Deep Learning - Parameters and Functions

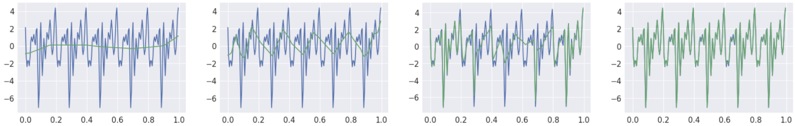
Spectral biases

Guido Montúfar
montufar@math.ucla.edu

48th Winter Conference in Statistics, March 2024, Hemavan

Benjamin Bowman

- "Spectral Bias Outside the Training Set for Deep Networks in the Kernel Regime"
- "Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks"



(a) Iteration 100 (b) Iteration 1000 (c) Iteration 10000 (d) Iteration 80000

Figure 1: Learned function (green) as training progresses[1].

[1] Rahaman et al. 2019.

Spectral biases

- For shallow univariate ReLU networks the dominant eigenfunctions of the Neural Tangent Kernel are smoother[2]
- ReLU nets in the kernel regime are biased towards smooth interpolants[3]
- "Spectral Bias" can be interpreted to mean bias towards learning the top eigenfunctions of the NTK
- By looking at empirical approximations to the eigenfunctions, spectral bias was demonstrated to hold on the training set[4]

---

[2] Basri et al. 2019, 2020.

[3] Jin and Montúfar 2023; Williams et al. 2019.

[4] Arora et al. 2019a; Basri et al. 2020; Cao et al. 2021.

Overview

- We provide quantitative bounds measuring the L^2 difference in function space between the trajectory of a

finite-width network trained on finitely many samples	idealized kernel dynamics of infinite width and infinite data

Overview

- We provide quantitative bounds measuring the L^2 difference in function space between the trajectory of a

finite-width network trained on finitely many samples	idealized kernel dynamics of infinite width and infinite data

- As an implication, eigenfunctions of the NTK integral operator (not just their empirical approximations) are learned at rates corresponding to their eigenvalues

Overview

- We provide quantitative bounds measuring the L^2 difference in function space between the trajectory of a

finite-width network trained on finitely many samples	idealized kernel dynamics of infinite width and infinite data

- As an implication, eigenfunctions of the NTK integral operator (not just their empirical approximations) are learned at rates corresponding to their eigenvalues
- The network inherits bias of the kernel at beginning of training even when the width only grows linearly with the training sample

NTK and convergence

- The NTK was introduced by Jacot, Gabriel, and Hongler 2018, and Du et al. 2018 used it implicitly to prove global convergence of GD in shallow ReLU network
- Since then, the NTK has been used to obtain global convergence for arbitrary labels in a series of works[5]
- For global convergence for arbitrary labels, a usual requirement is that the network width m is $\Omega(poly(n))$ or $\Omega(1/\epsilon)$
- If the target function aligns with the NTK model, for shallow nets this can be reduced to polylogarithmic (for the logistic loss) or linear (for the squared loss)[6]

[5] Allen-Zhu, Li, and Song 2019; Du et al. 2019; Du et al. 2018; Nguyen 2021; Nguyen and Mondelli 2020; Oymak and Soltanolkotabi 2020; Zou and Gu 2019; Zou et al. 2020.

[6] Bowman and Montúfar 2022a; E, Ma, and Wu 2020; Ji and Telgarsky 2020; Su and Yang 2019.

NTK spectrum and generalization

- The NTK tends to have skewed spectrum with a small number of large outlier eigenvalues[7]
- The spectrum of the NTK integral operator for ReLU networks has been shown to asymptotically follow a power law[8]
- Top eigenvectors of the NTK and low effective rank have appeared in generalization bounds and robustness[9]

[7] Arora et al. 2019a; Fan and Wang 2020; Karakida, Akaho, and Amari 2021; Li, Soltanolkotabi, and Oymak 2020; Murray et al. 2023; Oymak et al. 2020; Pennington and Bahri 2017; Pennington and Worah 2018; Yang and Salman 2019.

[8] Velikanov and Yarotsky 2021.

[9] Arora et al. 2019a; Li, Soltanolkotabi, and Oymak 2020; Oymak et al. 2020.

NTK eigenvector and eigenfunction convergence

- For infinitely wide networks the projections of the residual along eigenvectors of NTK decay linearly with rate of eigenvalues[10]
- We show a corresponding statement for the test residual instead of the empirical residual:

  Projections of the test residual along eigen*functions* of the NTK *integral operator* are learned at rates given by the eigenvalues.

