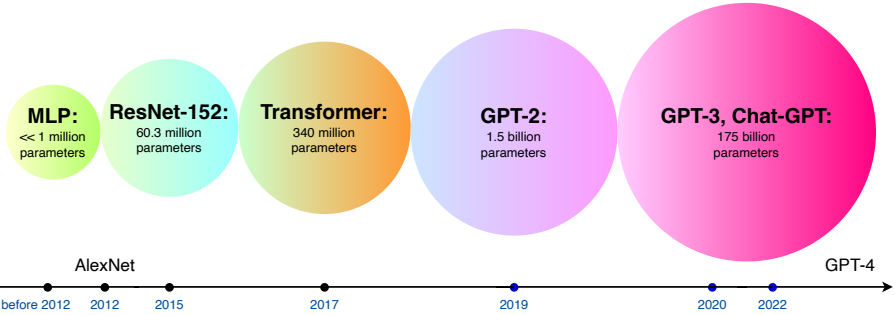
Over-parameterization in (two-layer) neural networks: double descent, function spaces, curse of dimensionality

Fanghui Liu

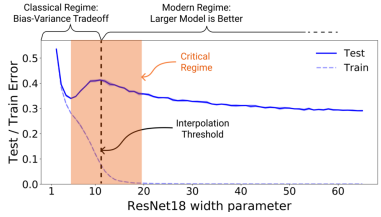
Department of Computer Science, University of Warwick, UK
Centre for Discrete Mathematics and its Applications (DIMAP), Warwick

Joint with
[Johan A.K. Suykens (KU Leuven), Volkan Cevher (EPFL)]
at LCSL (Laboratory for Computational and Statistical Learning), MaLGa

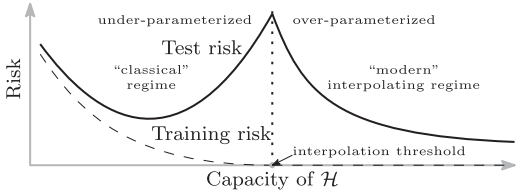
Over-parameterization: more parameters than training data



Surprises in modern neural networks: double descent

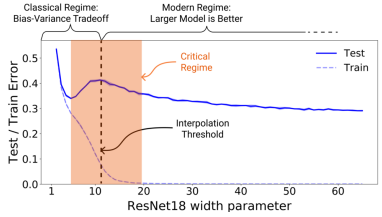


(a) Training and test error on ResNet18 [1]

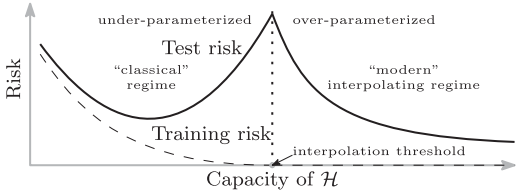


(b) Double descent [2] (Belkin, Hsu, Ma, Mandal, 2019).

Surprises in modern neural networks: double descent



(a) Training and test error on ResNet18 [1]

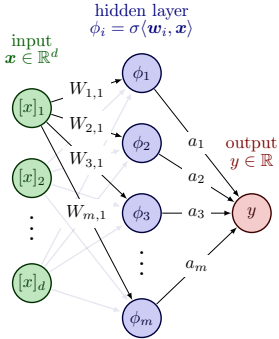


(b) Double descent [2] (Belkin, Hsu, Ma, Mandal, 2019).

Observations: beyond bias-variance trade-off

- 1) Peak at the interpolation thresholds
- 2) Monotonic decreasing in the overparameterized regime
- 3) Global minimum when #parameters is infinite

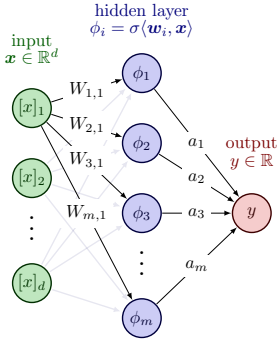
Background: Two-layer neural networks



$$f_m(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{m} a_i \phi(\boldsymbol{x}, \boldsymbol{w}_i), \quad \boldsymbol{\theta} := \{(a_i, \boldsymbol{w}_i)\}_{i=1}^m$$

- $\phi : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$, e.g., ReLU: $\phi(\boldsymbol{x}, \boldsymbol{w}) = \max(\langle \boldsymbol{x}, \boldsymbol{w}\rangle, 0)$

Background: Two-layer neural networks



$$f_m(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{m} a_i \phi(\boldsymbol{x}, \boldsymbol{w}_i), \quad \boldsymbol{\theta} := \{(a_i, \boldsymbol{w}_i)\}_{i=1}^m$$

- $\phi : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$, e.g., ReLU: $\phi(\boldsymbol{x}, \boldsymbol{w}) = \max(\langle \boldsymbol{x}, \boldsymbol{w}\rangle, 0)$
- Random features models (RFMs) [3]:
  - $\{\boldsymbol{w}_i\}_{i=1}^m \overset{iid}{\sim} \mu$ for a given $\mu \in \mathcal{P}(\mathcal{W})$
  - only train the second layer

random features mapping:

$$\varphi(\boldsymbol{x}) := \frac{1}{\sqrt{m}} \sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right) \quad W_{ij} \sim \mathcal{N}(0,1)$$

Recall RFMs in high-dimensional asymptotic setting (Mei and Montanari, 2022)

- high dimensional: $n, m, d \to \infty$, $m/d \to \psi_1$ and $n/d \to \psi_2$ as $d \to \infty$ with $\psi_1, \psi_2 \in (0, \infty)$

Recall RFMs in high-dimensional asymptotic setting (Mei and Montanari, 2022)

- high dimensional: $n, m, d \to \infty$, $m/d \to \psi_1$ and $n/d \to \psi_2$ as $d \to \infty$ with $\psi_1, \psi_2 \in (0, \infty)$
- random feature regression with $\widehat{\boldsymbol{a}}_\lambda = \arg\min_{\boldsymbol{a}} \widehat{\mathcal{E}}_\lambda(\boldsymbol{a})$

$$\widehat{\mathcal{E}}_\lambda(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle \boldsymbol{x}_i, \boldsymbol{w}_j\rangle)\right]^2 + \frac{\lambda m}{d}\|\boldsymbol{a}\|_2^2$$

Recall RFMs in high-dimensional asymptotic setting (Mei and Montanari, 2022)

