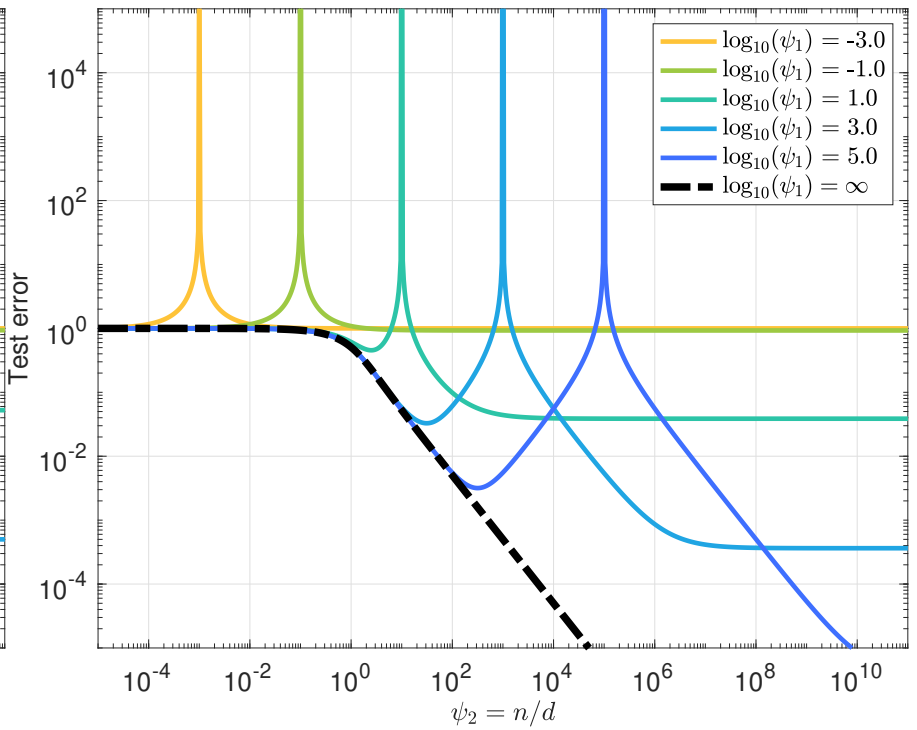
$\lambda > 0$

- ✓ Peak at the interpolation threshold
- ✓ Global minimum in the overparametrized regime
- ✓ Monotone decreasing in the overparametrized regime

Insigths

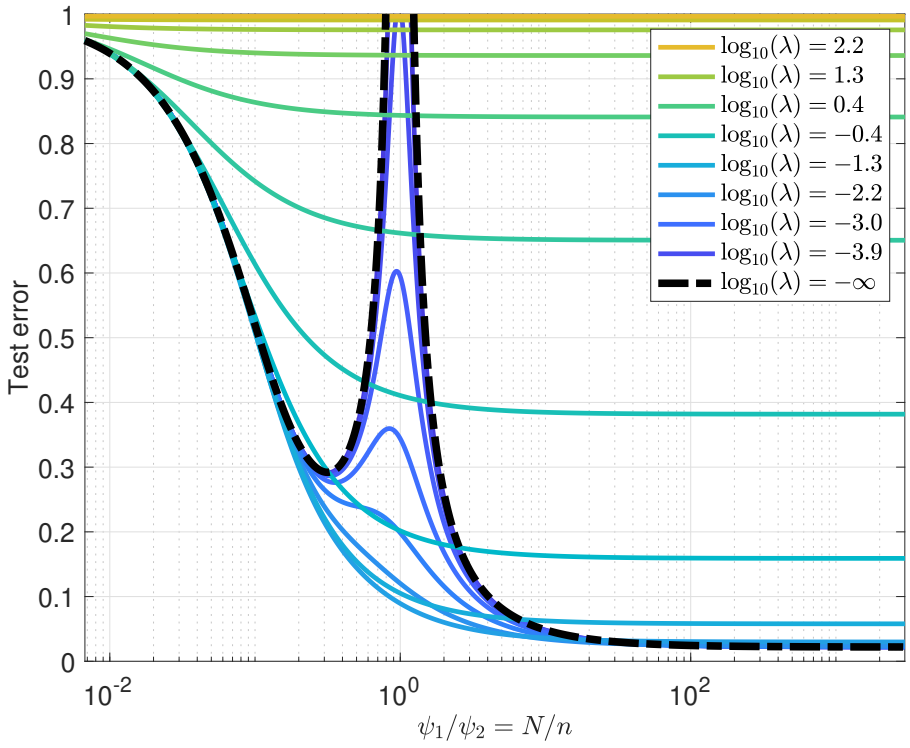


$\lambda = 0+$

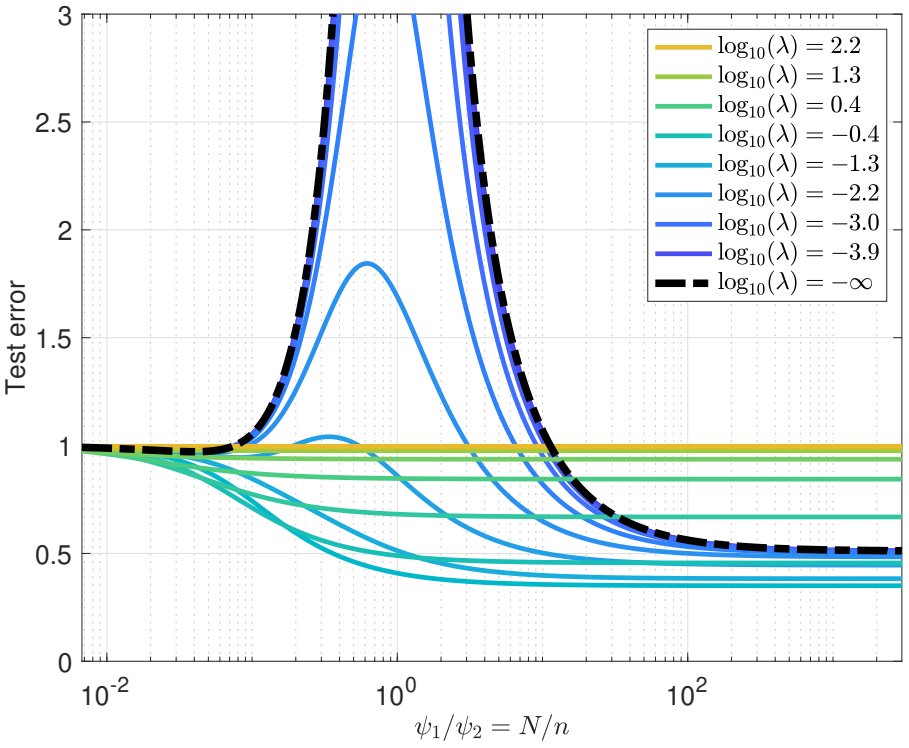


$\lambda = 0+$

- Singularity at the interpolation threshold
- Minimum risk at extreme overparametrization $N/n \to \infty$.