  Moreover, the result holds for networks that do not need to be under or extremely overparametrized and diverse architectures.

[10] Arora et al. 2019a; Basri et al. 2020; Cao et al. 2021; Luo et al. 2022.

Preliminaries

Settings

- Neural network: $f(x;\theta)$ taking inputs $x \in X \subset \mathbb{R}^d$, parameterized by $\theta \in \mathbb{R}^p$.
- Training data: $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$, $y_i = f^*(x_i)$.
- Residual error on training set: $\hat{r}(\theta) \in \mathbb{R}^n$, $\hat{r}(\theta)_i := f(x_i;\theta) - y_i$.
- Squared error loss:

$$\Phi(\theta) := \frac{1}{2n} \|\hat{r}(\theta)\|_2^2 = \frac{1}{2} \|\hat{r}(\theta)\|_{\mathbb{R}^n}^2$$

- Gradient flow:

$$\partial_t \theta_t = -\partial_\theta \Phi(\theta)$$

$\langle \bullet, \bullet \rangle$ and $\|\bullet\|_2$ Euclidean inner product and norm. $\langle \bullet, \bullet \rangle_{\mathbb{R}^n} = \frac{1}{n} \langle \bullet, \bullet \rangle$ and $\|\bullet\|_{\mathbb{R}^n} := \sqrt{\langle \bullet, \bullet \rangle_{\mathbb{R}^n}}$. Let $L^p(X, \nu)$ denote the L^p space over domain X with measure ν.

NTK definitions

- Analytical NTK:

$$K^\infty(x, x') := \mathbb{E}_{\theta_0 \sim \mu}\left[\langle \nabla_\theta f(x; \theta_0), \nabla_\theta f(x'; \theta_0)\rangle\right],$$

with expectation taken over the parameter initialization $\theta_0 \sim \mu$.

NTK definitions

- Analytical NTK:

$$K^{\infty}(x,x') := \mathbb{E}_{\theta_0 \sim \mu}\left[\langle \nabla_\theta f(x;\theta_0), \nabla_\theta f(x';\theta_0)\rangle\right],$$

  with expectation taken over the parameter initialization $\theta_0 \sim \mu$.
- Integral operator: The kernel K^{∞} induces

$$T_{K^{\infty}} : L^2(X,\rho) \to L^2(X,\rho); \quad g(x) \mapsto \int_X K^{\infty}(x,s) g(s) d\rho(s), \tag{1}$$

  where X is our input space and ρ is the input distribution.

NTK definitions

- Analytical NTK:

$$K^{\infty}(x,x') := \mathbb{E}_{\theta_0\sim\mu}\left[\langle \nabla_\theta f(x;\theta_0), \nabla_\theta f(x';\theta_0)\rangle\right],$$

  with expectation taken over the parameter initialization $\theta_0 \sim \mu$.

- Integral operator: The kernel K^{∞} induces

$$T_{K^{\infty}} : L^2(X,\rho) \to L^2(X,\rho); \quad g(x) \mapsto \int_X K^{\infty}(x,s)g(s)d\rho(s), \tag{1}$$

  where X is our input space and ρ is the input distribution.

- Spectral decomposition: By Mercer's theorem we have

$$K^{\infty}(x,x') = \sum_{i=1}^{\infty} \sigma_i \phi_i(x)\phi_i(x'),$$

  where $\{\phi_i\}$ is an orthonormal basis for $L^2(X,\rho)$ and $\{\sigma_i\}$ is a nonincreasing sequence of positive values.

Finite data and finite width

- Discretization: Training sample $x_1, \dots, x_n$ introduces

$$T_n : g(x) \mapsto \frac{1}{n} \sum_{i=1}^{n} K^\infty(x, x_i) g(x_i) = \int_X K^\infty(x, s) g(s) d\widehat{\rho}(s), \tag{2}$$

where $\widehat{\rho} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ is the empirical measure.

Finite data and finite width

- Discretization: Training sample $x_1, \dots, x_n$ introduces

$$T_n : g(x) \mapsto \frac{1}{n}\sum_{i=1}^{n} K^\infty(x, x_i) g(x_i) = \int_X K^\infty(x, s) g(s) d\widehat{\rho}(s), \tag{2}$$

  where $\widehat{\rho} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ is the empirical measure.

