Neural Tangent Kernel Analysis of Deep Narrow Neural Networks

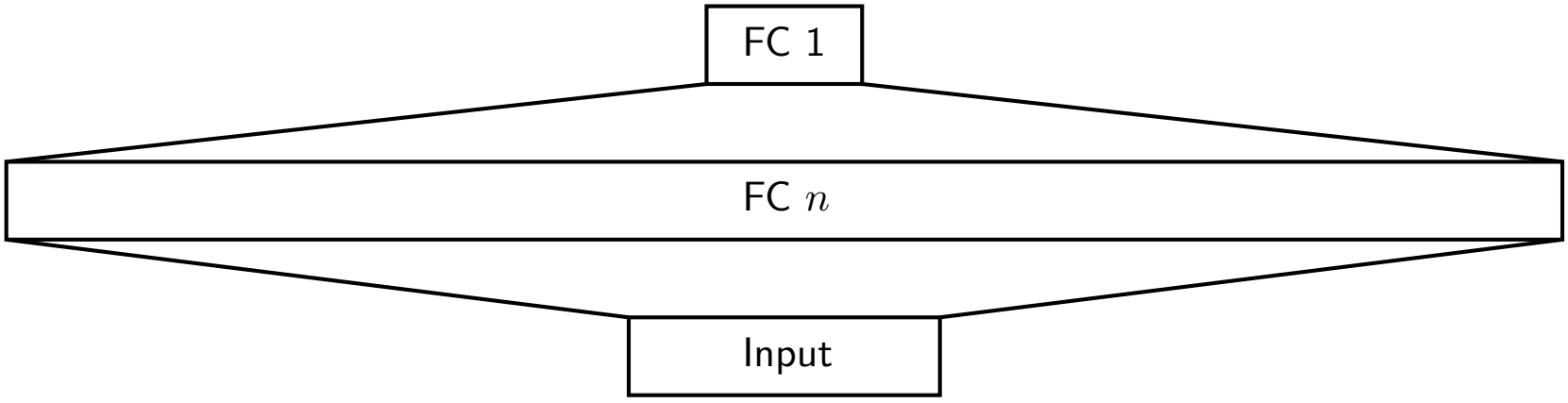
Jongmin Lee[1], Joo Young Choi[1], **Ernest K. Ryu**[1], and Albert No[2]

International Conference on Machine Learning, 2022

[1]Department of Mathematical Sciences, Seoul National University
[2]Department of Electronic and Electrical Engineering, Hongik University

Universal approximation theorem

Consider wide 2-layer neural networks.



Theorem (Informal)

Sufficiently wide 2-layer networks approximate any continuous function.

However, these are existence results. They says nothing about whether one can *learn* such approximations.

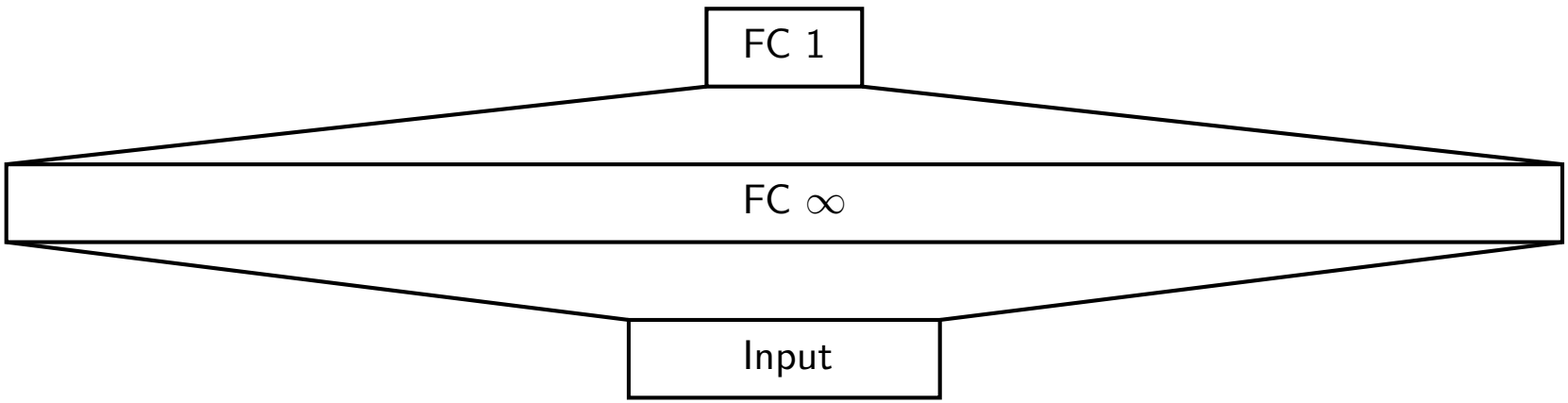
Cybenko, Approximation by superpositions of a sigmoidal function, *MCSS*, 1989.

