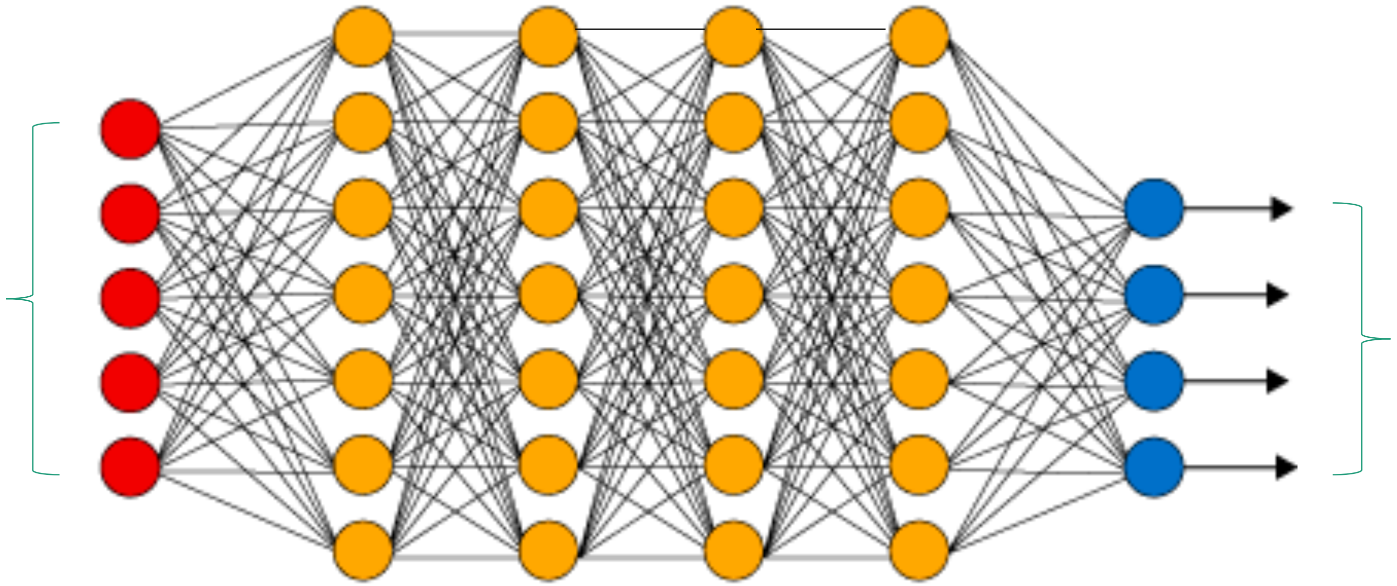
Width Provably Matters in Optimization for Deep Linear Neural Networks

Wei Hu
Princeton University

Joint work with Simon Du (CMU)

Deep Linear Neural Network

$x \in \mathbb{R}^{d_{\text{in}}}$



$y = W_L \cdots W_1 x \in \mathbb{R}^{d_{\text{out}}}$

W_1 W_2 $\cdots$ W_L

Training a Deep Linear Network

Training a Deep Linear Network

- Given training data $(x_1, y_1), \ldots, (x_n, y_n)$

Training a Deep Linear Network

- Given training data $(x_1, y_1), \ldots, (x_n, y_n)$
- Minimize the quadratic loss over training data

$$\ell(W_1, \ldots, W_L) = \frac{1}{2}\sum_{i=1}^{n} \|W_L \cdots W_1 x_i - y_i\|^2$$

Training a Deep Linear Network

- Given training data $(x_1, y_1), \dots, (x_n, y_n)$
- Minimize the quadratic loss over training data

$$\ell(W_1, \dots, W_L) = \frac{1}{2}\sum_{i=1}^{n} \|W_L \cdots W_1 x_i - y_i\|^2$$

- This work: gradient descent with standard independent random initialization on ℓ w.r.t. $W_1, \dots, W_L$

$$W_j(t+1) = W_j(t) - \eta \frac{\partial \ell}{\partial W_j}\big(W_1(t), \dots, W_L(t)\big)$$

Why Studying Deep Linear Networks?

Why Studying Deep Linear Networks?