- Time dependent NTK:

$$K_t(x, x') := \langle \nabla_\theta f(x; \theta_t), \nabla_\theta f(x'; \theta_t) \rangle$$

  has an associated time-dependent operator T_n^t

$$T_n^t g(x) := \frac{1}{n}\sum_{i=1}^{n} K_t(x, x_i) g(x_i) = \int_X K_t(x, s) g(s) d\widehat{\rho}(s). \tag{3}$$

Finite data and finite width

- Discretization: Training sample $x_1, \ldots, x_n$ introduces

$$T_n : g(x) \mapsto \frac{1}{n} \sum_{i=1}^{n} K^{\infty}(x, x_i) g(x_i) = \int_X K^{\infty}(x, s) g(s) d\widehat{\rho}(s), \tag{2}$$

  where $\widehat{\rho} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ is the empirical measure.
- Time dependent NTK:

$$K_t(x, x') := \langle \nabla_\theta f(x; \theta_t), \nabla_\theta f(x'; \theta_t) \rangle$$

  has an associated time-dependent operator T_n^t

$$T_n^t g(x) := \frac{1}{n} \sum_{i=1}^{n} K_t(x, x_i) g(x_i) = \int_X K_t(x, s) g(s) d\widehat{\rho}(s). \tag{3}$$

- Update rule: under gradient flow the residual is given by

$$\partial_t r_t(x) = -\frac{1}{n} \sum_{i=1}^{n} K_t(x, x_i) r_t(x_i) = -T_n^t r_t.$$

Idealized infinite width and infinite data

- Infinite width limit: Speaking loosely, as the network width tends to infinity the time-dependent NTK becomes constant so that

$$K_t(x, x') = K^\infty(x, x') \quad \text{and} \quad T_n^t = T_n$$

Idealized infinite width and infinite data

- Infinite width limit: Speaking loosely, as the network width tends to infinity the time-dependent NTK becomes constant so that

$$K_t(x, x') = K^\infty(x, x') \quad \text{and} \quad T_n^t = T_n$$

- Infinite data limit: Similarly, heuristically as $n \to \infty$ we have

$$T_n \to T_{K^\infty}$$

Idealized infinite width and infinite data

- Infinite width limit: Speaking loosely, as the network width tends to infinity the time-dependent NTK becomes constant so that

$$K_t(x, x') = K^{\infty}(x, x') \quad \text{and} \quad T_n^t = T_n$$

- Infinite data limit: Similarly, heuristically as $n \to \infty$ we have

$$T_n \to T_{K^{\infty}}$$

- In this idealized setting the update rule is $\partial_t r_t = -T_{K^{\infty}} r_t$, which has the solution $r_t = \exp(-T_{K^{\infty}} t) r_0$ defined via

$$\langle r_t, \phi_i \rangle_{\rho} = \exp(-\sigma_i t) \langle r_0, \phi_i \rangle_{\rho}. \tag{4}$$

Idealized infinite width and infinite data

- Infinite width limit: Speaking loosely, as the network width tends to infinity the time-dependent NTK becomes constant so that

$$K_t(x, x') = K^\infty(x, x') \quad \text{and} \quad T_n^t = T_n$$

- Infinite data limit: Similarly, heuristically as $n \to \infty$ we have

$$T_n \to T_{K^\infty}$$

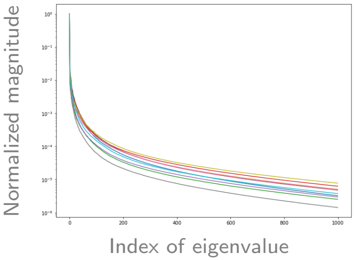
- In this idealized setting the update rule is $\partial_t r_t = -T_{K^\infty} r_t$, which has the solution $r_t = \exp(-T_{K^\infty} t) r_0$ defined via

$$\langle r_t, \phi_i \rangle_\rho = \exp(-\sigma_i t) \langle r_0, \phi_i \rangle_\rho. \tag{4}$$

- Thus in this idealized setting the network learns eigenfunctions ϕ_i at rates determined by their eigenvalues σ_i.

- The dependence of the convergence rate on σ_i is particularly relevant as the NTK tends to have a very skewed spectrum

LeNet-5 CNN NTK on MNIST



Shallow FC softplus NTK on CIFAR10

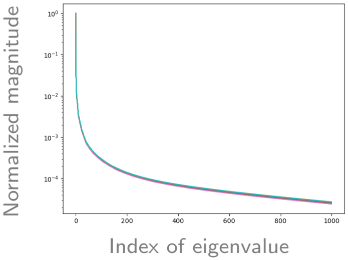


Figure 2: Normalized NTK spectrum λ_k/λ_1 on MNIST and CIFAR10 for two networks using 10 random parameter initializations and data batches.