- high dimensional: $n, m, d \to \infty$, $m/d \to \psi_1$ and $n/d \to \psi_2$ as $d \to \infty$ with $\psi_1, \psi_2 \in (0, \infty)$
- random feature regression with $\widehat{\boldsymbol{a}}_\lambda = \arg\min_{\boldsymbol{a}} \widehat{\mathcal{E}}_\lambda(\boldsymbol{a})$

$$\widehat{\mathcal{E}}_\lambda(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle \boldsymbol{x}_i, \boldsymbol{w}_j\rangle)\right]^2 + \frac{\lambda m}{d}\|\boldsymbol{a}\|_2^2$$

$$\mathcal{E}(\boldsymbol{a}, f_\rho) = \mathbb{E}_{\boldsymbol{x},y}\left[f_\rho(\boldsymbol{x}) - \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle \boldsymbol{x}_i, \boldsymbol{w}_j\rangle)\right]^2$$

Recall RFMs in high-dimensional asymptotic setting (Mei and Montanari, 2022)

- high dimensional: $n, m, d \to \infty$, $m/d \to \psi_1$ and $n/d \to \psi_2$ as $d \to \infty$ with $\psi_1, \psi_2 \in (0, \infty)$
- random feature regression with $\widehat{\boldsymbol{a}}_\lambda = \arg\min_{\boldsymbol{a}} \widehat{\mathcal{E}}_\lambda(\boldsymbol{a})$

$$\widehat{\mathcal{E}}_\lambda(\boldsymbol{a}) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle \boldsymbol{x}_i, \boldsymbol{w}_j\rangle)\right]^2 + \frac{\lambda m}{d}\|\boldsymbol{a}\|_2^2$$

$$\mathcal{E}(\boldsymbol{a}, f_\rho) = \mathbb{E}_{\boldsymbol{x},y}\left[f_\rho(\boldsymbol{x}) - \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle \boldsymbol{x}_i, \boldsymbol{w}_j\rangle)\right]^2$$

Theorem ([4] Mei and Montanari, 2022)

data $\{\boldsymbol{x}_i\}_{i=1}^n \sim$ *Unif*$(\mathbb{S}^{d-1}(\sqrt{d}))$, *label noise* $\mathbb{E}(\epsilon^2) = \tau^2$, *target function* $f_\rho(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x}\rangle$, *random features* $\{\boldsymbol{w}_j\}_{j=1}^m \overset{iid}{\sim}$ *Unif*$(\mathbb{S}^{d-1})$ *for* $G \sim \mathcal{N}(0,1)$, *define* $\mu_1 = \mathbb{E}(\sigma(G)G)$, $\mu_*^2 = \mathbb{E}[\sigma(G)^2] - (\mathbb{E}[\sigma(G)])^2 - (\mathbb{E}[\sigma(G)G])^2$, $\zeta = \mu_1/\mu_*$, *for any* $\lambda > 0$, *we have*

$$\mathcal{E}(\widehat{\boldsymbol{a}}_\lambda, f_\rho) = \|\beta\|_2^2 \mathtt{B}(\zeta, \psi_1, \psi_2, \lambda/\mu_*^2) + \tau^2 \mathtt{V}(\zeta, \psi_1, \psi_2, \lambda/\mu_*^2) + o_{d,\mathbb{P}}(1)\,.$$

observations 1), 2), 3) for double descent can be theoretically proved.

Questions on high dimensional kernel methods

high dimensional kernel methods: can only learn linear function! [5] (Ghorbani, Mei, Misiakiewicz, Montanari, 2021)

Questions on high dimensional kernel methods

high dimensional kernel methods: can only learn linear function! [5] (Ghorbani, Mei, Misiakiewicz, Montanari, 2021)

- asymptotic expansion under high dimensions [6] (El Karoui, 2010)
  under the setting of $n, d \to \infty$, $n/d \to \psi_1$ as $d \to \infty$ with $\psi_1 \in (0, \infty)$, we have

$$\|\boldsymbol{K} - (a\boldsymbol{X}\boldsymbol{X}^\top + b\boldsymbol{I})\|_2 \xrightarrow{\mathbb{P}} 0 \quad \text{when } d \to 0 \quad \text{for some parameters } a, b$$

Questions on high dimensional kernel methods

high dimensional kernel methods: can only learn linear function! [5] (Ghorbani, Mei, Misiakiewicz, Montanari, 2021)

- asymptotic expansion under high dimensions [6] (El Karoui, 2010)
  under the setting of $n, d \to \infty$, $n/d \to \psi_1$ as $d \to \infty$ with $\psi_1 \in (0, \infty)$, we have

$$\|\boldsymbol{K} - (a\boldsymbol{X}\boldsymbol{X}^\top + b\boldsymbol{I})\|_2 \xrightarrow{\mathbb{P}} 0 \quad \text{when } d \to 0 \quad \text{for some parameters } a, b$$

- $\|f^*\|_{\mathcal{H}} < \infty$?

Example (a linear function $f : \mathbb{S}^d \to \mathbb{R}$ such that $f(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x}$ for a certain $\boldsymbol{v} \in \mathbb{S}^d$)

- zero-order arc-cosine kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathbb{S}^d} 1_{\{\boldsymbol{\omega}^\top \boldsymbol{x} \geq 0\}} 1_{\{\boldsymbol{\omega}^\top \boldsymbol{x}' \geq 0\}} \mathrm{d}\mu(\boldsymbol{\omega})$
  $\Rightarrow \|f\|_{\mathcal{H}} = \frac{2d\pi}{d-1}\pi < 4\pi$ [7] (Bach 2017)

Questions on high dimensional kernel methods

high dimensional kernel methods: can only learn linear function! [5] (Ghorbani, Mei, Misiakiewicz, Montanari, 2021)

- asymptotic expansion under high dimensions [6] (El Karoui, 2010)
  under the setting of $n, d \to \infty$, $n/d \to \psi_1$ as $d \to \infty$ with $\psi_1 \in (0, \infty)$, we have

$$\|\boldsymbol{K} - (a\boldsymbol{X}\boldsymbol{X}^\top + b\boldsymbol{I})\|_2 \xrightarrow{\mathbb{P}} 0 \quad \text{when } d \to 0 \quad \text{for some parameters } a, b$$

- $\|f^*\|_{\mathcal{H}} < \infty$?

Example (a linear function $f : \mathbb{S}^d \to \mathbb{R}$ such that $f(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x}$ for a certain $\boldsymbol{v} \in \mathbb{S}^d$)

- zero-order arc-cosine kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathbb{S}^d} 1_{\{\boldsymbol{\omega}^\top \boldsymbol{x} \geq 0\}} 1_{\{\boldsymbol{\omega}^\top \boldsymbol{x}' \geq 0\}} \mathrm{d}\mu(\boldsymbol{\omega})$
  $\Rightarrow \|f\|_{\mathcal{H}} = \frac{2d\pi}{d-1}\pi < 4\pi$ [7] (Bach 2017)
- first-order arc-cosine kernel, we have $\|f\|_{\mathcal{H}} \asymp C\sqrt{d}$ for some constant C independent of d.