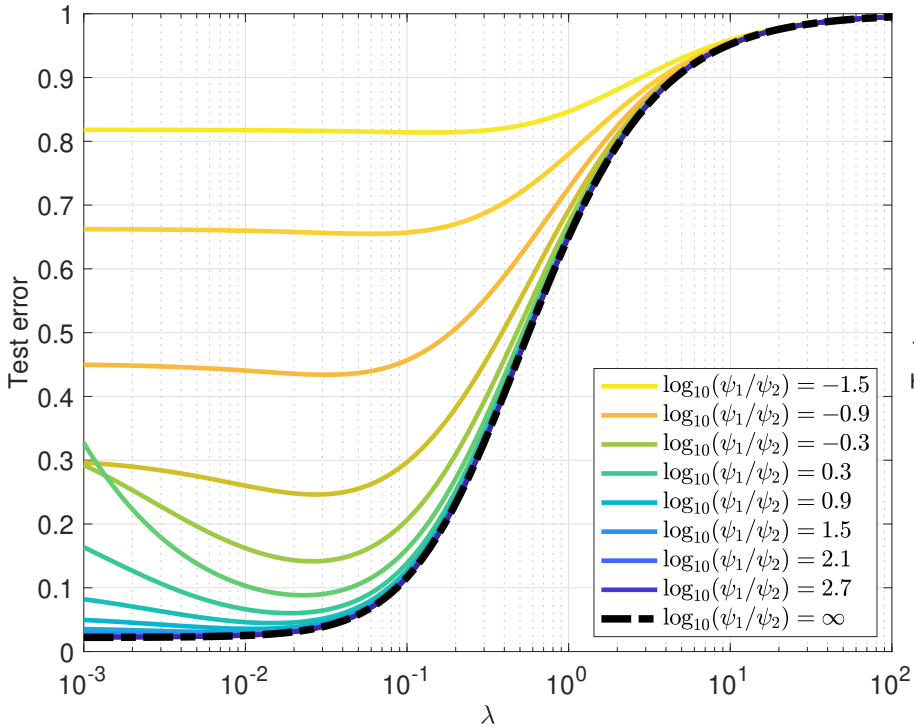


SNR= 5

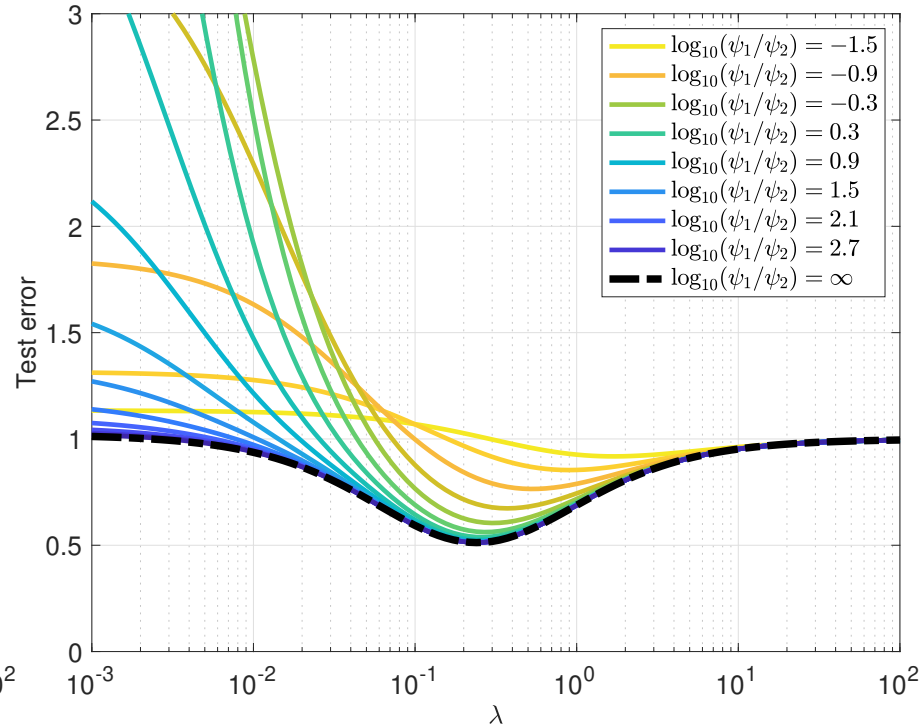


SNR= 1/5

- Same at $\lambda > 0$ fixed: Minimum at $N/n \to \infty$.



SNR= 5



SNR= 1/10

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

Beyond square loss?

Binary classification

- Data $(y_i, \boldsymbol{x}_i) \in \{+1, -1\} \times \mathbb{R}^d, \quad \boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$

$$\mathbb{P}(y_i = +1 | \boldsymbol{x}_i) = f_*(\langle \boldsymbol{\beta}_0, \boldsymbol{x}_i \rangle)$$

- Two-layer network

$$\hat{f}_a(x) = \text{sign}\left\{ \sum_{i=1}^{N} a_i \sigma(\langle w_i, x \rangle) \right\}$$

- Prediction error

$$\text{Err}_n \equiv \mathbb{P}(y^{\text{new}} \hat{f}_{\widehat{a}}(x^{\text{new}}) \le 0) \, ,$$

Binary classification

- Data $(y_i, \boldsymbol{x}_i) \in \{+1, -1\} \times \mathbb{R}^d, \quad \boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$

$$\mathbb{P}(y_i = +1 | \boldsymbol{x}_i) = f_*(\langle \boldsymbol{\beta}_0, \boldsymbol{x}_i \rangle)$$

- Two-layer network

$$\hat{f}_{\boldsymbol{a}}(\boldsymbol{x}) = \mathrm{sign}\left\{ \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \right\}$$

- Prediction error

$$\mathrm{Err}_n \equiv \mathbb{P}(y^{\mathrm{new}} \hat{f}_{\widehat{\boldsymbol{a}}}(\boldsymbol{x}^{\mathrm{new}}) \le 0) \, ,$$

Binary classification

- Data $(y_i, \boldsymbol{x}_i) \in \{+1, -1\} \times \mathbb{R}^d, \quad \boldsymbol{x}_i \sim \mathsf{N}(0, \boldsymbol{I}_d)$

$$\mathbb{P}(y_i = +1 | \boldsymbol{x}_i) = f_*(\langle \boldsymbol{\beta}_0, \boldsymbol{x}_i \rangle)$$

- Two-layer network

$$\hat{f}_{\boldsymbol{a}}(x) = \text{sign} \left\{ \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, x \rangle) \right\}$$

- Prediction error

$$\text{Err}_n \equiv \mathbb{P}\big(y^{\text{new}} \hat{f}_{\widehat{\boldsymbol{a}}}(\boldsymbol{x}^{\text{new}}) \le 0\big)\,,$$

Max-margin classification

Random features:

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

Maximize the margin:

$$\begin{aligned} \text{maximize} \quad & \min_{i \leq n} y_i \hat{f}_{\boldsymbol{a}}(\boldsymbol{x}_i) = \min_{i \leq n} y_i \langle \boldsymbol{a}, \sigma(\boldsymbol{W}\boldsymbol{x}_i) \rangle \,, \\ \text{subject to} \quad & \|\boldsymbol{a}\|_2 \leq 1 \,. \end{aligned}$$

Theoretical challenges

$$\max \Big\{ \min_{i \le n} \, y_i \langle \boldsymbol{a}, \sigma(\boldsymbol{W} \boldsymbol{x}_i) \rangle \; : \quad \|\boldsymbol{a}\| \le 1 \Big\}$$

- Can't use random matrix theory
- $\sigma(\boldsymbol{W} \boldsymbol{x}_i)$ has correlated entries
- Non-gaussian

Theoretical challenges

$$\max \Big\{ \min_{i \le n} \; y_i \langle \boldsymbol{a}, \sigma(\boldsymbol{W} \boldsymbol{x}_i) \rangle \; : \quad \|\boldsymbol{a}\| \le 1 \Big\}$$

- Can't use random matrix theory
- $\sigma(\boldsymbol{W} \boldsymbol{x}_i)$ has correlated entries
- Non-gaussian

Theoretical challenges

$$\max \Big\{ \min_{i \le n} \, y_i \langle \boldsymbol{a}, \sigma(\boldsymbol{W} \boldsymbol{x}_i) \rangle \; : \quad \|\boldsymbol{a}\| \le 1 \Big\}$$

- Can't use random matrix theory
- $\sigma(\boldsymbol{W} \boldsymbol{x}_i)$ has correlated entries
- Non-gaussian

Theoretical challenges

$$\max \Big\{ \min_{i \le n} \, y_i \langle \boldsymbol{a}, \sigma(\boldsymbol{W} \boldsymbol{x}_i) \rangle \; : \quad \|\boldsymbol{a}\| \le 1 \Big\}$$

- Can't use random matrix theory
- $\sigma(\boldsymbol{W} \boldsymbol{x}_i)$ has correlated entries
- Non-gaussian

Idea: 'Noisy linear features model'

Nonlinear features

$$\hat{f}_a(x_i) = \text{sign}(\langle a, u \rangle) \,,$$
$$u_{ij} = \sigma(\langle w_j, x_i \rangle) = b_1 \langle w_j, x_i \rangle + b_* \sigma_\perp(\langle w_j, x_i \rangle)$$

Noisy linear features

$$\hat{f}_a(x_i) = \text{sign}(\langle a, \tilde{u} \rangle) \,,$$
$$\tilde{u}_{ij} = b_1 \langle w_j, x_i \rangle + b_* z_{ij} \,, \qquad (z_{ij}) \sim_{iid} \mathsf{N}(0,1)$$

Gaussian, correlated

Idea: 'Noisy linear features model'

Nonlinear features

$$\hat{f}_a(x_i) = \text{sign}(\langle a, u \rangle)\,,$$
$$u_{ij} = \sigma(\langle w_j, x_i \rangle) = b_1 \langle w_j, x_i \rangle + b_* \sigma_\perp(\langle w_j, x_i \rangle)$$

Noisy linear features

$$\hat{f}_a(x_i) = \text{sign}(\langle a, \tilde{u} \rangle)\,,$$
$$\tilde{u}_{ij} = b_1 \langle w_j, x_i \rangle + b_* z_{ij}\,, \qquad (z_{ij}) \sim_{iid} \mathsf{N}(0,1)$$

Gaussian, correlated

Idea: 'Noisy linear features model'

Nonlinear features

$$\hat{f}_a(x_i) = \text{sign}(\langle a, u \rangle) \,,$$
$$u_{ij} = \sigma(\langle w_j, x_i \rangle) = b_1 \langle w_j, x_i \rangle + b_* \sigma_\perp(\langle w_j, x_i \rangle)$$

Noisy linear features

$$\hat{f}_a(x_i) = \text{sign}(\langle a, \tilde{u} \rangle) \,,$$
$$\tilde{u}_{ij} = b_1 \langle w_j, x_i \rangle + b_* z_{ij} \,, \qquad (z_{ij}) \sim_{iid} \mathsf{N}(0,1)$$

Gaussian, correlated

Idea: 'Noisy linear features model'

Nonlinear features

$$\hat{f}_{\boldsymbol{a}}(\boldsymbol{x}_i) = \text{sign}\big(\langle \boldsymbol{a}, \boldsymbol{u} \rangle\big)\,,$$
$$u_{ij} = \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle) = b_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + b_* \sigma_\perp(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle)$$

Noisy linear features

$$\hat{f}_{\boldsymbol{a}}(\boldsymbol{x}_i) = \text{sign}\big(\langle \boldsymbol{a}, \tilde{\boldsymbol{u}} \rangle\big)\,,$$
$$\tilde{u}_{ij} = b_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + b_* z_{ij}\,, \qquad (z_{ij}) \sim_{iid} \mathsf{N}(0,1)$$

Gaussian, correlated

Idea: 'Noisy linear features model'

Nonlinear features

$$\hat{f}_{\boldsymbol{a}}(\boldsymbol{x}_i) = \text{sign}\big(\langle \boldsymbol{a}, \boldsymbol{u} \rangle\big)\,,$$
$$u_{ij} = \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle) = b_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + b_* \sigma_\perp(\langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle)$$

Noisy linear features

$$\hat{f}_{\boldsymbol{a}}(\boldsymbol{x}_i) = \text{sign}\big(\langle \boldsymbol{a}, \tilde{\boldsymbol{u}} \rangle\big)\,,$$
$$\tilde{u}_{ij} = b_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + b_* z_{ij}\,, \qquad (z_{ij}) \sim_{iid} \mathsf{N}(0,1)$$

Gaussian, correlated

Asymptotic equivalence

Theorem (Mei, M, 2019)

Consider ridge regression in the proportional asymptotics $d \to \infty$, $N/d \to \psi_1$, $n/d \to \psi_2$.

