Efficient Computation of Deep, Nonlinear, Infinite-Width Neural Networks that Learn Features

Greg Yang, Michael Santacroce, and Edward Hu

ICLR 2022

π-limit

- From μ–limit[1]: current inf-width limits **are not learning features**

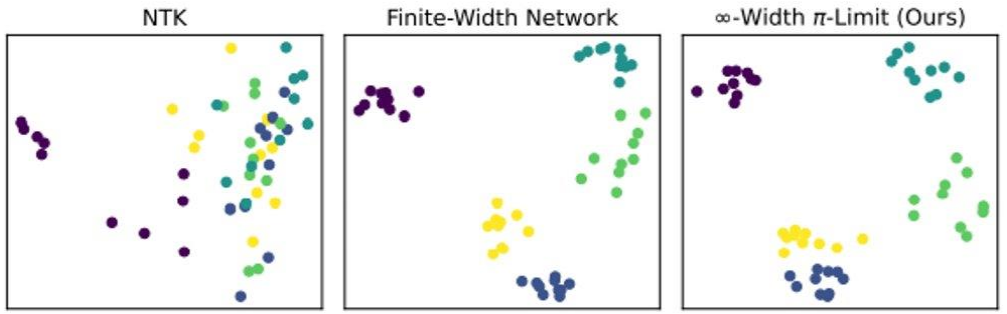


Figure 1: **PCA of representations of images from 5 classes (= 5 colors) in Omniglot test set.**

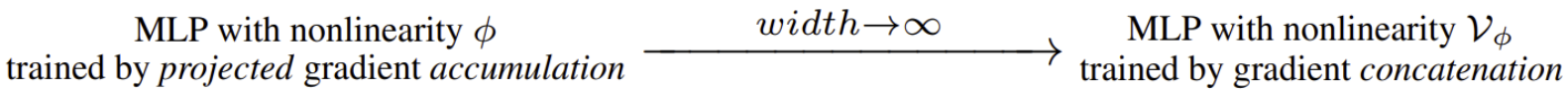
- **Goal**: computable, deep, nonlinear, feature-learning infinite-width limit

1. Feature Learning in Infinite-Width Neural Networks, *Greg Yang and Edward Hu*

π-limit

- **Why investigate feature learning?**
  - Practical finite networks depend on feature learning (i.e., language models)
  - Inf-width studies should build on limits that learn features

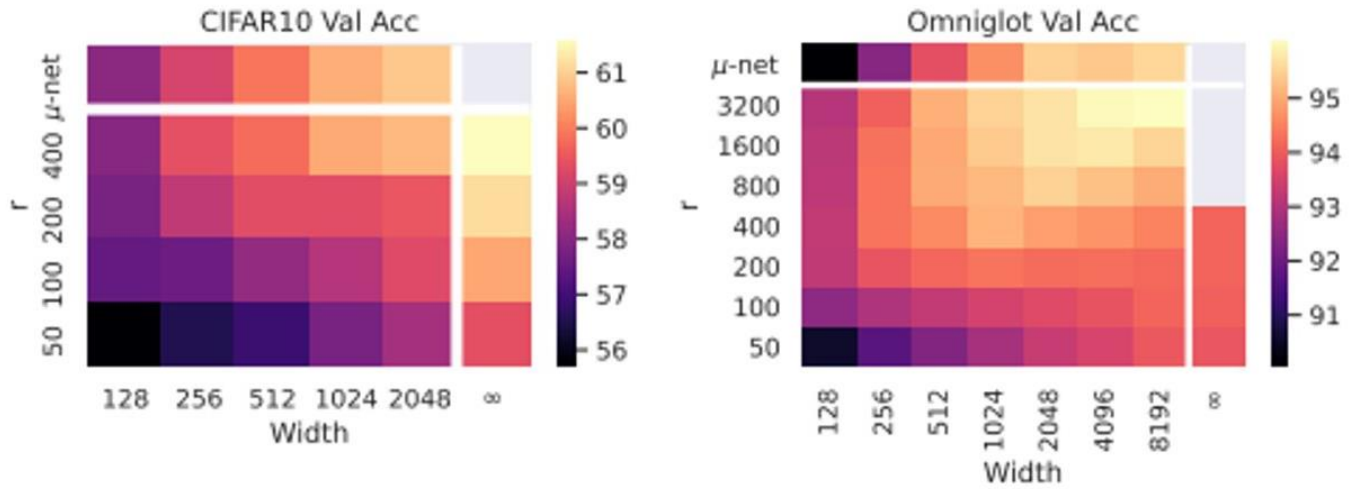
- **Key theory**: the inf-limit of *projected* gradient descent for *low rank* networks is efficiently computable (versus full gradient descent)



π-limit: Results

	NNGP	NTK	*NTK perf gap*	μ-Net	π-Net	π-Limit	π-Limit ImageNet Transfer
CIFAR10	58.92	59.63	$\longleftrightarrow$	61.31	60.64	**61.50**	**64.39**
Omniglot	43.80	51.72	$\longleftrightarrow$	91.22	**92.21**	**91.46**	-

- Big question: is low-rank projected learning worse than full-rank?



Computational Issues

- Finding full gradient descent limit has computational difficulties
- Gradient update causes one issue

$$f(\xi) \to \mathbb{E}(Z^{\sqrt{n}v} + \cdots)\phi(\xi Z^{\sqrt{n}u} + (\cdots)\phi'(\cdots))$$

- **Projecting gradient** solves this issue

$$f(\xi) \to \mathbb{E}(Z^{\sqrt{n}v} + \cdots)\phi(\tilde{c} Z^{\sqrt{n}u}) \qquad \theta \leftarrow \theta - \eta \Pi \nabla_\theta \mathcal{L}$$

Limit Formulation

- Expand to full network by projecting to **low rank subspace**

$$w^l \leftarrow \frac{1}{n}\Omega A^{l\top}\phi(B^l\Omega^\top) \in \mathbb{R}^{n\times n}$$

$$A^l, B^l \in \mathbb{R}^{M\times r}$$

$$\Omega \in \mathbb{R}^{n\times r}$$

- Optimizing hidden weights in (r x n) space
- Can **exactly compute** π-limit with just As and Bs

π-limit: Results

- Practical uses?
  - Memory grows Ω(T) during training, time Ω(T^2)
- Future Work?
  - Many topics for inf-width MLPs
  - Reduce memory, time requirements
  - Expand to more architectures