Spectral bias outside the training set

- We will see that the bias at the beginning of training can be described entirely through T_{K^∞} and its eigenfunctions.
- This depends only on the model architecture, parameter initialization distribution μ, and input distribution ρ.

- We consider deep networks of the form:

$$\begin{aligned} \alpha^{(0)} &:= x, \\ \alpha^{(l)} &:= \psi_l(\theta^{(l)}, \alpha^{(l-1)}), \quad l \in [L], \\ f(x;\theta) &:= \frac{1}{\sqrt{m_L}} v^T \alpha^{(L)}, \end{aligned}$$

- We assume each layer ψ_l has one of the following forms:

$$\begin{aligned} \text{Fully Connected} : \psi_l(\theta^{(l)}, \alpha^{(l-1)}) &= \omega\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} \alpha^{(l-1)}\right) \\ \text{Convolutional} : \psi_l(\theta^{(l)}, \alpha^{(l-1)}) &= \omega\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} * \alpha^{(l-1)}\right) \\ \text{Residual} : \psi_l(\theta^{(l)}, \alpha^{(l-1)}) &= \omega\left(\frac{1}{\sqrt{m_{l-1}}} W^{(l)} \alpha^{(l-1)}\right) + \alpha^{(l-1)} \end{aligned}$$

- We assume $\max m_l/m = O(1)$, $m = \min_l m_l$, and treat input dimension $d := m_0$, depth L, and filter sizes K as constant.

- Use antisymmetric initialization[11] with $\theta_0 \sim N(0, I)$
- This simultaneously ensures that the model is identically zero at initialization without changing the NTK at initialization

[11] Zhang et al. 2020.

Assumption 1

1. Twice continuously differentiable activation ω, Lipschitz ω, ω'
   (satisfied by most activations except ReLU)
2. Compact input domain X with strictly positive Borel measure ρ
   (sufficient condition for Mercer's theorem)
3. Target function f^* satisfies $\|f^*\|_{L^\infty(X,\rho)} = O(1)$
   (the target function is bounded)
4. Antisymmetric initilization so that $f(\bullet; \theta_0) \equiv 0$
   (probably not strictly necessary)

Theorem 1

- *Let $K(x, x')$ fixed continuous, symmetric, positive definite kernel*
- *Let $P_k : L^2(X, \rho) \to L^2(X, \rho)$ denote the orthogonal projection onto the span of the top k eigenfcts of the operator T_K*
- *Let $\sigma_k > 0$ denote the k-th eigenvalue of T_K*

Then $m = \tilde{\Omega}(T^4/\epsilon^2)$ and $n = \tilde{\Omega}(T^2/\epsilon^2)$ suffices to ensure with probability $1 - O(mn)\exp(-\Omega(\log^2 m)$ over the parameter initilization and the training samples that for all $t \leq T$ and $k \in \mathbb{N}$

$$\|P_k(r_t - \exp(-T_K t)r_0)\|^2_{L^2(X,\rho)} \leq \left[\frac{1-\exp(-\sigma_k t)}{\sigma_k}\right]^2 \cdot \left[4\,\|f^*\|^2_\infty \,\|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \epsilon\right]$$

and

$$\|r_t - \exp(-T_K t)r_0\|^2_{L^2(X,\rho)} \leq t^2 \cdot \left[4\,\|f^*\|^2_\infty \,\|K - K_0\|^2_{L^2(X^2,\rho\otimes\rho)} + \epsilon\right].$$

Interpretation

- Theorem 1 compares the dynamics of

$r_t(x) := f(x;\theta_t) - f^*(x)$	$\exp(-T_K t) r_0$
finite-width model trained on finitely many samples	idealized kernel method with infinite data

Interpretation

- Theorem 1 compares the dynamics of

$r_t(x) := f(x;\theta_t) - f^*(x)$	$\exp(-T_K t)r_0$
finite-width model trained on finitely many samples	idealized kernel method with infinite data

-

  $\exp(-T_K t)r_0$ learns projection along ϕ_i linearly at rate σ_i, by (4),

$$\langle r_t, \phi_i \rangle_\rho = \exp(-\sigma_i t)\langle r_0, \phi_i \rangle_\rho.$$

  Whenever the NTK at initialization K_0 concentrates around K, the residual r_t will inherit this bias of the kernel dynamics.