Then the nonlinear features model and the noisy linear features model are 'asymptotically equivalent:' they have asymptotically the same test error, training error, ...

Conjecture

Consider max-margin classification in the proportional asymptotics.

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

Asymptotic equivalence

Theorem (Mei, M, 2019)

Consider ridge regression in the proportional asymptotics $d \to \infty$, $N/d \to \psi_1$, $n/d \to \psi_2$.

Then the nonlinear features model and the noisy linear features model are 'asymptotically equivalent:' they have asymptotically the same test error, training error, ...

Conjecture

Consider max-margin classification in the proportional asymptotics.

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

Asymptotics of max-margin classification

Theorem (M, Ruan, Sogn, Yan, 2019)

Assume $\mathbb{P}(y_i = +1|\boldsymbol{x}_i) = f_0(\langle \boldsymbol{\beta}_0, \boldsymbol{x}_i\rangle)$, $\boldsymbol{x} \sim \mathsf{N}(0, \boldsymbol{I}_d)$, *and consider max-margin classification from noisy features* $\tilde{u}_{ij} = b_1\langle \boldsymbol{w}_j, \boldsymbol{x}_i\rangle + b_* z_{ij}$.

Let κ_n *be the margin and* Err_n *the test error. Then, we have*

$$\mathrm{Err}_n = \mathrm{Err}_*(\zeta, \psi_1, \psi_2, f_0) + o_P(1)\,, \quad \kappa_n = \kappa_*(\zeta, \psi_1, \psi_2, f_0) + o_P(1)\,.$$

where $\mathrm{Err}_*(\zeta, \psi_1, \psi_2, f_0)$, $\kappa_*(\zeta, \psi_1, \psi_2, f_0)$ *are explicitly given.*

- Paper: general Gaussian features.
- Statistical physics: Gardner 1988, Seung et al. 1992, ...

Asymptotics of max-margin classification

Theorem (M, Ruan, Sogn, Yan, 2019)

Assume $\mathbb{P}(y_i = +1|\boldsymbol{x}_i) = f_0(\langle \boldsymbol{\beta}_0, \boldsymbol{x}_i \rangle)$, $\boldsymbol{x} \sim \mathsf{N}(0, \boldsymbol{I}_d)$, *and consider max-margin classification from noisy features* $\tilde{u}_{ij} = b_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + b_* z_{ij}$.

Let κ_n *be the margin and* Err_n *the test error. Then, we have*

$$\mathrm{Err}_n = \mathrm{Err}_*(\zeta, \psi_1, \psi_2, f_0) + o_P(1)\,, \quad \kappa_n = \kappa_*(\zeta, \psi_1, \psi_2, f_0) + o_P(1)\,.$$

where $\mathrm{Err}_*(\zeta, \psi_1, \psi_2, f_0)$, $\kappa_*(\zeta, \psi_1, \psi_2, f_0)$ *are explicitly given.*

- Paper: general Gaussian features.
- Statistical physics: Gardner 1988, Seung et al. 1992, ...

Asymptotics of max-margin classification

Theorem (M, Ruan, Sogn, Yan, 2019)

Assume $\mathbb{P}(y_i = +1|\boldsymbol{x}_i) = f_0(\langle \boldsymbol{\beta}_0, \boldsymbol{x}_i \rangle)$, $\boldsymbol{x} \sim \mathsf{N}(0, \boldsymbol{I}_d)$, *and consider max-margin classification from noisy features* $\tilde{u}_{ij} = b_1 \langle \boldsymbol{w}_j, \boldsymbol{x}_i \rangle + b_* z_{ij}$.

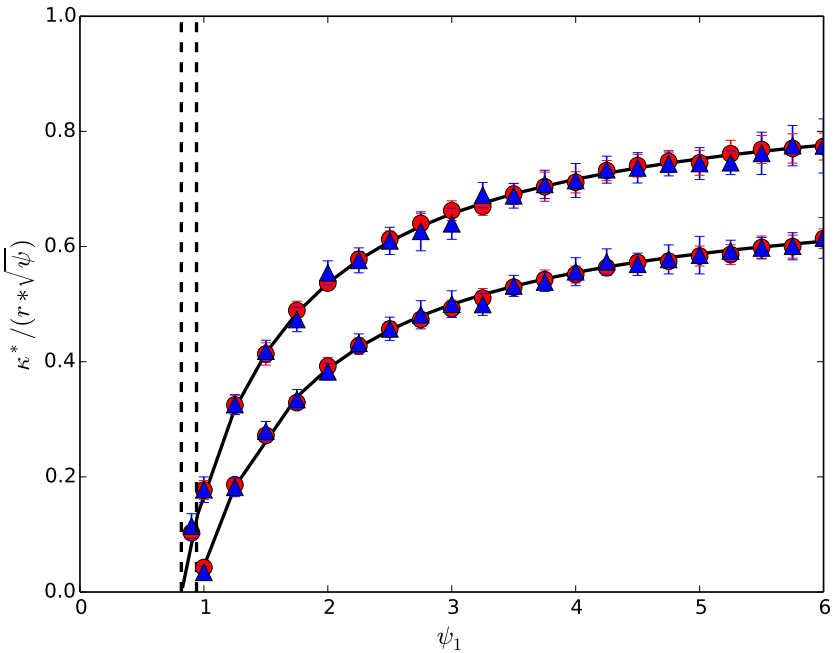
Let κ_n *be the margin and* Err_n *the test error. Then, we have*

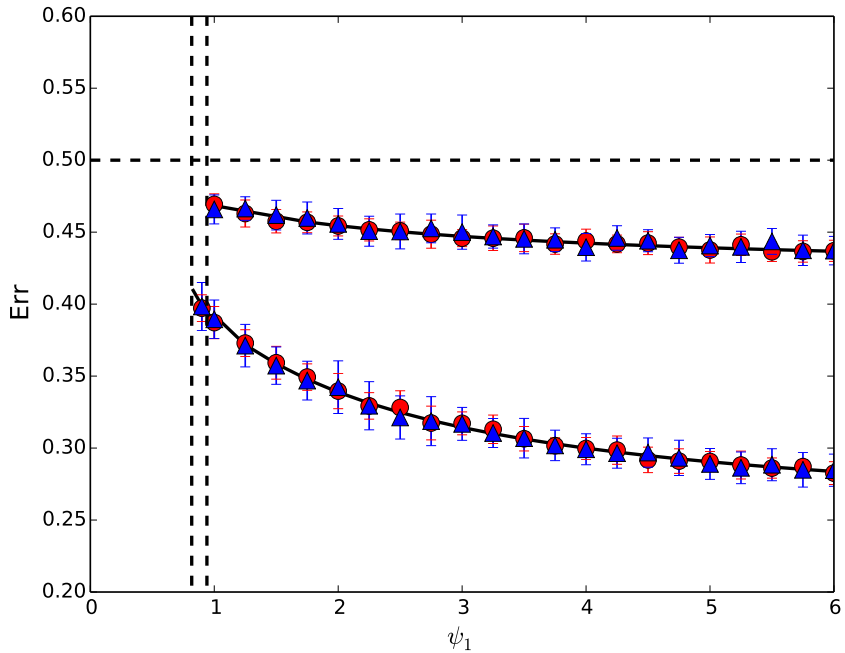
$$\mathrm{Err}_n = \mathrm{Err}_*(\zeta, \psi_1, \psi_2, f_0) + o_P(1)\,, \quad \kappa_n = \kappa_*(\zeta, \psi_1, \psi_2, f_0) + o_P(1)\,.$$

where $\mathrm{Err}_*(\zeta, \psi_1, \psi_2, f_0)$, $\kappa_*(\zeta, \psi_1, \psi_2, f_0)$ *are explicitly given.*

- Paper: general Gaussian features.
- Statistical physics: Gardner 1988, Seung et al. 1992, ...

Comparison





- $\psi_2 = n/d = 2,\ d = 200, 400,$ ReLU activations

What is not captured by linear models?

What is not captured by linear models?

- Approximation
- Low-dimensional structures
- A ‘simple’ example

[see 1904.12191, 1906.08899]

What is not captured by linear models?

