Deep Equals Shallow for ReLU Networks in Kernel Regimes

Alberto Bietti (NYU) Francis Bach (Inria)

ICLR 2021

Depth in neural networks

Deep learning practice

- Deeper is better!
- Folklore: key ingredient for the success of deep learning

Depth in neural networks

Deep learning practice

- Deeper is better!
- Folklore: key ingredient for the success of deep learning

Theory: depth separation

- Deeper is better! (*e.g.*, Eldan and Shamir, 2016; Telgarsky, 2016; Yarotsky, 2017)
- Deep nets can approximate certain functions f^* more efficiently than shallow nets
- **But**: **no known algorithms** to learn f^* from data

Depth in neural networks

Deep learning practice

- Deeper is better!
- Folklore: key ingredient for the success of deep learning

Theory: depth separation

- Deeper is better! (*e.g.*, Eldan and Shamir, 2016; Telgarsky, 2016; Yarotsky, 2017)
- Deep nets can approximate certain functions f^* more efficiently than shallow nets
- **But**: **no known algorithms** to learn f^* from data

Theory: approximation + optimization algorithms

- *"Kernel" regime*: **tractable** even for deep networks
- **This work**: role of depth in kernel regimes?

Kernel regimes for over-parameterized networks

Lazy training / kernel regime (Chizat et al., 2019; Jacot et al., 2018)

- θ stays close to initialization θ_0, model f_θ stays close to **linearized model**:

$$f_\theta(x) \approx f_{\theta_0}(x) + \langle \theta - \theta_0, \nabla_\theta f_\theta(x)|_{\theta=\theta_0} \rangle$$

- Optimization with width $m \to \infty$ $\approx$ kernel ridge regression with **neural tangent kernel**:

$$K_{NTK}(x, x') = \lim_{m \to \infty} \langle \nabla_\theta f_{\theta_0}(x), \nabla_\theta f_{\theta_0}(x') \rangle$$

Kernel regimes for over-parameterized networks

Lazy training / kernel regime (Chizat et al., 2019; Jacot et al., 2018)

- θ stays close to initialization θ_0, model f_θ stays close to **linearized model**:

$$f_\theta(x) \approx f_{\theta_0}(x) + \langle \theta - \theta_0, \nabla_\theta f_\theta(x)|_{\theta=\theta_0} \rangle$$

- Optimization with width $m \to \infty \approx$ kernel ridge regression with **neural tangent kernel**:

$$K_{NTK}(x, x') = \lim_{m\to\infty} \langle \nabla_\theta f_{\theta_0}(x), \nabla_\theta f_{\theta_0}(x') \rangle$$

Random features (Neal, 1996; Rahimi and Recht, 2007)

- Only train the last layer w of $f_{w,\theta_0}(x) = w^\top \phi_{\theta_0}(x)$

$$K_{RF}(x, x') = \lim_{m\to\infty} \langle \phi_{\theta_0}(x), \phi_{\theta_0}(x') \rangle$$

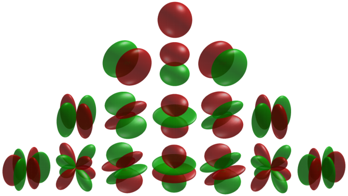
Approximation with dot-product kernels

Fully-connected networks $\implies$ dot-product kernels on the sphere

$$K(x, y) = \kappa(x^\top y), \quad x, y \in \mathbb{S}^{d-1}$$

Description of the RKHS (Mercer decomposition)

- **Rotation-invariant** kernel on the sphere
- $\Rightarrow$ RKHS description in the $L^2(\mathbb{S}^{d-1})$ basis of **spherical harmonics**



Approximation with dot-product kernels

Fully-connected networks $\Longrightarrow$ dot-product kernels on the sphere

$$K(x, y) = \kappa(x^\top y), \quad x, y \in \mathbb{S}^{d-1}$$

Description of the RKHS (Mercer decomposition)

- **Rotation-invariant** kernel on the sphere
- $\Rightarrow$ RKHS description in the $L^2(\mathbb{S}^{d-1})$ basis of **spherical harmonics**
- κ defines an integral operator on $L^2(\mathbb{S}^{d-1})$ with eigenvalues μ_k
- Decay of μ_k $\Leftrightarrow$ approximation properties in terms of regularity
- Slower decay $\Leftrightarrow$ "larger" RKHS

Kernels for deep ReLU networks

$$K(x, y) = \kappa(x^\top y), \quad x, y \in \mathbb{S}^{d-1}$$

- **Random features** (or NNGP/conjugate kernel)

$$\kappa_{RF}^{L}(u) = \underbrace{\kappa_1 \circ \cdots \circ \kappa_1}_{L-1 \text{ times}}(u)$$

- **Neural tangent kernel**

$$\kappa_{NTK}^{L}(u) = \kappa_{NTK}^{L-1}(u)\kappa_0(\kappa_{RF}^{L-1}(u)) + \kappa_{RF}^{L}(u)$$

(Cho and Saul, 2009; Daniely et al., 2016; Lee et al., 2018; Matthews et al., 2018; Jacot et al., 2018)

Kernels for deep ReLU networks

$$K(x, y) = \kappa(x^\top y), \quad x, y \in \mathbb{S}^{d-1}$$

- **Random features** (or NNGP/conjugate kernel)

$$\kappa_{RF}^{L}(u) = \underbrace{\kappa_1 \circ \cdots \circ \kappa_1}_{L-1 \text{ times}}(u)$$

- **Neural tangent kernel**

$$\kappa_{NTK}^{L}(u) = \kappa_{NTK}^{L-1}(u)\kappa_0(\kappa_{RF}^{L-1}(u)) + \kappa_{RF}^{L}(u)$$

- Decays of μ_k are known for two layers (Bach, 2017; Bietti and Mairal, 2019)

What about deep networks?