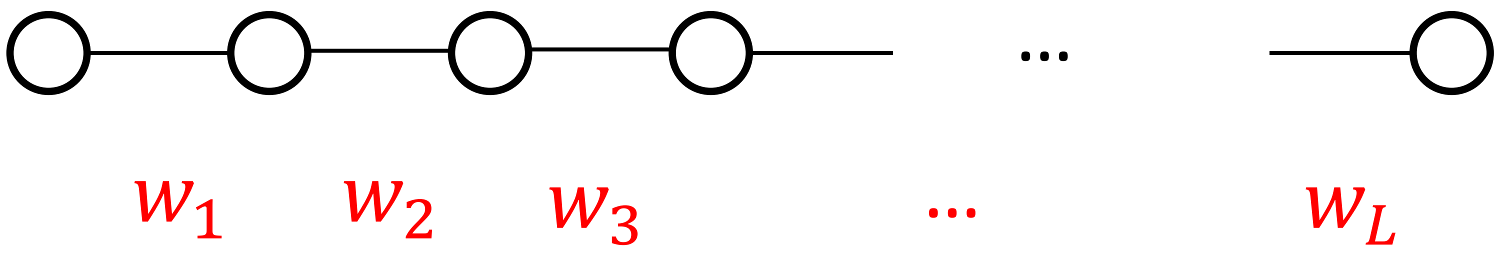
- Linear networks exhibit common challenges in optimization for deep learning
  - Non-convex
  - Non-strict saddle
  - Can have “vanishing gradient” or “exploding gradient”

Why Studying Deep Linear Networks?

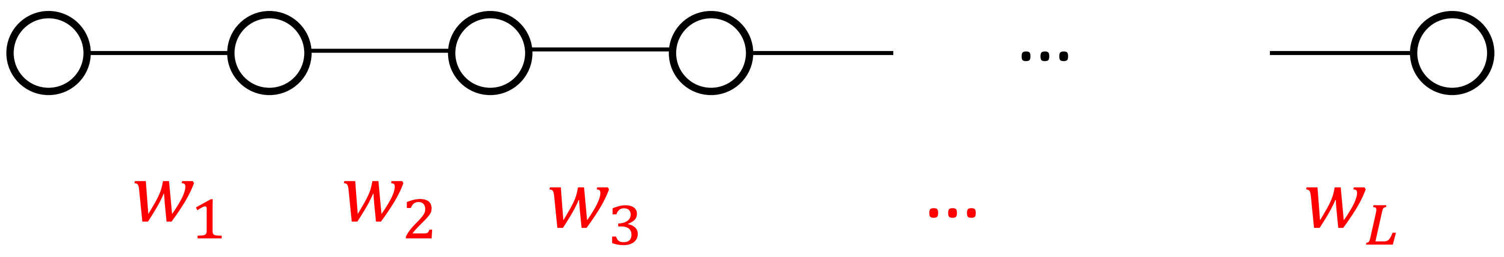
- Linear networks exhibit common challenges in optimization for deep learning
  - Non-convex
  - Non-strict saddle
  - Can have "vanishing gradient" or "exploding gradient"

- Deep linear networks may help generalization
  - [Lampinen, Ganguli, ICLR'19], [Arora, Cohen, H, Luo, 2019], [Gidel, Bach, Lacoste-Julien, 2019], etc.

Exponential Lower Bound for Narrow Linear Nets

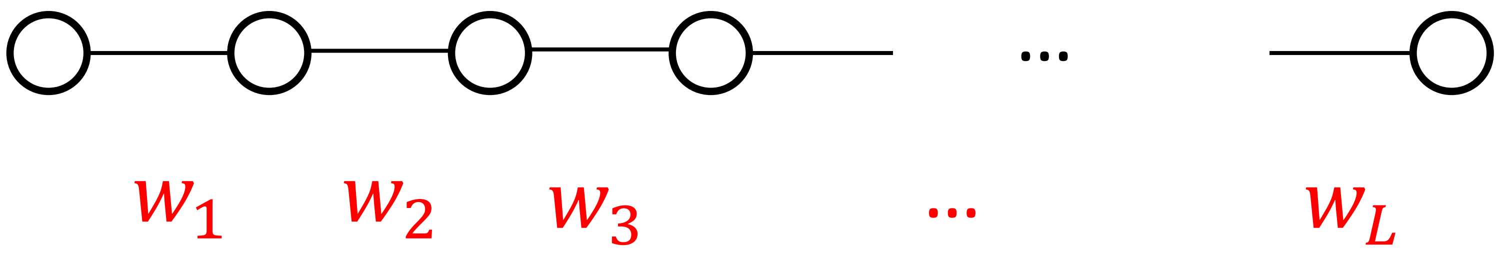


Exponential Lower Bound for Narrow Linear Nets



Theorem [Shamir, COLT'19]: GD with random initialization w.h.p. needs $2^{\Omega(L)}$ iterations to converge to global min