- Approximation
- Low-dimensional structures
- **A 'simple' example**

[see 1904.12191, 1906.08899]

Setting

- $\boldsymbol{x}_i \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$,

$$y_i = f_*(\boldsymbol{x}_i) \equiv b_0 + \langle \boldsymbol{x}_i, \boldsymbol{B}\boldsymbol{x}_i \rangle \quad \text{with} \quad \boldsymbol{B} \succeq 0.$$

- Optimum test error $n = \infty$

$$R_{\mathsf{M},N} = \min_{\hat{f} \in \mathcal{F}_{\mathsf{M},N}(\boldsymbol{W})} \mathbb{E}\{(f_*(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2\}, \quad \mathsf{M} \in \{\mathsf{RF}, \mathsf{NT}, \mathsf{NN}\}.$$

- Here $\sigma(x) = x^2$ (cf. paper for generalizations)

Setting

- $\boldsymbol{x}_i \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$,

$$y_i = f_*(\boldsymbol{x}_i) \equiv b_0 + \langle \boldsymbol{x}_i, \boldsymbol{B}\boldsymbol{x}_i \rangle \quad \text{with} \quad \boldsymbol{B} \succeq 0.$$

- Optimum test error $n = \infty$

$$R_{\mathsf{M},N} = \min_{\hat{f} \in \mathcal{F}_{\mathsf{M},N}(\boldsymbol{W})} \mathbb{E}\{(f_*(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2\}, \quad \mathsf{M} \in \{\mathsf{RF}, \mathsf{NT}, \mathsf{NN}\}.$$

- Here $\sigma(x) = x^2$ (cf. paper for generalizations)

Setting

- $\boldsymbol{x}_i \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$,

$$y_i = f_*(\boldsymbol{x}_i) \equiv b_0 + \langle \boldsymbol{x}_i, \boldsymbol{B}\boldsymbol{x}_i \rangle \quad \text{with} \quad \boldsymbol{B} \succeq 0.$$

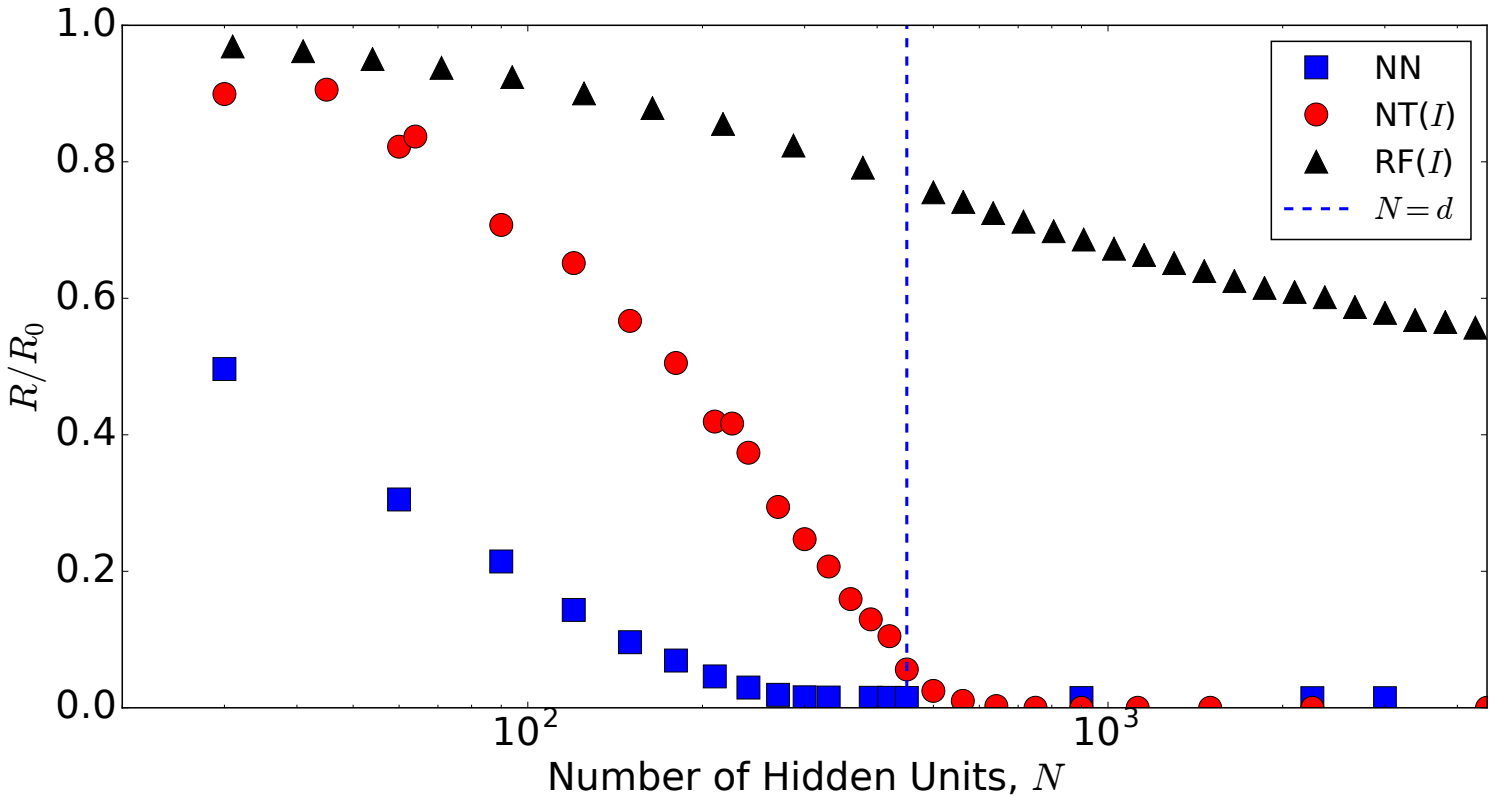
- Optimum test error $n = \infty$

$$R_{\mathsf{M},N} = \min_{\hat{f} \in \mathcal{F}_{\mathsf{M},N}(\boldsymbol{W})} \mathbb{E}\{(f_*(\boldsymbol{x}) - \hat{f}(\boldsymbol{x}))^2\}, \quad \mathsf{M} \in \{\mathsf{RF}, \mathsf{NT}, \mathsf{NN}\}.$$

- Here $\sigma(x) = x^2$ (cf. paper for generalizations)

Experiment

- $\boldsymbol{B} \in \mathbb{R}^{450\times 450}$, $\lambda_i(\boldsymbol{b}) \sim_{iid} \exp(1)$
- N varies in $\{30, \cdots, 4500\}$.



Quadratic Model - RF

RF Model

$f_{\mathsf{RF}}(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$ where $\boldsymbol{w}_i \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Gamma})$ and a_i are trained.

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

Assume $\sigma(x) = x^2 - 1$. Then we have, as $N, d \to \infty$ with $N/d \to \psi_1$:

$$R_{\mathsf{RF},N} = \|f_*\|_{L_2}^2 \left(1 - \frac{\psi_1 d \langle \boldsymbol{B}, \boldsymbol{\Gamma} \rangle^2}{\|\boldsymbol{B}\|_F^2 (1 + \psi_1 d \|\boldsymbol{\Gamma}\|_F^2)} + o_{d,\mathbb{P}}(1) \right) .$$

- The correlation between $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$ controls the risk.

Quadratic Model - RF

RF Model

$f_{\mathsf{RF}}(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$ where $\boldsymbol{w}_i \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Gamma})$ and a_i are trained.

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

Assume $\sigma(x) = x^2 - 1$. *Then we have, as* $N, d \to \infty$ *with* $N/d \to \psi_1$*:*

$$R_{\mathsf{RF},N} = \|f_*\|_{L_2}^2 \left(1 - \frac{\psi_1 \, d \langle \boldsymbol{B}, \boldsymbol{\Gamma} \rangle^2}{\|\boldsymbol{B}\|_F^2 (1 + \psi_1 \, d \|\boldsymbol{\Gamma}\|_F^2)} + o_{d,\mathbb{P}}(1) \right) .$$

- The correlation between $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$ controls the risk.

Quadratic Model - RF

RF Model

$f_{\mathsf{RF}}(\boldsymbol{x}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$ where $\boldsymbol{w}_i \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Gamma})$ and a_i are trained.

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

Assume $\sigma(x) = x^2 - 1$. *Then we have, as* $N, d \to \infty$ *with* $N/d \to \psi_1$*:*

$$R_{\mathsf{RF},N} = \|f_*\|_{L_2}^2 \left(1 - \frac{\psi_1 \, d \langle \boldsymbol{B}, \boldsymbol{\Gamma} \rangle^2}{\|\boldsymbol{B}\|_F^2 (1 + \psi_1 \, d \|\boldsymbol{\Gamma}\|_F^2)} + o_{d,\mathbb{P}}(1) \right) .$$

- The correlation between $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$ controls the risk.

Quadratic Model - RF

$$\lim_{\psi_1 \to \infty} \lim_{d \to \infty, N/d \to \psi_1} \frac{R_{\mathsf{RF},N}}{\mathbb{E}\{f_*^2\}} = \lim_{d \to \infty} \left(1 - \frac{\langle \boldsymbol{\Gamma}, \boldsymbol{B} \rangle^2}{\|\boldsymbol{\Gamma}\|_F^2 \|\boldsymbol{B}\|_F^2} \right) .$$

- Risk vanishes only if $\boldsymbol{\Gamma}$ is chosen perfectly and $\psi_1 \to \infty$.
- This result is true for any activation
- The asymptotic risk is independent of the non-linearity!

