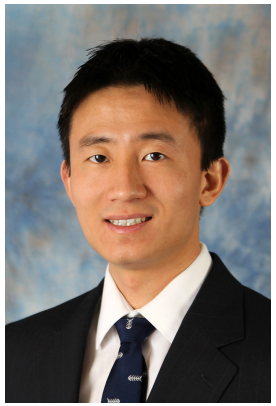
When Expressivity Meets Trainability: Fewer than n Neurons Can Work

NeurIPS 2021

Jiawei Zhang*

Yushun Zhang*

Mingyi Hong



Ruoyu Sun†

Zhi-Quan Luo

*: Equal contribution. Alphabetically ordered.

†: Corresponding author.

Motivation & Background

Training large networks is challenging.

Large neural networks require:

memory & computations

power consumption

Critical to resource constrained environments

embedded systems
e.g., mobile devices

real-time tasks
e.g., autonomous car

Figure from http://www.robots.ox.ac.uk/~namhoon/doc/slides-compression-ssp.pdf

Motivation

- Studying **small-sized** networks is still appealing.
  - Application: on-device AI, self-driving cars, etc.
  - Candidate strategy: Pruning, Quantization, **reducing the width**, etc.
  - In this work, we focus on **reducing the width** (training narrow nets).

Why do narrow nets performs badly?

- **However,** reducing the network **width** often leads to **worse** performance.
- What is the possible cause?
  - worse generalization power? (how the network performs on test sets)
  - weaker **expressivity**? (how large a dataset that a network can learn)
  - worse **trainability** ? (how effective a network can be optimized)
- We discuss the **expressivity** and **trainability** for **narrow** nets.

We ask two questions:

For the 1-hidden-layer network with width m < sample size n:

- (Q1): Can a narrow network have the strong **expressivity** to memorize n data samples?
  - When $m > n$: we naturally agree it is true.
  - When $m < n$: not clear.

- (Q2): If so, can a gradient-based method find a (near) **globally optimal** solution?
  - Cavate: bad basins (e.g. [Swirszcz et al.'16, Zhou et al.'17]),
  - GD iterates are hard to control.

Related works

- **Expressivity:** There exists a network to fit the data set (e.g. [Telgarsky'16, Zhang et al.'17, Park et al.'20])
  - [Shalev et al.'17] points out: these specially constructed networks CANNOT be found by gradient methods.

- **Trainability:** only for wide networks with width $O(\mathrm{ploy}(n))$

  (e.g. [Allen-Zhu et al.'19,Du et al.'19,Chizat et al.'18]).

- **When width $m < n$:** both are open questions.
- We suggest discussing these two topics together.
  (Attempted in [Daniely'19, Bubeck et al.'20, Zhou et al.'21.] But the settings & results are different.)

Our results (informal)

For 1-hidden-layer nets with width $m \geq \frac{2n}{d}$ (n: sample size, d: input dimension) : (when $d > 2,$ our results cover $m < n$.)

- **(A1) Expressivity:** there exists a global-min with zero empirical loss, i.e. the network can memorize n samples.

- **(A2) Trainability:** we propose a constrained problem where every KKT point has small loss.

Expressivity Analysis

Settings

- **Training set:** $\{x_i, y_i\}_{i=1}^n$

- **1-hidden-layer networks:**

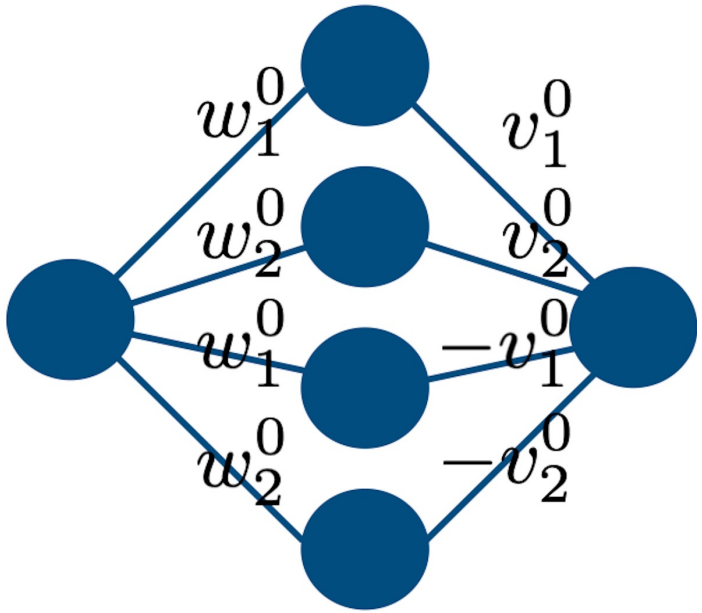
$$f(x_i;\theta) = \sum_{j=1}^m v_j \sigma\left(w_j^T x_i\right), \text{ where } \theta = \{w, v\}$$