Motivation

- high dimension vs. fixed dimension
- from asymptotic to non-asymptotic
- two-layer neural networks trained by SGD

Motivation

- high dimension vs. fixed dimension
- from asymptotic to non-asymptotic
- two-layer neural networks trained by SGD

○ Analysis

- SGD: implicit regularization $\rightarrow$ without λ
- dimension-free SGD bound
- multiple randomness sources
  - data sampling, label noise, Gaussian initialization, stochastic gradients

Motivation

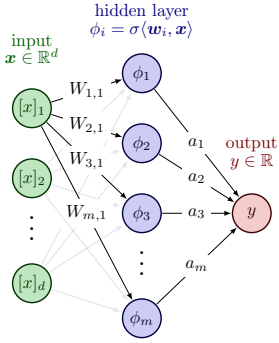
- high dimension vs. fixed dimension
- from asymptotic to non-asymptotic
- two-layer neural networks trained by SGD

○ Analysis

- SGD: implicit regularization $\rightarrow$ without λ
- dimension-free SGD bound
- multiple randomness sources
  - data sampling, label noise, Gaussian initialization, stochastic gradients

observations 1), 2), 3) can be still proved!

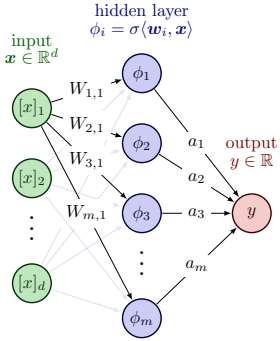
Problem settings: function space



random features mapping:

$$\varphi(\boldsymbol{x}) := \frac{1}{\sqrt{m}}\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right) \quad W_{ij} \sim \mathcal{N}(0,1)$$

Problem settings: function space



random features mapping:

$$\varphi(\boldsymbol{x}) := \frac{1}{\sqrt{m}}\sigma\left(\frac{\boldsymbol{W}\boldsymbol{x}}{\sqrt{d}}\right) \quad W_{ij} \sim \mathcal{N}(0,1)$$

function space

$$\mathcal{H} := \left\{ f \in L^2_{\rho_X} \,\middle|\, f(\boldsymbol{x}) = \langle \boldsymbol{a}, \varphi(\boldsymbol{x})\rangle \right\}, \quad \boldsymbol{W}_{ij} \sim \mathcal{N}(0,1)$$

covariance operator: $\Sigma_m := \mathbb{E}_{\boldsymbol{x}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]$

expected covariance operator: $\widetilde{\Sigma}_m := \mathbb{E}_{\boldsymbol{x},\boldsymbol{W}}[\varphi(\boldsymbol{x}) \otimes \varphi(\boldsymbol{x})]$

Problem settings: RFMs with the squared loss by SGD

Online SGD: one-pass, average output, adaptive step-size...

$$\boldsymbol{a}_t = \boldsymbol{a}_{t-1} + \gamma_t[y_t - \langle \boldsymbol{a}_{t-1}, \varphi(\boldsymbol{x}_t)\rangle]\varphi(\boldsymbol{x}_t), \qquad t = 1, 2, \dots n\,.$$

Problem settings: RFMs with the squared loss by SGD

Online SGD: one-pass, average output, adaptive step-size...

$$\boldsymbol{a}_t = \boldsymbol{a}_{t-1} + \gamma_t[y_t - \langle \boldsymbol{a}_{t-1}, \varphi(\boldsymbol{x}_t)\rangle]\varphi(\boldsymbol{x}_t), \qquad t = 1, 2, \dots n\,.$$

- averaged output: $\bar{\boldsymbol{a}}_n := \frac{1}{n}\sum_{t=0}^{n-1} \boldsymbol{a}_t \Longrightarrow \bar{f}_n = \langle \varphi(\cdot), \bar{\boldsymbol{a}}_n \rangle$
- adaptive step-size: $\gamma_t := \gamma_0 t^{-\zeta}, \zeta \in [0, 1)$

Problem settings: RFMs with the squared loss by SGD

Online SGD: one-pass, average output, adaptive step-size...

$$\boldsymbol{a}_t = \boldsymbol{a}_{t-1} + \gamma_t[y_t - \langle \boldsymbol{a}_{t-1}, \varphi(\boldsymbol{x}_t)\rangle]\varphi(\boldsymbol{x}_t), \qquad t = 1, 2, \dots n\,.$$

- averaged output: $\bar{\boldsymbol{a}}_n := \frac{1}{n}\sum_{t=0}^{n-1} \boldsymbol{a}_t \Longrightarrow \bar{f}_n = \langle \varphi(\cdot), \bar{\boldsymbol{a}}_n \rangle$
- adaptive step-size: $\gamma_t := \gamma_0 t^{-\zeta}, \zeta \in [0, 1)$

Averaged expected risk

- optimal solution: $f^* = \arg\min_{f\in\mathcal{H}} \|f - f_\rho\|^2_{L^2_{\rho_X}}$ with $\|f^*\|_{\mathcal{H}} < \infty$
- averaged excess risk: $\mathbb{E}\|\bar{f}_n - f^*\|^2_{L^2_{\rho_X}} = \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\langle \bar{f}_n - f^*, \Sigma_m(\bar{f}_n - f^*)\rangle$

Assumptions

Assumption (Basic assumptions)

- **non-asymptotic:** $\|\boldsymbol{x}\|_2^2 \leq \mathcal{O}(d)$, $\Sigma_d := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x} \otimes \boldsymbol{x}]$ *with* $\|\Sigma_d\|_2 < \infty$
- **boundedness of** f^***:** $\|f^*\|_{\mathcal{H}} < \infty$
- **activation function:** $\sigma(\cdot)$*: Lipschitz continuous*
- **label noise:** $\mathbb{E}(\varepsilon) = 0$ *and* $\mathbb{E}(\varepsilon^2) = \tau^2$

Assumptions

Assumption (Basic assumptions)

- **non-asymptotic:** $\|\boldsymbol{x}\|_2^2 \leq \mathcal{O}(d)$, $\Sigma_d := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x} \otimes \boldsymbol{x}]$ *with* $\|\Sigma_d\|_2 < \infty$
- **boundedness of** f^***:** $\|f^*\|_{\mathcal{H}} < \infty$
- **activation function:** $\sigma(\cdot)$*: Lipschitz continuous*
- **label noise:** $\mathbb{E}(\varepsilon) = 0$ *and* $\mathbb{E}(\varepsilon^2) = \tau^2$

