On the Double Descent of Random Features Models Trained with SGD

Fanghui Liu (*EPFL*), **Johan A.K. Suykens** (*KU Leuven*), **Volkan Cevher** (*EPFL*)

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

25th Nov. 2021

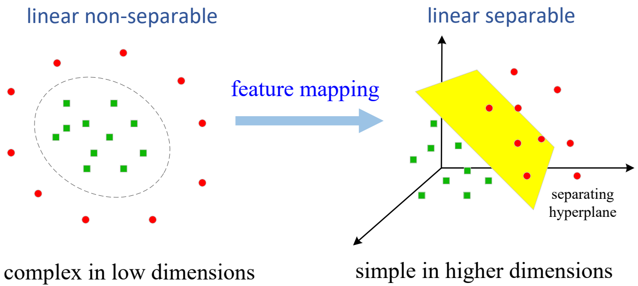
Outline

Research overview

Random features in double descent

Conclusion

Research Overview: Kernel approximation



Scalability of kernel methods: n-by-n kernel matrix.
Solution: approximate the kernel by a low-rank representation

- Nyström approximation: approximate the kernel matrix
- Random Fourier features[1]: approximate the kernel function

[1] Rahimi A, Recht B. Random features for large-scale kernel machines, NeurIPS2007. (the test-of-time award in NeurIPS2017)

Research Overview: Random Fourier features

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}} \approx \varphi^\top(\mathbf{x})\varphi(\mathbf{x}')\,,$$

where $\varphi(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^s$ is an **explicit** feature mapping

Bochner's theorem [1]

For a shift-invariant $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}')$ and positive definite kernel,

$$k(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^d} p(\omega) \exp\left(\mathrm{i}\omega^\top(\mathbf{x} - \mathbf{x}')\right) \mathrm{d}\omega$$

$$\approx \frac{1}{s}\sum_{j=1}^{s} \exp(\mathrm{i}\omega_j^\top \mathbf{x}) \exp(\mathrm{i}\omega_j^\top \mathbf{x}')^* = \varphi(\mathbf{x})^\top \varphi(\mathbf{x}')$$

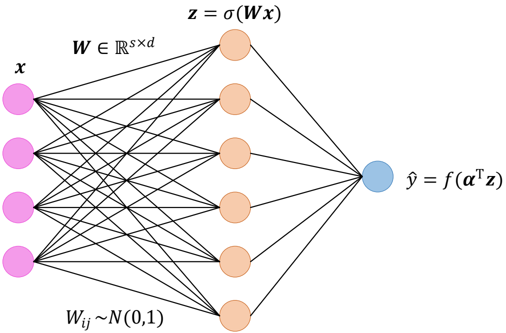
the explicit feature mapping:

$$\varphi(\mathbf{x}) := \frac{1}{\sqrt{s}}\left[\exp(-\mathrm{i}\omega_1^\top \mathbf{x}), \cdots, \exp(-\mathrm{i}\omega_s^\top \mathbf{x})\right]^\top.$$

Research Overview: Neural network view

RF model: a two-layer, (infinite)-width, fully-connected neural network

$$k\left(\mathbf{x},\mathbf{x}'\right) = \mathbb{E}_{\omega\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\sigma(\omega^\top \mathbf{x})\sigma(\omega^\top \mathbf{x}')]$$



- Gaussian kernel: $\sigma(x) = [\cos(x), \sin(x)]^\top$
- the 1st-order arc-cosine kernel: $\sigma(x) = \max\{0, x\}$
- soft-max in attention: $\sigma(x) = \exp(x)$

Research Overview: Applied to Linearized Attention in Transformers

self attention

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \underbrace{\text{softmax}(\mathbf{Q}\mathbf{K}^\top)}_{:=\mathbf{A}} \mathbf{V} \approx \mathbf{Q}'\mathbf{K}'^\top \mathbf{V}\,,$$

where $\mathbf{A}_{ij} = k(\mathbf{q}_i, \mathbf{k}_j) = \mathbb{E}[\sigma(\mathbf{q}_i)^\top \sigma(\mathbf{k}_j)]$

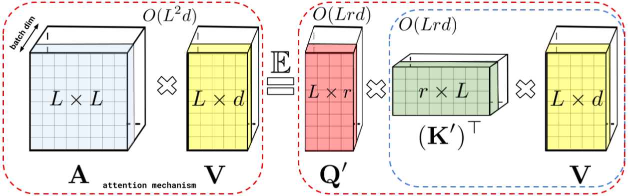


Figure: Approximation of self-attention. source: [2].