Interpretation

- Theorem 1 compares the dynamics of

$r_t(x) := f(x;\theta_t) - f^*(x)$	$\exp(-T_K t) r_0$
finite-width model trained on finitely many samples	idealized kernel method with infinite data

-

  $\exp(-T_K t) r_0$ learns projection along ϕ_i linearly at rate σ_i, by (4),

  $$\langle r_t, \phi_i \rangle_\rho = \exp(-\sigma_i t) \langle r_0, \phi_i \rangle_\rho.$$

  Whenever the NTK at initialization K_0 concentrates around K, the residual r_t will inherit this bias of the kernel dynamics.

- Furthermore, the bound for the projected differences is smaller when σ_k is larger. Therefore the bias appears more pronounced along eigendirections with large eigenvalues.

Consequences for the special case $K = K^\infty$

In infinite width limit, K_0 approaches K^∞ for general architectures[12]
The typical rate is $|K_0(x,x') - K^\infty(x,x')| = \tilde{O}(1/\sqrt{m})$ whp[13][14], so

Assumption 2

$m = \tilde{\Omega}(\epsilon^{-2})$ suffices to ensure that $\|K_0 - K^\infty\|^2_{L^2(X\times X,\rho\otimes\rho)} \le \epsilon$ holds whp $1-\delta(m)$ over the initialization θ_0, where $\delta(m) = o(1)$.

---

[12] Yang 2020.

[13] Du et al. 2019; Du et al. 2018; Huang and Yau 2020, for fixed x, x'.

[14] Bowman and Montúfar 2022a; Buchanan, Gilboa, and Wright 2021, uniformly over x, x'.

Consequences for the special case $K = K^\infty$

Corollary 2

Under Assumption 2, setting $K = K^\infty$, we have $m = \tilde{\Omega}(T^4/\epsilon^2)$ and $n = \tilde{\Omega}(T^2/\epsilon^2)$ suffices to ensure with probability $1 - O(mn)\exp(-\Omega(\log^2 m) - \delta(m)$ that for all $t \leq T$ and $k \in \mathbb{N}$

$$\|P_k(r_t - \exp(-T_{K^\infty} t) r_0)\|^2_{L^2(X,\rho)} \leq \left[\frac{1 - \exp(-\sigma_k t)}{\sigma_k}\right]^2 \cdot \epsilon$$

and

$$\|r_t - \exp(-T_{K^\infty} t) r_0\|^2_{L^2(X,\rho)} \leq t^2 \cdot \epsilon.$$

Consequences for the special case $K = K^\infty$

- Corollary 2 states that up to time T, $r_t \approx \exp(-T_{K^\infty} t) r_0$
- Given that K^∞ tends to have a highly skewed spectrum, the magnitude of σ_i is particularly relevant on the convergence rate
- The bound on projected difference is smaller when σ_k is large. Thus bias along top eigenfunctions is particularly pronounced

Observation 3

At the beginning of training the network learns projections along eigenfunctions of NTK integral operator T_{K^∞} at rates given by the eigenvalues; particularly so for eigenfcts with large eigenvalues.

Scaling wrt width and number of training samples

- As long as $n \leq m^{\alpha}$ for some $\alpha > 0$ the failure probability $O(mn)\exp(-\Omega(\log^2 m))$ goes to zero as $m \to \infty$.

  Thus once m and n are sufficiently large relative to T and ϵ, they can tend to infinity at any rate to achieve a high prob bound.
- m and n both have the same scaling $\tilde{\Omega}(\epsilon^{-2})$ with respect to ϵ

  Thus for fixed T we can send m, n to infinity at rate $m \sim n$ to get error $\epsilon \to 0$. This is significant as typical NTK analysis requires $m = \Omega(poly(n))$.

Observation 4

The network inherits the bias of the kernel at the beginning of training even when width m only grows linearly with the sample n.