Assumption (Fourth moment condition)

for any PSD operator A*, we assume*

$$\mathbb{E}_{\boldsymbol{W}}[\Sigma_m A \Sigma_m] \preccurlyeq r' \mathbb{E}_{\boldsymbol{W}}[\mathrm{Tr}(\Sigma_m A)\Sigma_m] \preccurlyeq r\mathrm{Tr}(\widetilde{\Sigma}_m A)\widetilde{\Sigma}_m \,.$$

Remark:

- the special case $A := I$ can be proved.
- holds for sub-Gaussian data.
- widely used in SGD analysis [8, 9, 10]

Main results: bias-variance decomposition

Define $\eta_t := f_t - f^*$, we have

$$\eta_t = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)](f_{t-1} - f^*) + \gamma_t \varepsilon_t \varphi(\boldsymbol{x}_t)\,,$$

Main results: bias-variance decomposition

Define $\eta_t := f_t - f^*$, we have

$$\eta_t = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)](f_{t-1} - f^*) + \gamma_t \varepsilon_t \varphi(\boldsymbol{x}_t)\,,$$

$$\eta_t^{\texttt{bias}} = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\texttt{bias}}, \quad \eta_0^{\texttt{bias}} = f^*\,,$$

Main results: bias-variance decomposition

Define $\eta_t := f_t - f^*$, we have

$$\eta_t = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)](f_{t-1} - f^*) + \gamma_t \varepsilon_t \varphi(\boldsymbol{x}_t)\,,$$

$$\eta_t^{\texttt{bias}} = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\texttt{bias}}, \quad \eta_0^{\texttt{bias}} = f^*\,,$$

$$\eta_t^{\texttt{var}} = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\texttt{var}} + \gamma_t \varepsilon_t \varphi(\boldsymbol{x}_t), \quad \eta_0^{\texttt{var}} = 0\,.$$

Main results: bias-variance decomposition

Define $\eta_t := f_t - f^*$, we have

$$\eta_t = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)](f_{t-1} - f^*) + \gamma_t \varepsilon_t \varphi(\boldsymbol{x}_t)\,,$$

$$\eta_t^{\texttt{bias}} = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\texttt{bias}}, \quad \eta_0^{\texttt{bias}} = f^*\,,$$

$$\eta_t^{\texttt{var}} = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\texttt{var}} + \gamma_t \varepsilon_t \varphi(\boldsymbol{x}_t), \quad \eta_0^{\texttt{var}} = 0\,.$$

Theorem (Bias-variance decomposition)

Under the above-mentioned assumptions, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ with $\zeta \in [0, 1)$ satisfies $\gamma_0 < C$, we have

$$\mathbb{E}\|\bar{f}_n - f^*\|_{L^2_{\rho_X}}^2 = \underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{W}}\langle \bar{\eta}_n^{\texttt{bias}}, \Sigma_m \bar{\eta}_n^{\texttt{bias}} \rangle}_{:=\texttt{Bias}} + \underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\langle \bar{\eta}_n^{\texttt{var}}, \Sigma_m \bar{\eta}_n^{\texttt{var}} \rangle}_{:=\texttt{Variance}}\,.$$

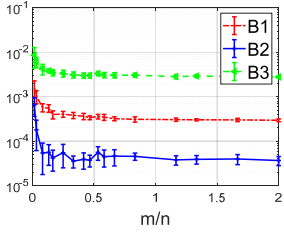
Main theorem

Theorem (Liu, Suykens, Volkan, 2022)

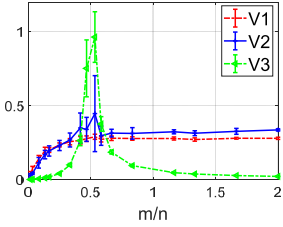
Under the above-mentioned assumptions, if the step-size $\gamma_t := \gamma_0 t^{-\zeta}$ *with* $\zeta \in [0, 1)$ *satisfies* $\gamma_0 < C$*, we have*

$$\texttt{Bias} \lesssim \gamma_0 r' n^{\zeta-1} \|f^*\|^2 \sim \mathcal{O}\left(n^{\zeta-1}\right).$$

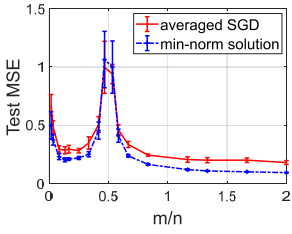
$$\texttt{Variance} \lesssim \gamma_0 r' \tau^2 \begin{cases} mn^{\zeta-1}, & \text{if } m \leqslant n \\ 1 + n^{\zeta-1} + \dfrac{n}{m}, & \text{if } m > n \end{cases}$$



(c) bias



(d) variance



(e) excess risk

Discussion

Constant step-size SGD doesn't hurt the convergence rate.

- under-parameterized regime (by taking $m = \mathcal{O}(\sqrt{n})$)

$$\mathbb{E}\|\bar{f}_n - f^*\|^2_{L^2_{\rho_X}} = \underbrace{\texttt{Bias}}_{\mathcal{O}(\frac{1}{n})} + \underbrace{\texttt{Variance}}_{\mathcal{O}(\frac{1}{\sqrt{n}})} \leq \mathcal{O}\left(\frac{1}{\sqrt{n}}\right),$$

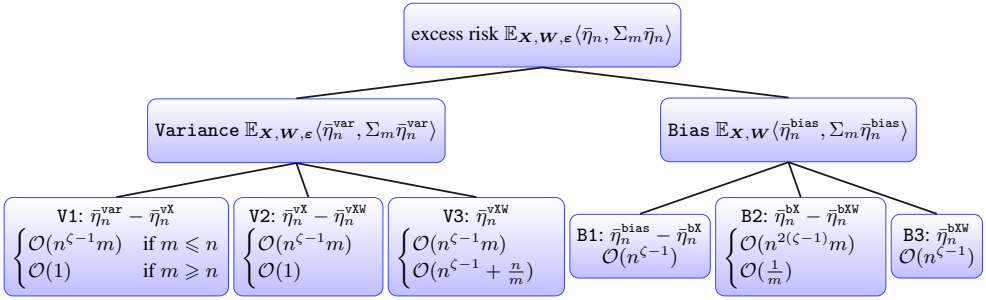
  matches [11] (Carratino, Rudi, Rosasco, 2018) under one-pass, one-batch, SGD...[1]

- over-parameterized regime: matches [12] (Belkin, Hsu, Xu, 2020)

- no lower bound: $\texttt{Bias} \leqslant 3(\texttt{B1} + \texttt{B2} + \texttt{B3})$ based on Minkowski inequality