Quadratic Model - RF

$$\lim_{\psi_1 \to \infty} \lim_{d \to \infty, N/d \to \psi_1} \frac{R_{\mathsf{RF},N}}{\mathbb{E}\{f_*^2\}} = \lim_{d \to \infty} \left(1 - \frac{\langle \boldsymbol{\Gamma}, \boldsymbol{B} \rangle^2}{\|\boldsymbol{\Gamma}\|_F^2 \|\boldsymbol{B}\|_F^2} \right) .$$

- Risk vanishes only if $\boldsymbol{\Gamma}$ is chosen perfectly and $\psi_1 \to \infty$.
- This result is true for any activation
- The asymptotic risk is independent of the non-linearity!

Quadratic Model - RF

$$\lim_{\psi_1 \to \infty} \lim_{d \to \infty, N/d \to \psi_1} \frac{R_{\mathsf{RF},N}}{\mathbb{E}\{f_*^2\}} = \lim_{d \to \infty} \left(1 - \frac{\langle \boldsymbol{\Gamma}, \boldsymbol{B} \rangle^2}{\|\boldsymbol{\Gamma}\|_F^2 \|\boldsymbol{B}\|_F^2} \right) .$$

- Risk vanishes only if $\boldsymbol{\Gamma}$ is chosen perfectly and $\psi_1 \to \infty$.
- This result is true for any activation
- The asymptotic risk is independent of the non-linearity!

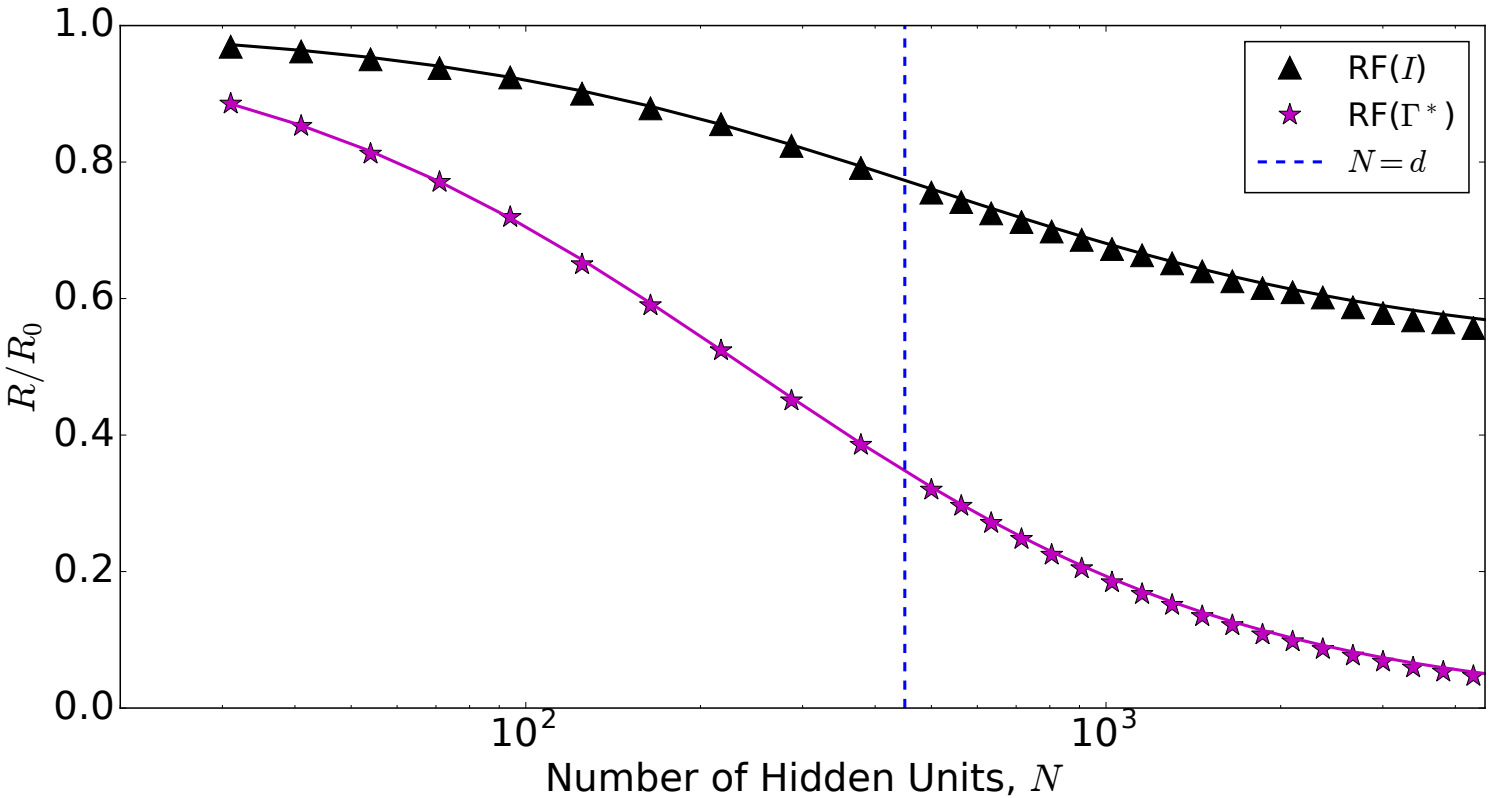
Quadratic Model - RF

$$\lim_{\psi_1 \to \infty} \lim_{d \to \infty, N/d \to \psi_1} \frac{R_{\mathsf{RF},N}}{\mathbb{E}\{f_*^2\}} = \lim_{d \to \infty} \left(1 - \frac{\langle \boldsymbol{\Gamma}, \boldsymbol{B} \rangle^2}{\|\boldsymbol{\Gamma}\|_F^2 \|\boldsymbol{B}\|_F^2} \right) .$$

- Risk vanishes only if $\boldsymbol{\Gamma}$ is chosen perfectly and $\psi_1 \to \infty$.
- This result is true for any activation
- The asymptotic risk is independent of the non-linearity!

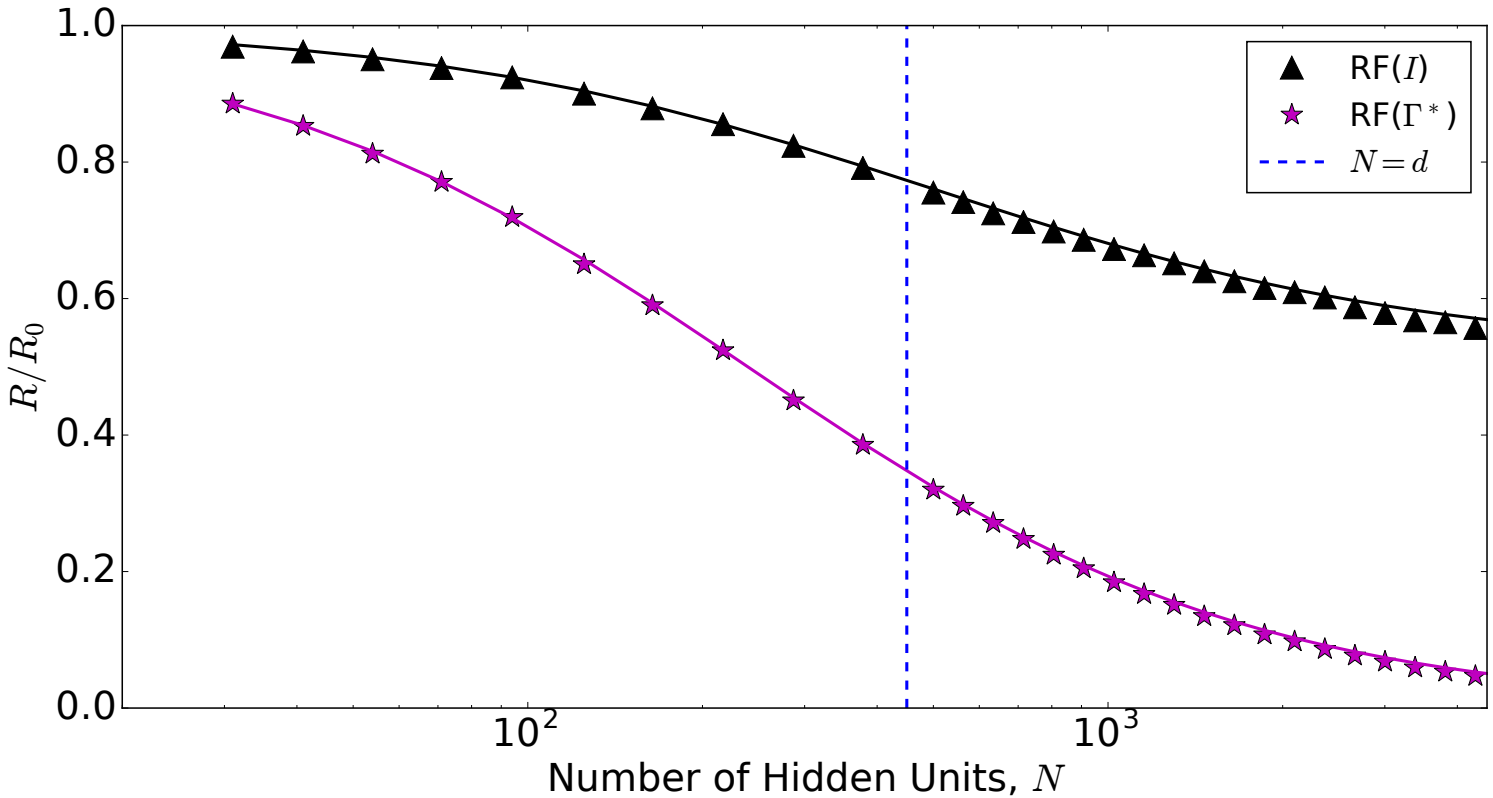
Quadratic Model - RF

- Naive RF does not learn an efficient representation of the data.