Exponential Lower Bound for Narrow Linear Nets

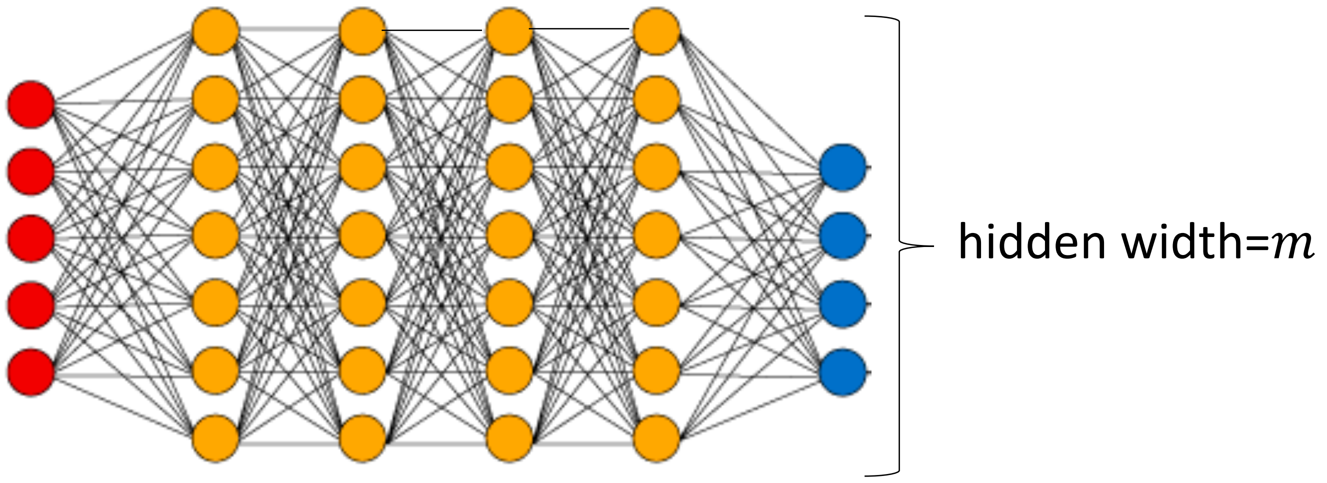


Theorem [Shamir, COLT'19]: GD with random initialization w.h.p. needs $2^{\Omega(L)}$ iterations to converge to global min

Questions: Can we get efficient convergence for wide linear nets? If so, how wide is enough?

Our Result

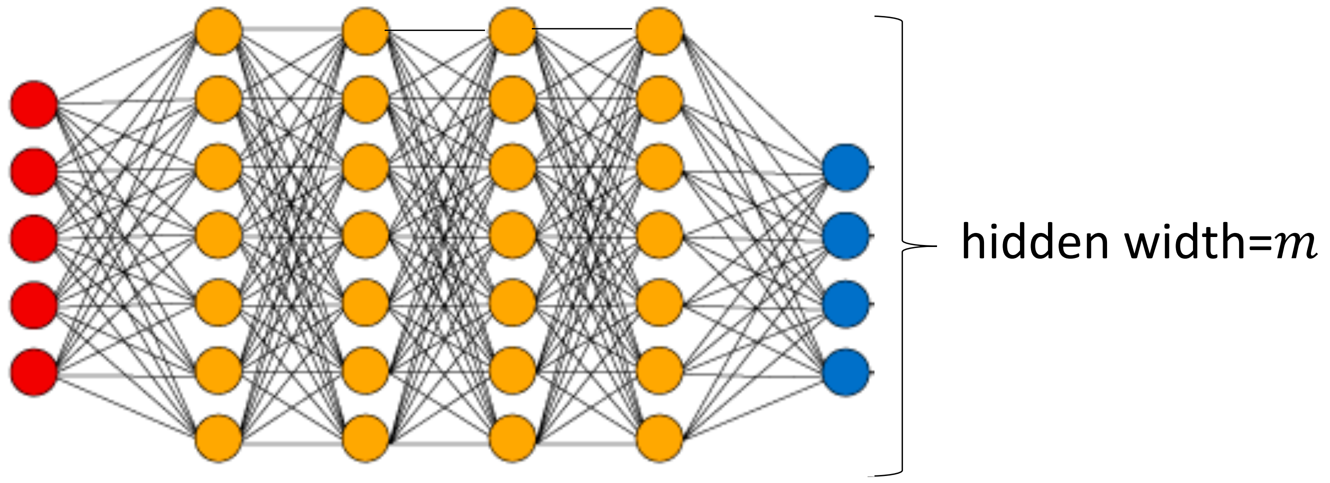
$$\ell(W_1, \ldots, W_L) = \frac{1}{2}\sum_{i=1}^{n} \|W_L \cdots W_1 x_i - y_i\|^2$$