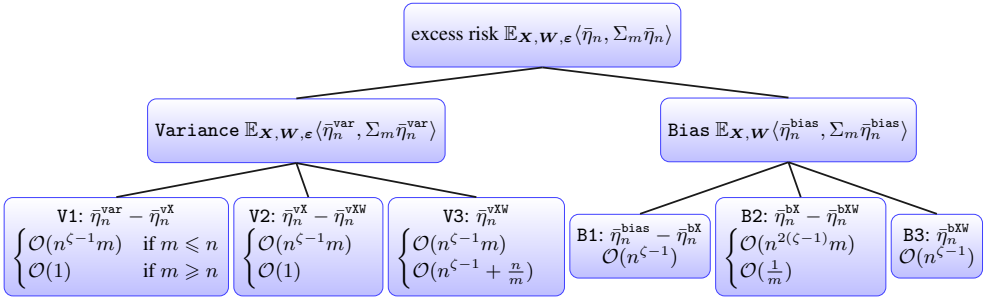
[1] but the selection on step-size is different

Proof framework: randomness decoupling



$$\mathtt{Bias}: \quad \eta_t^{\mathtt{bias}} = [I - \gamma_t \varphi(\boldsymbol{x}_t) \otimes \varphi(\boldsymbol{x}_t)] \eta_{t-1}^{\mathtt{bias}}$$

Proof framework: randomness decoupling

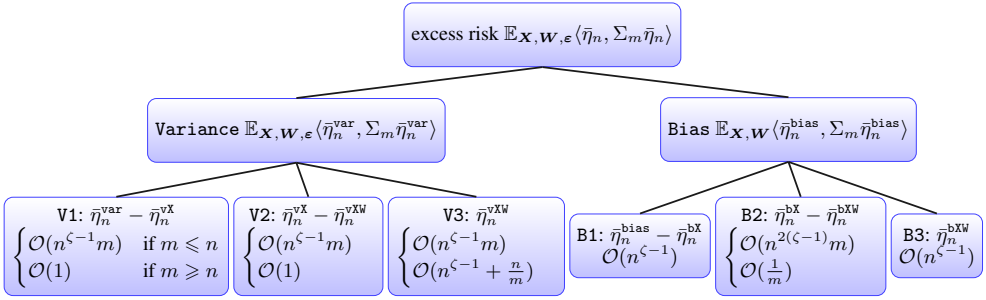


$$\mathtt{Bias}: \quad \eta_t^{\mathtt{bias}} = [I - \gamma_t\varphi(\boldsymbol{x}_t)\otimes\varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\mathtt{bias}}$$

Define "semi-stochastic" version: $\eta_t^{\mathtt{bX}} = (I - \gamma_t\Sigma_m)\eta_{t-1}^{\mathtt{bX}}, \quad \eta_t^{\mathtt{bXW}} = (I - \gamma_t\widetilde{\Sigma}_m)\eta_{t-1}^{\mathtt{bXW}},$

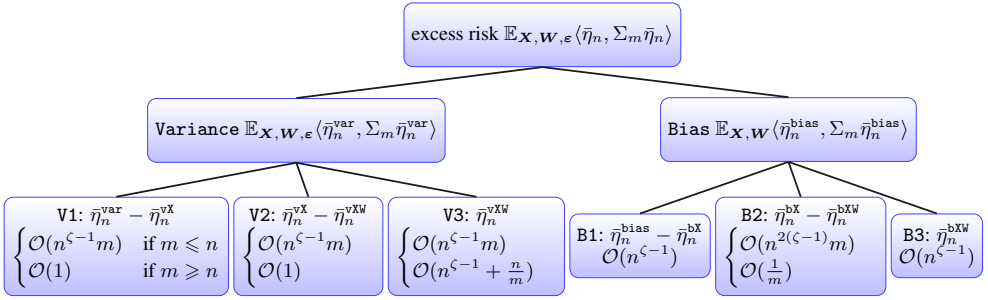
- $\mathtt{B1} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W}}\left[\langle\bar{\eta}_n^{\mathtt{bias}} - \bar{\eta}_n^{\mathtt{bX}}, \Sigma_m(\bar{\eta}_n^{\mathtt{bias}} - \bar{\eta}_n^{\mathtt{bX}})\rangle\right]$
- $\mathtt{B2} := \mathbb{E}_{\boldsymbol{W}}\left[\langle\bar{\eta}_n^{\mathtt{bX}} - \bar{\eta}_n^{\mathtt{bXW}}, \Sigma_m(\bar{\eta}_n^{\mathtt{bX}} - \bar{\eta}_n^{\mathtt{bXW}})\rangle\right]$
- $\mathtt{B3} := \langle\bar{\eta}_n^{\mathtt{bXW}}, \widetilde{\Sigma}_m\bar{\eta}_n^{\mathtt{bXW}}\rangle$

Proof framework



$$\mathtt{Variance}: \quad \eta_t^{\mathtt{var}} = [I - \gamma_t\varphi(\boldsymbol{x}_t)\otimes\varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\mathtt{var}} + \gamma_t\varepsilon_t\varphi(\boldsymbol{x}_t)$$

Proof framework



$$\mathtt{Variance}: \quad \eta_t^{\mathtt{var}} = [I - \gamma_t\varphi(\boldsymbol{x}_t)\otimes\varphi(\boldsymbol{x}_t)]\eta_{t-1}^{\mathtt{var}} + \gamma_t\varepsilon_t\varphi(\boldsymbol{x}_t)$$

Define "semi-stochastic" version: $\eta_t^{\mathtt{vX}} := (I - \gamma_t\Sigma_m)\eta_{t-1}^{\mathtt{vX}} + \gamma_t\varepsilon_t\varphi(\boldsymbol{x}_t)\,, \quad \eta_t^{\mathtt{vXW}} := (I - \gamma_t\widetilde{\Sigma}_m)\eta_{t-1}^{\mathtt{vXW}} + \gamma_t\varepsilon_t\varphi(\boldsymbol{x}_t)$

- $\mathtt{V1} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\Big[\langle\bar{\eta}_n^{\mathtt{var}} - \bar{\eta}_n^{\mathtt{vX}}, \Sigma_m(\bar{\eta}_n^{\mathtt{var}} - \bar{\eta}_n^{\mathtt{vX}})\rangle\Big]$
- $\mathtt{V2} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\Big[\langle\bar{\eta}_n^{\mathtt{vX}} - \bar{\eta}_n^{\mathtt{vXW}}, \Sigma_m(\bar{\eta}_n^{\mathtt{vX}} - \bar{\eta}_n^{\mathtt{vXW}})\rangle\Big]$
- $\mathtt{V3} := \mathbb{E}_{\boldsymbol{X},\boldsymbol{W},\boldsymbol{\varepsilon}}\langle\bar{\eta}_n^{\mathtt{vXW}}, \Sigma_m\bar{\eta}_n^{\mathtt{vXW}}\rangle \leq \frac{2}{n^2}\sum_{t=0}^{n-1}\sum_{k=t}^{n-1}\mathbb{E}_{\boldsymbol{W}}\left\langle\prod_{j=t}^{k-1}(I - \gamma_j\widetilde{\Sigma}_m)\Sigma_m, \underbrace{\mathbb{E}_{\boldsymbol{X},\boldsymbol{\varepsilon}}[\eta_t^{\mathtt{vXW}}\otimes\eta_t^{\mathtt{vXW}}]}_{:=C_t^{\mathtt{vXW}}}\right\rangle$