Quadratic Model - RF

- Naive RF does not learn an efficient representation of the data.



Quadratic Model - NT

NT Model

- $f_{\mathsf{NT}}(\boldsymbol{x}) = c + \sum_{i=1}^{N} \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$ where $\boldsymbol{w}_i \sim_{i.i.d} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$.

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

For $N, d \to \infty$ with $N/d \to \psi_1$

$$\frac{\mathbb{E}[R_{\mathsf{NT},N}]}{\mathbb{E}\{f_*^2\}} = \left\{ (1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+ \frac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} + o_d(1) \right\}$$

where the expectation is taken over $\boldsymbol{w}_i \sim_{i.i.d} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$.

Quadratic Model - NT

NT Model

- $f_{\mathsf{NT}}(\boldsymbol{x}) = c + \sum_{i=1}^{N} \sigma'(\langle \boldsymbol{w}_i, \boldsymbol{x} \rangle) \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$ where $\boldsymbol{w}_i \sim_{i.i.d} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$.

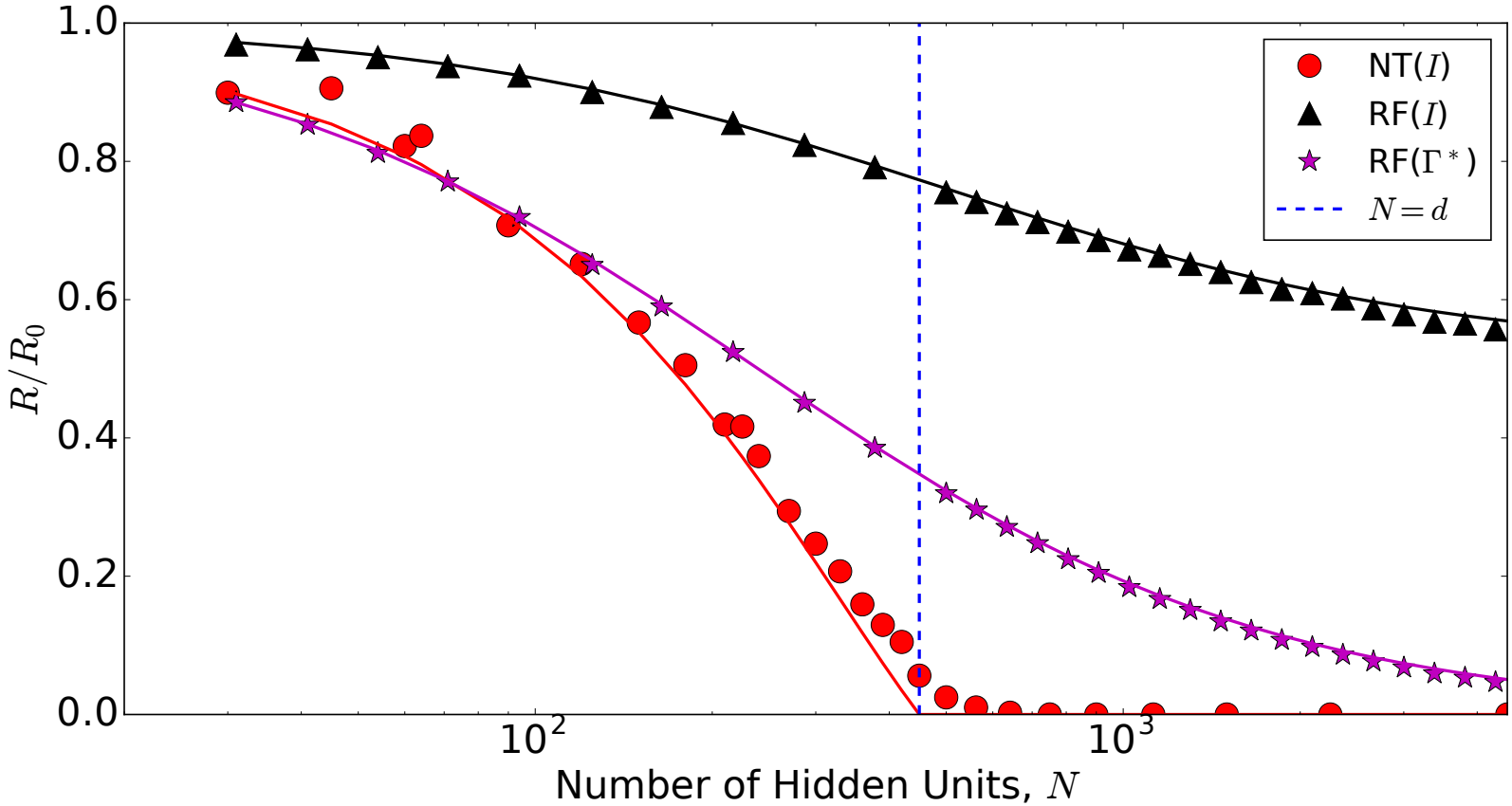
Theorem (Ghorbani, Mei, Misiakiewicz, M., 2019)

For $N, d \to \infty$ with $N/d \to \psi_1$

$$\frac{\mathbb{E}[R_{\mathsf{NT},N}]}{\mathbb{E}\{f_*^2\}} = \left\{ (1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+ \frac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} + o_d(1) \right\}$$

where the expectation is taken over $\boldsymbol{w}_i \sim_{i.i.d} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$.

Quadratic Model - NT



Quadratic Model - NT

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

For $N, d \to \infty$ *with* $N/d \to \psi_1$

$$\frac{\mathbb{E}[R_{\mathsf{NT},N}]}{\mathbb{E}\{f_*^2\}} = \left\{(1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+ \frac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} + o_d(1)\right\}$$

where the expectation is taken over $\boldsymbol{w}_i \sim_{i.i.d} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$.

- Does NT learn subspaces based on their importance in B?
- No! NT fits random directions (but more parameters).

Quadratic Model - NT

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

For $N, d \to \infty$ *with* $N/d \to \psi_1$

$$\frac{\mathbb{E}[R_{\mathsf{NT},N}]}{\mathbb{E}\{f_*^2\}} = \left\{(1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+ \frac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} + o_d(1)\right\}$$

where the expectation is taken over $\boldsymbol{w}_i \sim_{i.i.d} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$.

- Does NT learn subspaces based on their importance in $\boldsymbol{B}$?
- No! NT fits random directions (but more parameters).

Quadratic Model - NT

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

For $N, d \to \infty$ *with* $N/d \to \psi_1$

$$\frac{\mathbb{E}[R_{\mathsf{NT},N}]}{\mathbb{E}\{f_*^2\}} = \left\{(1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+ \frac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} + o_d(1)\right\}$$

where the expectation is taken over $\boldsymbol{w}_i \sim_{i.i.d} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$.

- Does NT learn subspaces based on their importance in $\boldsymbol{B}$?
- No! NT fits random directions (but more parameters).

Quadratic Model - NN

- Square non-linearity $\longrightarrow f_{\mathsf{NN}}(x; \boldsymbol{W}, c) = \sum_{i=1}^{N} \langle \boldsymbol{w}_i, x \rangle^2 + c.$
- $R(f_{\mathsf{NN}}) = R(\boldsymbol{W}, c) = \mathbb{E}\Big[\Big(\langle xx^{\mathsf{T}}, B - \boldsymbol{W}\boldsymbol{W}^{\mathsf{T}} \rangle + b_0 - c\Big)^2\Big].$
-

$$\boldsymbol{W}^*(\boldsymbol{W}^*)^T = \sum_{i=1}^{N \wedge d} \lambda_i(B) u_i(B) u_i(B)^T$$

$$\inf_{\boldsymbol{W}, c} R(\boldsymbol{W}, c) = \begin{cases} 2\sum_{i=N+1}^{d} \lambda_i(B)^2 & \text{for } N < d, \\ 0 & \text{for } N \geq d. \end{cases}$$

- Does gradient descent converge to this value?

Quadratic Model - NN

- Square non-linearity $\longrightarrow f_{\mathsf{NN}}(\boldsymbol{x}; \boldsymbol{W}, c) = \sum_{i=1}^{N} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle^2 + c.$
- $R(f_{\mathsf{NN}}) = R(\boldsymbol{W}, c) = \mathbb{E}\Big[\Big(\langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}, \boldsymbol{B} - \boldsymbol{W}\boldsymbol{W}^{\mathsf{T}} \rangle + b_0 - c\Big)^2\Big].$
-

$$\boldsymbol{W}^*(\boldsymbol{W}^*)^T = \sum_{i=1}^{N \wedge d} \lambda_i(\boldsymbol{B}) \boldsymbol{u}_i(\boldsymbol{B}) \boldsymbol{u}_i(\boldsymbol{B})^T$$

$$\inf_{\boldsymbol{W}, c} R(\boldsymbol{W}, c) = \begin{cases} 2\sum_{i=N+1}^{d} \lambda_i(\boldsymbol{B})^2 & \text{for } N < d, \\ 0 & \text{for } N \geq d. \end{cases}$$

- Does gradient descent converge to this value?