- **The empirical loss:**

$$\min_\theta \ell(\theta) := \ell(f) = \frac{1}{2}\sum_{i=1}^n \left(y_i - f(x_i;\theta)\right)^2$$

Settings

- We consider “Mirrored LeCun’s initialization” as follows.



- Property: the output will be 0. Its benefits are discussed in the paper.

Narrow nets have strong expressivity

When $m \geq \frac{2n}{d}$: (it covers $m < n$ when $d > 2$):

Theorem 1 (1st half, informal):

Consider Mirrored LeCun's initialization (MLI) $\theta_0 = (w_0, v_0)$, then for any small neighborhood around w^0, there exists a $\hat{\theta}$, s.t. $\ell(\hat{\theta})$=0.

- the network can memorize n samples.
- There exists at least one global-min near MLI.
- The proof is based on Inverse Function Theorem.

Nice local landscape around the global-min

When $m \geq \frac{2n}{d}$: (it covers $m < n$ when $d > 2$):

Theorem 1 (2nd half, informal):

Around Mirrored LeCun's initialization (MLI) , there exists a "nice region" where every stationary point is a global-min.

- This is the foundation of "trainability"
- Proof is based on the full-rankness of Jacobian.
- Caveat: GD iterates may leave the "nice region".

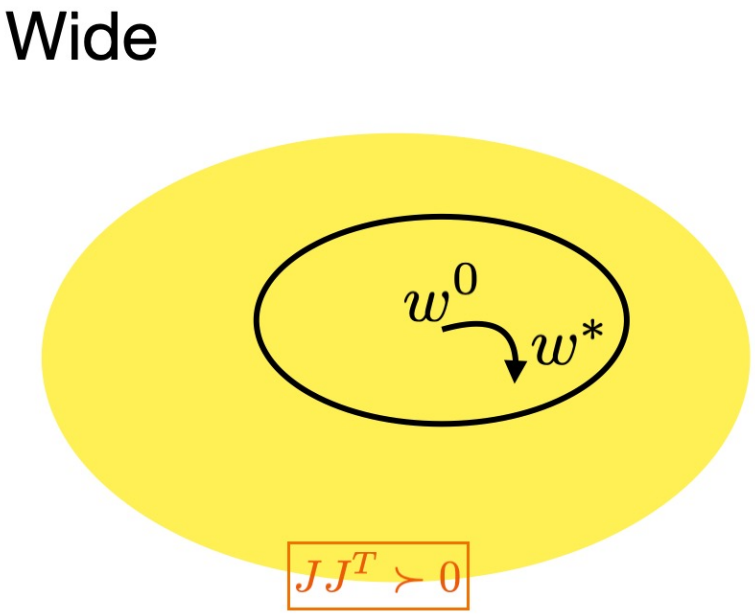
Trainability Analysis

How to find the global-min?

- **Theorem 1 tells us: Around the initialization MLI:**
  - a global-min with zero loss exists.
  - There is no bad local-min or saddles.