Proof framework: properties of covariance operators

Properties of $\widetilde{\Sigma}_m$

- the diagonal elements are the same $a := [\widetilde{\Sigma}_m]_{ii}, \forall i \in [m]$
- the non-diagonal elements are the same $b := [\widetilde{\Sigma}_m]_{ij}, \forall i, j \in [m], i \neq j$

$$\widetilde{\Sigma}_m = (a-b)\boldsymbol{I}_m + b\mathbf{1}\mathbf{1}^\top$$

- two distinct eigenvalues: $\widetilde{\lambda}_1 = a - b + bm \sim \mathcal{O}(1)$, $\widetilde{\lambda}_2 = \cdots = \widetilde{\lambda}_m = a - b \sim \mathcal{O}(1/m)$

Proof framework: properties of covariance operators

Properties of $\widetilde{\Sigma}_m$

- the diagonal elements are the same $a := [\widetilde{\Sigma}_m]_{ii}, \forall i \in [m]$
- the non-diagonal elements are the same $b := [\widetilde{\Sigma}_m]_{ij}, \forall i, j \in [m], i \neq j$

$$\widetilde{\Sigma}_m = (a-b)\boldsymbol{I}_m + b\mathbf{1}\mathbf{1}^\top$$

- two distinct eigenvalues: $\widetilde{\lambda}_1 = a - b + bm \sim \mathcal{O}(1)$, $\widetilde{\lambda}_2 = \cdots = \widetilde{\lambda}_m = a - b \sim \mathcal{O}(1/m)$

Example (ReLU activation)

- $(\widetilde{\Sigma}_m)_{ii} = \frac{1}{2md}\mathrm{Tr}(\Sigma_d)$
- $(\widetilde{\Sigma}_m)_{ij} = \frac{1}{2md\pi}\mathrm{Tr}(\Sigma_d)$

Proof framework: properties of covariance operators

Properties of $\widetilde{\Sigma}_m$

- the diagonal elements are the same $a := [\widetilde{\Sigma}_m]_{ii}, \forall i \in [m]$
- the non-diagonal elements are the same $b := [\widetilde{\Sigma}_m]_{ij}, \forall i, j \in [m], i \neq j$

$$\widetilde{\Sigma}_m = (a-b)\boldsymbol{I}_m + b\mathbf{1}\mathbf{1}^\top$$

- two distinct eigenvalues: $\widetilde{\lambda}_1 = a - b + bm \sim \mathcal{O}(1)$, $\widetilde{\lambda}_2 = \cdots = \widetilde{\lambda}_m = a - b \sim \mathcal{O}(1/m)$

Example (ReLU activation)

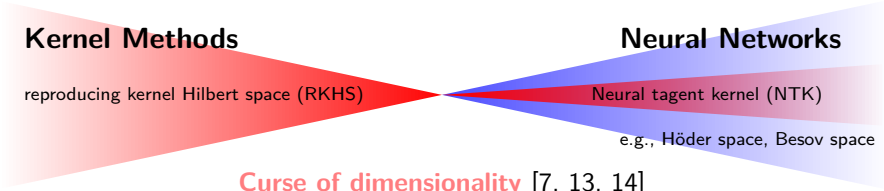
- $(\widetilde{\Sigma}_m)_{ii} = \frac{1}{2md}\mathrm{Tr}(\Sigma_d)$
- $(\widetilde{\Sigma}_m)_{ij} = \frac{1}{2md\pi}\mathrm{Tr}(\Sigma_d)$

sub-exponential random variables

$\|\Sigma_m\|_2$, $\|\Sigma_m - \widetilde{\Sigma}_m\|_2$, $\mathrm{Tr}(\Sigma_m)$, and $\left\|\widetilde{\Sigma}_m^{-1}\mathbb{E}_{\boldsymbol{W}}(\Sigma_m^2)\right\|_2$ with $\mathcal{O}(1)$ sub-exponential norm order

Function space: from kernel methods to neural networks

efficiently approximate non-smooth functions?

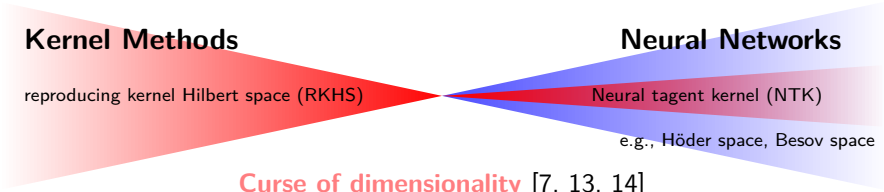


Curse of dimensionality [7, 13, 14]

[computational cost grows exponentially fast]

Function space: from kernel methods to neural networks

efficiently approximate non-smooth functions?



Curse of dimensionality [7, 13, 14]

[computational cost grows exponentially fast]

function space view

what is the suitable function space for neural networks?

RFMs: function spaces

Consider a RFM with infinite many features $f_a(\boldsymbol{x}) = \int_{\mathcal{W}} a(\boldsymbol{w})\phi(\boldsymbol{x}, \boldsymbol{w})\mathrm{d}\mu(\boldsymbol{w})$, define

$$\mathcal{F}_{p,\mu} := \{f_a : \|\boldsymbol{a}\|_{L^p(\mu)} < \infty\}, \quad \|f\|_{\mathcal{F}_{p,\mu}} := \inf_{f_a = f} \|\boldsymbol{a}\|_{L^p(\mu)}$$

RFMs: function spaces

Consider a RFM with infinite many features $f_a(\boldsymbol{x}) = \int_{\mathcal{W}} a(\boldsymbol{w})\phi(\boldsymbol{x}, \boldsymbol{w})\mathrm{d}\mu(\boldsymbol{w})$, define

$$\mathcal{F}_{p,\mu} := \{f_a : \|\boldsymbol{a}\|_{L^p(\mu)} < \infty\}, \quad \|f\|_{\mathcal{F}_{p,\mu}} := \inf_{f_a = f} \|\boldsymbol{a}\|_{L^p(\mu)}$$