Quadratic Model - NN

- Square non-linearity $\longrightarrow f_{\mathsf{NN}}(\boldsymbol{x}; \boldsymbol{W}, c) = \sum_{i=1}^{N} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle^2 + c.$
- $R(f_{\mathsf{NN}}) = R(\boldsymbol{W}, c) = \mathbb{E}\Big[\Big(\langle \boldsymbol{x}\boldsymbol{x}^\mathsf{T}, \boldsymbol{B} - \boldsymbol{W}\boldsymbol{W}^\mathsf{T} \rangle + b_0 - c\Big)^2\Big].$
-

$$\boldsymbol{W}^*(\boldsymbol{W}^*)^T = \sum_{i=1}^{N \wedge d} \lambda_i(\boldsymbol{B}) \boldsymbol{u}_i(\boldsymbol{B}) \boldsymbol{u}_i(\boldsymbol{B})^T$$

$$\inf_{\boldsymbol{W}, c} R(\boldsymbol{W}, c) = \begin{cases} 2\sum_{i=N+1}^{d} \lambda_i(\boldsymbol{B})^2 & \text{for } N < d, \\ 0 & \text{for } N \geq d. \end{cases}$$

- Does gradient descent converge to this value?

Quadratic Model - NN

- Square non-linearity $\longrightarrow f_{\mathsf{NN}}(\boldsymbol{x}; \boldsymbol{W}, c) = \sum_{i=1}^{N} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle^2 + c.$
- $R(f_{\mathsf{NN}}) = R(\boldsymbol{W}, c) = \mathbb{E}\Big[\Big(\langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}, \boldsymbol{B} - \boldsymbol{W}\boldsymbol{W}^{\mathsf{T}} \rangle + b_0 - c\Big)^2\Big].$
-

$$\boldsymbol{W}^*(\boldsymbol{W}^*)^T = \sum_{i=1}^{N \wedge d} \lambda_i(\boldsymbol{B}) \boldsymbol{u}_i(\boldsymbol{B}) \boldsymbol{u}_i(\boldsymbol{B})^T$$

$$\inf_{\boldsymbol{W}, c} R(\boldsymbol{W}, c) = \begin{cases} 2\sum_{i=N+1}^{d} \lambda_i(\boldsymbol{B})^2 & \text{for } N < d, \\ 0 & \text{for } N \geq d. \end{cases}$$

- Does gradient descent converge to this value?

Quadratic Model - NN

- Square non-linearity $\longrightarrow f_{\mathsf{NN}}(\boldsymbol{x}; \boldsymbol{W}, c) = \sum_{i=1}^{N} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle^2 + c.$
- $R(f_{\mathsf{NN}}) = R(\boldsymbol{W}, c) = \mathbb{E}\Big[\Big(\langle \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}, \boldsymbol{B} - \boldsymbol{W}\boldsymbol{W}^{\mathsf{T}} \rangle + b_0 - c\Big)^2\Big].$
-

$$\boldsymbol{W}^*(\boldsymbol{W}^*)^T = \sum_{i=1}^{N \wedge d} \lambda_i(\boldsymbol{B}) \boldsymbol{u}_i(\boldsymbol{B}) \boldsymbol{u}_i(\boldsymbol{B})^T$$

$$\inf_{\boldsymbol{W}, c} R(\boldsymbol{W}, c) = \begin{cases} 2\sum_{i=N+1}^{d} \lambda_i(\boldsymbol{B})^2 & \text{for } N < d, \\ 0 & \text{for } N \geq d. \end{cases}$$

- Does gradient descent converge to this value?

NN with Gradient Descent

- One-pass version of SGD,

$$(\boldsymbol{W}_{k+1}, c_{k+1}) = (\boldsymbol{W}_k, c_k) - \varepsilon \nabla_{\boldsymbol{W},c} \Big(f_*(\boldsymbol{x}_k) - \hat{f}(\boldsymbol{x}_k; \boldsymbol{W}, c) \Big)^2.$$

- Define

$$R(\ell, \varepsilon) \equiv R(\boldsymbol{W}_\ell, c_\ell).$$

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

We have

$$\lim_{t\to\infty} \lim_{\varepsilon\to 0} \mathbb{P}\Big(\Big| R(\ell = t/\varepsilon, \varepsilon) - \inf_{\boldsymbol{W},c} R(\boldsymbol{W}, c) \Big| \geq \delta \Big) = 0,$$

(probability is over the initialization $(\boldsymbol{W}_0, c_0)$ and the samples.)

NN with Gradient Descent

- One-pass version of SGD,

$$(\boldsymbol{W}_{k+1}, c_{k+1}) = (\boldsymbol{W}_k, c_k) - \varepsilon \nabla_{\boldsymbol{W},c} \Big(f_*(\boldsymbol{x}_k) - \hat{f}(\boldsymbol{x}_k; \boldsymbol{W}, c) \Big)^2.$$

- Define

$$R(\ell, \varepsilon) \equiv R(\boldsymbol{W}_\ell, c_\ell).$$

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

We have

$$\lim_{t\to\infty} \lim_{\varepsilon\to 0} \mathbb{P}\Big(\Big| R(\ell = t/\varepsilon, \varepsilon) - \inf_{\boldsymbol{W},c} R(\boldsymbol{W}, c) \Big| \ge \delta \Big) = 0,$$

(probability is over the initialization $(\boldsymbol{W}_0, c_0)$ and the samples.)

NN with Gradient Descent

- One-pass version of SGD,

$$(\boldsymbol{W}_{k+1}, c_{k+1}) = (\boldsymbol{W}_k, c_k) - \varepsilon \nabla_{\boldsymbol{W},c} \Big(f_*(\boldsymbol{x}_k) - \hat{f}(\boldsymbol{x}_k; \boldsymbol{W}, c) \Big)^2.$$

- Define

$$R(\ell, \varepsilon) \equiv R(\boldsymbol{W}_\ell, c_\ell).$$

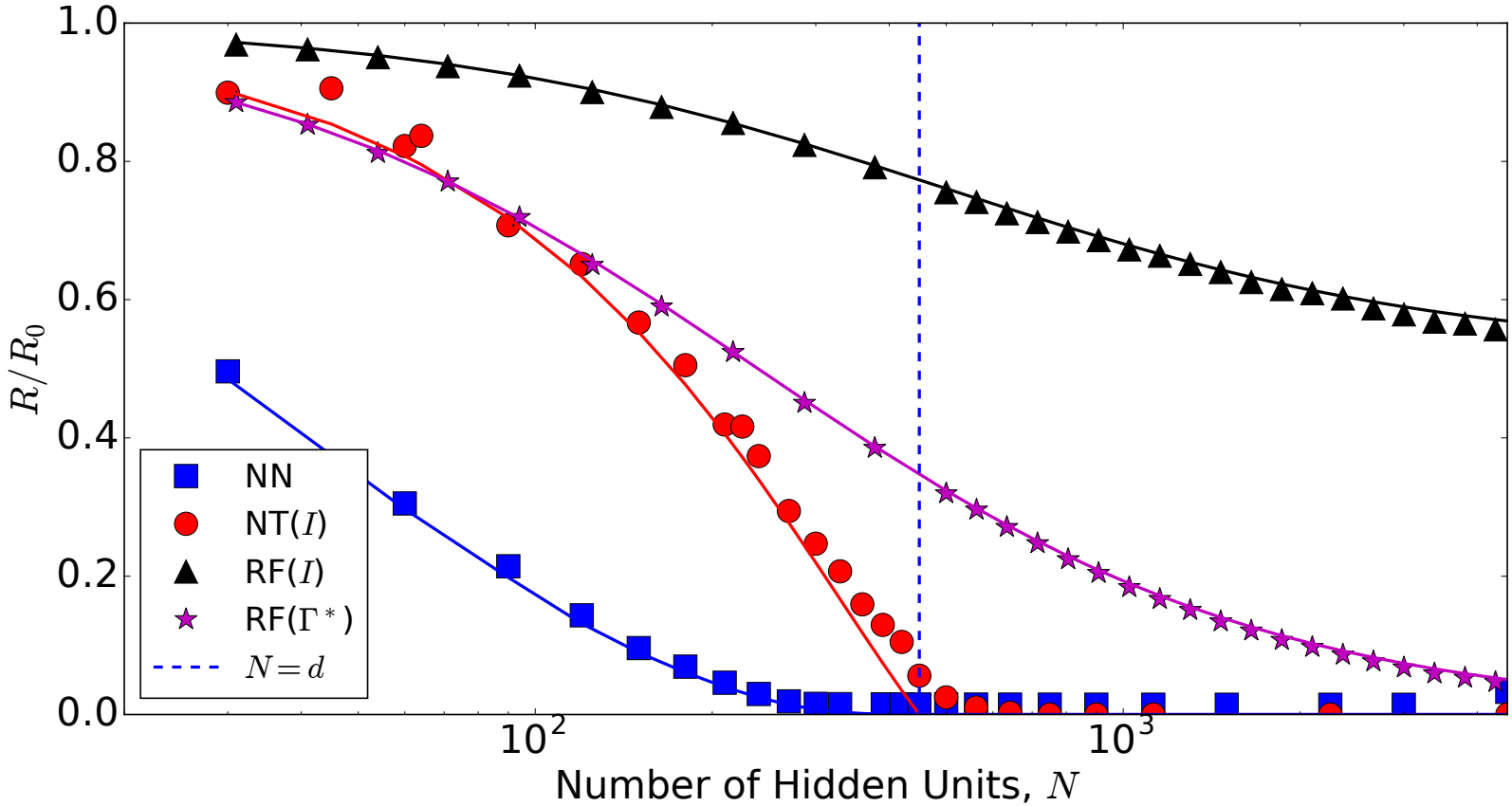
Theorem (Ghorbani, Mei, Misiakiewicz, M., 2019)

We have

$$\lim_{t\to\infty} \lim_{\varepsilon\to 0} \mathbb{P}\Big(\Big| R(\ell = t/\varepsilon, \varepsilon) - \inf_{\boldsymbol{W},c} R(\boldsymbol{W}, c) \Big| \geq \delta \Big) = 0,$$

(probability is over the initialization $(\boldsymbol{W}_0, c_0)$ and the samples.)

