Neural Tangent Kernel Beyond the Infinite-Width Limit: Effects of Depth and Initialization

Mariia Seleznova & Gitta Kutyniok

(Ludwig-Maximilians-Universität München)

ICML 2022

Jul 17th-23rd, 2022

Neural Tangent Kernel

Consider a neural network (NN) f trained on dataset $\mathcal{D}$ by gradient flow:

$$\dot{\mathbf{w}}^{(t)} = -\nabla_{\mathbf{w}}\mathcal{L}(\mathcal{D}) = -\sum_{(x_i,y_i)\in\mathcal{D}} \nabla_{\mathbf{w}} f(x_i)\frac{\partial\mathcal{L}(\mathcal{D})}{\partial f(x_i)},$$

where $\mathbf{w}$ is the vector of all the trainable parameters and $\mathcal{L}$ is the loss function.

Neural Tangent Kernel

Consider a neural network (NN) f trained on dataset $\mathcal{D}$ by gradient flow:

$$\dot{\mathbf{w}}^{(t)} = -\nabla_{\mathbf{w}}\mathcal{L}(\mathcal{D}) = -\sum_{(x_i,y_i)\in\mathcal{D}} \nabla_{\mathbf{w}} f(x_i) \frac{\partial \mathcal{L}(\mathcal{D})}{\partial f(x_i)},$$

where $\mathbf{w}$ is the vector of all the trainable parameters and $\mathcal{L}$ is the loss function. Then the dynamics of f is given by:

$$\dot{f}^{(t)}(x) = \nabla_{\mathbf{w}} f(x_i) \cdot \dot{\mathbf{w}}^{(t)} = -\sum_{(x_i,y_i)\in\mathcal{D}} \Theta(x, x_i) \frac{\partial \mathcal{L}(\mathcal{D})}{\partial f(x_i)}$$

Definition: *Neural tangent kernel (NTK)* of a NN with output function $f(\cdot)$ and trainable parameters $\mathbf{w}$ is given by

$$\Theta(x_i, x_j) := \nabla_{\mathbf{w}} f(x_i)^T \nabla_{\mathbf{w}} f(x_j), \quad x_i, x_j \in \mathcal{X}.$$

$\rightsquigarrow$ *The NTK captures the first-order approximation of NN's training!*

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Neural Tangent Kernel

Assume a NN $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ has depth L and layer widths $n_0, \ldots, n_L$.

In the infinite-width limit $n_\ell \to \infty, 1 \leq \ell < L$ [Jacot et al., 2018]:

- NTK is deterministic under random initialization:

$$\Theta^{(0)}(x_i, x_j) \to \mathbb{E}_{\mathbf{w}}[\Theta^{(0)}(x_i, x_j)] = \Theta^*(x_i, x_j),$$

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Neural Tangent Kernel

Assume a NN $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ has depth L and layer widths $n_0, \ldots, n_L$.

In the infinite-width limit $n_\ell \to \infty, 1 \leq \ell < L$ [Jacot et al., 2018]:

- NTK is deterministic under random initialization:

$$\Theta^{(0)}(x_i, x_j) \to \mathbb{E}_{\mathbf{w}}[\Theta^{(0)}(x_i, x_j)] = \Theta^*(x_i, x_j),$$

- NTK stays constant during training:

$$\Theta^{(t)}(x_i, x_j) \to \Theta^*(x_i, x_j).$$

Neural Tangent Kernel

Assume a NN $f : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ has depth L and layer widths $n_0, \ldots, n_L$.

In the infinite-width limit $n_\ell \to \infty, 1 \leq \ell < L$ [Jacot et al., 2018]:

- NTK is deterministic under random initialization:

$$\Theta^{(0)}(x_i, x_j) \to \mathbb{E}_{\mathbf{w}}[\Theta^{(0)}(x_i, x_j)] = \Theta^*(x_i, x_j),$$

- NTK stays constant during training:

$$\Theta^{(t)}(x_i, x_j) \to \Theta^*(x_i, x_j).$$

Thus, NNs dynamics is governed by a constant deterministic kernel in the infinite-width limit.

$\rightsquigarrow$ *Infinitely-wide NNs evolve as linear models with NTK kernel!*

Can we rely on the infinite-width limit?