Mean-field limit

Theorem (Informal)

The training dynamics of an infinitely wide 2-layer neural network is characterized by the solution of the PDE

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t \nabla \frac{\delta \mathcal{L}}{\delta \rho} \right)$$

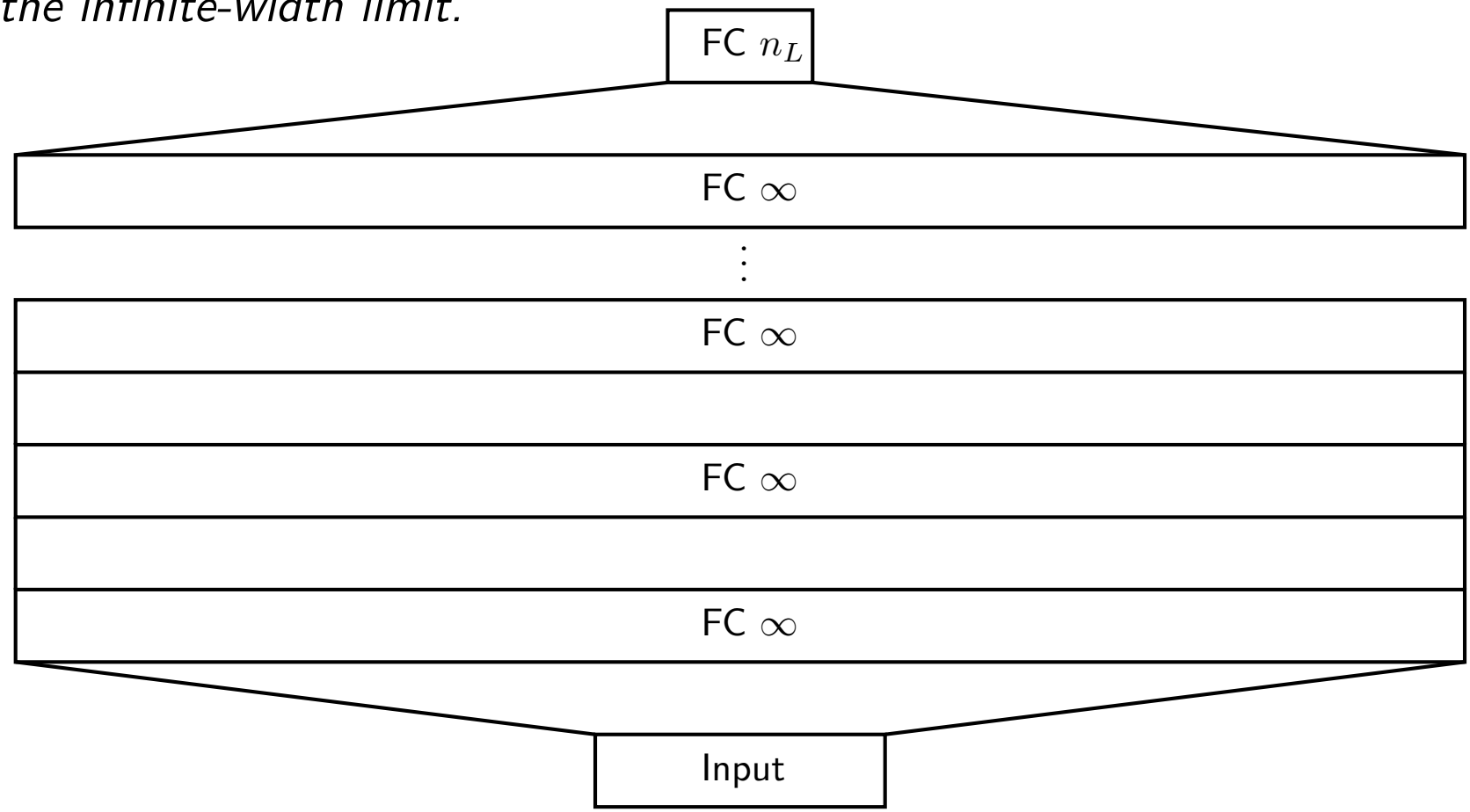


Concurrent works of [Chizat and Bach 2018], [Mei, Montanari, and Nguyen 2018], and [Rotskoff and Vanden-Eijnden 2018].

Neural tangent kernel

Theorem (Informal)

The training dynamics of a depth L neural network becomes "linear" in the infinite-width limit.



Jacot, Gabriel, and Hongler, Neural Tangent Kernel: Convergence and Generalization in Neural Networks, *NeurIPS*, 2018.

Deep narrow NN are universal approximators

Infinitely wide NN are provably trainable. What about deep NN?

Recently, universality results have been established for deep NN.

Theorem (Informal[3])

MLP with width $n_{\text{in}} + n_{\text{out}} + 1$ *and large depth can approximate any continuous function.*

However, these are existence results. They says nothing about whether one can *learn* such approximations.

[3] Kidger and Lyons, Universal approximation with deep narrow networks, *COLT*, 2020.

Goal: Trainability guarantees for deep neural networks

The tremendous recent progress does not sufficiently address the role of depth in deep learning.

We present the first trainability guarantee for infinitely deep but narrow neural networks using the NTK theory.

Trainability guarantee of deep narrow MLPs

Theorem (Main result, Informal)

Under a certain non-standard but implementable initialization,[4] *the training dynamics of the infinitely deep MLP with width* $n_{\text{in}} + n_{\text{out}} + 1$ *provably converges.*

Proof characterizes the NTK in the infinite-depth limit. Controlling the invariance of NTK is much more technical in the fininite-depth limit than in the infinite-width limit; analysis requries controlling an infinite product of matrices rather than an infinite sum of matrices.

[4] Construction inspired by:
Kidger and Lyons, Universal approximation with deep narrow networks, *COLT*, 2020.

Initialization scheme for MLP

We analyze MLP with depth $L \to \infty$ with a very particular initialization. Initialize weights as