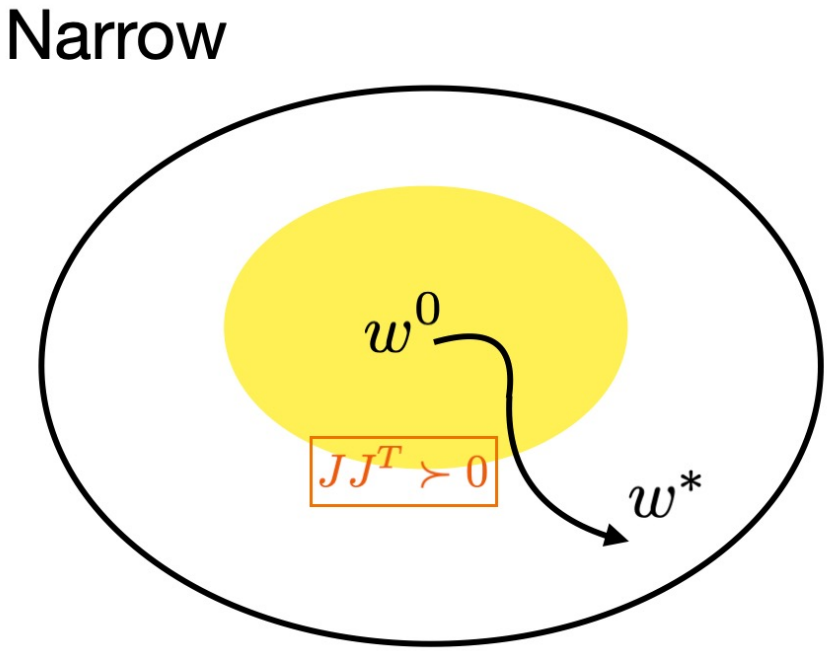
- **Main idea:** we want to keep the iterates around MLI, and search locally.

For wide nets, local search is natural.



- The GD iterates stay near initialization when $m = O(\text{ploy}(n))$.
- This is the key idea in NTK papers (e.g. [Du et al.'19])

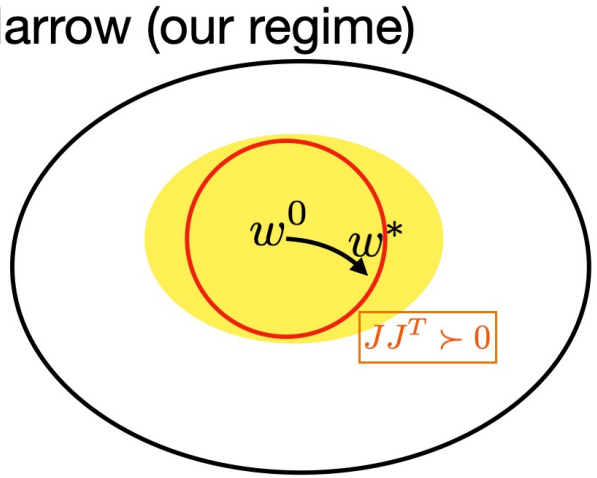
When width $m < n$: we cannot control the parameter movement



- It may hit a stationary point with singular Jacobian (with high loss).
- The traditional NTK story fails.

How to do local search for narrow nets?

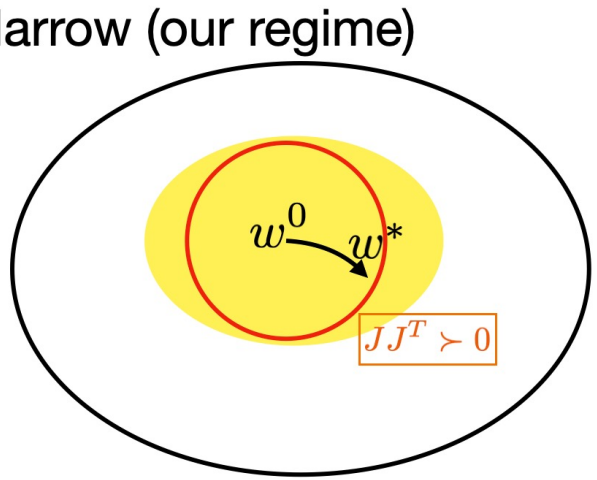
- An intuitive approach is to add constraint.



- 2 issues:
  - Perhaps there is no global-min inside the red ball.
  - Perhaps the algorithms will stop on the boundary (with large loss.)

How to do local search for narrow nets?

- An intuitive approach is to add constraint.



- 2 issues:
  - ~~Perhaps there is no global-min inside the red ball~~. By Thm1, a global-min exists in the red ball!
  - Perhaps the algorithms will stop on the boundary (with large loss.)

How to fix issue 2?

- We only need to change the output layer a bit (one line of code):
- Original form:

$$f\left(x_i; w; v\right) = \sum_{j=1}^{m} v_j \sigma\left(w_j^T x_i\right)$$

- New form:

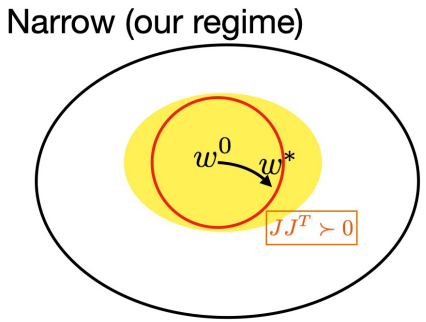
$$f(x; w, v) = \sum_{j=1}^{\frac{m}{2}} v_j \left(\sigma(w_j^T x) - \sigma(w_{j+\frac{m}{2}}^T x)\right).$$

- That is: we keep the pairwise pattern of v.