- Infinite-width NTK is *label-agnostic* and *does not learn features*.
  $\Rightarrow$ cannot provide optimal representation system for a task.

Can we rely on the infinite-width limit?

- Infinite-width NTK is *label-agnostic* and *does not learn features*.
  $\Rightarrow$ cannot provide optimal representation system for a task.
- In the NTK limit, the *depth-to-width ratio* tends to zero:

$$L\text{– fixed}, \;\; n^{\ell} \to \infty \Rightarrow \frac{L}{n^{\ell}} \to 0, \quad 1 \leq \ell \leq L-1$$

  $\Rightarrow$ this limit only models shallow networks.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Can we rely on the infinite-width limit?

- Infinite-width NTK is *label-agnostic* and *does not learn features*.
  $\Rightarrow$ cannot provide optimal representation system for a task.
- In the NTK limit, the *depth-to-width ratio* tends to zero:

$$L\text{– fixed}, \;\; n^{\ell} \to \infty \Rightarrow \frac{L}{n^{\ell}} \to 0, \quad 1 \leq \ell \leq L-1$$

  $\Rightarrow$ this limit only models shallow networks.
- Infinite-width approximations often get worse as the *depth* increases [Li et al., 2021, Hanin and Nica, 2020, Hu and Huang, 2021].

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Can we rely on the infinite-width limit?

- Infinite-width NTK is *label-agnostic* and *does not learn features*.
  $\Rightarrow$ cannot provide optimal representation system for a task.
- In the NTK limit, the *depth-to-width ratio* tends to zero:
  $$L\text{– fixed}, \;\; n^{\ell} \to \infty \Rightarrow \frac{L}{n^{\ell}} \to 0, \;\;\; 1 \leq \ell \leq L-1$$
  $\Rightarrow$ this limit only models shallow networks.
- Infinite-width approximations often get worse as the *depth* increases [Li et al., 2021, Hanin and Nica, 2020, Hu and Huang, 2021].
- Empirical performance of the NTK and finite NNs differs [Fort et al., 2020, Lee et al., 2020].

Can we rely on the infinite-width limit?

- Infinite-width NTK is *label-agnostic* and *does not learn features*.
  $\Rightarrow$ cannot provide optimal representation system for a task.
- In the NTK limit, the *depth-to-width ratio* tends to zero:
  $$L\text{– fixed}, \ \ n^\ell \to \infty \Rightarrow \frac{L}{n^\ell} \to 0, \ \ 1 \leq \ell \leq L-1$$
  $\Rightarrow$ this limit only models shallow networks.
- Infinite-width approximations often get worse as the *depth* increases [Li et al., 2021, Hanin and Nica, 2020, Hu and Huang, 2021].
- Empirical performance of the NTK and finite NNs differs [Fort et al., 2020, Lee et al., 2020].

$\rightsquigarrow$ *It is not clear when the NTK regime explains NNs' behavior!*

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Our setting

We study the NTK of fully-connected ReLU NNs with:

- Comparable depth and width: $\frac{L}{n_\ell} =: \lambda_\ell > 0, \ 1 \leq \ell \leq L-1.$
- Initialization given by: $\mathbf{W}_{ij}^\ell \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{n_{\ell-1}}\right), \quad \mathbf{b}_i^\ell = 0.$

Our setting

We study the NTK of fully-connected ReLU NNs with:

- Comparable depth and width: $\frac{L}{n_\ell} =: \lambda_\ell > 0,\ 1 \leq \ell \leq L-1.$
- Initialization given by: $\mathbf{W}_{ij}^\ell \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{n_{\ell-1}}\right), \quad \mathbf{b}_i^\ell = 0.$

Phase transition at initialization [Poole et al., 2016]**:**

- *Chaotic phase:* If $\sigma_w^2 > 2$, gradients norm increases with depth.
- *Ordered phase:* If $\sigma_w^2 < 2$, gradients norm decreases.
- *«Edge of chaos» (EOC):* $\sigma_w^2 \approx 2$ allows deeper signal propagation.