- m: width of every hidden layer

Our Result

$$\ell(W_1, \dots, W_L) = \frac{1}{2}\sum_{i=1}^{n}\|W_L \cdots W_1 x_i - y_i\|^2$$



- m: width of every hidden layer

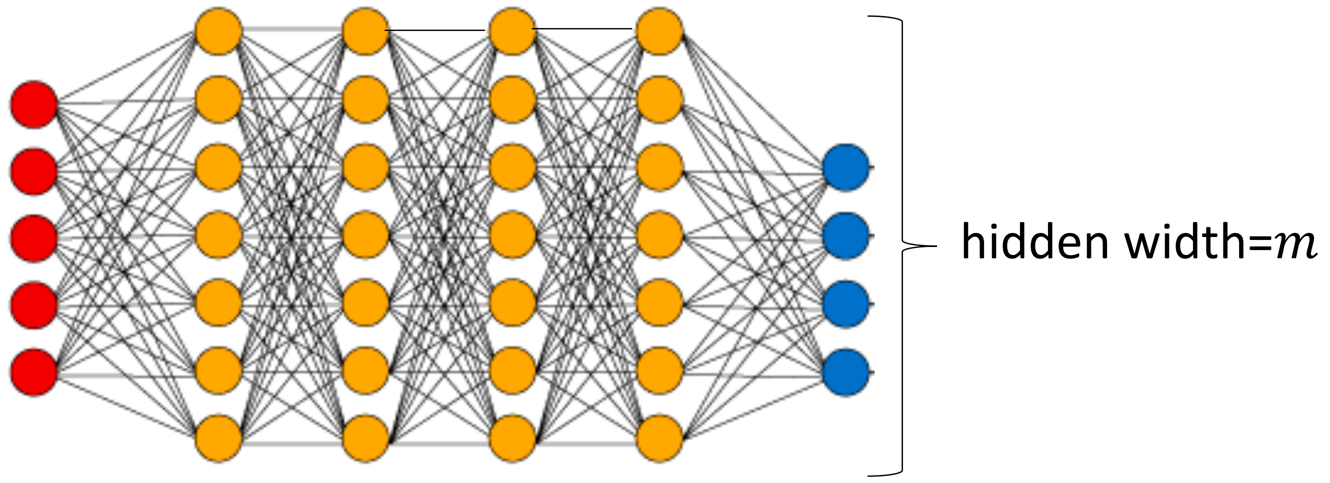
Main Theorem: if $m \geq \widetilde{\Omega}(L)$, then GD with random init converges to global min at a linear rate w.h.p., i.e.

$$\mathrm{loss}(t) - \mathrm{OPT} \leq e^{-\Omega(t)}(\mathrm{loss}(0) - \mathrm{OPT})$$

Our Result

$$\ell(W_1, \ldots, W_L) = \frac{1}{2}\sum_{i=1}^{n} \|W_L \cdots W_1 x_i - y_i\|^2$$

- m: width of every hidden layer



Main Theorem: if $m \geq \tilde{\Omega}(L)$, then GD with random init converges to global min at a linear rate w.h.p., i.e.

$$\text{loss}(t) - \text{OPT} \leq e^{-\Omega(t)}(\text{loss}(0) - \text{OPT})$$

Width provably matters

narrow network → $\exp(L)$ time

wide network → $\text{poly}(L)$ time

Comparison with Previous Work

Paper	Init	Opt soln	Data	Global convergence?
[Bartlett, Helmbold, Long, ICML'18]	identity	PD or close to identity	whitened	no
[Arora, Cohen, Golowich, H, ICLR'19]	balanced	full rank	whitened	no
This paper	random	any	any	yes

Poster: tonight #94