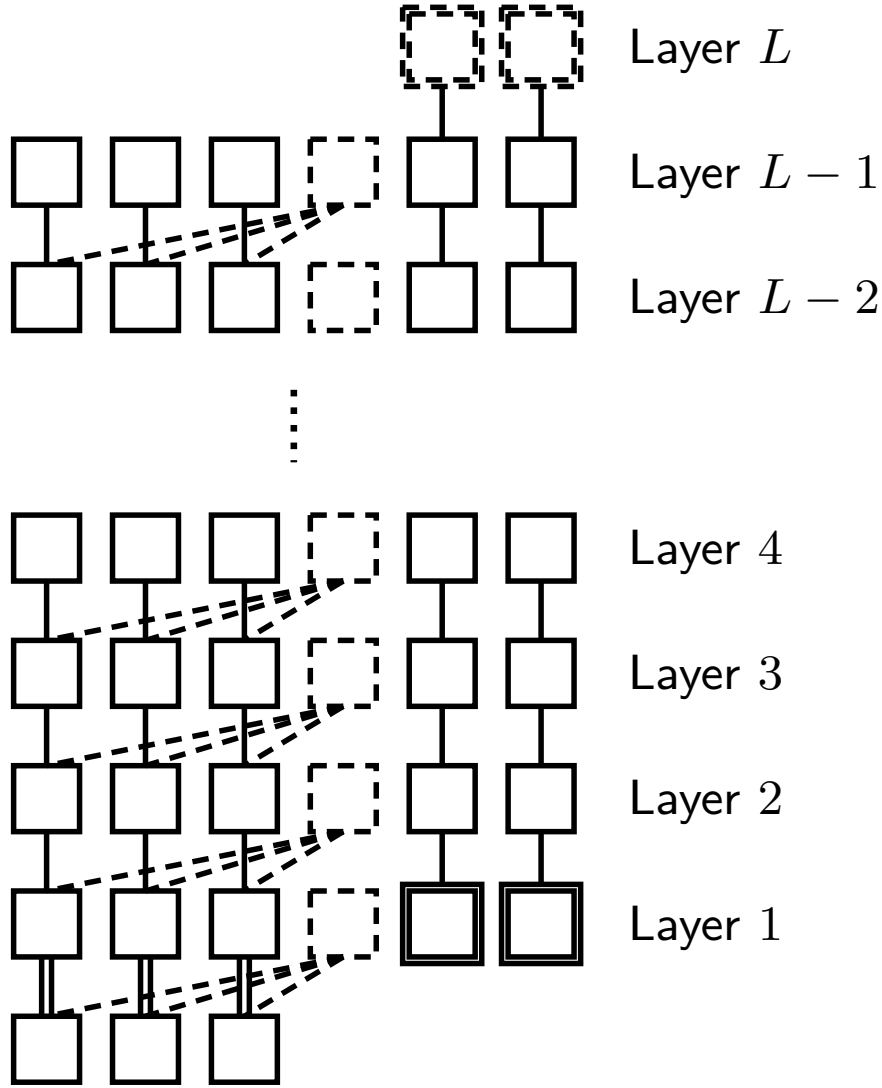
$$W^1 = \begin{bmatrix} C_L I_{d_{\text{in}}} \\ u^1 \\ 0_{d_{\text{out}} \times d_{\text{in}}} \end{bmatrix}, \qquad W^l = \begin{bmatrix} I_{d_{\text{in}}} & 0_{d_{\text{in}} \times 1} & 0_{d_{\text{in}} \times d_{\text{out}}} \\ u^l & 0_{1 \times 1} & 0_{1 \times d_{\text{out}}} \\ 0_{d_{\text{out}} \times d_{\text{in}}} & 0_{d_{\text{out}} \times 1} & I_{d_{\text{out}}} \end{bmatrix},$$

$$W^L = \begin{bmatrix} 0_{d_{\text{out}} \times d_{\text{in}}} & 0_{d_{\text{out}} \times 1} & I_{d_{\text{out}}} \end{bmatrix},$$

for $2 \le l \le L-1$, where $u_i^l \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{d_{\text{in}}}\rho^2)$ for $1 \le l \le L-1$ and $C_L > 0$ is a scalar growing sufficiently fast such that $L^2/C_L \to 0$

$$b^1 = \begin{bmatrix} 0_{d_{\text{in}} \times 1} \\ v^1 \\ C_L \mathbb{1}_{d_{\text{out}}} \end{bmatrix}, \quad b^l = \begin{bmatrix} 0_{d_{\text{in}} \times 1} \\ v^l \\ 0_{d_{\text{out}} \times 1} \end{bmatrix}, \quad b^L = \begin{bmatrix} -C_L \mathbb{1}_{d_{\text{out}}} \end{bmatrix}$$

for $2 \le l \le L-1$, where $v^l \overset{iid}{\sim} \mathcal{N}(0, C_L^2 \beta^2)$ for $1 \le l \le L-1$.