Our setting

We study the NTK of fully-connected ReLU NNs with:

- Comparable depth and width: $\frac{L}{n_\ell} =: \lambda_\ell > 0,\ 1 \le \ell \le L-1.$
- Initialization given by: $\mathbf{W}_{ij}^\ell \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{n_{\ell-1}}\right), \quad \mathbf{b}_i^\ell = 0.$

Phase transition at initialization [Poole et al., 2016]**:**

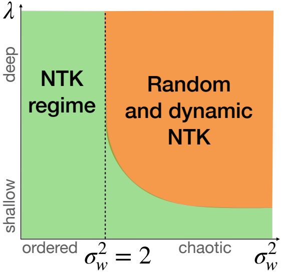
- *Chaotic phase:* If $\sigma_w^2 > 2$, gradients norm increases with depth.
- *Ordered phase:* If $\sigma_w^2 < 2$, gradients norm decreases.
- *«Edge of chaos» (EOC):* $\sigma_w^2 \approx 2$ allows deeper signal propagation.

Related work:

- [Hanin and Nica, 2020] showed that the NTK of ReLU NNs with $\lambda > 0$ is random and dynamic for $\sigma_w^2 = 2$ (EOC).
- [Xiao et al., 2020, Hayou et al., 2019] studied the effects of the phase transition on the infinite-width NTK.

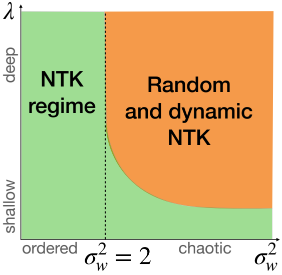
Contributions

- Show that properties of the NTK depend significantly on *depth-to-width* ratio λ and *initialization* variance σ_w^2.



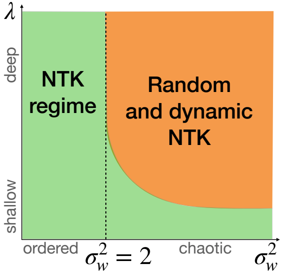
Contributions

- Show that properties of the NTK depend significantly on *depth-to-width* ratio λ and *initialization* variance σ_w^2.
- Namely, the NTK regime can approximate only wide and *shallow ReLU networks* ($\lambda \approx 0$) or *deep networks* ($\lambda \gg 0$) in the ordered phase.



Contributions

- Show that properties of the NTK depend significantly on *depth-to-width* ratio λ and *initialization* variance σ_w^2.
- Namely, the NTK regime can approximate only wide and *shallow ReLU networks* ($\lambda \approx 0$) or *deep networks* ($\lambda \gg 0$) in the ordered phase.

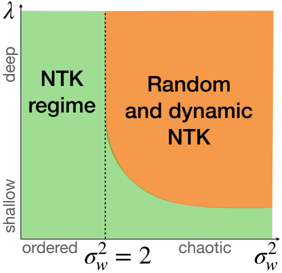


- Characterize the NTK variability in the *infinite-depth-and-width* limit in all three phases, as well as *finite-width* approximations.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Contributions

- Show that properties of the NTK depend significantly on *depth-to-width* ratio λ and *initialization* variance σ_w^2.
- Namely, the NTK regime can approximate only wide and *shallow ReLU networks* ($\lambda \approx 0$) or *deep networks* ($\lambda \gg 0$) in the ordered phase.

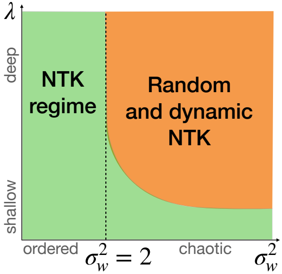


- Characterize the NTK variability in the *infinite-depth-and-width* limit in all three phases, as well as *finite-width* approximations.
- Study the first gradient descent step of the NTK in the infinite-depth-and-width limit.

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Contributions

- Show that properties of the NTK depend significantly on *depth-to-width* ratio λ and *initialization* variance σ_w^2.
- Namely, the NTK regime can approximate only wide and *shallow ReLU networks* ($\lambda \approx 0$) or *deep networks* ($\lambda \gg 0$) in the ordered phase.