- RFMs $\equiv$ kernel methods by taking $p = 2$ using Representer theorem [15]
  - function space: reproducing kernel Hilbert space $\mathcal{H}_{k_\mu} = \mathcal{F}_{2,\mu}$

$$\hat{k}_m(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{m}\sum_{i=1}^{m} \phi(\boldsymbol{x}, \boldsymbol{w}_i)\phi(\boldsymbol{x}', \boldsymbol{w}_i) \;\to\; k_\mu(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathcal{W}} \phi(\boldsymbol{x}, \boldsymbol{w})\phi(\boldsymbol{x}', \boldsymbol{w})\mathrm{d}\mu(\boldsymbol{w})$$

RFMs: function spaces

Consider a RFM with infinite many features $f_a(\boldsymbol{x}) = \int_{\mathcal{W}} a(\boldsymbol{w})\phi(\boldsymbol{x}, \boldsymbol{w})\mathrm{d}\mu(\boldsymbol{w})$, define

$$\mathcal{F}_{p,\mu} := \{f_a : \|\boldsymbol{a}\|_{L^p(\mu)} < \infty\}, \quad \|f\|_{\mathcal{F}_{p,\mu}} := \inf_{f_a = f} \|\boldsymbol{a}\|_{L^p(\mu)}$$

- RFMs $\equiv$ kernel methods by taking $p = 2$ using Representer theorem [15]
  - function space: reproducing kernel Hilbert space $\mathcal{H}_{k_\mu} = \mathcal{F}_{2,\mu}$

$$\hat{k}_m(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{m}\sum_{i=1}^{m} \phi(\boldsymbol{x}, \boldsymbol{w}_i)\phi(\boldsymbol{x}', \boldsymbol{w}_i) \;\to\; k_\mu(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathcal{W}} \phi(\boldsymbol{x}, \boldsymbol{w})\phi(\boldsymbol{x}', \boldsymbol{w})\mathrm{d}\mu(\boldsymbol{w})$$

- RFMs $\not\equiv$ kernel methods if $p < 2$
  function space: $\mathcal{F}_{\infty,\mu} \subseteq \mathcal{F}_{p,\mu} \subseteq \mathcal{F}_{q,\mu} \subseteq \mathcal{F}_{1,\mu}$ if $p \geq q$

Learning beyond RKHS

learning in $\mathcal{F}_{p,\mu}$ with $p \in (1,2)$ by RFMs

- good approximation if $m \geq \Omega(n^2 \vee n^{\frac{2p-1}{2p-2}})$ under interpolation [14] (Celentano, Misiakiewicz, Montanari, 2021)
- duality between approximation and generalization [16] (Chen, Long, Wu, 2023)

Learning beyond RKHS

learning in $\mathcal{F}_{p,\mu}$ with $p \in (1,2)$ by RFMs

- good approximation if $m \geq \Omega(n^2 \vee n^{\frac{2p-1}{2p-2}})$ under interpolation [14] (Celentano, Misiakiewicz, Montanari, 2021)
- duality between approximation and generalization [16] (Chen, Long, Wu, 2023)

Questions

- can we do it more efficiently? from uniform sampling to data-dependent sampling
- how to do SGD beyond RKHS?

Learning beyond RKHS

learning in $\mathcal{F}_{p,\mu}$ with $p \in (1, 2)$ by RFMs

- good approximation if $m \geq \Omega(n^2 \vee n^{\frac{2p-1}{2p-2}})$ under interpolation [14] (Celentano, Misiakiewicz, Montanari, 2021)
- duality between approximation and generalization [16] (Chen, Long, Wu, 2023)

Questions

- can we do it more efficiently? from uniform sampling to data-dependent sampling
- how to do SGD beyond RKHS?

learning in $\mathcal{F}_{p,\mu}$ with $p = 1$ by RFMs

- curse of dimensionality: approximation requires $\Omega(\exp(d))$ random features [14] (Celentano, Misiakiewicz, Montanari, 2021)

Learning beyond RKHS

learning in $\mathcal{F}_{p,\mu}$ with $p \in (1,2)$ by RFMs

- good approximation if $m \geq \Omega(n^2 \vee n^{\frac{2p-1}{2p-2}})$ under interpolation [14] (Celentano, Misiakiewicz, Montanari, 2021)
- duality between approximation and generalization [16] (Chen, Long, Wu, 2023)

Questions

- can we do it more efficiently? from uniform sampling to data-dependent sampling
- how to do SGD beyond RKHS?

learning in $\mathcal{F}_{p,\mu}$ with $p = 1$ by RFMs

- curse of dimensionality: approximation requires $\Omega(\exp(d))$ random features [14] (Celentano, Misiakiewicz, Montanari, 2021)

beyond RKHS but still not data adaptive!

From RKHS to Barron space

Definition (Barron space [17] (E, Ma, Wu, 2021))

For any $1 \leq p \leq \infty$, we have

$$\mathcal{B} = \cup_{\mu \in \mathcal{P}(\mathcal{W})} \mathcal{F}_{p,\mu}\,, \quad \|f\|_{\mathcal{B}} = \inf_{\mu \in \mathcal{P}(\mathcal{W})} \|f\|_{\mathcal{F}_{p,\mu}}$$

From RKHS to Barron space

Definition (Barron space [17] (E, Ma, Wu, 2021))

For any $1 \leq p \leq \infty$, we have

$$\mathcal{B} = \cup_{\mu \in \mathcal{P}(\mathcal{W})} \mathcal{F}_{p,\mu}\,, \quad \|f\|_{\mathcal{B}} = \inf_{\mu \in \mathcal{P}(\mathcal{W})} \|f\|_{\mathcal{F}_{p,\mu}}$$

Remark: ◦ Two-layer neural networks: data-adaptive kernel $\mathcal{B} = \cup_{\mu \in \mathcal{P}(\mathcal{W})} \mathcal{H}_{k_\mu}$
◦ equivalent to path norm $\|\boldsymbol{\Theta}\|_{\mathcal{P}} := \frac{1}{m} \sum_{k=1}^{m} |a_k| \|\boldsymbol{w}_k\|_1$
◦ parameter space vs. measure space
e.g., [7] (Bach, 2017), [18] (Bartolucci, Vito, Rosasco, Vigogna, 2022).