Trainability results

When $m \geq \frac{2n}{d}$: (it covers $m < n$ when $d > 2$):

Theorem 2 (informal): With the new proposed output layer and MLI, we propose a constrained problem which keeps $\|w - w^0\|_F \leq \epsilon$



Then all KKT points are near-global optimal.

- i.e., $\ell(w^*, v^*) = O(\epsilon^2)$, where ϵ is the constraint size.
- NO bad local-min on the boundary! (Proof is based on local geometry analysis)

Experiments

Experiments

- **We propose a new training method:**
- Mirrored initialization + pairwise output layer + constrained problem (with PGD).

- **Empirical performance of our method:**
- **Training:** our method can memorize random CIFAR-10.
- **Test:** our method generalizes well on R-ImageNet.

- **Ablation studies:**
- The narrow nets are hard to train using unconstrained SGD.
- It is necessary to change the algorithm.

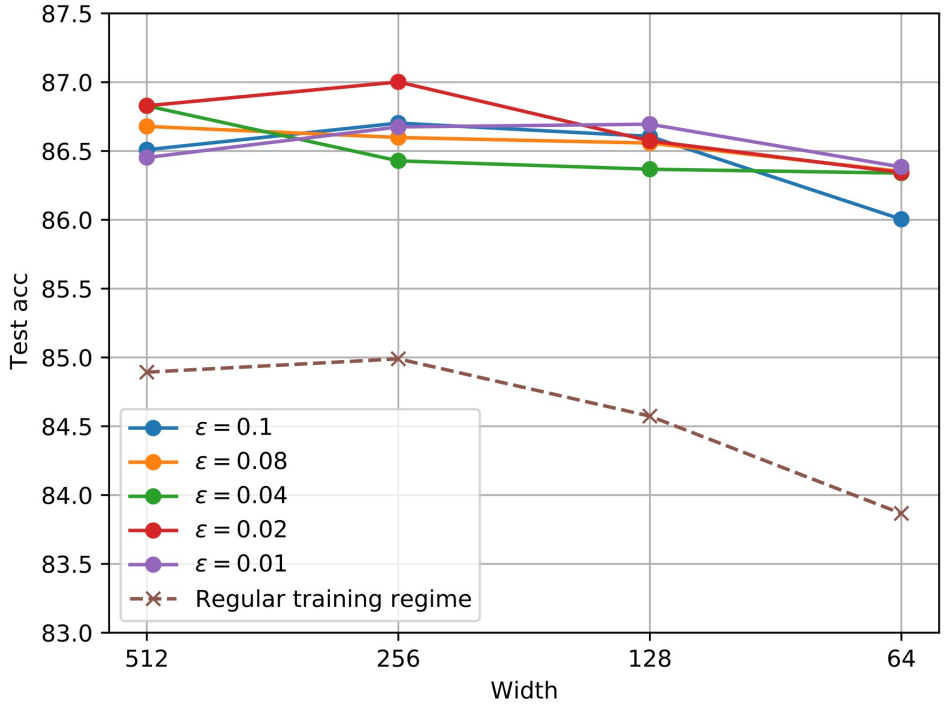
Training performance on random data

Table 3: Results on the random-labeled CIFAR-10

Width	Epoch	Activation	Train acc	Test acc
1024	1000	ReLU	0.9931	0.1011
2048	1000	ReLU	0.9984	0.1022
4096	1000	ReLU	0.9998	0.0962
1024	1000	Tanh	0.9872	0.0991
2048	1000	Tanh	0.9927	0.1024
4096	1000	Tanh	0.9938	0.0962

- **Using our training method:** 1-hidden-layer nets can memorize random-labeled CIFAR-10