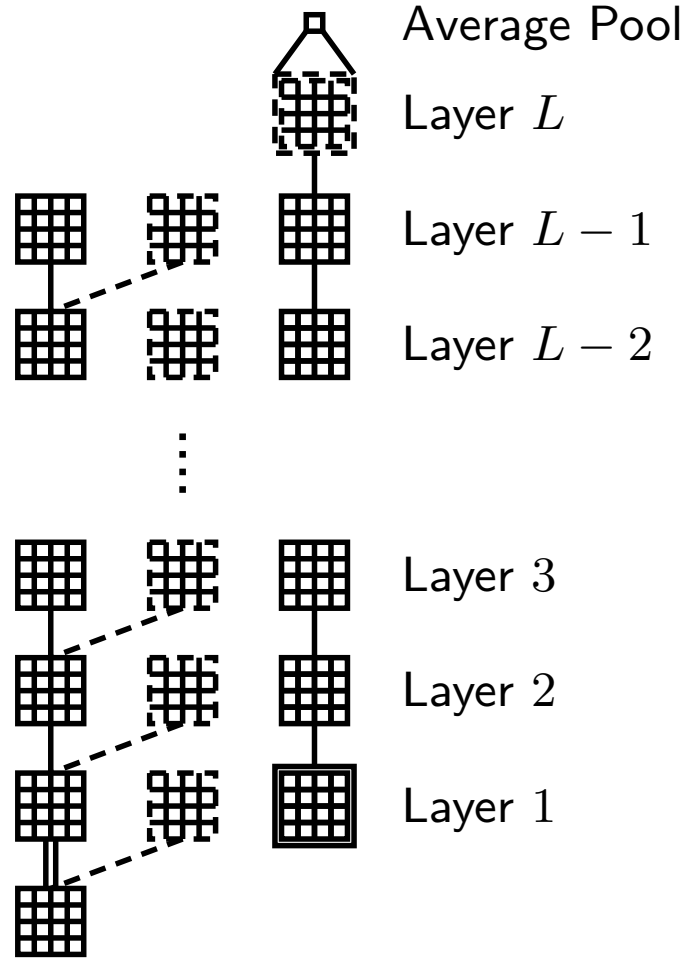
Initialization of deep MLP with $d_{in} = 3$ and $d_{out} = 2$. Intermediate layers have width $d_{in} + 1 + d_{out}$. Line styles indicate types of weight initializations (solid:1, double:C_L, dash:Gaussian, none:0). Box styles indicate types of bias initializations (solid:0, dash:Gaussian, double:C_L, double-dash:$-C_L$).

Initialization scheme for CNN

We analyze CNN with depth $L \to \infty$ with a very particular initialization.

$$
\begin{aligned}
w^1_{1,1,:,:} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & C_L & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad w^1_{2,1,:,:} = \begin{bmatrix} u^1_{1,1} & u^1_{1,2} & u^1_{1,3} \\ u^1_{2,1} & u^1_{2,2} & u^1_{2,3} \\ u^1_{3,1} & u^1_{3,2} & u^1_{3,3} \end{bmatrix}, \\
w^1_{3,1,:,:} &= 0_{3\times 3} \\
w^l_{1,:,:,:} &= \boldsymbol{\iota}_{3\times 3}, 0_{3\times 3}, 0_{3\times 3} \\
w^l_{2,:,:,:} &= \begin{bmatrix} u^l_{1,1} & u^l_{1,2} & u^l_{1,3} \\ u^l_{2,1} & u^l_{2,2} & u^l_{2,3} \\ u^l_{3,1} & u^l_{3,2} & u^l_{3,3} \end{bmatrix}, 0_{3\times 3}, 0_{3\times 3} \\
w^l_{3,:,:,:} &= 0_{3\times 3}, 0_{3\times 3}, \boldsymbol{\iota}_{3\times 3} \\
w^L_{1,:,:,:} &= 0_{3\times 3}, 0_{3\times 3}, \boldsymbol{\iota}_{3\times 3}
\end{aligned}
$$

for $2 \le l \le L-1$, where $u^l_{i,j} \overset{iid}{\sim} \mathcal{N}(0, \rho^2)$ for $1 \le l \le L-1$ and $C_L > 0$ is a scalar growing as a function of L at a rate satisfying $L^2/C_L \to 0$.