- soft-max in attention: $\exp(\mathbf{x}^\top \mathbf{x}') = \mathbb{E}_{\omega \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)} \left[\exp\left(\omega^\top \mathbf{x} - \frac{\|\mathbf{x}\|_2^2}{2} \right) \exp\left(\omega^\top \mathbf{x}' - \frac{\|\mathbf{x}'\|_2^2}{2} \right) \right]$

Research Overview: Taxonomy

i) Monte Carlo sampling
- variance reduction
- acceleration

ii) Quasi-Monte Carlo sampling
- QMC
- SSF
- MM

iii) Quadrature rules
- GQ, SGQ
- SSR

data-dependent
- random features selection/learning
- leverage score

$$
\begin{aligned}
k(\boldsymbol{x}-\boldsymbol{x}') &= \int_{\mathbb{R}^d} p(\boldsymbol{\omega}) \exp\left(\mathrm{i}\boldsymbol{\omega}^\top(\boldsymbol{x}-\boldsymbol{x}')\right)\mathrm{d}\boldsymbol{\omega} \\
&= \int_{[0,1]^d} \exp\left(\mathrm{i}(\boldsymbol{x}-\boldsymbol{x}')^\top \Phi^{-1}(\boldsymbol{t})\right)\mathrm{d}\boldsymbol{t} \\
&= \prod_{j=1}^d \left(\int_{-\infty}^{\infty} p_j(\omega^{(j)}) \exp\left(\mathrm{i}\omega^{(j)}(x^{(j)} - x'^{(j)})\right)\mathrm{d}\omega^{(j)}\right) \\
&= (2\pi)^{-d/2} \int_{U_d} \int_0^\infty e^{-\frac{r^2}{2}} |r|^{d-1} g(r\boldsymbol{u}) \mathrm{d}r\mathrm{d}\boldsymbol{u} \\
&= \int_{\mathbb{R}^d} q(\boldsymbol{\omega}) \frac{p(\boldsymbol{\omega})}{q(\boldsymbol{\omega})} \exp\left(\mathrm{i}\boldsymbol{\omega}^\top(\boldsymbol{x}-\boldsymbol{x}')\right)\mathrm{d}\boldsymbol{w}
\end{aligned}
$$

Figure: Taxonomy of random features based algorithms[2].

[2] Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan A.K. Suykens. *Random Features for Kernel Approximation: A Survey on Algorithms, Theory, and Beyond.* TPAMI2021.

Outline

Research overview

Random features in double descent

Conclusion

Background: Double descent

over-parameterized models, e.g., neural networks, random features

- high dimensions: large n and d
- abnormal phenomena: training error can be zero but still generalize well

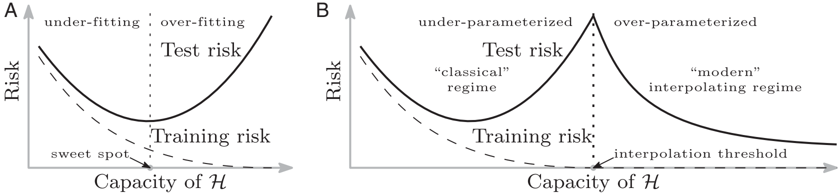


Figure: Bias-variance trade-off [3] (Belkin et al. PNAS2019).

Research Overview: Motivation

- interplay between optimization and excess risk: trained by SGD
- bias-variance decomposition for understanding multiple randomness sources

	data assumption	solution	result
(Hastie et al., 2019)	Gaussian	closed-form	variance $\nearrow\searrow$
(Ba et al., 2020)	Gaussian	GD	variance $\nearrow\searrow$
(Mei & Montanari, 2019)	i.i.d on sphere	closed-form	variance, bias $\nearrow\searrow$
(d'Ascoli et al., 2020a)	Gaussian	closed-form	refined [2]
(Gerace et al., 2020)	Gaussian	closed-form	$\nearrow\searrow$
(Adlam & Pennington, 2020)	Gaussian	closed-form	refined
(Dhifallah & Lu, 2020)	Gaussian	closed-form	$\nearrow\searrow$
(Hu & Lu, 2020)	Gaussian	closed-form	$\nearrow\searrow$
(Liao et al., 2020)	general	closed-form	$\nearrow\searrow$
(Lin & Dobriban, 2021)	isotropic features with finite moments	closed form	refined
(Li et al., 2021)	correlated features with polynomial decay on Σ_d	closed form	interpolation learning
Ours	(at least) sub-exponential data	SGD	variance $\nearrow\searrow$, bias $\searrow$

[1] A refined decomposition on variance is conducted by sources of randomness on data sampling, initialization, label noise to possess each term (d'Ascoli et al., 2020b) or their full decomposition in (Adlam & Pennington, 2020; Lin & Dobriban, 2021).