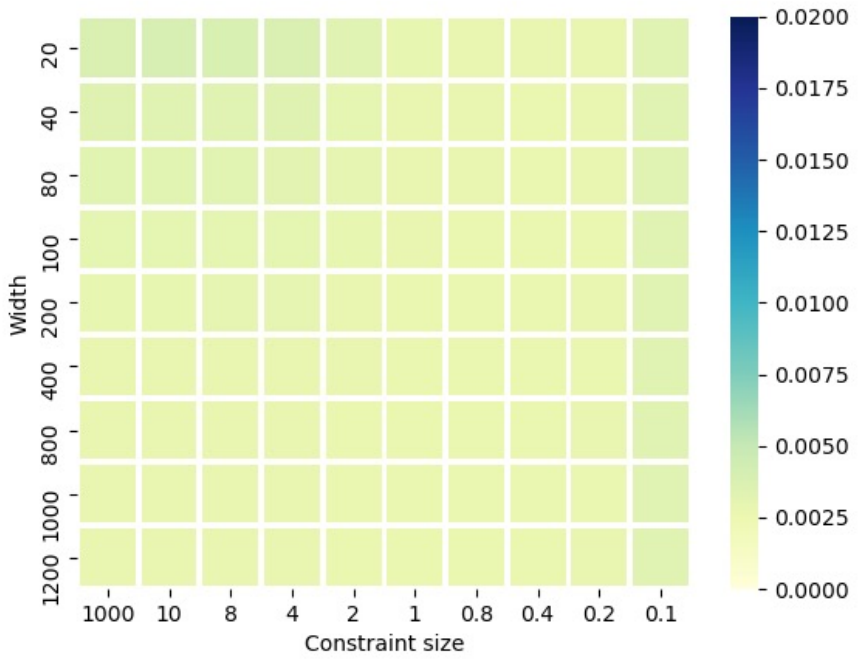
How about generalization?



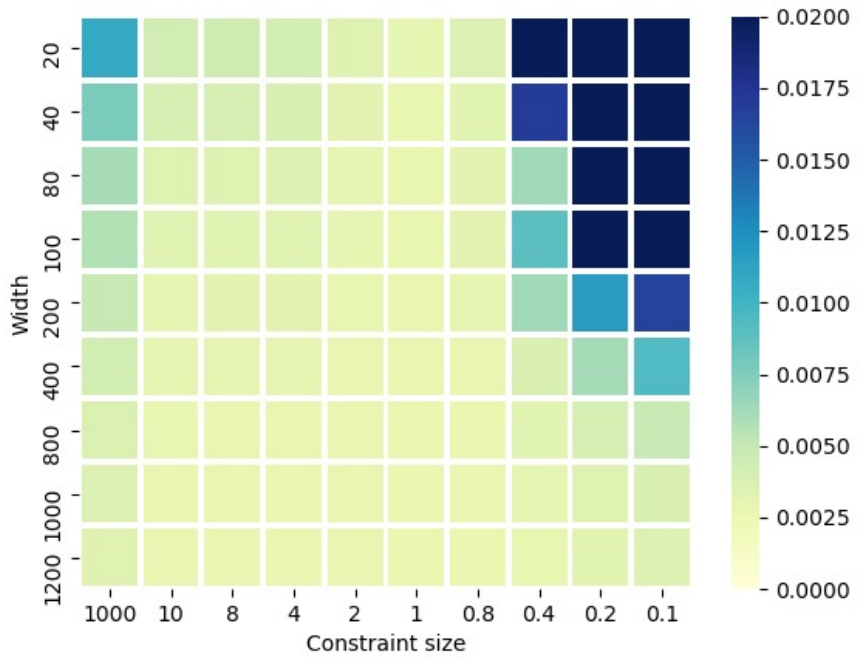
- On Restricted-ImageNet, our training regime (with PGD) outperforms SGD-based training in ‘test acc’, especially in narrow cases.
- More experiments on MNIST, CIFAR10, CIFAR100 can be seen in the paper.

Ablation studies on synthetic data: training error

GD&PGD+Mirrored init+parwised output layer



GD&PGD



- Unconstrained GD fails for narrow nets.
- Directly adding constraint will not help:
  We need Mirrored LeCun's initialization + changes of the output layer.

Conclusions

Conclusion

We shed new light on narrow nets training.

For 1-hidden-layer nets with width $m \geq \frac{2n}{d}$ (when $d > 2$, our results cover $m < n$.) :

- **(A1) Expressivity:** there exists a global-min with zero empirical loss, i.e. the network can memorize n samples.
- **(A2) Trainability:** we propose a constrained problem where every KKT point has small loss.
- **Empirically:** our training method promotes the training & test performance