Scaling with respect to stopping time

As $t \geq \log(\frac{\|f^*\|_{L^\infty(X,\rho)}}{\epsilon})\frac{1}{\sigma_k}$ suffices for $\|P_k \exp(-T_{K^\infty} t) r_0\|_{L^2(X,\rho)} \leq \epsilon$,

Corollary 5

Under Assumption 2, $T = \tilde{\Omega}(1/\sigma_k)$ and $\epsilon > 0$, we have that $m = \tilde{\Omega}({\sigma_k}^{-8}/\epsilon^2)$ and $n = \tilde{\Omega}(\sigma_k^{-6}/\epsilon^2)$ suffices to ensure that with probability at least $1 - O(mn)\exp(-\Omega(\log^2(m)) - \delta(m)$

$$\|P_k r_T\|^2_{L^2(X,\rho)} \leq \epsilon$$

and in particular

$$\frac{1}{2}\|r_T\|^2_{L^2(X,\rho)} \leq \tilde{O}(\epsilon) + \|(I - P_k) r_0\|^2_{L^2(X,\rho)} .$$

Scaling with respect to stopping time

- Corollary 5 says $T = \tilde{\Omega}(1/\sigma_k)$ is long enough to ensure that the network has learned the top k eigenfunctions to ϵ accuracy provided that $m = \tilde{\Omega}(\sigma_k^{-8}\epsilon^{-2})$ and $n = \tilde{\Omega}(\sigma_k^{-6}\epsilon^{-2})$.
- We also have a bound on the test error $\frac{1}{2}\|r_t\|^2_{L^2(X,\rho)}$.
  From ASI, $\|(I-P_k)r_0\|^2_{L^2(X,\rho)} = \|(I-P_k)f^*\|^2_{L^2(X,\rho)}$.
  For general f^*, this can decay arbitrary slowly wrt k.

---

[15] One can show $\|\exp(-T_{K^\infty} t)r_0\|^2_{L^2(X,\rho)} = O(\frac{\|f^*\|^2_{\mathcal{H}}}{t})$

[16] Velikanov and Yarotsky 2021, $\|\exp(-T_{K^\infty} t)r_0\|^2_{L^2(X,\rho)} \sim Ct^{-\xi}$.

Scaling with respect to stopping time

- Corollary 5 says $T = \tilde{\Omega}(1/\sigma_k)$ is long enough to ensure that the network has learned the top k eigenfunctions to ϵ accuracy provided that $m = \tilde{\Omega}(\sigma_k^{-8}\epsilon^{-2})$ and $n = \tilde{\Omega}(\sigma_k^{-6}\epsilon^{-2})$.
- We also have a bound on the test error $\frac{1}{2}\|r_t\|^2_{L^2(X,\rho)}$. From ASI, $\|(I - P_k)r_0\|^2_{L^2(X,\rho)} = \|(I - P_k)f^*\|^2_{L^2(X,\rho)}$. For general f^*, this can decay arbitrary slowly wrt k.

To get a learning guarantee:

- When f^* is in the RKHS of K^∞, one can[15] choose $T \sim \epsilon^{-1}$ to bring the test error to ϵ provided $m, n = \tilde{\Omega}(poly(\epsilon^{-1}))$.
- One can identify cases where a power law holds[16]. Then choose $T \sim \epsilon^{-1/\xi}$ to get a guarantee provided $m, n = \tilde{\Omega}(poly(\epsilon^{-1}))$.

---

[15] One can show $\|\exp(-T_{K^\infty} t)r_0\|^2_{L^2(X,\rho)} = O(\frac{\|f^*\|^2_{\mathcal{H}}}{t})$

[16] Velikanov and Yarotsky 2021, $\|\exp(-T_{K^\infty} t)r_0\|^2_{L^2(X,\rho)} \sim Ct^{-\xi}$.

- Linearization: There are results[17] which compare $f(x;\theta)$ to its linearization $f_{lin}(x;\theta) := \langle \nabla_\theta f(x;\theta_0), \theta - \theta_0 \rangle + f(x;\theta_0)$ in the regime $m = \Omega(poly(n))$, in which case the loss converges to zero and the parameter changes $\|\theta_t - \theta_0\|_2$ are bounded.

  By contrast we avoid $m = \Omega(poly(n))$ by using a stopping time.

[17] Arora et al. 2019b; Jin and Montúfar 2023; Lee et al. 2019.

Comparison to other works

- Spectral bias on empirical: There are results[17] similar to Th 1 and Cor 2 but which roughly replace T_{K^∞} with Gram matrix on training data $(G^\infty)_{i,j} = K^\infty(x_i, x_j)$ and ρ with $\hat{\rho} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$.

  Arora et al. 2019a; Basri et al. 2020 operate in the regime $m = \Omega(poly(n))$ and as a benefit do not need a stopping time.