Problem settings: Random features regression model

data: $y = f_\rho(\mathbf{x}) + \varepsilon$

- training data: $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \sim \rho$
  Assumption: sub-exponential data and $\|\mathbf{x}\|_2^2 \sim \mathcal{O}(d)$
- target function: $f_\rho(\mathbf{x}) = \int_Y y \, \mathrm{d}\rho(y \mid \mathbf{x})$
- noise: $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{E}(\varepsilon^2) = \tau^2$

function space

define the random features mapping $\varphi(\mathbf{x}) := \frac{1}{\sqrt{m}}\sigma(\mathbf{W}\mathbf{x}/\sqrt{d})$,

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

covariance operator: $\Sigma_m := \int_X [\varphi(\mathbf{x}) \otimes \varphi(\mathbf{x})] \mathrm{d}\rho_X(\mathbf{x})$

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

Problem settings: averaged SGD under adaptive step-size setting

$$\theta_t = \theta_{t-1} + \gamma_t [y_t - \langle \theta_{t-1}, \varphi(\mathbf{x}_t) \rangle] \varphi(\mathbf{x}_t), \qquad t = 1, 2, \dots n \,,$$

- averaged output: $\bar{\theta}_n := \frac{1}{n} \sum_{t=0}^{n-1} \theta_t \Longrightarrow \bar{f}_n = \langle \varphi(\cdot), \bar{\theta}_n \rangle$
- adaptive step-size: $\gamma_t := \gamma_0 t^{-\zeta}, \zeta \in [0, 1)$
- optimal solution: $f^* = \arg\min_{f \in \mathcal{H}} \|f - f_\rho\|^2_{L^2_{\rho_X}}$
- averaged excess risk: $\mathbb{E}\|\bar{f}_n - f^*\|^2_{L^2_{\rho_X}} = \mathbb{E}_{\mathbf{X},\mathbf{W},\varepsilon} \langle \bar{f}_n - f^*, \Sigma_m (\bar{f}_n - f^*) \rangle$

Assumptions

- **boundedness of f^*:** $\|f^*\|_{\mathcal{H}} < \infty$
- **high dimension:** $c \leqslant \{d/n, m/n\} \leqslant C$, $\|\mathbf{x}\|_2^2 \sim \mathcal{O}(d)$, $\Sigma_d := \mathbb{E}_{\mathbf{x}}[\mathbf{x} \otimes \mathbf{x}]$ with $\|\Sigma_d\|_2 < \infty$
- **activation function:** $\sigma(\cdot)$: Lipschitz continuous
- **noise condition**: $\Xi := \mathbb{E}_{\mathbf{x}}[\varepsilon^2 \varphi(\mathbf{x}) \otimes \varphi(\mathbf{x})] \preccurlyeq \tau^2 \Sigma_m$.
  uniformly bounded noise, sub-Gaussian noise
- **fourth moment condition:**
  for any PSD operator A, we have $\mathbb{E}_{\mathbf{W}}[\Sigma_m A \Sigma_m] \preccurlyeq r' \mathbb{E}_{\mathbf{W}}[\mathrm{Tr}(\Sigma_m A)\Sigma_m] \preccurlyeq r \mathrm{Tr}(\widetilde{\Sigma}_m A)\widetilde{\Sigma}_m$.
  1) The special case $A := I$ can be proved.
  2) holds for sub-Gaussian/exponential data.