- Characterize the NTK variability in the *infinite-depth-and-width* limit in all three phases, as well as *finite-width* approximations.
- Study the first gradient descent step of the NTK in the infinite-depth-and-width limit.
- Discuss *structure of the NTK matrix* and its training dynamics outside of the NTK regime.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Variability of the NTK at initialization

Theorem (Seleznova & Kutyniok, 2022)

For NNs of constant width M the following holds for the NTK dispersion:

1. *In the* **chaotic phase** *the NTK dispersion grows exponentially with λ:*

$$\frac{\mathbb{E}[\Theta^2(x,x)]}{\mathbb{E}^2[\Theta(x,x)]} \xrightarrow[\substack{M\to\infty, L\to\infty,\\ L/M\to\lambda\in\mathbb{R}}]{} \frac{1}{2\lambda}e^{5\lambda}\left(1-\frac{1}{4\lambda}(1-e^{-4\lambda})\right).$$

2. *At the* **EOC** *the NTK dispersion grows exponentially with a slower rate:*

$$\frac{\mathbb{E}[\Theta^2(x,x)]}{\mathbb{E}^2[\Theta(x,x)]} \to \frac{1}{(1+\alpha_0)^2}\left[\frac{1}{2\lambda}e^{5\lambda}\left(1-\frac{1}{4\lambda}(1-e^{-4\lambda})\right)+g(\lambda,\alpha_0)\right].$$

3. *In the* **ordered phase** *the variance is zero:* $\frac{\mathbb{E}[\Theta^2(x,x)]}{\mathbb{E}^2[\Theta(x,x)]} \xrightarrow[\substack{M\to\infty, L\to\infty,\\ L/M\to\lambda\in\mathbb{R}}]{} 1.$

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Variability of the NTK at initialization

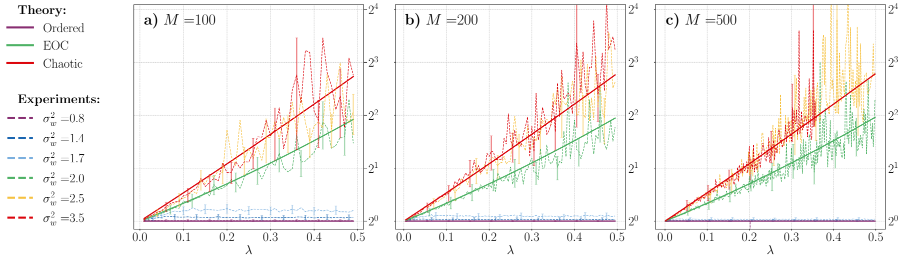


Figure: $\mathbb{E}[\Theta^2(x,x)]/\mathbb{E}^2[\Theta(x,x)]$ ratio for constant-width ReLU NNs.

$\rightsquigarrow$ *We can estimate the dispersion of a given NN!*

More results in the paper:

- Finite-width approximations of the NTK moments
- Changes of the NTK in the first GD step
- Bound on the dispersion of non-diagonal NTK elements
- ...

Thank you for your attention!

References I

Fort, S., Dziugaite, G. K., Paul, M., Kharaghani, S., Roy, D. M., and Ganguli, S. (2020).
Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel.
In *Advances in Neural Information Processing Systems*.

Hanin, B. and Nica, M. (2020).
Finite depth and width corrections to the neural tangent kernel.
In *International Conference on Learning Representations, ICLR*.

Hayou, S., Doucet, A., and Rousseau, J. (2019).
Mean-field behaviour of neural tangent kernel for deep neural networks.
CoRR, abs/1905.13654.

References II

Hu, Z. and Huang, H. (2021).
On the random conjugate kernel and neural tangent kernel.
In *International Conference on Machine Learning*, pages 4359–4368. PMLR.

Jacot, A., Hongler, C., and Gabriel, F. (2018).
Neural tangent kernel: Convergence and generalization in neural networks.
In *Advances in Neural Information Processing Systems*, pages 8580–8589.

Lee, J., Schoenholz, S. S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. (2020).
Finite versus infinite neural networks: an empirical study.
In *Advances in Neural Information Processing Systems*.

References III

Li, M. B., Nica, M., and Roy, D. M. (2021).
The future is log-gaussian: ResNets and their infinite-depth-and-width limit at initialization.
CoRR, abs/2106.04013.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. (2016).
Exponential expressivity in deep neural networks through transientchaos.
In *Advances in Neural Information Processing Systems*, pages 3360–3368.

Xiao, L., Pennington, J., and Schoenholz, S. (2020).
Disentangling trainability and generalization in deep neural networks.
In *International Conference on Machine Learning*. PMLR.