Initialization of deep CNN with 4×4 input. The 3 grids per row represent the 3 channels per layer, and the box at the top represents the scalar output of the final average pool. Line styles indicate types of weight initializations (solid:$\mathbf{diag}(0, 1, 0)$, double:$\mathbf{diag}(0, C_L, 0)$, dash:Gaussian, none:$0_{3\times 3}$). Box styles indicate types of bias initializations (solid:0, dash:Gaussian, double:C_L, double-dash:$-C_L$).

Trainability guarantee of deep narrow CNNs

Initialize the biases as follows:

$$(b_1^1, b_2^1, b_3^1) = (0, v^1, C_L)$$
$$(b_1^l, b_2^l, b_3^l) = (0, v^l, 0)$$
$$b^L = -C_L$$

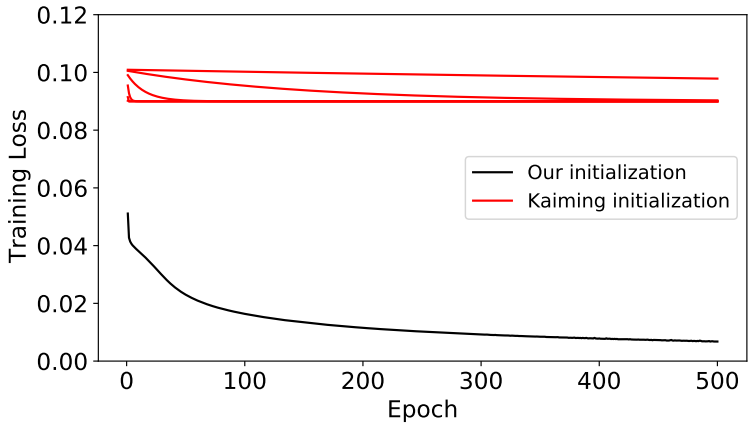
for $2 \leq l \leq L-1$, where $v^l \overset{iid}{\sim} \mathcal{N}(0, C_L^2 \beta^2)$ for $1 \leq l \leq L-1$.

Theorem (Main result, Informal)

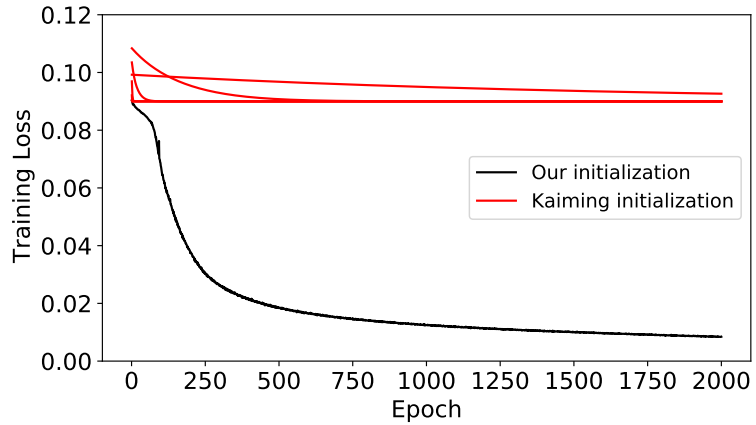
Under said initialization, the training dynamics of the infinitely deep CNN with width $n_{\text{in}} + n_{\text{out}} + 1$ *provably converges.*

Proof follows from analogous argument as the MLP case.

Experiments



(Left) Depth 1000 MLP



(Right) Depth 1000 CNN

Depth 1000 MLPs and CNNs with MNIST are trainable with our proposed initialization but not with the standard Kaiming He initialization. For Kaiming initialization, we show trials with learning rates 1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 0.01, 0.1, and 1.

Conclusion

Deep ReLU MLPs and CNNs with particular initializations (but standard architecture) are provably trainable in the infinite-depth limit.