Properties of covariance operators

$\sigma(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ Lipschitz continuous
covariance operator $\Sigma_m := \mathbb{E}_{\mathbf{x}}[\varphi(\mathbf{x}) \otimes \varphi(\mathbf{x})]$
expected covariance operator $\widetilde{\Sigma}_m := \mathbb{E}_{\mathbf{x},\mathbf{W}}[\varphi(\mathbf{x}) \otimes \varphi(\mathbf{x})]$

eigenvalue of $\widetilde{\Sigma}_m$

the same diagonal/non-diagonal elements: $\mathcal{O}(1/m)$
two distinct eigenvalues: $\widetilde{\lambda}_1 \sim \mathcal{O}(1)$, $\widetilde{\lambda}_2 \sim \mathcal{O}(1/m)$

sub-exponential random variables

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

Bias-variance decomposition

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

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

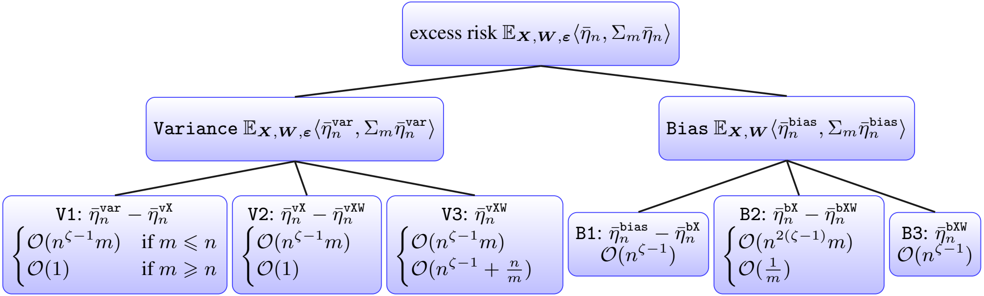
$$\eta_t^{\texttt{bias}} = [I - \gamma_t \varphi(\mathbf{x}_t) \otimes \varphi(\mathbf{x}_t)]\eta_{t-1}^{\texttt{bias}}, \quad \eta_0^{\texttt{bias}} = f^*\,,$$

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

Bias-variance decomposition

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

Proof framework

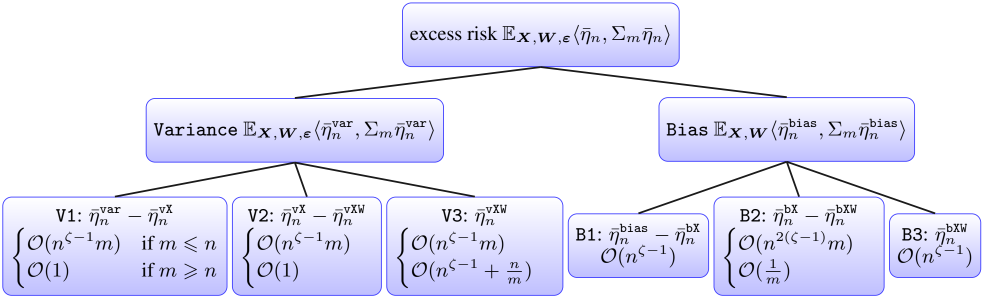


$$\mathtt{Bias}: \quad \eta_t^{\mathtt{bias}} = [I - \gamma_t \varphi(\mathbf{x}_t) \otimes \varphi(\mathbf{x}_t)]\eta_{t-1}^{\mathtt{bias}}$$

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

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

Proof framework



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

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

- $\texttt{V1} := \mathbb{E}_{\mathbf{X},\mathbf{W},\varepsilon}\left[\langle \bar{\eta}_n^{\texttt{var}} - \bar{\eta}_n^{\texttt{vX}}, \Sigma_m(\bar{\eta}_n^{\texttt{var}} - \bar{\eta}_n^{\texttt{vX}})\rangle\right]$
- $\texttt{V2} := \mathbb{E}_{\mathbf{X},\mathbf{W},\varepsilon}\left[\langle \bar{\eta}_n^{\texttt{vX}} - \bar{\eta}_n^{\texttt{vXW}}, \Sigma_m(\bar{\eta}_n^{\texttt{vX}} - \bar{\eta}_n^{\texttt{vXW}})\rangle\right]$
- $\texttt{V3} := \mathbb{E}_{\mathbf{X},\mathbf{W},\varepsilon}\langle \bar{\eta}_n^{\texttt{vXW}}, \Sigma_m \bar{\eta}_n^{\texttt{vXW}} \rangle$