Comparison


1.0
0.8
0.6
0.4
0.2
0.0
R/R_0
NN
NT(I)
RF(I)
RF(Γ^*)
$N = d$
10^2
10^3
Number of Hidden Units, N

Comparison

$$\frac{R_{\mathsf{M},N}}{\mathbb{E}\{f_*^2\}} \approx \begin{cases} 1 - \dfrac{\psi_1}{1+\psi_1}\dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{RF}, \\ (1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+ \dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NT}, \\ 1 - \dfrac{\sum_{i=1}^{d\wedge N}\lambda_i(\boldsymbol{B})^2}{\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NN}. \end{cases}$$

- Naive RF, NT do not learn good representations of the data.
- $\exists B$ arbtrarily large gap between NN and NT.
- How general are these phenomena?

Comparison

$$\frac{R_{\mathsf{M},N}}{\mathbb{E}\{f_*^2\}} \approx \begin{cases} 1 - \dfrac{\psi_1}{1+\psi_1}\dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{RF}, \\ (1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+\dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NT}, \\ 1 - \dfrac{\sum_{i=1}^{d\wedge N}\lambda_i(\boldsymbol{B})^2}{\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NN}. \end{cases}$$

- Naive RF, NT do not learn good representations of the data.
- $\exists B$ arbtrarily large gap between NN and NT.
- How general are these phenomena?

Comparison

$$\frac{R_{\mathsf{M},N}}{\mathbb{E}\{f_*^2\}} \approx \begin{cases} 1 - \dfrac{\psi_1}{1+\psi_1}\dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{RF}, \\ (1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+\dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NT}, \\ 1 - \dfrac{\sum_{i=1}^{d\wedge N}\lambda_i(\boldsymbol{B})^2}{\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NN}. \end{cases}$$

- Naive RF, NT do not learn good representations of the data.
- $\exists B$ arbtrarily large gap between NN and NT.
- How general are these phenomena?

Comparison

$$\frac{R_{\mathsf{M},N}}{\mathbb{E}\{f_*^2\}} \approx \begin{cases} 1 - \dfrac{\psi_1}{1+\psi_1}\dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{RF}, \\ (1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+\dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NT}, \\ 1 - \dfrac{\sum_{i=1}^{d\wedge N}\lambda_i(\boldsymbol{B})^2}{\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NN}. \end{cases}$$

- Naive RF, NT do not learn good representations of the data.
- $\exists B$ arbtrarily large gap between NN and NT.
- How general are these phenomena?

Comparison

$$\frac{R_{\mathsf{M},N}}{\mathbb{E}\{f_*^2\}} \approx \begin{cases} 1 - \dfrac{\psi_1}{1+\psi_1}\dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{RF}, \\ (1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+ \dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NT}, \\ 1 - \dfrac{\sum_{i=1}^{d\wedge N}\lambda_i(\boldsymbol{B})^2}{\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NN}. \end{cases}$$

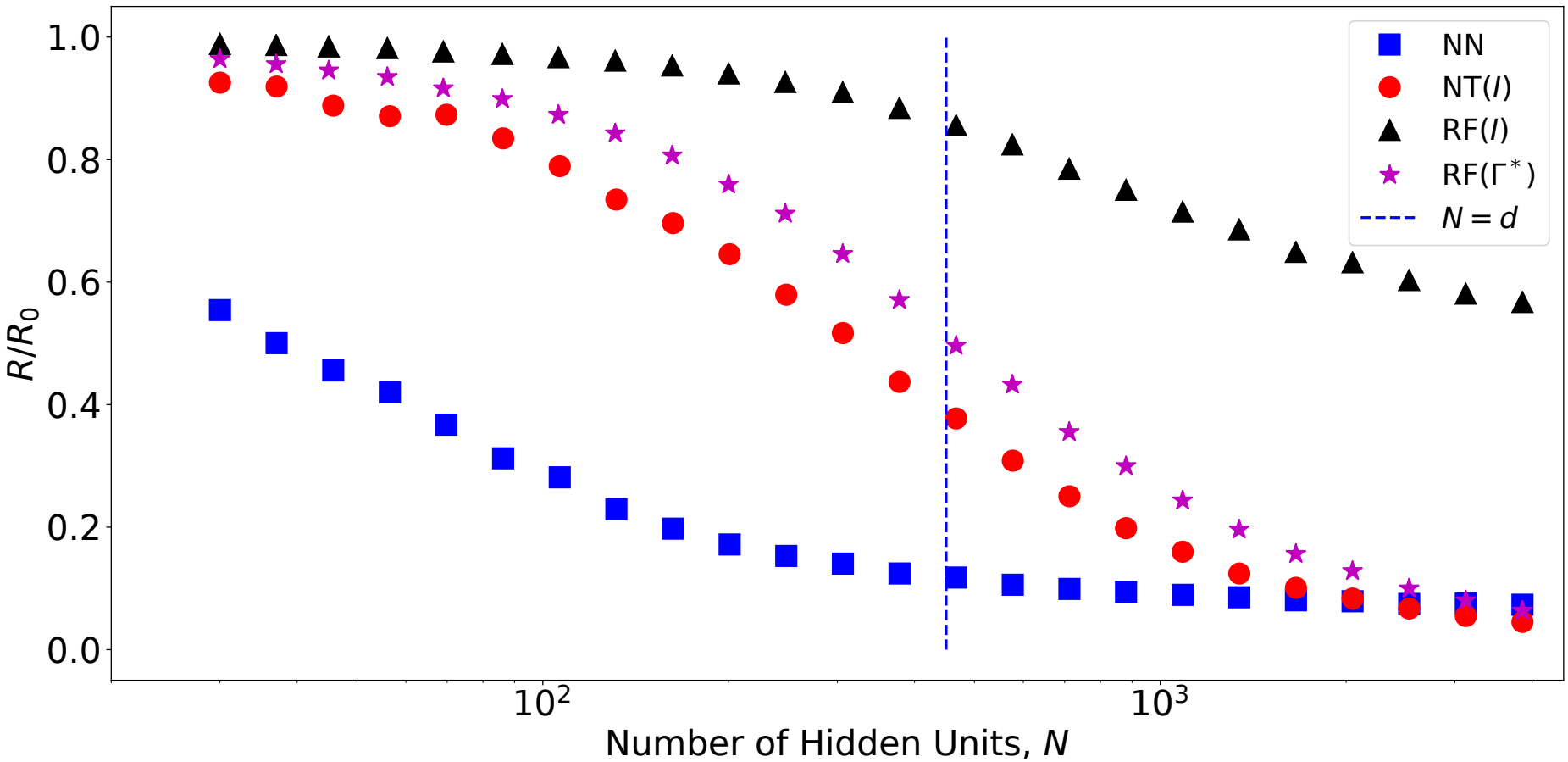
- Naive RF, NT do not learn good representations of the data.
- $\exists \boldsymbol{B}$ arbtrarily large gap between NN and NT.
- How general are these phenomena?

Comparison

$$\frac{R_{\mathsf{M},N}}{\mathbb{E}\{f_*^2\}} \approx \begin{cases} 1 - \dfrac{\psi_1}{1+\psi_1}\dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{RF}, \\ (1-\psi_1)_+^2 + \psi_1(1-\psi_1)_+ \dfrac{\mathrm{Tr}(\boldsymbol{B})^2}{d\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NT}, \\ 1 - \dfrac{\sum_{i=1}^{d\wedge N}\lambda_i(\boldsymbol{B})^2}{\|\boldsymbol{B}\|_F^2} & \text{for } \mathsf{M} = \mathsf{NN}. \end{cases}$$

- Naive RF, NT do not learn good representations of the data.
- $\exists \boldsymbol{B}$ arbtrarily large gap between NN and NT.
- How general are these phenomena?

Comparison with ReLU



Conclusion

Conclusion

- Thermodynamic/wide limit: a useful concept
- Different regimes (depends on training)
- Linearized regime captures certain phenomena

Thanks!

Conclusion

- Thermodynamic/wide limit: a useful concept
- Different regimes (depends on training)
- Linearized regime captures certain phenomena

Thanks!