  Cao et al. 2021 instead requires $m = \Omega(\max\{\sigma_k^{-14}, \epsilon^{-6}\})$ where σ_k is the cutoff eigenvalue.

[17] Arora et al. 2019a; Basri et al. 2020; Cao et al. 2021.

Comparison to other works

- Underparameterized Bowman and Montúfar 2022a obtained a version of Cor 2 for an underparameterized shallow net. They require $m = \tilde{\Omega}(\epsilon^{-1} T^2)$ and $n = \tilde{\Omega}(\epsilon^{-1} p T^2)$ and thus $n \gg p$.

  We removed the dependence of n on p and demonstrated the result for general deep architectures at the expense of slightly worse scaling with respect to T and ϵ.

Summary

- Quantitative bounds on the L^2 difference in function space between a finite-width network trained on finite samples and the corresponding kernel method with infinite width and data.
- The network inherits the bias of the kernel at the beginning of training even when the width scales linearly with the sample size.
- Bias is not only over training data but over entire input space.

Interesting future work:

- Investigate if flat minima manifesting a low-effective-rank FIM after training can be related to the behavior of the network on out-of-sample data after training.

References I

Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song (2019). "A Convergence Theory for Deep Learning via Over-Parameterization". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 242–252. URL: https://proceedings.mlr.press/v97/allen-zhu19a.html.

Arora, Sanjeev et al. (2019a). "Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 322–332. URL: https://proceedings.mlr.press/v97/arora19a.html.

References II

Arora, Sanjeev et al. (2019b). "On Exact Computation with an Infinitely Wide Neural Net". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/dbc4d84bfcfe2284ba11beffb853a8c4-Paper.pdf.

Basri, Ronen et al. (2019). "The Convergence Rate of Neural Networks for Learned Functions of Different Frequencies". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/5ac8bb8a7d745102a978c5f8ccdb61b8-Paper.pdf.

Basri, Ronen et al. (2020). "Frequency Bias in Neural Networks for Input of Non-Uniform Density". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 685–694. URL: https://proceedings.mlr.press/v119/basri20a.html.

- Bowman, Benjamin and Guido Montúfar (2022a). "Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=VLgmhQDVBV.
- — (2022b). "Spectral Bias Outside the Training Set for Deep Networks in the Kernel Regime". In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., pp. 30362–30377. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c4006ff54a7bbda74c09bad6f7586f5b-Paper-Conference.pdf.
- Buchanan, Sam, Dar Gilboa, and John Wright (2021). "Deep Networks and the Multiple Manifold Problem". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=O-6Pm_d_Q-.

References IV

Cao, Yuan et al. (Aug. 2021). "Towards Understanding the Spectral Bias of Deep Learning". In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Main Track. International Joint Conferences on Artificial Intelligence Organization, pp. 2205–2211. DOI: 10.24963/ijcai.2021/304. URL: https://doi.org/10.24963/ijcai.2021/304.

Du, Simon et al. (2019). "Gradient Descent Finds Global Minima of Deep Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 1675–1685. URL: https://proceedings.mlr.press/v97/du19c.html.

Du, Simon S et al. (2018). "Gradient descent provably optimizes over-parameterized neural networks". In: *arXiv preprint arXiv:1810.02054*.

References V

E, Weinan, Chao Ma, and Lei Wu (2020). "A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics". In: *Science China Mathematics* 63, pp. 1235–1258.

Fan, Zhou and Zhichao Wang (2020). "Spectra of the Conjugate Kernel and Neural Tangent Kernel for Linear-Width Neural Networks". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546.

Huang, Jiaoyang and Horng-Tzer Yau (2020). "Dynamics of Deep Neural Networks and Neural Tangent Hierarchy". In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 4542–4551. URL: https://proceedings.mlr.press/v119/huang20l.html.

Jacot, Arthur, Franck Gabriel, and Clement Hongler (2018). "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Ji, Ziwei and Matus Telgarsky (2020). "Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HygegyrYwH.

Jin, Hui and Guido Montúfar (2023). "Implicit Bias of Gradient Descent for Mean Squared Error Regression with Two-Layer Wide Neural Networks". In: *Journal of Machine Learning Research* 24.137, pp. 1–97. URL: http://jmlr.org/papers/v24/21-0832.html.