Results: error bounds

Theorem

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*

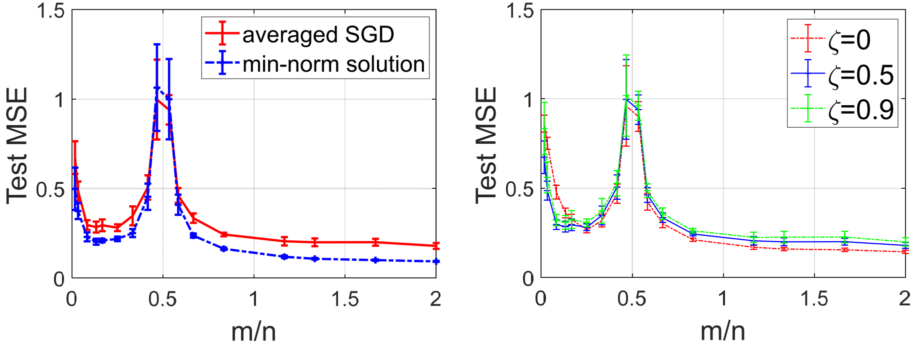
$$\mathtt{Bias} \lesssim \frac{\gamma_0 r' n^{\zeta - 1}}{\sqrt{\mathbb{E}[1 - \gamma_0 r' \operatorname{Tr}(\Sigma_m)]^4}} \|f^*\|^2 \sim \mathcal{O}\left(n^{\zeta-1}\right) .$$

$$\mathtt{Variance} \lesssim \frac{\gamma_0 r' \tau^2}{\sqrt{\mathbb{E}[1 - \gamma_0 r' \operatorname{Tr}(\Sigma_m)]^2}} \begin{cases} m n^{\zeta - 1}, & \text{if } m \leqslant n \\ \gamma_0 \tau^2, & \text{if } m > n \end{cases}$$

$$\sim \begin{cases} \mathcal{O}\left(m n^{\zeta - 1}\right), & \text{if } m \leqslant n \\ \mathcal{O}(1), & \text{if } m > n . \end{cases}$$

Experiments on MNIST

Gaussian kernel $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2d}\right)$



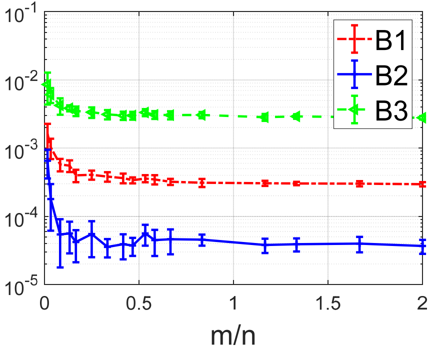
(a) SGD vs. min-norm solution

(b) step-size

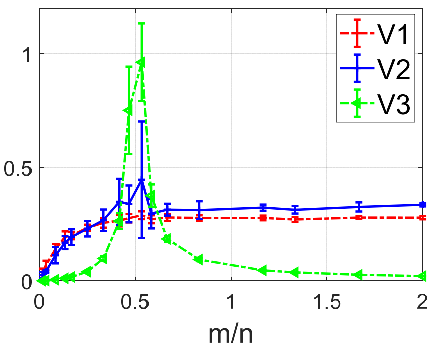
Figure: Test MSE (mean±std.) of RF regression as a function of the ratio m/n on MNIST data set (digit 3 vs. 7) for $d = 784$ and $n = 600$.

Validation for bias and variance

- noise: $\varepsilon \sim \mathcal{N}(0,1)$
- Σ_m, $\widetilde{\Sigma}_m$: sample covariance matrices with Monte Carlo sampling



(a) Bias



(b) Variance

Outline

Take-away message

- high dimensional random features model trained by SGD
- findings
  - expected covariance operator $\widetilde{\Sigma}_m$ has only two distinct eigenvalues
  - bias-variance decomposition: multiple randomness sources
  - monotonic decreasing bias and unimodal variance
  - optimization effect on excess risk: constant step-size SGD vs. min-norm solution

Future works:

- SGD: implicit bias/regularization
- function space, high dimensions

Thanks for your attention!

Q & A

my homepage http://lfhsgre.org for more information!

NEW: ERC Advanced Grant E-DUALITY

Exploring duality for future data-driven modelling

References I

[1] Salomon Bochner.
Harmonic Analysis and the Theory of Probability.
Courier Corporation, 2005.

[2] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, and Weller Adrian.
Rethinking attention with performers.
In *International Conference on Learning Representations*, 2021.

[3] 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.