From RKHS to Barron space

Definition (Barron space [17] (E, Ma, Wu, 2021))

For any $1 \leq p \leq \infty$, we have

$$\mathcal{B} = \cup_{\mu \in \mathcal{P}(\mathcal{W})} \mathcal{F}_{p,\mu}\,, \quad \|f\|_{\mathcal{B}} = \inf_{\mu \in \mathcal{P}(\mathcal{W})} \|f\|_{\mathcal{F}_{p,\mu}}$$

Remark: ◦ Two-layer neural networks: data-adaptive kernel $\mathcal{B} = \cup_{\mu \in \mathcal{P}(\mathcal{W})} \mathcal{H}_{k_\mu}$
◦ equivalent to path norm $\|\boldsymbol{\Theta}\|_{\mathcal{P}} := \frac{1}{m} \sum_{k=1}^{m} |a_k| \|\boldsymbol{w}_k\|_1$
◦ parameter space vs. measure space
e.g., [7] (Bach, 2017), [18] (Bartolucci, Vito, Rosasco, Vigogna, 2022).

Optimization in Barron spaces is difficult: curse of dimensionality!

	approximation	generalization	optimization
RKHS	CoD	-	-
Barron spaces	$\mathcal{O}(m^{-\frac{2d}{d+3}})$	$\Theta(n^{-\frac{d+3}{2d+3}})$?	CoD

- [19] (Siegel, Xu, 2022) on metric entropy

$$\epsilon^{-\frac{2d}{d+3}}{}_d \lesssim \log \mathscr{N}_2(\mathcal{G}_1, \epsilon) \lesssim_d \epsilon^{-\frac{2d}{d+3}} .$$

	approximation	generalization	optimization
RKHS	CoD	-	-
Barron spaces	$\mathcal{O}(m^{-\frac{2d}{d+3}})$	$\Theta(n^{-\frac{d+3}{2d+3}})$?	CoD

- [19] (Siegel, Xu, 2022) on metric entropy

$$\epsilon^{-\frac{2d}{d+3}_d} \lesssim \log \mathscr{N}_2(\mathcal{G}_1, \epsilon) \lesssim_d \epsilon^{-\frac{2d}{d+3}} \leqslant 6144 d^5 \epsilon^{-\frac{2d}{d+2}} \quad \text{[Ours]}$$

What is the suitable function space beyond RKHS, that can be learned both statistically and computationally for NNs?

What is the suitable function space beyond RKHS, that can be learned both statistically and computationally for NNs?

- *Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond.* (Liu, Huang, Chen, Suykens, TPAMI2021).
- IEEE ICASSP 2023 Tutorial - "Neural networks: the good, the bad, and the ugly"
- CVPR 2023 Tutorial - "Deep learning theory for computer vision"

What is the suitable function space beyond RKHS, that can be learned both statistically and computationally for NNs?

- *Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond.* (Liu, Huang, Chen, Suykens, TPAMI2021).
- IEEE ICASSP 2023 Tutorial - "Neural networks: the good, the bad, and the ugly"
- CVPR 2023 Tutorial - "Deep learning theory for computer vision"

Thanks for your attention!

Q & A

my homepage www.lfhsgre.org for more information!

References I

[1] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever.
Deep double descent: Where bigger models and more data hurt.
In *International Conference on Learning Representations*, 2019.
(Cited on pages 3 and 4.)

[2] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal.
Reconciling modern machine-learning practice and the classical bias–variance trade-off.
the National Academy of Sciences, 116(32):15849–15854, 2019.
(Cited on pages 3 and 4.)

[3] Ali Rahimi and Benjamin Recht.
Random features for large-scale kernel machines.
In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007.
(Cited on pages 5 and 6.)

[4] Song Mei and Andrea Montanari.
The generalization error of random features regression: Precise asymptotics and the double descent curve.
Communications on Pure and Applied Mathematics, 75(4):667–766, 2022.
(Cited on pages 7, 8, 9, and 10.)

References II

[5] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari.
Linearized two-layers neural networks in high dimension.
Annals of Statistics, 49(2):1029–1054, 2021.
(Cited on pages 11, 12, 13, and 14.)

[6] Noureddine El Karoui.
The spectrum of kernel random matrices.
Annals of Statistics, 38(1):1–50, 2010.
(Cited on pages 11, 12, 13, and 14.)

[7] Francis Bach.
Breaking the curse of dimensionality with convex neural networks.
Journal of Machine Learning Research, 18(1):629–681, 2017.
(Cited on pages 11, 12, 13, 14, 38, 39, 47, 48, and 49.)

[8] Francis Bach and Eric Moulines.
Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$.
Advances in Neural Information Processing Systems, 26:773–781, 2013.
(Cited on pages 23 and 24.)

References III

[9] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, Venkata Krishna Pillutla, and Aaron Sidford.
A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares).
arXiv preprint arXiv:1710.09430, 2017.
(Cited on pages 23 and 24.)

[10] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade.
Benign overfitting of constant-stepsize sgd for linear regression.
In *Conference on Learning Theory*, 2021.
(Cited on pages 23 and 24.)

[11] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco.
Learning with SGD and random features.
In *Advances in Neural Information Processing Systems*, pages 10212–10223, 2018.
(Cited on page 30.)

[12] Mikhail Belkin, Daniel Hsu, and Ji Xu.
Two models of double descent for weak features.
SIAM Journal on Mathematics of Data Science, 2(4):1167–1180, 2020.
(Cited on page 30.)

References IV

[13] Gilad Yehudai and Ohad Shamir.
On the power and limitations of random features for understanding neural networks.
In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2019.
(Cited on pages 38 and 39.)

[14] Michael Celentano, Theodor Misiakiewicz, and Andrea Montanari.
Minimum complexity interpolation in random features models.
arXiv preprint arXiv:2103.15996, 2021.
(Cited on pages 38, 39, 43, 44, 45, and 46.)

[15] Ali Rahimi and Benjamin Recht.
Uniform approximation of functions with random bases.
In *Annual Allerton Conference on Communication, Control, and Computing*, pages 555–561. IEEE, 2008.
(Cited on pages 40, 41, and 42.)

[16] Hongrui Chen, Jihao Long, and Lei Wu.
A duality framework for generalization analysis of random feature models and two-layer neural networks.
arXiv preprint arXiv:2305.05642, 2023.
(Cited on pages 43, 44, 45, and 46.)

References V

[17] Weinan E, Chao Ma, and Lei Wu.
The barron space and the flow-induced function spaces for neural network models.
Constructive Approximation, pages 1–38, 2021.
(Cited on pages 47, 48, and 49.)

[18] Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna.
Understanding neural networks with reproducing kernel Banach spaces.
Applied and Computational Harmonic Analysis, 2023.
(Cited on pages 47, 48, and 49.)

[19] Jonathan W Siegel and Jinchao Xu.
Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks.
arXiv preprint arXiv:2101.12365, 2021.
(Cited on pages 50 and 51.)