Karakida, Ryo, Shotaro Akaho, and Shun-ichi Amari (July 2021). "Pathological Spectra of the Fisher Information Metric and Its Variants in Deep Neural Networks". In: *Neural Computation* 33.8. _eprint: https://direct.mit.edu/neco/article-pdf/33/8/2274/1930880/neco_a_01411.pdf, pp. 2274–2307. ISSN: 0899-7667. DOI: 10.1162/neco_a_01411. URL: https://doi.org/10.1162/neco_a_01411.

Lee, Jaehoon et al. (2019). "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf.

Li, Mingchen, Mahdi Soltanolkotabi, and Samet Oymak (2020). "Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 4313–4324. URL: https://proceedings.mlr.press/v108/li20j.html.

Luo, Tao et al. (2022). "On the Exact Computation of Linear Frequency Principle Dynamics and Its Generalization". In: *SIAM Journal on Mathematics of Data Science* 4.4, pp. 1272–1292. DOI: 10.1137/21M1444400. eprint: https://doi.org/10.1137/21M1444400. URL: https://doi.org/10.1137/21M1444400.

Murray, Michael et al. (2023). "Characterizing the spectrum of the NTK via a power series expansion". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Tvms8xrZHyR.

Nguyen, Quynh (2021). "On the Proof of Global Convergence of Gradient Descent for Deep ReLU Networks with Linear Widths". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8056–8062. URL: https://proceedings.mlr.press/v139/nguyen21a.html.

Nguyen, Quynh and Marco Mondelli (2020). "Global Convergence of Deep Networks with One Wide Layer Followed by Pyramidal Topology". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 11961–11972. URL: https://proceedings.neurips.cc/paper/2020/file/8abfe8ac9ec214d68541fcb888c0b4c3-Paper.pdf.

Oymak, Samet and Mahdi Soltanolkotabi (2020). "Toward Moderate Overparameterization: Global Convergence Guarantees for Training Shallow Neural Networks". In: *IEEE Journal on Selected Areas in Information Theory* 1.1, pp. 84–105.

Oymak, Samet et al. (2020). *Generalization Guarantees for Neural Networks via Harnessing the Low-rank Structure of the Jacobian*. URL: https://openreview.net/forum?id=ryl5CJSFPS.

Pennington, Jeffrey and Yasaman Bahri (2017). "Geometry of Neural Network Loss Surfaces via Random Matrix Theory". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 2798–2806. URL: https://proceedings.mlr.press/v70/pennington17a.html.

Pennington, Jeffrey and Pratik Worah (2018). "The Spectrum of the Fisher Information Matrix of a Single-Hidden-Layer Neural Network". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/18bb68e2b38e4a8ce7cf4f6b2625768c-Paper.pdf.

Rahaman, Nasim et al. (2019). “On the Spectral Bias of Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, pp. 5301–5310. URL: http://proceedings.mlr.press/v97/rahaman19a.html.

Su, Lili and Pengkun Yang (2019). “On Learning Over-parameterized Neural Networks: A Functional Approximation Perspective”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/253f7b5d921338af34da817c00f42753-Paper.pdf.

Velikanov, Maksim and Dmitry Yarotsky (2021). "Explicit loss asymptotics in the gradient descent training of neural networks". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 2570–2582. URL: https://proceedings.neurips.cc/paper/2021/hash/14faf969228fc18fcd4fcf59437b0c97-Abstract.html.

Williams, Francis et al. (2019). "Gradient dynamics of shallow univariate relu networks". In: *Advances in Neural Information Processing Systems*, pp. 8378–8387.

Yang, Greg (2020). *Tensor Programs II: Neural Tangent Kernel for Any Architecture*. DOI: 10.48550/ARXIV.2006.14548. URL: https://arxiv.org/abs/2006.14548.

Yang, Greg and Hadi Salman (2019). *A Fine-Grained Spectral Perspective on Neural Networks*. DOI: 10.48550/ARXIV.1907.10599. URL: https://arxiv.org/abs/1907.10599.

Zhang, Yaoyu et al. (2020). "A type of generalization error induced by initialization in deep neural networks". In: *Proceedings of The First Mathematical and Scientific Machine Learning Conference*. Vol. 107. Proceedings of Machine Learning Research. PMLR, pp. 144–164. URL: https://proceedings.mlr.press/v107/zhang20a.html.

Zou, Difan and Quanquan Gu (2019). "An Improved Analysis of Training Over-parameterized Deep Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/6a61d423d02a1c56250dc23ae7ff12f3-Paper.pdf.

Zou, Difan et al. (2020). "Gradient descent optimizes over-parameterized deep ReLU networks". In: *Machine learning* 109, 467–492.