(Cho and Saul, 2009; Daniely et al., 2016; Lee et al., 2018; Matthews et al., 2018; Jacot et al., 2018)

Main result: deep equals shallow

Theorem (Eigenvalue decay from differentiability)

If κ has the following expansions for $t > 0$, with $\nu > 0$ and p_1, p_{-1} polynomials

$$\kappa(1-t) = p_1(t) + c_1 t^{\nu} + o(t^{\nu})$$
$$\kappa(-1+t) = p_{-1}(t) + c_{-1} t^{\nu} + o(t^{\nu}),$$

Then the eigenvalues μ_k decay as $k^{-d-2\nu+1}$.

Main result: deep equals shallow

Theorem (Eigenvalue decay from differentiability)

If κ has the following expansions for $t > 0$, with $\nu > 0$ and p_1, p_{-1} polynomials

$$\kappa(1 - t) = p_1(t) + c_1 t^{\nu} + o(t^{\nu})$$
$$\kappa(-1 + t) = p_{-1}(t) + c_{-1} t^{\nu} + o(t^{\nu}),$$

Then the eigenvalues μ_k decay as $k^{-d-2\nu+1}$.

Consequences

- For ReLU networks of any depth, $\nu = 3/2$ (RF) or $\nu = 1/2$ (NTK)

$\Rightarrow$ **same decay for deep and shallow networks**

Main result: deep equals shallow

Theorem (Eigenvalue decay from differentiability)

If κ has the following expansions for $t > 0$, with $\nu > 0$ and p_1, p_{-1} polynomials

$$\kappa(1-t) = p_1(t) + c_1 t^\nu + o(t^\nu)$$
$$\kappa(-1+t) = p_{-1}(t) + c_{-1} t^\nu + o(t^\nu),$$

Then the eigenvalues μ_k decay as $k^{-d-2\nu+1}$.

Consequences

- For ReLU networks of any depth, $\nu = 3/2$ (RF) or $\nu = 1/2$ (NTK)

  $\Rightarrow$ **same decay for deep and shallow networks**
- Precise decays for other kernels (Laplace, γ-exponential, RF with step activations)

Main result: deep equals shallow

Theorem (Eigenvalue decay from differentiability)

If κ has the following expansions for $t > 0$, with $\nu > 0$ and p_1, p_{-1} polynomials

$$\kappa(1-t) = p_1(t) + c_1 t^{\nu} + o(t^{\nu})$$
$$\kappa(-1+t) = p_{-1}(t) + c_{-1} t^{\nu} + o(t^{\nu}),$$

Then the eigenvalues μ_k decay as $k^{-d-2\nu+1}$.

Consequences

- For ReLU networks of any depth, $\nu = 3/2$ (RF) or $\nu = 1/2$ (NTK)

  $\Rightarrow$ **same decay for deep and shallow networks**
- Precise decays for other kernels (Laplace, γ-exponential, RF with step activations)
- Benefits of depth may arise from: (i) architecture; (ii) non-kernel regimes

Main result: deep equals shallow

Theorem (Eigenvalue decay from differentiability)

If κ has the following expansions for $t > 0$, with $\nu > 0$ and p_1, p_{-1} polynomials

$$\kappa(1-t) = p_1(t) + c_1 t^{\nu} + o(t^{\nu})$$
$$\kappa(-1+t) = p_{-1}(t) + c_{-1} t^{\nu} + o(t^{\nu}),$$

Then the eigenvalues μ_k decay as $k^{-d-2\nu+1}$.

Consequences

- For ReLU networks of any depth, $\nu = 3/2$ (RF) or $\nu = 1/2$ (NTK)

  $\Rightarrow$ **same decay for deep and shallow networks**
- Precise decays for other kernels (Laplace, γ-exponential, RF with step activations)
- Benefits of depth may arise from: (i) architecture; (ii) non-kernel regimes

Thanks!

References I

F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research (JMLR)*, 18(1):629–681, 2017.

A. Bietti and J. Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

A. Daniely, R. Frostig, and Y. Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *Conference on Learning Theory (COLT)*, 2016.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

References II

J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

A. Matthews, M. Rowland, J. Hron, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

R. M. Neal. *Bayesian learning for neural networks*. Springer, 1996.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

M. Telgarsky. Benefits of depth in neural networks. In *Conference on Learning Theory (COLT)*, 2016.

D. Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.